Journal of Biopharmaceutical Statistics, 16: 711–731, 2006

ISSN: 1054-3406 print/1520-5711 online
DOI: 10.1080/10543400600860576

Taylor & Francis
Taylor & Francis Group

MULTIPLE COMPARISONS AND MULTIPLE CONTRASTS IN RANDOMIZED DOSE-RESPONSE TRIALS—CONFIDENCE INTERVAL ORIENTED APPROACHES

Ludwig A. Hothorn
Institute for Biostatistics, Leibniz University of Hannover, Hannover, Germany

According to the ICH E9 recommendation, the evaluation of randomized dose-finding trials focuses on the graphical presentation of different kinds of simultaneous confidence intervals:* i) *superiority of at least one dose vs. placebo with and without the assumption of order restriction,* ii) *noninferiority of at least one dose vs. active control,* iii) *identification of the minimum effective dose,* iv) *identification of the peak dose,* v) *identification of the maximum safe dose for a safety endpoint, and* vi) *estimation of simultaneous confidence intervals for "many-to-one-by-condition interaction contrasts." Moreover, global tests for a monotone trend or a trend with a possible downturn effect are discussed. The basic approach involved obtaining multiple contrasts for different problem-related contrast definitions. For all approaches, definitions of relevance margins for superiority or noninferiority are needed. Because consensus on margins only exists for selected therapeutic areas and the definition of absolute thresholds may be difficult, simultaneous confidence intervals for ratio to placebo were also used. All approaches are demonstrated in an example-based manner using the R-packages multcomp (difference), for hypotheses based on difference, and mratios (ratio), for hypotheses based on ratios.

Key Words: Multiple contrasts; Randomized dose-finding trials; Simultaneous confidence intervals; Trend test.

1. INTRODUCTION

According to the ICH E4 guideline, the design of dose-finding trials can be classified into titration, crossover and parallel groups. Only the randomized parallel group design including a placebo or/and an active control (reference drug) and some (e.g., two to four) fixed dose groups will be considered here: $[P, D_1, \ldots, D_k, A]$. Manifold objectives can be intended:

i) the proof of efficacy by demonstrating a global dose-response relationship and investigation of the shape and location of the dose-response curve according to ICH E9 (U.S. Food and Drug Administration, 1998);
ii) the investigation of the shape and location of the dose-response curve according to ICH E9 (U.S. Food and Drug Administration, 1998);

Received September 2005; Accepted December 2005
Address correspondence to Ludwig A. Hothorn, Institut fuer Biostatistik, Leibniz University of Hannover, Herrenhaeuser Str. 2, D-30419 Hannover, Germany; Fax: +49 511 762 4966; E-mail: hothorn@biostat.uni-hannover.de

iii) the determination of a maximal dose beyond which additional benefit would be unlikely to occur according to ICH E9 (U.S. Food and Drug Administration, 1998);
iv) the identification of the smallest dose with a discernible useful effect according to ICH E4 (Center for Drug Evaluation and Research, 1994), that is, the minimum effective dose (Committee for Proprietary Medicinal Products, 2002);
v) the relationship of drug dosage to clinical beneficial or undesirable effects (Center for Drug Evaluation and Research, 1994);
vi) finding a dose that is noninferior to the active control (Committee for Proprietary Medicinal Products, 2002); and so on.

Two important recommendations for the choice of evaluation were formulated: i) "the applications of procedures to estimate the relationship between dose and response, including the construction of confidence intervals and the use of graphical methods, is as important as the use of statistical tests" (ICH E9) and, in the CPMP (Committee for Proprietary Medicinal Products, 2002) recommendation, ii) "for therapeutic dose response studies that aim at identifying one or several doses of an investigational drug for its recommended use in a specific patient population, the control of the family-wise type I error in a strong sense is mandatory (CDMP)." Therefore, this paper focuses on procedures for simultaneous confidence intervals, which control the family-wise type I error, and related graphical presentations to demonstrate the dose-response relationship for particular purposes:

i) superiority of at least one dose vs. placebo without and with the assumption of order restriction,
ii) noninferiority of at least one dose vs. active control,
iii) identification of the minimum effective dose (MED),
iv) identification of the peak dose (PD) (i.e., the maximal dose beyond which additional benefit would be unlikely to occur),
v) identification of the maximum safe dose for a safety endpoint (MAXSD),
vi) estimation of simultaneous confidence intervals for "many-to-one-by-condition interaction contrasts,"
vii) global tests for monotone trend or trend with a possible downturn effect.

The methodological focus is on multiple contrast tests and procedures that include multiple comparison procedures. The different approaches will be explained by means of simple data examples evaluated by *R* programs, version 2.0.1 (*R* Development Core Team, 2004) mainly based on the *R*-packages multcomp (Bretz et al., 2003) and mratios (Dilba and Schaarschmidt, 2005) available via CRAN. A similar SAS macro for simultaneous confidence intervals for differences is available, but not demonstrated explicitly here (Westfall et al., 1999).

The paper is organized as follows. First the demonstration of superiority of at least one dose for the difference or ratio to placebo without the assumption of order restriction is described, followed by the analogous demonstration of noninferiority. Hereby will be the simultaneous demonstration of superiority of at least one dose vs. placebo and noninferiority of at least one dose vs. active control discussed only briefly because up to now only marginal confidence intervals are available. The second topic is testing a global trend by multiple contrast tests under monotonicity assumption, including possible downturns at higher doses. The same

technique of multiple contrasts will be used for simultaneous confidence intervals for the demonstration of superiority of at least one dose vs. placebo with the order restriction. In a third topic, confidence interval–based approaches for identification of the minimal effective dose, the peak dose and the maximal safe dose for a safety endpoint are described. Finally, simultaneous confidence intervals for interaction contrasts for dose-finding trials with a secondary factor will be demonstrated.

2. DEMONSTRATION OF SUPERIORITY OF AT LEAST ONE DOSE VS. PLACEBO WITHOUT ORDER RESTRICTION

A randomized one-way layout $[P, D_1, \ldots, D_k]$ (where k is commonly small, often between 2 and 4, in real trials) with approximately normally distributed single primary efficacy endpoint y_{ij} $(i \in (P, 1, \ldots, k); j = 1, \ldots, n_i)$ and homogeneous variances are assumed. The objective is to demonstrate the superiority of at least one dose in comparison to placebo. First, a procedure for simultaneous confidence intervals without order restriction will be proposed. The disadvantage of such a procedure is smaller power and no direct possibility of claiming "dose-response relationship"; the advantage is the simplicity and the robustness against possible nonmonotonicity at high(er) dose(s). Either simultaneous confidence intervals for the difference or the ratio to placebo can be used. The characteristic of the difference approach is its scale-variant interpretation in the clinical measurement unit; that of the ratio approach is its scale-invariant interpretation in percentage change. Therefore, the kind of interpretation should primarily determine the use of either the difference or the ratio approach. Unfortunately, in clinical trials literature, nearly only the difference approach is used. Three other criteria for the choice between the difference and the ratio approaches exist: i) if placebo's mean becomes too small (in combination with high variance and small sample size), the estimation of multiple ratios becomes unstable; ii) the ratio approach is restricted to positive values; and iii) power is larger for some configurations (Dilba et al., 2006b).

The interpretation of the confidence intervals due to superiority vs. placebo is simple: for endpoints where increasing values are clinically beneficial, the lower one-sided limits of the several doses should be larger than 0 (difference approach) or larger than 1 (ratio approach); for endpoints where decreasing values are clinically beneficial, the upper one-sided limits of the several doses should be smaller than 0 (for the difference) or smaller than 1 (for the ratio). For the evaluation of superiority in clinical trials, the use of one-sided 97.5% confidence limits or related two-sided 95% confidence limits are common.

2.1. Simultaneous Confidence Interval for the Difference

The parametric simultaneous one-sided lower (upper) confidence limits for the difference of i dose groups vs. placebo can be estimated using the well-known Dunnett (1955) procedure:

$$\left[\bar{y}_i - \bar{y}_P - St_{k,df,R,1-\alpha}\sqrt{1/n_i + 1/n_P}; \infty\right]$$

where S denotes the root of the common variance estimator, $t_{k,df,R,1-\alpha}$ denotes the upper $(1-\alpha)$ quantile of the multivariate t-distribution with common degree of

freedom df, k denotes the number of dose groups, and R denotes the correlation matrix with the elements for two contrasts with coefficients a_i and b_i:

$$\rho_{a,b} = \sum_{i=P}^{k} a_i b_i / n_i \Big/ \sqrt{\left(\sum_{i=P}^{k} a_i^2 / n_i\right)\left(\sum_{i=P}^{k} b_i^2 / n_i\right)}.$$

An Example. In a dose-response trial of an angina drug with the endpoint change from pretreatment of pain-free walking in minutes (Westfall et al., 1999, p. 164), four dose groups are compared with a placebo (abbreviated as 0). Larger values are clinically beneficial, and approximate Gaussian distribution as well as variance homogeneity will be assumed, which seems to be acceptable (see the boxplots in Fig. 1a).

The *R*-code for the estimation of simultaneous confidence limits is quite simple and we use the function simint.

```
library(multcomp)
simint(response~dose, data = angina, alternative = "greater",
       type = "Dunnett", conf.level = 0.975)
```

From Fig. 1b it can be seen that only the lower limits for (dose3–dose0) and (dose4–dose0) are larger than zero whereby only the lower limit of the highest dose seems, with 7.1 min of pain-free walking, to be clinically superior. From the

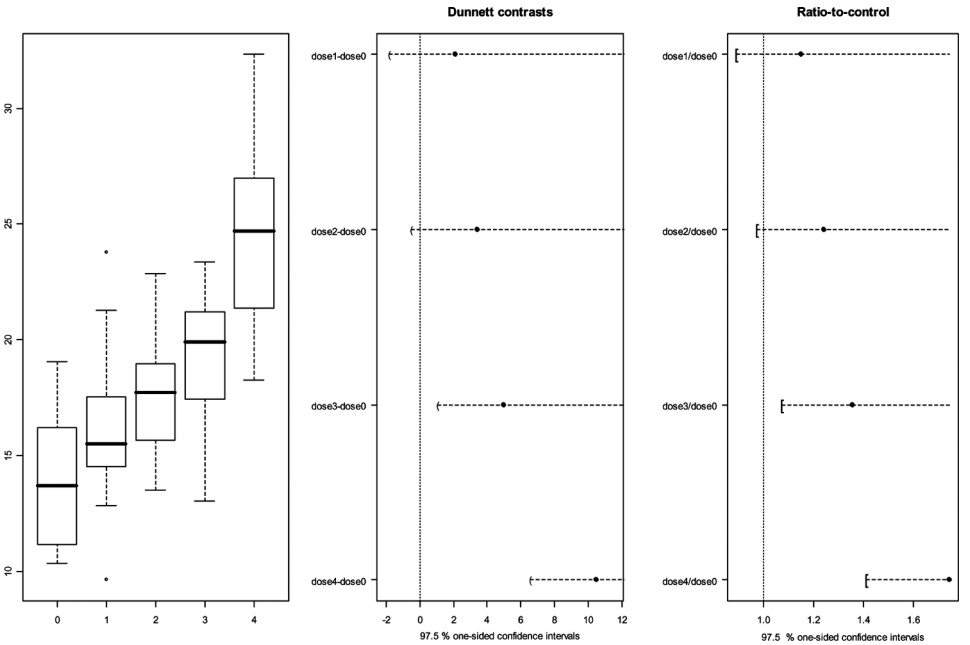


Figure 1a,b,c Boxplots and simultaneous confidence intervals for the difference and ratio to placebo for the angina trial.

combination of the boxplots and the simultaneous confidence intervals in Fig. 1a and Fig. 1b, a monotone dose-response relationship (without inferential reasoning) can be nicely interpreted.

2.2. Simultaneous Confidence Intervals for Ratios

The parametric simultaneous one-sided lower confidence limits for ratios of i dose groups vs. placebo can be estimated for the multiple ratios $\gamma_i = \mu_i/\mu_P$ by

$$\left[\frac{\hat{\gamma}_i - \sqrt{g[\hat{\gamma}_i^2 + (1-g)n_P/n_i]}}{1-g}; \infty\right]$$

(Dilba et al., 2004) with

$$g = S^2 t_{k,df,R,1-\alpha}^2 / n_P \bar{y}_P^2.$$

However, since the elements of the correlation matrix R for two contrasts a and b,

$$\rho_{a,b} = \gamma_a \gamma_b \Big/ \sqrt{(\gamma_a^2 + n_P/n_a)(\gamma_b^2 + n_P/n_b)},$$

depend on the a priori unknown γ_i, we can use a simple plug-in approach where the unknown ratios are estimated by $\hat{\gamma}_i = \bar{y}_i/\bar{y}_P$. According to Hauschke and Kieser (2001), this approach is superior to the conservative but simple Bonferroni approach as long as $\bar{y}_i\sqrt{n_i}/\bar{y}_P\sqrt{n_P} > 1$ (see Dilba et al., 2006a). The one-sided lower simultaneous limits for the angina data are presented in Fig. 1c. Using the library(mratios) function (sci.ratio), the *R*-code is quite simple to use. With both the simint and sci.ratio functions, an object can be defined and the confidence limits can simply be plotted using the plot(object) function.

```
CIrat = sci.ratio(response~dose, data = angina, alternative =
                  "greater", conf.level = 0.975, control.T =
                  "0", method = "Plug")
```

The confidence limits for the difference and ratio to placebo in Fig. 1b and Fig. 1c show a similar pattern; the major difference is that, for the comparison of the highest dose vs. placebo in Fig. 1b, the distance to 0 is 7.1 min pain-free walking, while, in Fig. 1c, the distance to 100% is 141%. A further argument to prefer the ratio approach is higher power in comparison with the difference approach for relative margins smaller than 1: the proof of superiority when smaller endpoints are beneficial and the proof of noninferiority when larger endpoints are beneficial (for details, see Dilba, 2005).

3. DEMONSTRATION OF NONINFERIORITY OF AT LEAST ONE DOSE VS. ACTIVE CONTROL WITHOUT ORDER RESTRICTION

In a randomized one-way layout $[D_1, \ldots, D_k, A]$, several doses are compared vs. an active control A on at least noninferiority without order restriction. For an increasing endpoint, a dose group i is noninferior if the $(1-\alpha)$ upper confidence

limit for the difference $(\mu_A - \mu_i)$ is smaller than a noninferiority margin φ (with $\varphi < 1$) or, for the ratio μ_A/μ_i, is smaller than ψ (with $\psi < 1$). For a decreasing endpoint, the reverse holds true. Therefore the distinction between proof on superiority and noninferiority is simple. For an increasing endpoint: either using the lower one-sided simultaneous confidence limit for $(\bar{y}_i - \bar{y}_A)$ with the decision on noninferiority if the lower limit is larger than φ (with $\varphi < 1$) and/or the decision on superiority if the lower limit is larger than 1 or an additionally predefined superiority margin; or using the upper confidence limit for $(\bar{y}_A - \bar{y}_i)$ with the decision on noninferiority if the upper limit is smaller than φ' (with $\varphi' < 1$) and/or the decision on superiority if the upper limit is smaller than 1 or a predefined superiority margin (analogously for the ratio).

An Example. A single, a double, and a quadruple administration of the same dose of a new drug are compared with two active drugs according to cholesterol reduction Westfall et al. (1999, p. 153), where smaller values are clinically beneficial and approximate Gaussian distribution as well as variance homogeneity will be assumed (see the boxplots in Fig. 2a) (to keep the example simple, the second active drug E was ignored in the original data).

Noninferiority can be concluded whether, for this decreasing endpoint, the $(1 - \alpha)$ lower confidence limit for $(\mu_A - \mu_i)$ or μ_A/μ_i is larger than the noninferiority margins. The *R*-function simint calculates the interval for many-to-one comparisons $(\mu_i - \mu_A)$ (using option base = 4 to select drug D as a control) but the lower one-sided interval for $(\mu_A - \mu_i)$ is needed; therefore, the one-sided upper interval for $(\mu_i - \mu_A)$ was chosen. The data set cholesterol was modified into mychol with problem-adequate factor labels. It should be noted that, for the noninferiority problem, one-sided 95% confidence levels are appropriate.

```
data(cholesterol)
mychol = cholesterol[1:40,] # ignore drug E
mychol$trt = factor(trt, labels = c("one", "double",
                    "quad", "active"))
CIchol = simint(response~trt, data = mychol, alternative =
                "less", type = "Dunnett", conf.level = 0.95,
                base = 4)
```

From Fig. 2b, it can be derived that single and double administration are not only noninferior but also superior because the upper bounds are negative. The quadruple administration of the same dose is noninferior because 0.06 is very small relative to a not defined noninferiority margin for cholesterol reduction.

The *R*-function sci.ratio calculates the interval for many-to-one comparisons μ_i/μ_A, and again the 95% one-sided upper intervals were calculated for claiming noninferiority vs. the active control.

```
CIrchol = sci.ratio(response~trt, data = mychol,
                    alternative = "less", conf.level = 0.95,
                    control.T = "active")
```

Analogously, it can be derived from Fig. 2c that single and double administration are not only noninferior but also superior because the upper limits

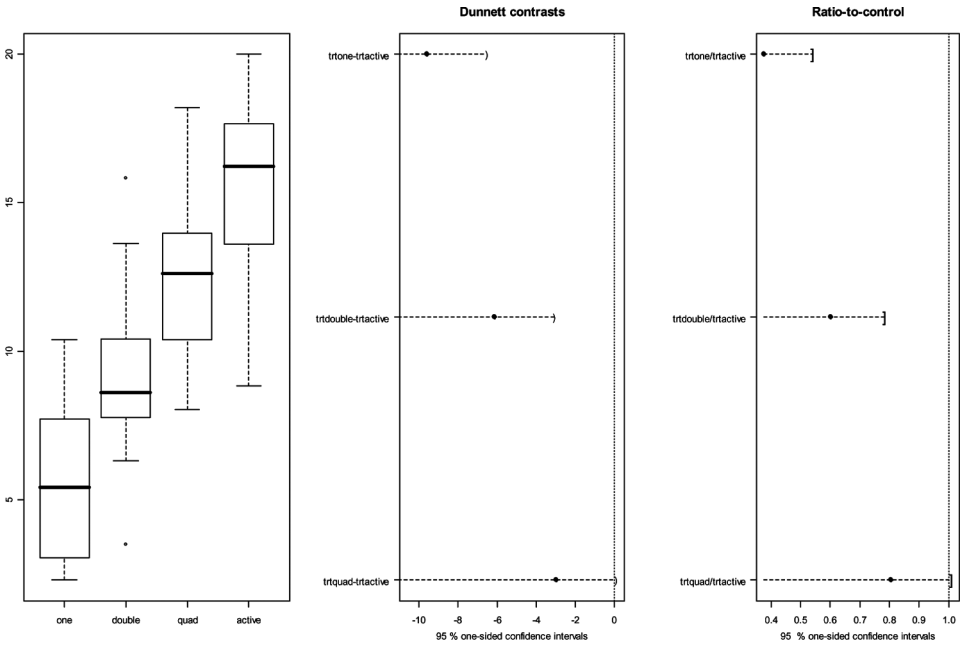


Figure 2a,b,c Boxplot for a noninferiority trial and simultaneous confidence limits for the difference and the ratio.

are clearly less than 1, and the quadruple administration of the same dose can be assumed to be noninferior because the limit of 1.01 is near 1.

4. SIMULTANEOUS DEMONSTRATION OF SUPERIORITY OF AT LEAST ONE DOSE VS. PLACEBO AND NONINFERIORITY OF AT LEAST ONE DOSE VS. ACTIVE CONTROL

In principle, a therapeutic window should be estimated; it should be lower bounded by the minimum effective dose for the single primary efficacy endpoint and upper bounded by the maximum safe dose for the single safety endpoint. Related individual approaches for the identification of the MED and the MAXSD will be described in Sections 8 and 10. In a few clinical dose-finding trials a single safety endpoint is a priori known, and a dose is defined within the therapeutic window if it is both superior with respect to placebo and (at least) noninferior with respect to active control. A related decision-based approach using the intersection-union test principle was published (Bauer et al., 1998) but up to now no confidence interval–oriented approach is available.

5. TESTING THE GLOBAL DOSE-RESPONSE RELATIONSHIP ASSUMING MONOTONICITY—AN INTRODUCTION TO MULTIPLE CONTRAST TESTS

A basic objective of the evaluation of a dose-finding trial is to demonstrate a global trend. Testing the global dose-response relationship under a monotonicity

assumption can be performed by an order-restricted test. Many related proposals can be found in the literature, but only few trend tests are sensitive to any shape of the dose-response relationship, which is clearly a priori unknown—it is an outcome of the clinical trial. No uniformly most powerful trend test exists. If the shape is linear, the most powerful test is a linear trend test; if the shape is concave, the most powerful test is the Helmert contrast test. The likelihood-ratio test under total order restriction (Bartholomew, 1961) provides the highest "average" power among the present trend tests. This statistic can be seen as an ANOVA-F-test analogue under total order restriction for the ratio of the between groups sum of squares for the maximum likelihood estimates $\hat{\mu}_i$ under simple order restriction and the total sum of squares:

$$\overline{E}_k^2 = \sum_{i=P}^{k} n_i \left(\hat{\mu}_i - \bar{\bar{y}}\right)^2 \Bigg/ \sum_{i=P}^{k} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{\bar{y}}\right)^2,$$

where $\bar{\bar{y}} = \sum_{i=P}^{k} n_i \bar{y}_i / N$ denotes the total mean. This test was rarely used in the analysis of clinical dose-finding trials because it is difficult to evaluate even its null distribution (for unbalanced designs, a related SAS/IML program is available Bretz and Seidel, 2000).

A simple nearly "average" powerful alternative is the multiple contrast test. The global alternative, $H_1^{\text{global}} : \mu_P \leq \mu_1 \leq \cdots \leq \mu_k$, can be decomposed in $(2^k - 1)$ elementary alternatives with particular patterns of equalities and inequalities. For example, for $k = 2$, three elementary alternatives exist: a concave $[\mu_P = \mu_1 < \mu_2]$ alternative, a convex $[\mu_P < \mu_1 = \mu_2]$ alternative, and a linear $[\mu_P < \mu_1 < \mu_2]$ alternative with related power-optimal specific contrasts $C_{\text{concave}} = -\bar{y}_P - \bar{y}_1 + 2\bar{y}_2$, $C_{\text{convex}} = -2\bar{y}_P + \bar{y}_1 + \bar{y}_2$, and $C_{\text{linear}} = -\bar{y}_P + 0 * \bar{y}_1 + \bar{y}_2$. For such a single contrast, a contrast test can be derived by studentized standardization

$$T^{\text{single contrast}} = \sum_{i=P}^{k} c_i \bar{y}_i \Bigg/ S \sqrt{\sum_{i=P}^{k} c_i^2 / n_i},$$

which is univariate t-distributed, whereby the condition $\sum_i c_i = 0$ guarantees the level α under the null hypothesis. Only one selected individual alternative can be valid, and therefore only one single contrast test has the maximum value, which yields to a maximum test $T^{\text{multiple contrast}} = \max\left\{T_1^{\text{single contrast}}, K, T_q^{\text{single contrast}}\right\}$ based on q single-contrast tests. The joint distribution of the $T_l^{\text{single contrast}}$ is a central q-variate t-distribution with common degrees of freedom df and the correlation matrix R under the global null hypothesis H_0. The elements of R consist of correlations between each two of the q contrast vectors; for example, for two contrasts with coefficients a_i and b_i, the correlation is $\rho_{a,b}$ (see Section 2.1). The null hypothesis, H_0, is therefore rejected if and only if $T^{\text{multiple contrast}}$ exceeds a critical value $c_{q,R,df,1-\alpha}$. Under the global alternative H_1^{global}, the $T_l^{\text{single contrast}}$ follow jointly a noncentral multivariate t-distribution with the parameters R, df, and the noncentrality parameter

$$\zeta = \left\{ \sum_{i=P}^{k} c_{ij} \mu_i \Bigg/ \sigma \sqrt{\sum_{i=P}^{k} c_i^2 / n_i} \right\} \quad (1 \leq i \leq q)$$

Table 1 Selected contrast coefficients for the balanced design with three dosage groups and a placebo under order restriction

Type of contrasts	Number of contrasts	Alternative	Contrast c
Isotonic	$2^k - 1$	$\mu_P < \mu_1 = \mu_2 = \mu_3$	{−3 1 1 1}
Bretz (1999)		$\mu_P < \mu_1 = \mu_2 < \mu_3$	{−1 −1 1 1}
		$\mu_P = \mu_1 = \mu_2 < \mu_3$	{−1 −1 −1 3}
		$\mu_P < \mu_1 < \mu_2 < \mu_3$	{−3 −1 1 3}
		$\mu_P = \mu_1 < \mu_2 < \mu_3$	{−1 −1 0 2}
		$\mu_P < \mu_1 = \mu_2 < \mu_3$	{−1 0 0 1}
		$\mu_P < \mu_1 < \mu_2 = \mu_3$	{−2 0 1 1}
Marcus	$k(k + 1)/2$	$\mu_P < \mu_1 = \mu_2 = \mu_3$	{−3 1 1 1}
(Marcus, 1976)		$\mu_P < \mu_1 < \mu_2 = \mu_3$	{−2 0 1 1}
		$\mu_P < \mu_1 = \mu_2 < \mu_3$	{−1 −1 1 1}
		$\mu_P < \mu_1 = \mu_2 < \mu_3$	{−1 0 0 1}
		$\mu_P = \mu_1 < \mu_2 < \mu_3$	{−1 −1 0 2}
		$\mu_P = \mu_1 = \mu_2 < \mu_3$	{−1 −1 −1 3}
Change point	$k - 1$	$\mu_P < \mu_1 = \mu_2 = \mu_3$	{−3 1 1 1}
(Hirotsu, 1997)		$\mu_P = \mu_1 < \mu_2 = \mu_3$	{−1 −1 1 1}
		$\mu_P = \mu_1 = \mu_2 < \mu_3$	{−1 −1 −1 3}
Up/down	2	$\mu_P < \mu_1 = \mu_2 = \mu_3$	{−3 1 1 1}
(Neuhaeuser and Hothorn, 1997; Stewart and Ruberg, 2000)		$\mu_P = \mu_1 = \mu_2 < \mu_3$	{−1 −1 −1 3}
Single linear	1	$\mu_P < \mu_1 < \mu_2 < \mu_3$	{−3 −1 1 3}

(Bretz and Hothorn, 2003). Several proposals for the choice of contrasts were published; see Table 1 for a selection of $k = 3$ multiple contrasts.

A clear distinction between a trend test and simultaneous confidence intervals, even for the same contrast coefficients, is that the trend test is a global test with a decision for or against the null hypothesis or an adjusted p-value, and the confidence intervals estimate individual contrasts simultaneously:

$$\sum_{i=P}^{k} c_{ij}\mu_i \in \left[\sum_{i=P}^{k} c_{ij}\bar{y}_i - c_{q,R,df,1-\alpha} S \sqrt{\sum_{i=P}^{k} c_i^2/n_i}, \infty\right).$$

6. TESTING A GLOBAL DOSE-RESPONSE RELATIONSHIP WITH POSSIBLE NONMONOTONICITY AT HIGHER DOSES

Assuming a monotone-ordered alternative in the case of real data with a clear downturn effect at high doses may lead to a seriously biased result. Sometimes, we are interested in testing exactly those particular nonmonotonic trends with known or even unknown peak points (see Marcus and Genizi, 1994) or we want to test the monotonic part only (ignoring the downturn part) using the so-called downturn-protected trend test (Bretz and Hothorn, 2001). Hereby, one alternative from k elementary alternatives with the peak point at the ith dose is possible. Therefore, we take the maximum over these k alternatives, and, within each an appropriate

multiple contrast for any elementary shape, we find that

$$T^{\text{protected test}} = \max_{i=1,\ldots,k} \max_{l=1,\ldots,q} \left\{ T_{il}^{\text{single contrast}} \right\}.$$

For example, for $k = 3$ and elementary change point alternatives, the contrast matrix is

$\bar{y}_P$	$\bar{y}_1$	$\bar{y}_2$	$\bar{y}_3$
-3	1	1	1
-1	-1	1	1
-1	-1	-1	3
-2	1	1	0
-1	-1	2	0
-1	-1	0	0

An Example. Extreme nonmonotonic Ames mutagenicity data were published by Combes (1997); see the boxplot in Fig. 3a for the number of revertants for a control (d0) and five doses (d1 through d5).

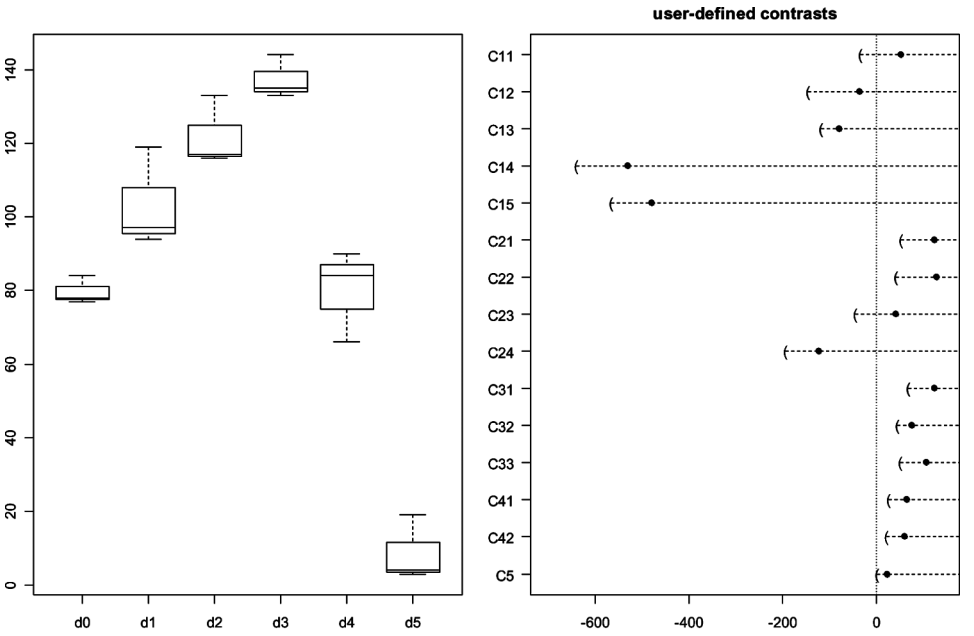
The *R*-code uses simint, where user-defined contrasts for downturn-protected multiple change-point contrasts are needed.

```
C11 = c(-5, 1, 1, 1, 1, 1)
C12 = c(-4, -4, 2, 2, 2, 2)
C13 = c(-1, -1, -1, 1, 1, 1)
C14 = c(-2, -2, -2, -2, 4, 4)
C15 = c(-1, -1, -1, -1, -1, 5)
C21 = c(-4, 1, 1, 1, 1, 0)
C22 = c(-3, -3, 2, 2, 2, 0)
C23 = c(-2, -2, -2, 3, 3, 0)
C24 = c(-1, -1, -1, -1, 4, 0)
C31 = c(-3, 1, 1, 1, 0, 0)
C32 = c(-1, -1, 1, 1, 0, 0)
C33 = c(-1, -1, -1, 3, 0, 0)
C41 = c(-2, 1, 1, 0, 0, 0)
C42 = c(-1, -1, 2, 0, 0, 0)
C5 = c(-1, 1, 0, 0, 0, 0)

Cdown = rbind(C11, C12, C13, C14, C15, C21, C22, C23, C24, C31,
              C32, C33, C41, C42, C5)

CIdown = simint(rev~typ, data = ames, alternative = "greater",
                cmatrix = Cdown, conf.level = 0.95)
```

A monotonic trend test is not significant (e.g., the adjusted *p*-value based on Marcus contrasts is 0.180), but downturn-protected confidence intervals based on step-contrast show an interesting pattern (see Fig. 3b). The largest distance of the lower limit of a simultaneous confidence interval reveals contrast C31 (i.e., for the



(a) (b)

Figure 3a,b Boxplots for the number of revertants and simultaneous intervals for possibly nonmonotonic contrasts.

peak point d3, the Helmert contrast indicates a monotone concave trend up to d3) (Hothorn, 2004).

7. DEMONSTRATION OF THE SUPERIORITY OF AT LEAST ONE DOSE VS. PLACEBO WITH THE ORDER RESTRICTION ASSUMPTION

In many clinical dose-finding trials, we can a priori strongly assume that, with increasing doses, the endpoint increases monotonically (or, alternatively, decreases monotonically); that is, the specific restricted alternative can be formulated as $\mu_P \leq \mu_1 \leq \cdots \leq \mu_k$ (where at least $\mu_P < \mu_k$ holds true). Global trend tests were described in Section 5, but only some procedures for simultaneous confidence intervals based on Williams (1972), Marcus (1976), and change-point contrasts (Hirotsu, 1997) exist (for contrast definitions, see Table 1 in Section 5). Unlike in the global trend test, a global adjusted p-value will not be estimated; instead, individual contrasts will be estimated simultaneously (see also Channon and McEntegart, 2001). The contrasts for the control and the four dose groups are given explicitly in Table 2.

An Example. For the angina trial, the simultaneous confidence intervals for the Williams, Marcus, and Hirotsu approaches are presented in Fig. 4.

The *R*-code for these multiple contrasts is simple using the *R*-function simint.

```
CIWil = simint(response~dose, data = angina, alternative =
                "greater", type = "Williams", conf.level = 0.975)
```

Table 2 Balanced Williams contrasts and change-point contrasts for the angina data example

	Williams contrast					Change-point contrast				
	μ_P	μ_1	μ_2	μ_3	μ_4	μ_P	μ_1	μ_2	μ_3	μ_4
C1	−1	0	0	0	1	−4	1	1	1	1
C2	−2	0	0	1	1	−3	−3	2	2	2
C3	−3	0	1	1	1	−2	−2	−2	3	3
C4	−4	1	1	1	1	−1	−1	−1	−1	4

```
CIMarc = simint(response~dose, data = angina, alternative =
                "greater", type = "Marcus", conf.level = 0.975)
CIHiro = simint(response~dose, data = angina, alternative =
                 "greater", type = "Changepoint",
                 conf.level = 0.975)
```

Because at least one lower confidence limit is larger than 0 (for either contrast method), a global dose-response trend exists. In the change-point approach, the lower limit for contrast $C4$ is 4.8 for the mean of all lower doses (including the zero dose placebo) minus the mean of the highest dose; this indicates the existence of a remarkable change point between dose 3 and dose 4 (see the related boxplot in Fig. 1a in Section 3).

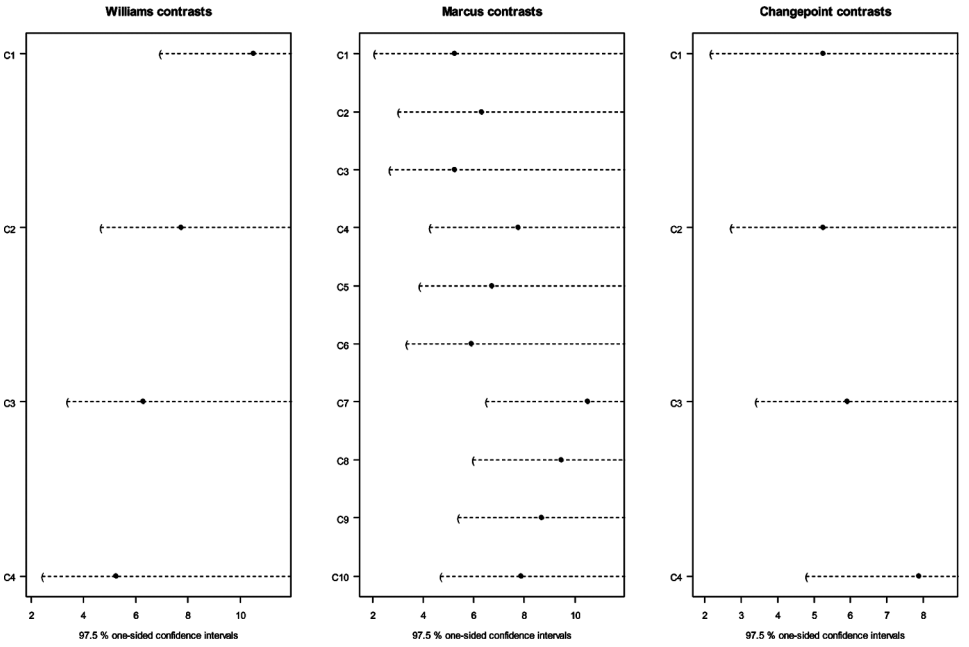


Figure 4a,b,c Simultaneous confidence intervals for the three types of order-restricted multiple contrasts for the angina trial.

8. IDENTIFICATION OF THE MINIMAL EFFECTIVE DOSE

Both the ICH E4 (Center for Drug Evaluation and Research, 1994) guideline and the CPMP (Committee for Proprietary Medicinal Products, 2002) paper recommend the identification of the minimal effective dose (MED) in randomized clinical dose-finding trials. This is the smallest dose with a significant increase (decrease) with respect to the placebo where all higher doses must be significant to the placebo, too. Several papers exist on the identification of the MED (e.g., Jan and Chen, 2004; Tamhane et al., 1996) using contrasts or order-restricted tests. However, all these approaches are decision approaches only; simultaneous confidence intervals are not available. Recently, Bretz et al. (2003) published simultaneous confidence intervals for the identification of the MED based on multiple ratios to control which interval presents a generalization of the difference approach of Hsu and Berger (1999). These approaches inherently need the a priori definition of a superiority margin θ (respective δ) (i.e., the smallest dose that is not only significantly but also clinically relevant and different from the placebo will be estimated). The approach is based on the assumption that all higher doses must be significantly different from the placebo and the partitioning principle (Finner and Strassburger, 2002). First, the marginal (unadjusted) one-sided limits for the difference or the ratio to placebo are calculated, commonly for clinical trials at the 2.5% level. For MED identification using simultaneous confidence limits, a top-down stepwise decision is performed where the sequence stops when the lower marginal limit is smaller than θ (respective δ) (for increasing endpoints, or when the upper limit is larger than the margins for decreasing endpoints). In that dose, the simultaneous limit is equal to the marginal limit. As long as the decisions are significant, the simultaneous limit is equal to θ (respective δ), but not if all doses were significant; here, all limits are equal to the minimum marginal limit. For the angina trial in Table 3, the raw confidence intervals for both the difference and the ratio to placebo and their simultaneous intervals are presented. Let us assume that $\theta = 5.0$ min pain-free walking and $\delta = 120\%$.

A clear disadvantage of this approach is that, for all doses higher than MED (and for the MED itself), the confidence limits are noninformative because they are equal to the superiority margin. A definition of a superiority margin is only available for a few therapeutic areas. In trials without such a definition, we can "play" with empirically chosen margins (e.g., one just in the superiority space $\delta = 101\%$, the other irrelevantly smaller than the marginal limits). Table 4 demonstrates this approach for ratio to placebo (0) comparisons for "informative" MED confidence limits (where the MED is marked bold).

Table 3 Identification of the MED by simultaneous confidence intervals

	Difference to placebo			Ratio to placebo		
Comparison	$CI^{\text{difference,lower}}_{\text{marginal},i,1-\alpha}$	$CI^{\text{difference,lower}}_{\text{simultan.},1-\alpha}$	Decision	$CI^{\text{ratio,lower}}_{\text{marginal},i,1-\alpha}$	$CI^{\text{ratio,lower}}_{\text{simultan.},1-\alpha}$	Decision
"1 vs. P"	−1.02			0.97		
"2 vs. P"	0.28	Stop		1.05	Stop	
"3 vs. P"	1.88	1.88		1.15	1.15	
"4 vs. P"	7.38	5.0	MED	1.51	1.20	MED

Table 4 Simultaneous lower confidence limits for the ratio to placebo for several chosen superiority margins δ

Comparison	$CI^{\text{difference,lower}}_{\text{marginal},i,1-\alpha}$	$\delta = 100.1\%$	$\delta = 104.99\%$	$\delta = 114.99\%$	$\delta = 150.99\%$
"1 vs. P"	0.97	0.97	0.97		
"2 vs. P"	1.05	**1.001**	**1.0499**	1.05	
"3 vs. P"	1.15	1.001	1.0499	**1.1499**	1.12
"4 vs. P"	1.51	1.001	1.0499	1.1499	**1.5099**

This approach estimates dose 4 as the MED for an empirical superiority threshold of 1.5099, dose 3 as the MED for an empirical superiority for a threshold of 1.1499, dose 2 as the MED for an empirical superiority for a threshold of 1.0499, but it never estimates dose 1 as the MED. Table 4 may be easier to interpret than the original approach presented in Table 3.

9. IDENTIFICATION OF THE PEAK DOSE

One objective of randomized dose-finding trials is to determine the maximal dose beyond which additional benefits would be unlikely to occur (U.S. Food and Drug Administration, 1998) to avoid overdosing, which is here denoted as the peak dose (PD). Inferential decision procedures based on point-zero null hypotheses can not be recommended; because clinical relevance should be taken into account, confidence intervals should be used. Rarely, a priori relevance margins are known and therefore, a posteriori, the estimates for the limits of simple marginal confidence intervals can be used as clinical relevance measures. In particular, information on both the MED and the PD should be obtained. Hereby either confidence intervals for the difference or for the ratio can be used primarily depending on the appropriateness of a scale-variant or scale-invariant interpretation. Because the MED identification is based on the principle of the intersection-union test (i.e., on level α intervals), the PD is based on the union-intersection test principle but with a negative correlation; that is, simultaneous intervals based simply on level $\alpha/(k-1)$ marginal intervals can be used (Hsu and Berger, 1999; Liu et al., 2000).

For simplicity, we assume a single efficacy endpoint where higher values are more efficient and there are monotonically increasing expected values. The MED is the smallest dose with a relevant increasing effect over the placebo, so the lower $(1-\alpha)$ confidence limits for $(\mu_i - \mu_P)$ or μ_i/μ_P should be plotted. Under the monotonicity assumption, the PD is the maximum relevant increment, so the lower $1-\alpha/(k-1)$ confidence limits for $(\mu_i - \mu_{i-1})$ or μ_i/μ_{i-1} $i = 1, \ldots, k$ should be plotted. Assuming that the maximum increment occurs between $(l-1)$ and l, this PD estimate is biased if $\bar{y}_{l-1} < \bar{y}_i$, $i = l-2, \ldots, P$. This can be used to check that group $(l-1)$ is noninferior (or even superior) to all lower-dose groups and the placebo by marginal $(1-\alpha)$ upper limits $(\mu_{l-1} - \mu_i)$ larger than $-\theta$ respective μ_{l-1}/μ_i, $i = l-2, \ldots, P$ larger than δ (with $\delta < 1$).

An Example. For the angina trial, the simultaneous confidence intervals for incremental contrasts $(\mu_i - \mu_{i-1})$ are presented in Fig. 5.

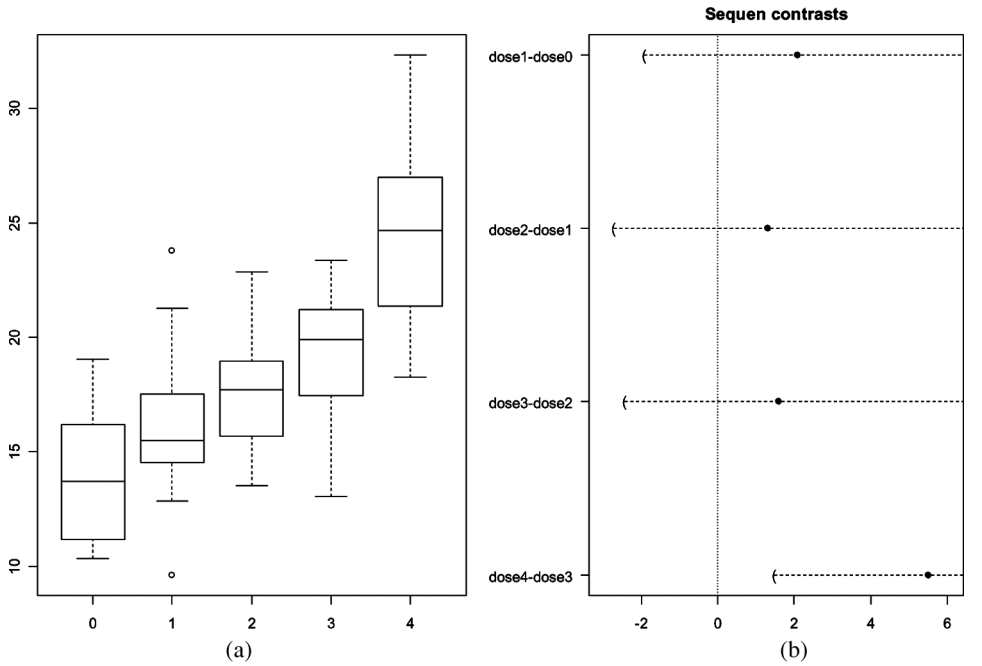


Figure 5a,b Simultaneous confidence intervals for incremental contrasts for the angina trial.

The *R*-code for incremental contrasts is simple using the *R*-function simint.

```
CImax = simint(response dose, data = angina, alternative =
               "greater", type = "Sequen", conf.level = 0.975)
```

Because only the upper limit for the increment $(\mu_4 - \mu_3)$ is larger than zero, the highest dose (dose 4) is also the PD. Notice that the above confidence intervals are marginal only for the identification of a peak dose; simultaneous intervals are the objective of research.

10. IDENTIFICATION OF THE MAXIMAL SAFE DOSE

Similar to the identification of the MED for the efficacy endpoint, the maximal safe dose (MAXSD) should be identified for a possible safety endpoint. Again, some proposals are available (e.g., Bauer et al., 2001; Hothorn and Hauschke, 2000; Tamhane and Logan, 2002, 2004), but they do not estimate the simultaneous confidence intervals (see Hsu and Berger, 1999). The identification of the MAXSD using simultaneous confidence intervals is analogous to the identification of the MED using a bottom-up stepwise-decision approach.

An Example. In a toxicological study, Chen et al. (1996) reported the body weight changes vs. baseline in female rats after 14 days of treatment with 0 (control), 100, 200, 500, and 750 mg/kg Aconiazide given to 10 animals each. The mean values

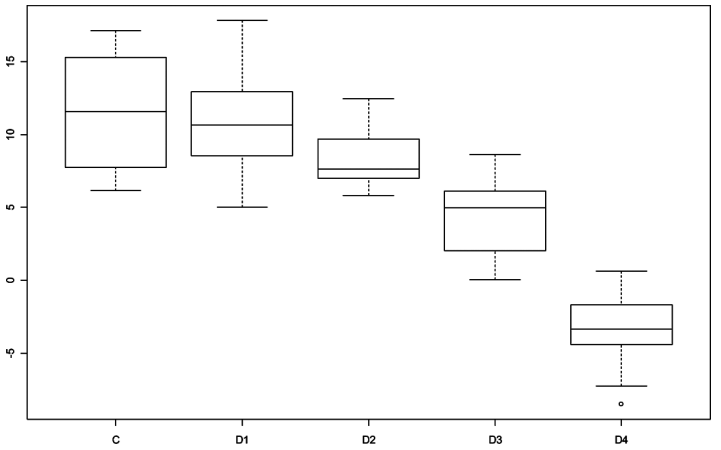


Figure 6a,b Boxplot for the body weight changes of a toxicological study.

(standard deviations) are 10.4 (4.25), 10.33 (2.67), 8.48 (1.88), 4.43 (3.29), and −4.50 (2.93) g. One-sided decreasing body weight changes are considered, hence one-sided upper confidence intervals are considered for MAXSD identification.

Because the raw data were not available, normal distributed random variables were generated accordingly. The calculation of unadjusted confidence intervals for both difference and ratio values using the common degree of freedom and the mean square estimate of a k-sample layout can be simply performed by the *R*-functions confint and sci.ratio using the method = "Unadj" option.

```
unadjust = lm(bdw~group, data = bodyw)
confint(unadjust, level = 0.90) # one-sided by 2*alpha level
library(mratios)
sci.ratio(bdw group, data = bodyw, alternative = "less",
          conf.level = 0.95, control.T = "C", method = "Unadj")
```

Starting with the comparison "100 vs. 0" and assuming an absolute body weight decrease of −2.0 g, and that a percentage decrease of 80% is still safe, the MAXSD is 200 mg with the related simultaneous confidence limits in Table 5.

Table 5 Identification of the MAXSD by simultaneous confidence intervals (n.e. = not estimable due to negative values)

	Difference to control (0)			Ratio to control (0)		
Comparison	$CI^{\text{difference,upper}}_{\text{marginal},i,1-\alpha}$	$CI^{\text{difference,upper}}_{\text{simultan.},1-\alpha}$	Decision	$CI^{\text{ratio,upper}}_{\text{marginal},i,1-\alpha}$	$CI^{\text{ratio,upper}}_{\text{simultan.},1-\alpha}$	Decision
"100 vs. 0"	1.51	2.0		1.14	0.80	
"200 vs. 0"	1.31	2.0	**MAXSD**	0.87	0.80	**MAXSD**
"500 vs. 0"	−5.00	-5.0		0.54	0.54	
"750 vs. 0"	−12.97	Stop		n.e.	Stop	

11. INTERACTION CONTRASTS

Some dose-finding trials are performed in a randomized two-way layout with a secondary factor (e.g., center, gender, condition). Simultaneous confidence intervals for the

$$\frac{L(L-1)}{2}\,\frac{(k+1)k}{2}$$

tetrad pair-wise interaction contrasts are appropriate with $l = 1, \ldots, L$ centers and $i = P, 1, \ldots, k$ doses using the union-intersection test

$$\max\left\{\left|(\mu_{li} - \mu_{li'}) - (\mu_{l'i} - \mu_{l'i'})\right|\right\}$$

$l = 1, \ldots, L-1$; $l' = 2, \ldots, L$; $i = P, 1, \ldots, k-1$; $i' = 1, \ldots, k$ according to Hochberg and Tamhane (1987). Bradu and Gabriel (1974) proposed a conservative Bonferroni approach for these tetrad contrasts, while Hochberg and Tamhane (1987) used the quantile for the range statistics. However, the distribution under the null hypothesis is available only for the global balanced case (i.e., the same sample sizes in all groups and all centers). This is an unrealistic restriction for practical studies. Westfall et al. (1999) propose a simulation-based SAS algorithm for these all-pairs tetrade contrasts for any sample size. But tetrade contrasts consider the differences between all centers and between all dose groups. The experimental question in clinical dose-finding trials is frequently how the doses compare to the control only. Therefore, the following "many-to-one-by-condition interaction contrast" approach is proposed by using an union-intersection test for

$$\max\left\{\left|(\mu_{li} - \mu_{lP}) - (\mu_{l'i} - \mu_{l'P})\right|\right\}$$

$l = 1, \ldots, L-1$; $l' = 2, \ldots, L$; $i = 1, \ldots, k$. This test is different from the all-pairs tetrade test: the narrower null hypothesis of the many-to-one-by-center interaction difference is assumed to be equal (Hothorn, 2004). Moreover, all pair-wise tetrade contrasts and their simultaneous confidence intervals are inherently two-sided, but sometimes one-sided intervals are appropriate (e.g., when proving noninferiority).

An Example. In an *in vivo* study, the implantation quality was investigated in five animals for three doses and a control with the standard (s) and a new introduced material (n) as a second randomized factor. Multiple endpoints were available; an amount of a selected tissue is used here as a single endpoint (see the boxplot in Fig. 7a). The objective for the new material is that it will at least be noninferior relative to the standard for the comparisons of dose vs. control.

The *R*-code for tetrade contrasts is using the *R*-function simint, where, for many-to-one-by-center interaction contrasts, a user-defined contrast matrix is needed.

```
CItetra = simint(tissue~dose:material, data = tis,
                 alternative = "two.sided", type = "Tetrade",
                 nlevel = c(4, 2), conf.level = 0.95)
C12 = c(1, -1, -1, 1, 0, 0, 0, 0)
C13 = c(1, -1, 0, 0, -1, 1, 0, 0)
```

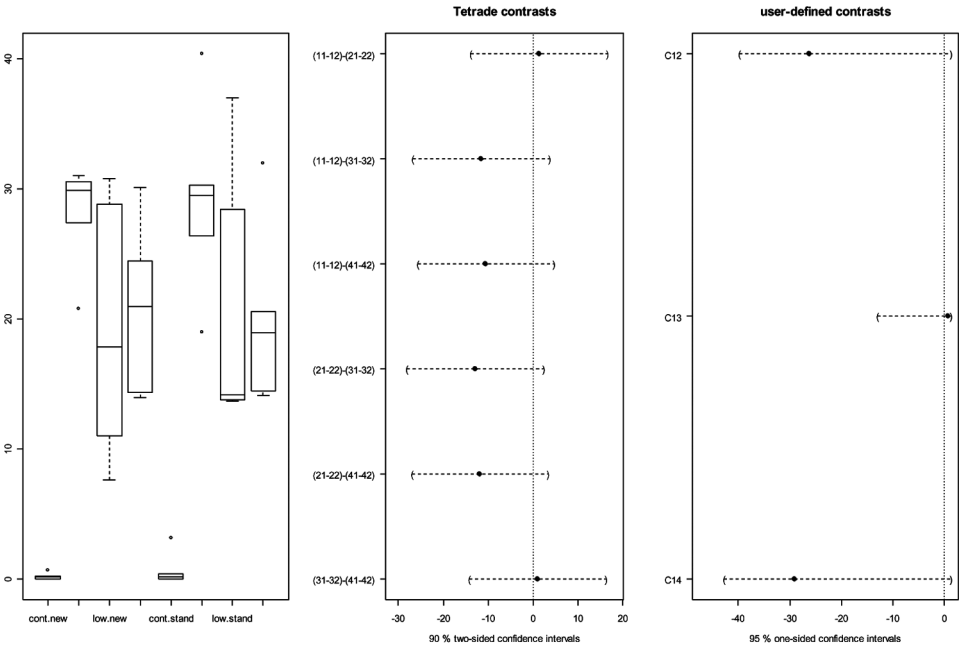


Figure 7a,b,c Boxplots for the tissue amount and simultaneous tetrade and many-to-one-by-condition interaction contrast confidence intervals.

```
C14 = c(1, -1, 0, 0, 0, 0, -1, 1)
Cm21 = rbind(C12, C13, C14)
CIm21 = simint(tissue~dose:material, data = tis, alternative =
                "greater", cmatrix = Cm21, conf.level = 0.95)
```

For the proof of no dose-by-condition interactions, two-sided $(1 - 2\alpha)$ tetrade intervals were used. From Fig. 7b, it can be seen that all tetrade contrast intervals are within about $\pm 30\%$ tissue amount of each other (i.e., some similarity of the dose effects for new and standard materials can be seen). From the interaction "doses vs. control"-by-conditions, the one-sided lower confidence limits seem to make a claim of noninferiority difficult because the lower limit for contrast $C14$, $(\mu_{\text{new},3} - \mu_{\text{new},P}) - (\mu_{\text{standard},3} - \mu_{\text{standard},P})$, is rather large at -43. Alternatively, a dose-increments-by-center-interactions approach $(\mu_{li} - \mu_{li+1}) - (\mu_{l'i} - \mu_{l'i+1})$ can be used (Hothorn, 2003).

12. DISCUSSION

The evaluation of randomized dose-finding trials can be performed using rather different approaches (e.g., selecting the "best" dose-response model according to Bretz et al., 2005). According to the recommendation of the ICH E9 guidance, this paper focuses on the graphical presentation of different kinds of simultaneous confidence intervals. Therefore, it represents a distinguished alternative to the

recently published decision approach using the closed testing procedure by Kong et al. (2005).

The basic approach behind our simultaneous confidence intervals for multiple contrasts is based on maximum tests, which are multivariate t-distributed with a selected correlation matrix. The *R*-package multcomp offers a related public-domain software solution with predefined or user-defined contrast matrices. The main advantage of confidence intervals is that they are clinically interpretable in terms of their statistical significance and clinical relevance. Hereby, however, not only a type I error level must be chosen in advance, but also relevance margin for superiority or noninferiority. Consensus on these margins only exists for a few therapeutic areas, and the definition of absolute (i.e., the scale variant) thresholds may be difficult. Therefore, simultaneous confidence intervals for ratio to placebo and ratio to active control were proposed alternatively. The style of this paper avoids an overly formal description of the contrast-based intervals; instead, it presents example-based explanations together with the related *R*-code that was used.

Finally, the decision-makers—the clinicians—will decide on the appropriateness of such an approach for problem-adequate reasoning. This paper will support the necessary discussion.

Many aspects were not mentioned due to space limitations, including nonparametric approaches for an endpoint with any distribution, approaches for differences and ratios of proportions, (see Bretz and Hothorn, 2003) and dose-finding trials with a primary and a secondary endpoint (Hothorn and Wassmer, 2003).

ACKNOWLEDGMENTS

I would like to thank Dr. Naitee Ting, Pfizer, for inviting me to contribute in the special issue and Dr. Frank Bretz, Novartis, for providing helpful discussions. I am grateful to the two anonymous reviewers for their comments that substantially improved this article.

REFERENCES

Bartholomew, D. J. (1961). Ordered tests in the analysis of variance. *Biometrika* 48:325–332.

Bauer, P., Roehmel, J., Maurer, W., Hothorn, L. A., Lehmacher, W. (1998). Testing strategies in multi-dose experiments including active control. *Stat. Med.* 17:2133–2146.

Bauer, P., Brannath, W., Posch, M. (2001). Multiple testing for identifying effective and safe treatments. *Biometr. J* 43:605–616.

Bradu, D., Gabriel, K. R. (1974). Simultaneous statistical inference on interaction in 2-way analysis of variance. *J. Am. Stat. Assoc.* 69:428–436.

Bretz, F. (1999). *Powerful Modifications of Williams Test on Trend*. Ph.D. thesis, University of Hannover. (Available at http://www.bioinf.uni-hannover.de)

Bretz, F., Hothorn, L. A. (2000). A powerful alternative to Williams test with application to toxicological dose-response relationships of normally distributed data. *Environm. Ecol. Stat.* 7:135–154.

Bretz, F., Seidel, D. (2000). SAS/IML programs for exact calculations of orthant probabilities for arbitrary dimensions. *Comput. Stat. Data Anal.* 33:220–221.

Bretz, F., Hothorn, L. A. (2001). Testing dose-response relationships with a priori unknown possibly nonmonotone shapes. *J. Biopharm. Stat.* 11:193–207.

Bretz, F., Hothorn, L. A. (2003). Statistical analysis of monotone or non-monotone dose-response data from in vitro toxicological assays. *ATLA-Altern. Lab. Anim.* 31:81–96 Suppl. 1.

Bretz, F., Hothorn, L. A., Hsu, J. C. (2003). Identifying effective and/or safe doses by stepwise confidence intervals for ratios. *Stat. Med.* 22:847–858.

Bretz, F., Hothorn, T. Westfall, P. (2002). On multiple comparisons in R. *R-news* 3:14–17.

Bretz, F., Pinheiro, I. C., Branson, M. (2005). Combining multiple comparisons and modeling techniques in dose-response studies. *Biometrics* 61:738–748.

Center for Drug Evaluation and Research (1994). *Guideline for Industry: Dose-Response Information to Support Drug Registration*. Rockville, MD: U.S. Food and Drug Administration. (Available at http://www.fda.gov/cder/guidance/iche4.pdf)

Channon, E. J., McEntegart, D. J. (2001). Confidence intervals and *p*-values for Williams' and other step-down multiple comparison tests against control. *J. Biopharm Stat.* 11:45–63.

Chen, J. J., Kodell, R. L., Gaylor, D. W. (1996). Risk assessment for nonquantal effects. In: Fan, A. M., Chang, L. W., eds. *Toxicology and Risk Assessment*. New York: Marcel Dekker, pp. 503–513.

Combes, R. D. (1997). Statistical analysis of dose-response data from in vitro assays: an illustration using Salmonella mutagenicity data. *Toxicol. in Vitro* 11:683–687.

Committee for Proprietary Medicinal Products (2002). *Points to Consider on Multiplicity Issues in Clinical Trials*. London: The European Agency for the Evaluation of Medicinal Products.

Dilba, G. (2005). Simultaneous Inference for Ratio of Location Parameters. Ph.D. thesis, University of Hannover.

Dilba, G., Schaarschmidt, F. (2005). *The R Package mratios Constructing Simultaneous Confidence Intervals for Ratio-to-Control*. Biostatistics Unit, University of Hannover. (Available at http://www.bioinf.uni-hannover.de).

Dilba, G., Bretz, F., Guiard, V., Hothorn, L. A. (2004). Simultaneous confidence intervals for ratios with application to the comparison of several treatments with a control. *Methods Inf. Med.* 43:465–469.

Dilba, G., Bretz, F., Guiard, V. (2006a). Simultaneous confidence sets and confidence intervals for multiple ratios. *J. Stat. Plann. Inf.* 36:2640–2658.

Dilba, G., Bretz, F., Hothorn, L. A., Guiard, V. (2006b). Power and sample size computations for simultaneous tests for non-inferiority and equivalence trials with inferences based on the ratio of means. *Stat. Med.* 25:1131–1147.

Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *J. Amer. Statist. Assoc.* 50:1096–1121.

Fieller, E. (1954). Some problems in interval estimation. *J. R. Stat. Soc*. B16:175–185.

Finner, H., Strassburger, K. (2002). The partitioning principle: a powerful tool in multiple decision theory. *Ann. Stat.* 30:1194–1213.

Hauschke, D., Kieser, M. (2001). Multiple testing to establish non-inferiority of k treatments with a reference based on the ratio of two means. *Drug Inform. J.* 35:1247–1251.

Hirotsu, C. (1997). Isotonic inference with particular interest in application to clinical trials. In: Kitsos, C. P., Edler, L., eds. *Industrial Statistics*. Heidelberg, Germany: Physica-Verlag, pp. 233–241.

Hochberg, Y., Tamhane, A. C. (1987). *Probability and Mathematical Statistics: Multiple Comparison Procedures*. New York: John Wiley & Sons.

Hothorn, L. A. (2003). Statistics of interlaboratory in vitro toxicological studies. *ATLA-Altern. Lab. Anim.* 31:43–63, Suppl. 1.

Hothorn, L. A. (2004). A robust statistical procedure for evaluating genotoxicity data. *Environmetrics* 15:635–641.

Hothorn, L. A., Bretz, F. (2000). One-sided simultaneous confidence intervals for effective dose steps in unbalanced designs. *Biometr. J.* 42:995–1006.

Hothorn, L. A., Hauschke, D. (2000). Identifying the maximum safe dose: a multiple testing approach. *J. Biopharm. Stat.* 10:15–30.

Hothorn, L. A., Wassmer, G. (2003). Analyzing randomized dose finding studies with a primary and a secondary endpoint. *J. Biopharm. Stat.* 13:301–305.

Hsu, J. C., Berger, R. L. (1999). Stepwise confidence intervals without multiplicity adjustment for dose-response and toxicity studies. *J. Am. Stat. Assoc.* 94:468–482.

Jan, S. L., Chen, Y. I. (2004). Nonparametric procedures for simultaneous identification of the minimum effective dose in each of several groups. *J. Biopharm. Stat.* 14:781–789.

Kong, L., Koch, G., Liu, T. (2005). Performance of some multiple testing procedures to compare three doses of a test drug and placebo. *Pharm. Stat.* 4:25–35.

Liu, W., Miwa, T., Hayter, A. J. (2000). Simultaneous confidence interval estimation for successive comparisons of ordered treatment effects. *J. Stat. Plann. Inf.* 88:75–86.

Marcus, R. (1976). The power of some tests of the equality of normal means against an ordered alternative. *Biometrika* 63:177–183.

Marcus, R., Genizi, A. (1994). Simultaneous confidence intervals for umbrella contrasts of normal means. *Comput. Stat. Data Anal.* 17:393–407.

Neuhaeuser, M., Hothorn, L. A. (1997). Trend tests for dichotomous endpoints with application in carcinogenicity studies. *Drug Inform. J.* 30:463–469.

R Development Core Team (2004). *The R Project for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. (available at http://www.R-project.org).

Robertson, T., Wright, F. T., Dykstra, R. L. (1988). *Order Restricted Statistical Inference*. New York: John Wiley & Sons.

Stewart, W. H., Ruberg, S. J. (2000). Detecting dose response with contrasts. *Stat. Med.* 19:913–921.

Tamhane, A. C., Logan, B. R. (2002). Multiple test procedures for identifying the minimum effective and maximum safe doses of a drug. *J. Am. Stat. Assoc.* 97:293–301.

Tamhane, A. C., Logan, B. R. (2004). Finding the maximum safe dose level for heteroscedastic data. *J. Biopharm. Stat.* 14:843–856.

Tamhane, A. C., Hochberg, Y., Dunnett, C. W. (1996). Multiple test procedures for dose finding. *Biometrics* 52:21–37.

Tamhane, A. C., Dunnett, C. W., Green, J. W., Wetherington, J. D. (2001). Multiple test procedures for identifying the maximum safe dose. *J. Am. Stat. Assoc.* 96:835–843.

U.S. Food and Drug Administration (1998) *Guidance for Industry: E9 Statistical Principles for Clinical Trials*. Rockville, MD: FDA. (available at http://www.fda.gov/cder/guidance/ICH_E9-fnl.PDF).

Westfall, P. H., Young, S. S. (1993). *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. New York: John Wiley & Sons.

Westfall, P. H., Tobias, R. D., Rom, D. (1999). *Multiple comparisons and multiple tests*. Cary, NC: SAS.

Williams, D. A. (1972). The comparison of several dose levels with a zero dose control. *Biometrics* 28:519–531.