Leibniz
Universität Hannover

Reports of the Institute of Biostatistics

No 01 / 2007

Leibniz University of Hannover

Natural Sciences Faculty

Titel: *IUT for multiple endpoints*

Authors: *Mario Hasler*

1 Introduction

Some of the focus in new drug development has been shifted to develop new medicines which may not necessarily be more effective but have some other advantages compared to currently marked drugs, like reducing toxicity. An application, e.g., is to show safety of a new treatment on multiple endpoints compared to a reference. A rigorous claiming is to declare global safety if and only if each endpoint is safe. Two-sided hypotheses are appropriate for most endpoints because a direction of a harm effect is not known a priori. This is, each endpoint both must not undershoot a given lower limit of the reference and must not overshoot a given upper limit of this reference, respectively. Because it is often hard to fix uniform absolute safety thresholds jointly for all endpoints, ratios (not differences) to control shall be considered, too. The equivalence thresholds must be set a priori. But they are relative, e.g. in percent, giving an easy interpretation. For example, the new treatment will be declared as safe if, for each endpoint, not undershooting a lower limit of 80% of the reference and not overshooting an upper limit of 125% of this reference, respectively.

Much work has been done on the assessment of bioequivalence or therapeutic equivalence between two treatments on a univariate endpoint. But there is limited research on the assessment of equivalence on multiple endpoints. The traditional way to treat this problem, the intersection-union-test (IUT), is known to be conservative in many situations. Against the background of this problem, the question arises whether there are tests not having this weak point. In fact, there are some improved tests based on the IUT but most of them only hold for special cases. On the other hand, different approaches exist, like the Hotelling's T^2-test, using a square sum test statistic for the differences in the means to show equivalence on multiple endpoints. A short recommendation in literature is: Bloch *et al.* [2], Berger and Hsu [1], Casella and Berger [3], Hochberg and Tamhane [5], Wu *et al.* [8]. These tests either do not exploit the complete type I error - they have level α, not size α - or they are not applicable for ratios.

Like the union-intersection-test (UIT) for which a multivariate t-distribution can be derived for the global test statistic, the idea was to do the same for the intersection-union-test (IUT). The traditional IUT becomes less conservative for high correlations and, hence, very conservative for lower or negative. A multivariate approach, taking correlations into account, was assumed to avoid this handicap. The expected advantage was to get a size-α test this way.

2 Union-intersection and intersection-union method

2.1 Union-intersection method

The union-intersection method (UI) of test construction might be useful when the null hypothesis can be conveniently expressed as an intersection of a family of hypotheses, this is,

$$H_0 = \bigcap_{i=1}^{k} H_{0i}.$$

Suppose that a suitable test is available for each $H_{0i} : \theta \in \Theta_i$ versus $H_{1i} : \theta \in \Theta_i^c$. We can then write

$$H_0 : \theta \in \bigcap_{i=1}^{k} \Theta_i.$$

Say the rejection region for the test of H_{0i} is $\{x : T_i(x) \in R_i\}$. Hence, according to Roy (1953), the rejection region for the union-intersection test of H_0 is

$$\bigcup_{i=1}^{k} \{x : T_i(x) \in R_i\}.$$

This means that the global null hypothesis H_0 is rejected if and only if at least one of its component local null hypotheses H_{0i} is rejected. I.e., a new drug is tested and said to be hazardous if at least one endpoint is hazardous.

Depending on the test direction the local rejection region for each of the individual tests may be

$$\{x : T_i(x) > c\}.$$

with a common c for each individual test. The global rejection region of the UIT is

$$\bigcup_{i=1}^{k} \{x : T_i(x) > c\} = \{x : \max_{i=1,\ldots,k} T_i(x) > c\}.$$

Thus, the test statistic for testing H_0 is

$$T(x) = \max_{i=1,\ldots,k} T_i(x).$$

For the inverse test direction, the local rejection region for each of the H_{0i} is

$$\{x : T_i(x) < c\}.$$

Analogical considerations lead to the test statistic

$$T(x) = \min_{i=1,\ldots,k} T_i(x).$$

2.2 Intersection-union method

In contrast to the union-intersection method (IU) of test construction the intersection-union method is useful if the null hypothesis can be conveniently expressed as an union of a family of hypotheses, this is,

$$H_0 = \bigcup_{i=1}^{k} H_{0i}.$$

Again, supposing that a suitable test is available for each $H_{0i} : \theta \in \Theta_i$ versus $H_{1i} : \theta \in \Theta_i^c$ we can then write

$$H_0 : \theta \in \bigcup_{i=1}^{k} \Theta_i.$$

The rejection region for the test of H_{0i} is $\{x : T_i(x) \in R_i\}$. Hence, the rejection region for the intersection-union test of H_0 is

$$\bigcap_{i=1}^{k} \{x : T_i(x) \in R_i\}.$$

This means that the global null hypothesis H_0 is rejected if and only if each of its component local null hypotheses H_{0i} is rejected. I.e., a new drug is tested and said to be safe if each endpoint is safe.

Theorem: Let α_i be the size of the test of H_{0i} with rejection region R_i ($i = 1, \ldots, k$). Then the IUT with rejection region $R = \bigcap_{i=1}^{k} R_i$ is a level-α test, that is, its size is at most α with

$$\alpha = \max_{i=1,\ldots,k} \alpha_i.$$

Proof: Let $\theta \in \bigcup_{i=1}^{k} \Theta_i$. Then $\theta \in \Theta_i$ for some i and

$$P_\theta(X \in R) \leq P_\theta(X \in R_i) = \alpha_i \leq \alpha.$$

q.e.d.

Suppose the test direction for which the local rejection region for each of the individual tests is

$$\{x : T_i(x) > c\}$$

with a common c for each individual test. Then the global rejection region of the IUT is

$$\bigcap_{i=1}^{k} \{x : T_i(x) > c\} = \{x : \min_{i=1,\ldots,k} T_i(x) > c\}.$$

And thus, the test statistic for testing H_0 is

$$T(x) = \min_{i=1,\ldots,k} T_i(x).$$

Again, the inverse test direction leads to the local rejection region for each of the H_{0i}, this is now,

$$\{x : T_i(x) < c\}.$$

And we obtain the test statistic

$$T(x) = \max_{i=1,\ldots,k} T_i(x).$$

3 Test procedure

3.1 Assumptions

For $i = 1, ..., k$ and $j = 1, ..., n_X$, let X_{ij} denote the outcomes for k endpoints of an experimental treatment. Suppose that these random variables follow a k-variate normal distribution with mean vector $\mu_X = (\mu_{X1}, \dots \mu_{Xk})'$ and unknown covariance matrix $\mathbf{\Sigma}_X$. In the same manner, let the outcomes Y_{ij} of a reference treatment be k-variate normal distributed with parameters $\mu_Y = (\mu_{Y1}, \dots \mu_{Yk})'$ and $\mathbf{\Sigma}_Y$. Suppose that X_{ij} and Y_{ij} are mutually independent and $\mathbf{\Sigma}_X = \mathbf{\Sigma}_Y = \mathbf{\Sigma}$. In this way, the experimental and the reference treatment are presumed to have the same variation per each single endpoint. Let $\bar{X} = (\bar{X}_1, \dots, \bar{X}_k)'$, $\bar{Y} = (\bar{Y}_1, \dots, \bar{Y}_k)'$ and $\hat{\mathbf{\Sigma}}_X$, $\hat{\mathbf{\Sigma}}_Y$ be the sample mean vectors and the sample covariance matrices for both treatments, respectively, with

$$\bar{X}_i = \frac{1}{n_X} \sum_{j=1}^{n_X} X_{ij}, \quad \bar{Y}_i = \frac{1}{n_Y} \sum_{j=1}^{n_Y} Y_{ij}.$$

The pooled sample covariance matrix $\hat{\mathbf{\Sigma}}$ is given by

$$\hat{\mathbf{\Sigma}} = \frac{(n_X - 1)\hat{\mathbf{\Sigma}}_X + (n_Y - 1)\hat{\mathbf{\Sigma}}_Y}{n_X + n_Y - 2}$$

with the elements

$$\hat{\sigma}_{ij} = \widehat{Cov}_{ij} = \frac{(n_X - 1)\widehat{Cov}(X_i, X_j) + (n_Y - 1)\widehat{Cov}(Y_i, Y_j)}{n_X + n_Y - 2} \quad (1 \leq i, j \leq k)$$

where $\widehat{Cov}(X_i, X_j)$ and $\widehat{Cov}(Y_i, Y_j)$ are the estimates for the covariances of the several endpoints. This does not mean the same weighting as Bloch et al. [2] do. But this denotation results in the fact that the diagonal elements then are

$$\hat{\sigma}_{ii} = \hat{\sigma}_i^2 = \frac{(n_X - 1)S_{X_i}^2 + (n_Y - 1)S_{Y_i}^2}{n_X + n_Y - 2} \quad (i = 1, \dots, k)$$

with

$$S_{X_i}^2 = \frac{1}{n_X - 1} \sum_{j=1}^{n_X} (X_{ij} - \bar{X}_i)^2, \quad S_{Y_i}^2 = \frac{1}{n_Y - 1} \sum_{j=1}^{n_Y} (Y_{ij} - \bar{Y}_i)^2$$

which are necessary in the following test procedure. From the pooled sample covariance matrix $\hat{\mathbf{\Sigma}}$, we then derive the estimation of the common correlation matrix of the data $\hat{\mathbf{R}}$.

The object is to compare the new experimental treatment with the reference, and to consider it to be safe if each endpoint is safe. This means an intersection-union test. We first observe the one-sided, later on, the equivalence test problem.

3.2 Test for differences in means

The new experimental treatment is declared to be safe if and only if each endpoint does not undershoot a given fixed limit of the reference. This results in the component local tests

$$H_{0i} : \mu_{Xi} - \mu_{Yi} \leq \delta_i \quad \text{vs.} \quad H_{1i} : \mu_{Xi} - \mu_{Yi} > \delta_i \tag{1}$$

with a relevant threshold δ_i. The global null hypothesis of the underlying intersection-union test (IUT) is

$$H_0 = \bigcup_{i=1}^{k} H_{0i}.$$

Figure 1 shows the parameter space of a test for the case of $k = 2$ endpoints. The rejection region for the test of H_0 is

$$\bigcap_{i=1}^{k} \{x_i, y_i : T_i(x_i, y_i) > c\}.$$

with the *t*-test statistics

$$T_i = \frac{\bar{X}_i - \bar{Y}_i - \delta_i}{\hat{\sigma}_i \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}}, \tag{2}$$

a common quantile c for each individual test and the pooled estimators $\hat{\sigma}_i^2$ for σ_i^2. Under the marginal assumptions of H_{0i}, that is, $\mu_{Xi} - \mu_{Yi} - \delta_i = 0$, the test statistics T_i are *t*-distributed with $n_X + n_Y - 2$ degrees of freedom. The global rejection region of the IUT is

$$\bigcap_{i=1}^{k} \{x_i, y_i : T_i(x_i, y_i) > c\} = \{x_i, y_i : \min_{i=1,\ldots,k} \{T_i(x_i, y_i)\} > c\}.$$

And thus, the test statistic for testing H_0 is

$$T(x, y) = \min_{i=1,\ldots,k} \{T_i(x_i, y_i)\}. \tag{3}$$

Under the marginal assumptions of all H_{0i} (the intersection of them), the test statistics T_i approximatively follow a joint k-variate t-distribution with $n_X + n_Y - 2$ degrees of freedom and a correlation matrix depending on the data's correlation matrix, $\mathbf{R}$. But because the global null hypothesis is a union - and not an intersection - of its local hypotheses, the margin of this global null hypothesis is not unique which would be necessary for deriving a joint k-variate t-distribution under H_0. So, we, indeed, have to take quantiles $c = t_{\nu,1-\alpha}$ of a univariate t-distribution. The decision rule is to reject H_0 and to conclude global safety if

$$T(x, y) > t_{\nu,1-\alpha}. \tag{4}$$

If safety is declared if and only if each endpoint does not overshoot a given fixed limit of the reference, the component local tests are

$$H_{0i} : \mu_{Xi} - \mu_{Yi} \geq \delta_i \quad \text{vs.} \quad H_{1i} : \mu_{Xi} - \mu_{Yi} < \delta_i \tag{5}$$

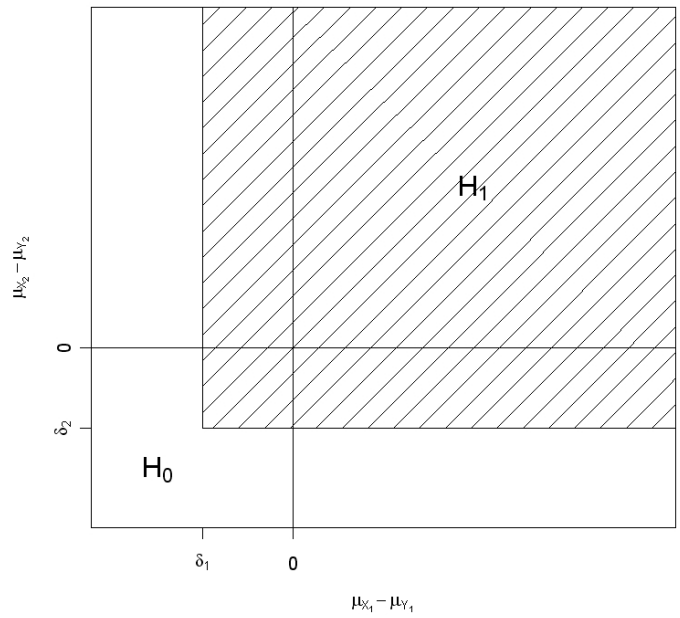

Figure 1: Parameter space of the test by differences for non-inferiority with $k = 2$ endpoints.

with a relevant threshold δ_i. Figure 2 shows the parameter space of a test for the case of $k = 2$ endpoints. The rejection region for the test of H_0 is

$$\bigcap_{i=1}^{k} \{x_i, y_i : T_i(x_i, y_i) < c\}$$

with the *t*-test statistics according to Equation (2). The global rejection region of the IUT is

$$\bigcap_{i=1}^{k} \{x_i, y_i : T_i(x_i, y_i) < c\} = \{x_i, y_i : \max_{i=1,\ldots,k} \{T_i(x_i, y_i)\} < c\}.$$

The test statistic for testing H_0 is now

$$T(x, y) = \max_{i=1,\ldots,k} \{T_i(x_i, y_i)\}. \tag{6}$$

The decision rule now is to reject H_0 if

$$T(x, y) < t_{\nu,\alpha} \tag{7}$$

which corresponds with $T(x, y) < -t_{\nu,1-\alpha}$.

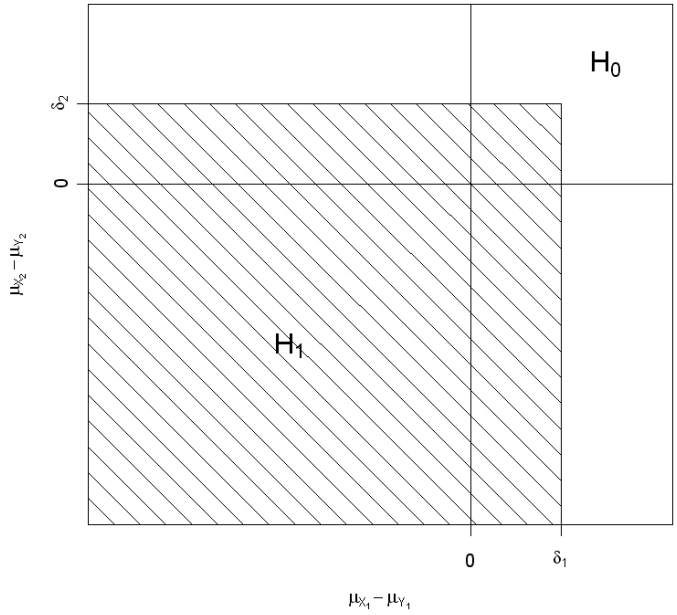

Figure 2: Parameter space of the test by differences for non-superiority with $k = 2$ endpoints.

Now, the new experimental treatment is declared to be safe if and only if each endpoint both does not undershoot a given fixed lower limit of the reference and does not overshoot a given fixed upper limit of the reference, respectively. This results in the component local tests for

$$\begin{aligned} H_{0i} &: \mu_{Xi} - \mu_{Yi} \leq \delta_i^{(1)} \quad \text{or} \quad \mu_{Xi} - \mu_{Yi} \geq \delta_i^{(2)} \quad \text{vs.} \\ H_{1i} &: \mu_{Xi} - \mu_{Yi} > \delta_i^{(1)} \quad \text{and} \quad \mu_{Xi} - \mu_{Yi} < \delta_i^{(2)} \end{aligned} \tag{8}$$

with relevant thresholds $\delta_i^{(1)} < \delta_i^{(2)}$. The global null hypothesis of the underlying intersection-union test is

$$H_0 = \bigcup_{i=1}^{k} H_{0i} = \bigcup_{i=1}^{k} \{H_{0i}^{(1)} \cup H_{0i}^{(2)}\}$$

with

$$H_{0i}^{(1)} : \mu_{Xi} - \mu_{Yi} \leq \delta_i^{(1)} \quad \text{and} \quad H_{0i}^{(2)} : \mu_{Xi} - \mu_{Yi} \geq \delta_i^{(2)}.$$

The global test on equivalence is an IUT because the null hypothesis can be expressed as a union of a family of hypotheses. Each local test itself is an IUT, too, because made up of two one-sided tests with contrary direction. In rewriting H_0 by

$$H_0 = \bigcup_{i=1}^{k} H_{0i}^{(1)} \cup \bigcup_{i=1}^{k} H_{0i}^{(2)} = H_0^{(1)} \cup H_0^{(2)},$$

we reorganize the test problem. $H_0^{(1)}$ and $H_0^{(2)}$ represent two one-sided IUT now with contrary direction we have already focused. The test for the global H_0 is still an IUT because the null hypothesis is again a union of two hypotheses. Figure 3 shows the parameter space of a test for the case of $k = 2$ endpoints. The rejection region for the test of H_0 is

$$\bigcap_{i=1}^{k}\{x_i, y_i : T_i^{(1)}(x_i, y_i) > c^{(1)}\} \cap \bigcap_{i=1}^{k}\{x_i, y_i : T_i^{(2)}(x_i, y_i) < c^{(2)}\}.$$

with the *t*-test statistics

$$T_i^{(1)} = \frac{\bar{X}_i - \bar{Y}_i - \delta_i^{(1)}}{\hat{\sigma}_i\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}}, \quad T_i^{(2)} = \frac{\bar{X}_i - \bar{Y}_i - \delta_i^{(2)}}{\hat{\sigma}_i\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}}, \tag{9}$$

the quantiles $c^{(1)}$ and $c^{(2)}$ for the individual test and the pooled estimators $\hat{\sigma}_i^2$ for σ_i^2. Under the marginal assumptions of $H_{0i}^{(1)}$, the test statistics $T_i^{(1)}$ are *t*-distributed with $n_X + n_Y - 2$ degrees of freedom. Under the marginal assumptions of $H_{0i}^{(2)}$, the test statistics $T_i^{(2)}$ are *t*-distributed with $n_X + n_Y - 2$ degrees of freedom. From the considerations above, it follows that the rejection region for this IUT is

$$\{x_i, y_i : \min_{i=1,\ldots,k} T_i^{(1)}(x_i, y_i) > c^{(1)}\} \cap \{x_i, y_i : \max_{i=1,\ldots,k} T_i^{(2)}(x_i, y_i) < c^{(2)}\}.$$

We now rewrite the test hypotheses of Equation (8) as follows,

$$\begin{aligned} &H_{0i} : \mu_{Xi} - \mu_{Yi} \leq \delta_i^{(1)} \quad \text{or} \quad \mu_{Yi} - \mu_{Xi} \leq -\delta_i^{(2)} \quad \text{vs.} \\ &H_{1i} : \mu_{Xi} - \mu_{Yi} > \delta_i^{(1)} \quad \text{and} \quad \mu_{Yi} - \mu_{Xi} > -\delta_i^{(2)}. \end{aligned} \tag{10}$$

All the considerations above stay the same but the pair of test statistics according to Equation (9) changes into

$$T_i^{(1)} = \frac{\bar{X}_i - \bar{Y}_i - \delta_i^{(1)}}{\hat{\sigma}_i\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}}, \quad \tilde{T}_i^{(2)} = \frac{\bar{Y}_i - \bar{X}_i + \delta_i^{(2)}}{\hat{\sigma}_i\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}}, \tag{11}$$

The test statistics $T_i^{(2)}$ and $\tilde{T}_i^{(2)}$ now have converse test directions and hence, $T_i^{(1)}$ and $\tilde{T}_i^{(2)}$ have the same. Herewith, the rejection region can be transformed into

$$\{x_i, y_i : \min_{i=1,\ldots,k} T_i^{(1)}(x_i, y_i) > c^{(1)}\} \cap \{x_i, y_i : \min_{i=1,\ldots,k} \tilde{T}_i^{(2)}(x_i, y_i) > -c^{(2)}\}.$$

As mentioned above, the test for the global H_0 is an IUT because the null hypothesis is an union of two hypotheses. But the local null hypotheses $H_{0i}^{(1)}$ and $H_{0i}^{(2)}$ exclude each other. When $H_{0i}^{(1)}$ is true then $H_{0i}^{(2)}$ can not. Hence, we can not assume both the marginal assumptions of $H_{0i}^{(1)}$ and $H_{0i}^{(2)}$. There is no unique margin for the global null hypothesis. The following relations can be shown,

$$E(T_i^{(1)}|\partial H_{0i}^{(2)}) = \frac{\delta_i^{(2)} - \delta_i^{(1)}}{\sigma_i\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}}, \tag{12}$$

$$E(\tilde{T}_i^{(2)}|\partial H_{0i}^{(1)}) = \frac{\delta_i^{(2)} - \delta_i^{(1)}}{\sigma_i\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}}. \tag{13}$$

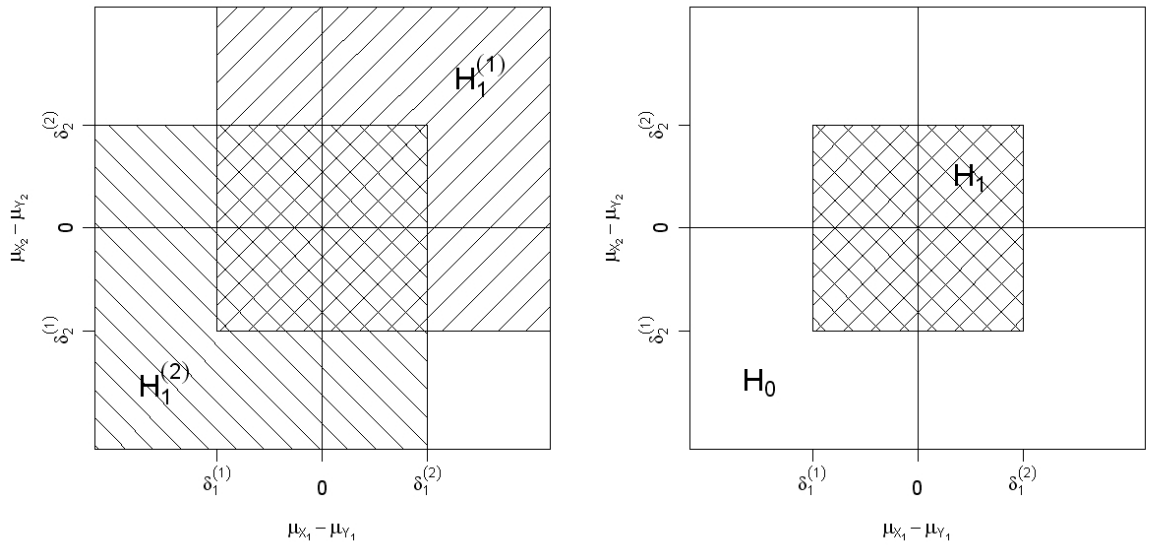

Figure 3: Parameter space of a test by differences on equivalence for $k = 2$ endpoints, the alternative hypothesis H_1 is an intersection of two one-sided alternative hypotheses $H_1^{(1)}$ and $H_1^{(2)}$.

These relations are easy to see in writing the test statistics $T_i^{(1)}$ and $-T_i^{(2)}$ in terms of each other,

$$\begin{aligned} T_i^{(1)} &= \frac{\bar{X}_i - \bar{Y}_i - \delta_i^{(1)} + \delta_i^{(2)} - \delta_i^{(2)}}{\hat{\sigma}_i\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} = \frac{-\left(\bar{Y}_i - \bar{X}_i + \delta_i^{(2)}\right)}{\hat{\sigma}_i\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} + \frac{\delta_i^{(2)} - \delta_i^{(1)}}{\hat{\sigma}_i\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} \\ &= -\tilde{T}_i^{(2)} + \frac{\delta_i^{(2)} - \delta_i^{(1)}}{\hat{\sigma}_i\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}}, \\ \tilde{T}_i^{(2)} &= \frac{\bar{Y}_i - \bar{X}_i + \delta_i^{(2)} + \delta_i^{(1)} - \delta_i^{(1)}}{\hat{\sigma}_i\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} = \frac{-\left(\bar{X}_i - \bar{Y}_i - \delta_i^{(1)}\right)}{\hat{\sigma}_i\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} + \frac{\delta_i^{(2)} - \delta_i^{(1)}}{\hat{\sigma}_i\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} \\ &= -T_i^{(1)} + \frac{\delta_i^{(2)} - \delta_i^{(1)}}{\hat{\sigma}_i\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}}. \end{aligned}$$

Therefore, under the marginal assumption of $H_{0i}^{(1)}$, the test statistic $\tilde{T}_i^{(2)}$ follows a non-central univariate t-distribution with $n_X + n_Y - 2$ degrees of freedom, non-centrality parameter

$$\theta_i = \frac{\delta_i^{(2)} - \delta_i^{(1)}}{\sigma_i\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}}. \tag{14}$$

The test statistic $T_i^{(1)}$ follows the same distribution but under the marginal assumption of $H_{0i}^{(2)}$. For this reason, we need two test statistics for testing H_0, namely

$$T^{(1)}(x,y) = \min_{i=1,\ldots,k} \left\{ T_i^{(1)}(x_i, y_i) \right\}, \quad T^{(2)}(x,y) = \min_{i=1,\ldots,k} \left\{ \tilde{T}_i^{(2)}(x_i, y_i) \right\}. \tag{15}$$

Again, under the marginal assumptions of all $H_{0i}^{(1)}$ (the intersection of them), the test statistics T_i approximatively follow a joint k-variate t-distribution with $n_X + n_Y - 2$ degrees of freedom and a correlation matrix depending on the data's one, $\mathbf{R}$. But because of the said reasons, one can not derive a joint k-variate t-distribution under H_0. So, we, have to take quantiles $c = t_{\nu,1-\alpha}$ of a univariate t-distribution. The decision rule is to reject $H_0^{(1)}$ if

$$T^{(1)}(x,y) > t_{\nu,1-\alpha}.$$

In the same manner, the decision rule is to reject $H_0^{(2)}$ if

$$T^{(2)}(x,y) > t_{\nu,1-\alpha}.$$

Safety can only be concluded if both

$$T^{(1)}(x,y) > t_{\nu,1-\alpha} \quad \text{and} \quad T^{(2)}(x,y) > t_{\nu,1-\alpha}. \tag{16}$$

3.3 Test for ratios of means

Most of the results of the test for differences in means holds for the case of ratios, too. So, (1) chances into

$$H_{0i} : \frac{\mu_{Xi}}{\mu_{Yi}} \leq \psi_i \quad \text{vs.} \quad H_{1i} : \frac{\mu_{Xi}}{\mu_{Yi}} > \psi_i \tag{17}$$

with a relevant threshold ψ_i. Figure 4 shows the parameter space of a test for the case of $k = 2$ endpoints. The local ratio-test statistics are

$$T_i = \frac{\bar{X}_i - \psi_i \bar{Y}_i}{\hat{\sigma}_i \sqrt{\frac{1}{n_X} + \frac{\psi_i^2}{n_Y}}}. \tag{18}$$

The test statistic for testing H_0 is

$$T(x,y) = \min_{i=1,\ldots,k} \left\{ T_i(x_i, y_i) \right\}. \tag{19}$$

The decision rule is to reject H_0 and to conclude global safety if

$$T(x,y) > t_{\nu,1-\alpha}. \tag{20}$$

Correspondingly, (5) chances into

$$H_{0i} : \frac{\mu_{Xi}}{\mu_{Yi}} \geq \psi_i \quad \text{vs.} \quad H_{1i} : \frac{\mu_{Xi}}{\mu_{Yi}} < \psi_i. \tag{21}$$

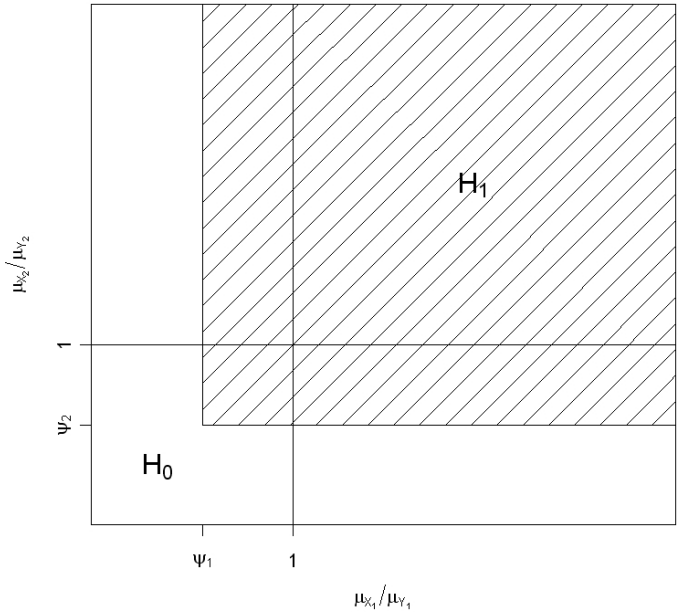

Figure 4: Parameter space of the test by ratios for non-inferiority with $k = 2$ endpoints.

Figure 5 shows the parameter space of a test for the case of $k = 2$ endpoints. The local ratio-test statistics are the same as in Equation (18). The test statistic for testing H_0 is now

$$T(x, y) = \max_{i=1,\ldots,k} \{T_i(x_i, y_i)\}. \tag{22}$$

The decision rule is to reject H_0 if

$$T(x, y) < t_{\nu,\alpha} \tag{23}$$

which corresponds with $T(x, y) < -t_{\nu,1-\alpha}$.

When the new experimental treatment is declared to be safe if and only if each endpoint both does not undershoot a given relative lower limit of the reference and does not overshoot a given relative upper limit of the reference, respectively, (8) chances into

$$\begin{aligned} H_{0i} &: \frac{\mu_{Xi}}{\mu_{Yi}} \leq \psi_i^{(1)} \quad \text{or} \quad \frac{\mu_{Xi}}{\mu_{Yi}} \geq \psi_i^{(2)} \quad \text{vs.} \\ H_{1i} &: \frac{\mu_{Xi}}{\mu_{Yi}} > \psi_i^{(1)} \quad \text{and} \quad \frac{\mu_{Xi}}{\mu_{Yi}} < \psi_i^{(2)} \end{aligned} \tag{24}$$

with relevant thresholds $\psi_i^{(1)} < \psi_i^{(2)}$. Figure 6 shows the parameter space of a test for the case of $k = 2$ endpoints. The local ratio-test statistics are

$$T_i^{(1)} = \frac{\bar{X}_i - \psi_i^{(1)} \bar{Y}_i}{\hat{\sigma}_i \sqrt{\frac{1}{n_X} + \frac{\psi_i^{(1)^2}}{n_Y}}}, \quad T_i^{(2)} = \frac{\bar{X}_i - \psi_i^{(2)} \bar{Y}_i}{\hat{\sigma}_i \sqrt{\frac{1}{n_X} + \frac{\psi_i^{(2)^2}}{n_Y}}}. \tag{25}$$

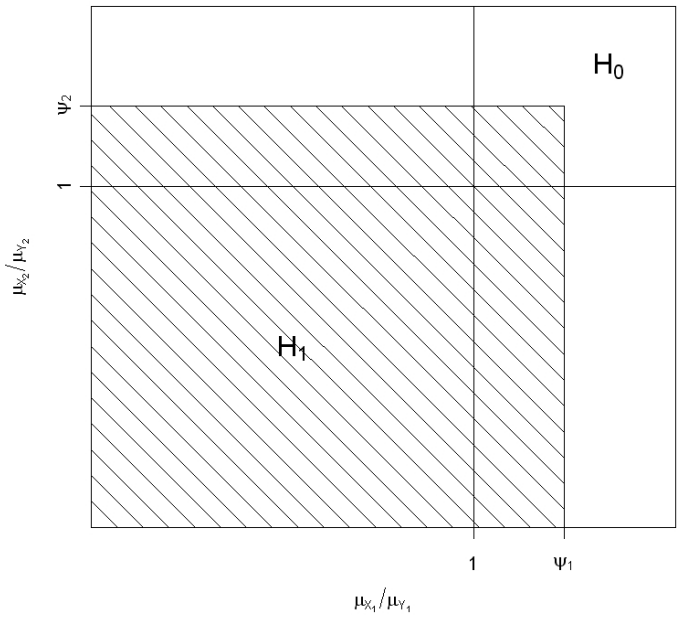

Figure 5: Parameter space of the test by ratios for non-superiority with $k = 2$ endpoints

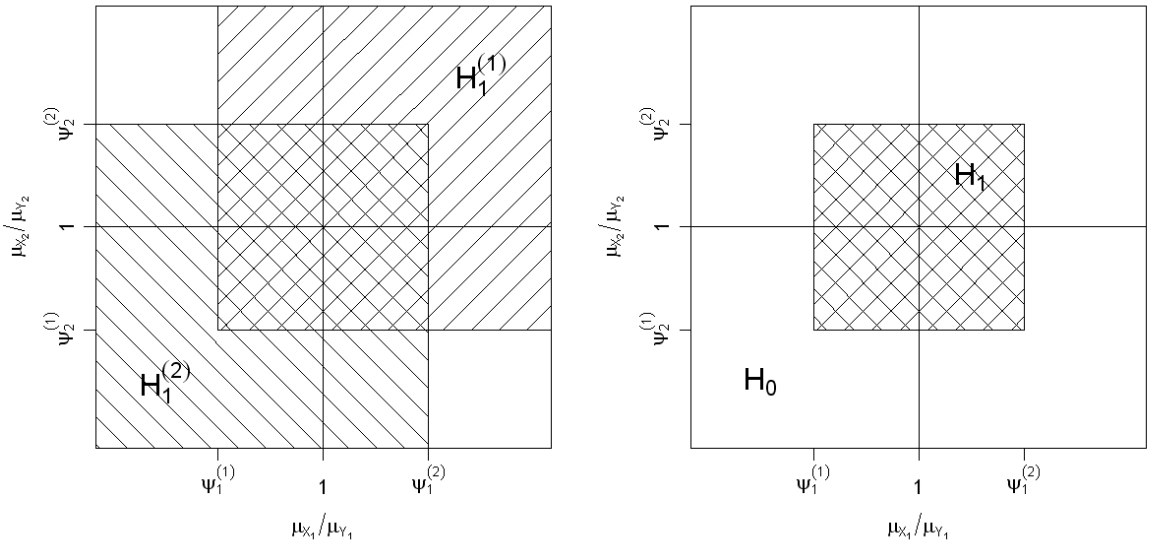

Figure 6: Parameter space of a test by ratios on equivalence for $k = 2$ endpoints, the alternative hypothesis H_1 is an intersection of two one-sided alternative hypotheses $H_1^{(1)}$ and $H_1^{(2)}$.

They will be rewritten into

$$T_i^{(1)} = \frac{\bar{X}_i - \psi_i^{(1)}\bar{Y}_i}{\hat{\sigma}_i\sqrt{\frac{1}{n_X} + \frac{\psi_i^{(1)^2}}{n_Y}}}, \quad \tilde{T}_i^{(2)} = \frac{\bar{Y}_i - \frac{1}{\psi_i^{(2)}}\bar{X}_i}{\hat{\sigma}_i\sqrt{\frac{1}{n_Y} + \frac{1}{\psi_i^{(2)^2} n_X}}}, \tag{26}$$

having the same test directions now. The following relations can be shown,

$$E(T_i^{(1)}|\partial H_{0i}^{(2)}) = \frac{\left(1 + \frac{1}{\psi_i^{(2)}}\right)\mu_{Xi} - \left(1 + \psi_i^{(1)}\right)\mu_{Yi}}{\sigma_i\sqrt{\frac{1}{n_X} + \frac{\psi_i^{(1)^2}}{n_Y}}}, \tag{27}$$

$$E(\tilde{T}_i^{(2)}|\partial H_{0i}^{(1)}) = \frac{\left(1 + \psi_i^{(1)}\right)\mu_{Yi} - \left(1 + \frac{1}{\psi_i^{(2)}}\right)\mu_{Xi}}{\sigma_i\sqrt{\frac{1}{n_Y} + \frac{1}{\psi_i^{(2)^2} n_X}}}. \tag{28}$$

These relations are again easy to see in writing the test statistics $T_i^{(1)}$ and $-T_i^{(2)}$ in terms of each other,

$$\begin{aligned}
T_i^{(1)} &= \frac{\bar{X}_i - \psi_i^{(1)}\bar{Y}_i + \bar{Y}_i - \frac{1}{\psi_i^{(2)}}\bar{X}_i - \bar{Y}_i + \frac{1}{\psi_i^{(2)}}\bar{X}_i}{\hat{\sigma}_i\sqrt{\frac{1}{n_X} + \frac{\psi_i^{(1)^2}}{n_Y}}} \\
&= \frac{\bar{X}_i - \psi_i^{(1)}\bar{Y}_i - \bar{Y}_i + \frac{1}{\psi_i^{(2)}}\bar{X}_i}{\hat{\sigma}_i\sqrt{\frac{1}{n_X} + \frac{\psi_i^{(1)^2}}{n_Y}}} + \frac{\bar{Y}_i - \frac{1}{\psi_i^{(2)}}\bar{X}_i}{\hat{\sigma}_i\sqrt{\frac{1}{n_X} + \frac{\psi_i^{(1)^2}}{n_Y}}} \frac{\hat{\sigma}_i\sqrt{\frac{1}{n_Y} + \frac{1}{\psi_i^{(2)^2} n_X}}}{\hat{\sigma}_i\sqrt{\frac{1}{n_Y} + \frac{1}{\psi_i^{(2)^2} n_X}}} \\
&= \frac{\bar{X}_i - \psi_i^{(1)}\bar{Y}_i - \bar{Y}_i + \frac{1}{\psi_i^{(2)}}\bar{X}_i}{\hat{\sigma}_i\sqrt{\frac{1}{n_X} + \frac{\psi_i^{(1)^2}}{n_Y}}} + \tilde{T}_i^{(2)}\sqrt{\frac{\frac{1}{n_Y} + \frac{1}{\psi_i^{(2)^2} n_X}}{\frac{1}{n_X} + \frac{\psi_i^{(1)^2}}{n_Y}}}, \\
\tilde{T}_i^{(2)} &= \frac{\bar{Y}_i - \frac{1}{\psi_i^{(2)}}\bar{X}_i + \bar{X}_i - \psi_i^{(1)}\bar{Y}_i - \bar{X}_i + \psi_i^{(1)}\bar{Y}_i}{\hat{\sigma}_i\sqrt{\frac{1}{n_Y} + \frac{1}{\psi_i^{(2)^2} n_X}}} \\
&= \frac{\bar{Y}_i - \frac{1}{\psi_i^{(2)}}\bar{X}_i - \bar{X}_i + \psi_i^{(1)}\bar{Y}_i}{\hat{\sigma}_i\sqrt{\frac{1}{n_Y} + \frac{1}{\psi_i^{(2)^2} n_X}}} + \frac{\bar{X}_i - \psi_i^{(1)}\bar{Y}_i}{\hat{\sigma}_i\sqrt{\frac{1}{n_Y} + \frac{1}{\psi_i^{(2)^2} n_X}}} \frac{\hat{\sigma}_i\sqrt{\frac{1}{n_X} + \frac{\psi_i^{(1)^2}}{n_Y}}}{\hat{\sigma}_i\sqrt{\frac{1}{n_X} + \frac{\psi_i^{(1)^2}}{n_Y}}} \\
&= \frac{\bar{Y}_i - \frac{1}{\psi_i^{(2)}}\bar{X}_i - \bar{X}_i + \psi_i^{(1)}\bar{Y}_i}{\hat{\sigma}_i\sqrt{\frac{1}{n_Y} + \frac{1}{\psi_i^{(2)^2} n_X}}} + T_i^{(1)}\sqrt{\frac{\frac{1}{n_X} + \frac{\psi_i^{(1)^2}}{n_Y}}{\frac{1}{n_Y} + \frac{1}{\psi_i^{(2)^2} n_X}}}.
\end{aligned}$$

Therefore, under the marginal assumption of $H_{0i}^{(1)}$, the test statistic $\tilde{T}_i^{(2)}$ follows a non-central univariate t-distribution with $n_X + n_Y - 2$ degrees of freedom, non-centrality parameter

$$\theta_i^{(2)} = \frac{\left(1 + \psi_i^{(1)}\right)\mu_{Yi} - \left(1 + \frac{1}{\psi_i^{(2)}}\right)\mu_{Xi}}{\sigma_i \sqrt{\frac{1}{n_Y} + \frac{1}{\psi_i^{(2)^2} n_X}}}. \tag{29}$$

Under the marginal assumption of $H_{0i}^{(2)}$, the test statistic $T_i^{(1)}$ follows the same distribution but with non-centrality parameter

$$\theta_i^{(1)} = \frac{\left(1 + \frac{1}{\psi_k^{(2)}}\right)\mu_{Xk} - \left(1 + \psi_k^{(1)}\right)\mu_{Yk}}{\sigma_k \sqrt{\frac{1}{n_X} + \frac{\psi_k^{(1)^2}}{n_Y}}}. \tag{30}$$

For this reason, we now have two test statistics for testing H_0 as follows,

$$T^{(1)}(x,y) = \min_{i=1,\ldots,k} \left\{ T_i^{(1)}(x_i, y_i) \right\}, \quad T^{(2)}(x,y) = \min_{i=1,\ldots,k} \left\{ \tilde{T}_i^{(2)}(x_i, y_i) \right\}. \tag{31}$$

The decision rule is to reject H_0 and to conclude safety if both

$$T^{(1)}(x,y) > t_{\nu,1-\alpha} \quad \text{and} \quad T^{(2)}(x,y) > t_{\nu,1-\alpha}. \tag{32}$$

3.4 α-simulations

Simulation studies were performed for 2, 4, 8 and 20, 40, 80 endpoints with several means and variances. For each fixed number of endpoints $k \in \{2, 4, 8, 20, 40, 80\}$, different grades of correlation were considered: maximal negative correlation, correlation 0, correlation 0.5 and maximal correlation. For each fixed k and grade of correlation, the endpoints were equicorrelated, this is $\rho_{ij} = \rho$ for all $1 \leq i \neq j \leq k$. Note that the negative correlations in the left column are bounded below by $\rho_{min} = -\frac{1}{k-1}$. 100000 simulation runs were taken for 2, 4, 8 endpoints, 10000 for 20, 40, 80 endpoints. Each simulation result was obtained using a program code in the statistic software R [7] and applying the package mvtnorm by Genz and Bretz [4].

The aim to show that using related quantiles of a k-variate t-distribution to obtain a size-α test, was not achieved. Indeed, this approach leads to an exact size α but only for the intersection of all margins of local null hypotheses $\bigcap_{i=1}^{k} \partial H_{0i}$. Quantiles of a univariate t-distribution result in very conservative decisions for that situation. But for the case of interest (the union $\bigcup_{i=1}^{k} \partial H_{0i} = \partial H_0$), the univariate method keeps the α-level conservatively, while the multivariate quite fails. An example is: 4 endpoints with correlation 0, one-sided testing (non-inferiority), balanced sample size 100, coefficient of variation 0.25, $\mu_Y = (0.1, 1, 10, 100)'$, $\mu_X = (0.079, 1.0, 10, 100)'$, $\delta = (-0.02, -0.20, -2.00, -20.00)'$. That means that μ_{X1} is inferior (unsafe), the others are non-inferior (safe). Because not each endpoint is safe, global safety could not be declared. The related type I errors are: 0.37 (multivariate method) and 0.03 (traditional univariate IUT).

4 Discussion

One possibility to show bioequivalence or therapeutic equivalence between two treatments on multiple endpoints is using an IUT for either differences in means or ratios of them. This then yields a global tests which rejects if and only if each local test rejects. E.g., a new experimental treatment is declared to be safe if each endpoint is safe, and safety is defined in not under-/ overshooting a given fixed limit of a reference. The IUT is known to be very conservative in many situations. One reason is that it does not take any correlations into account. Each endpoint will be tested separately using quantiles or p-values from univariate t-distributions. Another reason is the nature of the margin of null hypothesis, ∂H_0. So, the aim was to extend the IUT to a multivariate approach like the UIT using a multivariate t-distribution instead of a, say Bonferroni, adjustment. The conclusion so far is that there is no such easy equivalent multivariate-t approach for the IUT. The studied one does not keep the α-level for the complete space of the null hypotheses.

Another noteworthy fact is that an IUT for showing equivalence between two treatments on multiple endpoints always comes to a global decision. All endpoints together are equivalent or not. If not, omitting the hazardous endpoints does not synonymously mean the equivalence of the remaining ones. To demonstrate equivalence on a subset of endpoints (at least $1, \ldots, k-1$ of k) and to identify those, the procedure of Quan *et al.* [6] is an appropriate solution, for example.

References

[1] R.L. Berger and J.C. Hsu. Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science*, 1(4):283–319, 1996.

[2] D.A. Bloch, T.L. Lai, and P. Tubert-Bitter. One-sided tests in clinical trials with multiple endpoints. *Biometrics*, 57:1039–1047, 2001.

[3] G. Casella and R.L. Berger. *Statistical Inference*. Duxbury, Thomson Learning, 2002.

[4] A. Genz, F. Bretz, and R port by T. Hothorn. *mvtnorm: Multivariate Normal and t Distribution*, 2006. R package version 0.7-5.

[5] Y. Hochberg and A. C. Tamhane. *Multiple Comparison Procedures*. John Wiley and Sons, Inc., New York, 1987.

[6] Hui Quan, Jim Bolognese, and Weiying Yuan. Assessment of equivalence on multiple endpoints. *Statistics in Medicine*, 20:3159–3173, 2001.

[7] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2007. ISBN 3-900051-07-0.

[8] Y. Wu, M.G. Genton, and L.A. Stefanski. A multivariate two-sample mean test for small sample size and missing data. *Biometrics*, 62:877–885, 2006.