Leibniz
Universität Hannover

# Reports of the Institute of Biostatistics

No 03 / 2008

Leibniz University of Hannover

Natural Sciences Faculty

Titel: *Equivalence for multiple endpoints*

Authors: *Mario Hasler*

# 1 Data situation

For $i = 1, \ldots, k$ let $X_{ij}$ denote the $j$th (of $n_X$) observation on the $i$th endpoint under treatment $X$ and $Y_{ij}$ denote the $j$th (of $n_Y$) observation on the $i$th endpoint under treatment $Y$. Suppose these random variables to be mutually independent and follow $k$-variate normal distributions with mean vectors $\mu_X = (\mu_{X1}, \ldots \mu_{Xk})'$, $\mu_Y = (\mu_{Y1}, \ldots \mu_{Yk})'$ and unknown covariance matrices $\mathbf{\Sigma}_X$, $\mathbf{\Sigma}_Y$. Presume the treatments to have the same variation per each single endpoint, this is, $\mathbf{\Sigma}_X = \mathbf{\Sigma}_Y = \mathbf{\Sigma}$.

# 2 Methods

The underlying concept is related to Quan et al. [Quan et al. (2001)]. Differences in means between the treatments $X$ and $Y$ for the $k$ endpoints are considered. The object is to show equivalence - and hence safety - for as many endpoints as possible. The test hypotheses are:

$$\begin{aligned} H_{0i}: \quad & \mu_{Xi} - \mu_{Yi} \leq \delta_i^{lower} \quad \text{or} \quad \mu_{Xi} - \mu_{Yi} \geq \delta_i^{upper} \quad \text{(harmful)} \quad \text{vs.} \\ H_{Ai}: \quad & \mu_{Xi} - \mu_{Yi} > \delta_i^{lower} \quad \text{and} \quad \mu_{Xi} - \mu_{Yi} < \delta_i^{upper} \quad \text{(harmless)} \\ & \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (1 \leq i \leq k). \end{aligned}$$

with absolute thresholds $\delta_i^{upper}$ and $\delta_i^{lower}$. Therefore, the following step-up procedure will be applied. In a first step, calculate $(1-2\alpha)\times 100\%$ confidence intervals (CI) or related $p$-values for all $k$ endpoints. If each CI is within fixed equivalence ranges, all endpoints are equivalent and safe. The same conclusion can be obtained if each $p$-value is smaller than $\alpha$. The procedure then stops. If not, all endpoints failing this demand - say $m$ - are not equivalent and hence, unsafe. The remaining $(k - m)$ not decided endpoints are taken for next steps of the procedure to show potential safety for these ones. Calculate $(1 - 2\alpha/(m + 1)) \times 100\%$ CI for the said enpoints. Alternatively, take the former $p$-values again and compare them now with level $\alpha/(m + 1)$. The procedure ends with not later than the $k$-th step where the possibly last undecided endpoint comes to a conclusion using a $(1 - 2\alpha/k) \times 100\%$ CI or a level $\alpha/k$ for $p$-values.
The above procedure can also be extended to ratios of means. The test hypotheses then are:

$$\begin{aligned} H_{0i}: \quad & \mu_{Xi}/\mu_{Yi} \leq \theta_i^{lower} \quad \text{or} \quad \mu_{Xi}/\mu_{Yi} \geq \theta_i^{upper} \quad \text{(harmful)} \quad \text{vs.} \\ H_{Ai}: \quad & \mu_{Xi}/\mu_{Yi} > \theta_i^{lower} \quad \text{and} \quad \mu_{Xi}/\mu_{Yi} < \theta_i^{upper} \quad \text{(harmless)} \\ & \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (1 \leq i \leq k). \end{aligned}$$

with relative thresholds $\theta_i^{upper}$ and $\theta_i^{lower}$. Note that the related confidence intervals only exist if the $\mu_{Yi}$ are significantly larger than 0.

# 3 Simulation study

All the results are obtained by 10000 simulation runs with the same starting seed (seed 10) using a program code in the statistic software R and with package `mvtnorm`. For both differences and ratios of means, global and local control of the familywise error rate (FWER) are investigated for the tests based on $p$-values (not for the confidence intervals). For global control, the data are simulated under the

marginal assumptions of the null hypothesis. To get an impression of the local control, the treatments' last endpoint (of 4), the last two (of 8),. . . , and the last 5 (of 20) have the same mean. Only the FWER for the non-equal endpoints are focused. The following settings are investigated:
For global control of the FWER, differences:

- two treatments $X, Y$
- several number of endpoints and arbitrarily chosen thresholds:
  - 4 endpoints:
    $\mu_X = (0.08, 0.8, 12, 120),\ \mu_Y = (0.1, 1, 10, 100),\ \text{diag}\Sigma = 0.25\mu_Y$
    $\delta^{upper} = (0.02, 0.20, 2.00, 20.00),\ \delta^{lower} = (-0.02, -0.20, -2.00, -20.00)$
  - 8 endpoints with parameters like for four endpoints, all components 2-fold
  - 12 endpoints with parameters like for four endpoints, all components 3-fold
  - 16 endpoints with parameters like for four endpoints, all components 4-fold
  - 20 endpoints with parameters like for four endpoints, all components 5-fold
- several equicorrelations: $\rho^{min}$, 0, 0.5, 1
- fix sample size 20 for each endpoint of each treatment
- $\alpha = 0.05$

For local control of the FWER, differences:

- same setting as for global, differences, but
- last endpoint (of 4), last two (of 8),. . . , last 5 (of 20) are equivalent
  - 4 endpoints: $\mu_X = (0.08, 0.8, 12, 100)$
  - . . .

For global control of the FWER, ratios:

- same setting as for global, differences, but
- $\mu_X = (0.08, 0.8, 12.5, 125)$, $\theta^{upper} = (1.25, 1.25, 1.25, 1.25),\ \theta^{lower} = (0.8, 0.8, 0.8, 0.8)$

For local control of the FWER, ratios:

- same setting as for local, differences, but
- $\mu_X = (0.08, 0.8, 12.5, 100)$, $\theta^{upper} = (1.25, 1.25, 1.25, 1.25),\ \theta^{lower} = (0.8, 0.8, 0.8, 0.8)$

The step-up procedure by Quan et al. [Quan et al. (2001)] does not take any correlation of the endpoints into account. Therefore, decisions tend to be conservative with increasing correlations. This conservativeness is additionally compounded for an increasing number of endpoints. This can be seen in Tables 1, 2, 3 and 4. The ratio based tests there are always less conservative than the difference based ones. One reason is the little higher upper margins (125% instead of related 120%) for the ratio test.

| Endpoints | Correlations | | | |
|---|---|---|---|---|
| | $\rho^*$ | 0 | 0.5 | 1 |
| 4 | 0.0383 | 0.0398 | 0.0390 | 0.0331 |
| 8 | 0.0226 | 0.0233 | 0.0215 | 0.0134 |
| 12 | 0.0200 | 0.0180 | 0.0180 | 0.0068 |
| 16 | 0.0138 | 0.0125 | 0.0109 | 0.0044 |
| 20 | 0.0119 | 0.0121 | 0.0097 | 0.0034 |

Table 1: Simulated global FWER of the test on equivalence (differences, two-sided) for several numbers of endpoints, and equicorrelations; $\alpha = 0.05$.

| Endpoints | Correlations | | | |
|---|---|---|---|---|
| | $\rho^*$ | 0 | 0.5 | 1 |
| 4 | 0.0295 | 0.0296 | 0.0313 | 0.0223 |
| 8 | 0.0196 | 0.0217 | 0.0219 | 0.0123 |
| 12 | 0.0145 | 0.0152 | 0.0169 | 0.0068 |
| 16 | 0.0126 | 0.0135 | 0.0122 | 0.0030 |
| 20 | 0.0095 | 0.0092 | 0.0094 | 0.0025 |

Table 2: Simulated local FWER of the test on equivalence (differences, two-sided) for several numbers of endpoints, and equicorrelations; $\alpha = 0.05$.

| Endpoints | Correlations | | | |
|---|---|---|---|---|
| | $\rho^*$ | 0 | 0.5 | 1 |
| 4 | 0.0450 | 0.0462 | 0.0441 | 0.0353 |
| 8 | 0.0343 | 0.0335 | 0.0326 | 0.0171 |
| 12 | 0.0378 | 0.0342 | 0.0310 | 0.0107 |
| 16 | 0.0303 | 0.0271 | 0.0230 | 0.0074 |
| 20 | 0.0255 | 0.0259 | 0.0228 | 0.0065 |

Table 3: Simulated global FWER of the test on equivalence (ratios, two-sided) for several numbers of endpoints, and equicorrelations; $\alpha = 0.05$.

| Endpoints | Correlations | | | |
|---|---|---|---|---|
| | $\rho^*$ | 0 | 0.5 | 1 |
| 4 | 0.0353 | 0.0359 | 0.0351 | 0.0246 |
| 8 | 0.0295 | 0.0321 | 0.0302 | 0.0152 |
| 12 | 0.0283 | 0.0263 | 0.0264 | 0.0109 |
| 16 | 0.0268 | 0.0280 | 0.0233 | 0.0058 |
| 20 | 0.0244 | 0.0257 | 0.0217 | 0.0056 |

Table 4: Simulated local FWER of the test on equivalence (ratios, two-sided) for several numbers of endpoints, and equicorrelations; $\alpha = 0.05$.

# References

[Quan et al. (2001)] Quan H, Bolognese J, and Yuan W (2001). Assessment of equivalence on multiple endpoints. Statistics in Medicine, 20: 3159-3173.