

Evaluation of clinical dose finding studies using a non-parametric Shirley-type trend test for relative effects

Ludwig A. Hothorn¹ and Frank Konietschke²

May 16, 2010

¹Institut für Biostatistik, Leibniz-Universität Hannover
Herrenhäuser Strasse 2, D-30419 Hannover, Germany
Telephone: ++49-511-7625566
Fax: ++49-511-7624966
Email: hothorn@biostat.uni-hannover.de

²Abteilung für Medizinische Statistik, Universität Göttingen
Humboldtallee 32, D-37073 Germany
Telephone: ++49-551-39 4989
Fax: ++49-551-39 4995
Email: Frank.Konietschke@medizin.uni-goettingen.de

Abstract

Key Words: trend test, relative effect size, Shirley-type test, simultaneous confidence intervals

1 The problem

Double-blinded, randomized Phase IIb - III clinical dose finding studies are commonly performed as an one-way layout with the levels placebo, low, medium and high dose. Hereby either a single efficacy endpoint or multiple primary, co-primary or secondary endpoints are considered [13]. As a motivating example the dose-response-study of a new drug to treat angina pectoris is used here [16]. The efficacy endpoints is the duration of pain-free walking (relative to the baseline value) after treatment with placebo or four doses (denoted as 0, 1, 2, 3, 4). Hereby indicate large values positive effects. There are at least two experimental questions: i) exists a monotonic dose-response relationship at all? and ii) which is the minimal effective dose group? Many model-based and model-free approaches - or combinations of both- exist, see for an overview [2]. Hereby, mostly decision approaches based on p-values are used as special cases of the closure testing principle [6]. However, the ICH E9 guidance [10] for statistical evaluation of randomized clinical trials recommends instead of decision approaches or p-values: *Estimates of treatment effects should be accompanied by confidence intervals, whenever possible....* Therefore the **first question** arises which effect size is appropriate for randomized clinical dose finding trials. Widespread the difference of expected values $\mu_i - \mu_{Placebo}$ is used, where i is the index of the dose groups. For studies with individuals as randomized units, such as volunteers in randomized clinical trials recently [3] criticized the use of population means differences as effect size: *I once asked a well-published medical researcher what $p < 0.05$ meant to him. He said: It means that everyone on X did better than everyone on Y . Instead we need ... a measure of how often a random subject receiving treatment X will outperform a random subject receiving treatment Y , typically expressed as $P(X > Y)$.* The relative effect size was introduced as measure for *stochastic order*: $p_{01} = P(X_{01} < X_{11})$ (for continuous data) [4]. The generalization for non-continuous data (ties, scores data,...): p_{01} as *ordinal effect size measure* [14] and is defined:

$$p_{01} = \int F_0 dF_1 = P(X_{01} < X_{11}) + 0.5P(X_{01} = X_{11}). \quad (1)$$

This effect size is a probability that a randomly selected volunteer in placebo reveals a smaller response value than a randomly selected patient in the dose group. The decision to larger values is, if $p_{01} > 0.5$, to smaller values, if $p_{01} < 0.5$, and under H_0 if $p_{01} = 0.5$. Related confidence intervals are available [4].

The **second question** arises which trend test approach makes confidence intervals available - both for claiming a global trend and for selection appropriate doses? The trend test

according to Williams (1971, 1972) [17] denotes a parametric test procedure for detecting arbitrarily trend patterns, not just a linear trend. Let $\bar{X}_P$ denote the sample mean of the placebo group, s^2 the common variance estimator and let $\hat{\mu}_i$ for $s = i, \dots, k$, denote the maximum likelihood estimators for the treatment effects μ_i under the constraint $\mu_P \leq \dots \leq \mu_k$. The test statistic is given by the pairwise t-type statistic $\bar{t}_s = \frac{\hat{\mu}_P - \bar{X}_i}{s\sqrt{\frac{1}{n_P} + \frac{1}{n_i}}}$.

The Williams' test can be approximated by a multiple contrast test [1] where its distribution is available both under the null and alternative hypotheses according to central and non-central joint multivariate t- distribution with correlation matrix $\mathbf{R}$, particularly for the general unbalanced design. Generally multiple contrast tests reveal both multiplicity-adjusted p-values and simultaneous confidence intervals [8].

The **third question** arises for a non-parametric version of the Williams-type trend test. The trend test according to Shirley (1977) [15] belongs to the widely used tests for comparing dose groups with a control when data are not normally distributed. No particular shape of the dose-response relationship should be used as an assumption, such as linearity for the widely-used Jonckheere trend test [11]. Variance heterogeneity occurs particularly in dose-finding trials where larger effects are sometimes connected with higher variability, i.e. the **fourth question** for a non-parametric Behrens-Fisher solution arises. Lastly the **fifth question** for availability of related software arises.

2 A non-parametric Shirley-type test for relative effects

The original Shirley-test [15] represents an asymptotic non-parametric multiple test procedure for the hypothesis $H_0^F : F_0 = \dots = F_k$ formulated for the distribution functions against the ordered alternative $H_1^F : F_0 \leq \dots \leq F_k$ with at least one- anyone- inequality $F_i < F_s, i \neq s$. It controls the FWER strongly. However, the distribution of the rank means is unknown under the alternative, no simultaneous confidence intervals are available for a general unbalanced design.

Therefore a Shirley-type test for relative effects and its simultaneous confidence intervals are used [9].

Rather easily two-sample relative effects p_{0s} can be estimated with pairwise rankings of the observations, i.e., the two samples $X_{01}, \dots, X_{0n_0}$ and $X_{s1}, \dots, X_{sn_s}$ are pairwise ranked. Let $R_{sj}^{(0s)}$ denote the rank of X_{sj} among all $n_0 + n_s$ observations within the samples 0 and s . Moreover, let $\bar{R}_{s\cdot}^{(0s)} = \frac{1}{n_s} \sum_{j=1}^{n_s} R_{sj}^{(0s)}$ denote the mean of these ranks. Now, the rank means can be used to estimate p_{0s} by

$$\hat{p}_{0s} = \frac{1}{n_0} \left(\bar{R}_{s\cdot}^{(0s)} - \frac{n_s + 1}{2} \right).$$

Linear combinations of the estimators $\hat{p}_{0s}$ can be used for the estimation of the treatment

effects:

$$\begin{aligned}
\hat{p}_1 &= \hat{p}_{0k} \\
\hat{p}_2 &= \frac{n_{k-1}}{n_{k-1} + n_k} \hat{p}_{0(k-1)} + \frac{n_k}{n_{k-1} + n_k} \hat{p}_{0k} \\
&\vdots \\
\hat{p}_q &= \frac{n_1}{n_1 + \dots + n_k} \hat{p}_{01} + \dots + \frac{n_k}{n_1 + \dots + n_k} \hat{p}_{0k}.
\end{aligned}$$

Asymptotically $\sqrt{N}(\hat{p}_1 - p_1, \dots, \hat{p}_q - p_q)'$ follows a central multivariate normal distribution with expectation $\mathbf{0}$ and covariance matrix $\mathbf{V}_N$, see for details [12]. Related approximate $(1 - \alpha)100\%$ one-sided lower simultaneous confidence limits are:

$$\left[\hat{p}_\ell - t_{q,1-\alpha}(\nu, \mathbf{R}) \sqrt{S_\ell}; \right], \ell = 1, \dots, q, \quad (2)$$

with the $(1 - \alpha)$ quantile $t_{q,1-\alpha}(\nu, \mathbf{R})$ of the related q -variate $t(\mathbf{0}, \nu, \mathbf{R})$ -distribution. Here, the degree of freedom ν and the correlation matrix $\mathbf{R}$ are estimated with pairwise rankings from the data. These simultaneous confidence intervals may not be range preserving, i.e., the lower bounds can be negative and the upper bounds can take on larger values than 1. Therefore, range preserving confidence intervals can be easily constructed by using the delta method and the probit transformation [12].

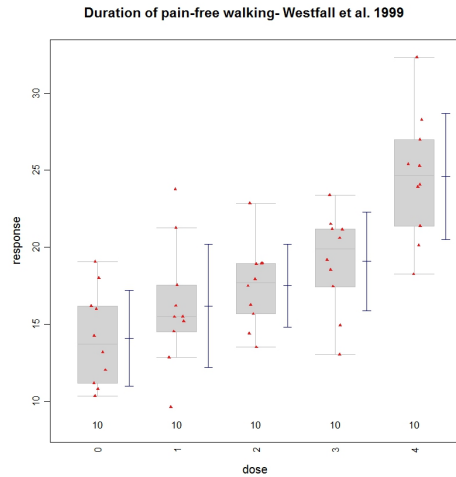
Interpretation is relatively simple. A global trend occur when ever a lower confidence limit for any contrast is larger than 0.5 for those endpoints where higher efficacy occurs at larger values (vice versa). According to the closure principle [7] is the minimal effective dose the smallest dose which is contained in a significant contrast. Using the lower confidence limits biologically-relevant decisions can be drawn when using a non-zero relevance threshold.

3 Software

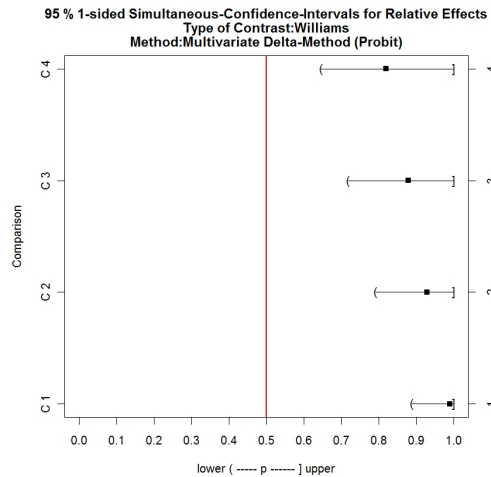
In the freely available software package R the package *nparcomp* is available. It contains two main functions: the function *npar.t.test* and the function *nparcomp*. The function *npar.t.test* focuses on two independent samples, and computes confidence intervals for relative effects. The function *nparcomp* computes simultaneous confidence intervals and adjusted p-values for relative effects in user-defined contrast settings, such as according to Shirley. The main outcomes are one-sided or two-sided confidence intervals using multivariate normal-approximation, multivariate t_ν -approximation, logit-approximation or probit-approximation and their visualization.

4 Evaluation of the example

The box-plots in Figure 1 indicate some non-normal distribution by the differences between the medians and the arithmetic means as well as variance heterogeneity, although a balanced design is used.



Therefore for this dose-finding trial the new Shirley-type trend approach is appropriate. Since larger values indicate higher efficacy, the one-sided lower simultaneous confidence limits are estimated, see Figure 2.



We see a global trend since all lower limits do not contain the value of H_0 of 0.5. The most distant limits belongs to the contrast $C1$, i.e. the comparison of placebo with the high dose only. If we would assume an at least $\eta = 80\%$ benefit for a volunteer in a dose group, the minimal effective and relevant dose is the second highest dose $D3$. For this balanced design the contrast $C2$ is defined by $\hat{p}_{C2} = \frac{1}{2}\hat{p}_{03} + \frac{1}{2}\hat{p}_{04}$. Some arguments are against the use of pooled contrasts for estimation the minimum effect dose. Alternatively the related Dunnett-type contrasts for just comparing the placebo versus the individual dose groups

can be used [5], where the lower limit for dose D_3 is 62% which is an argument to use the highest dose is the minimum efficacious dose based on a relevance threshold of $\eta = 80\%$. The related R code is quite simple:

```
library(mratios)
data(angina)
angina$Dose <- as.factor(angina$dose)
library(npacomp)
npacomp(response ~ dose , data=angina, asy.method = "probit",
  type = "Williams", alternative = "lower", plot.simci = TRUE)
npacomp(response ~ dose , data=angina, asy.method = "probit",
  type = "Dunnett", alternative = "lower", plot.simci =FALSE)
```

5 Discussion

At least six arguments favor the analysis of real-data dose-finding studies by the Shirley-type procedure for relative effects: i) use an effect size which allows individualized claim of efficacy, ii) use confidence limits which overcomes the limited interpretation of p-value and allows the interpretation in both statistical significance and biological relevance, iii) use multiple contrast tests for an order restriction which allows global trend and local dose-response shape interpretation without stepwise procedure and without assumption of the particular function of the dose-response relationship, iv) use of multiple contrast tests which allows easy switching from order restriction to many-to-comparison and even from non-inferiority to superiority (not shown here, see [7]), v) does not assume neither normal distribution nor variance homogeneity, and vi) related software is free available. These six advantages are in opposite to at least two disadvantages: difficult extrapolation to non-experimental dose-ranges and minor loss of power compared with stepwise decision procedures.

References

- [1] Bretz, F. (2006). An extension of the Williams trend test to general unbalanced linear models. *Computational Statistics & Data Analysis* **50**, 1735 - 1748
- [2] Bretz, F. Pinheiro, J.C., Branson, M. (2006). Combining multiple comparisons and modeling techniques in dose-response studies. *Biometrics* **61**, 738 - 748
- [3] Browne, R. H. (2010). The t-Test p Value and Its Relationship to the Effect Size and $P(X > Y)$. *The American Statistician* **64**, DOI: 10.1198/tast.2010.08261
- [4] Brunner, E., Munzel, U. (2000). The Nonparametric Behrens-Fisher Problem: Asymptotic Theory and a Small-Sample Approximation. *Biometrical Journal* **1**, 17 - 21
- [5] Dunnett, C. W. (1955). A Multiple Comparison Procedure For Comparing Several Treatments With A Control. *Journal Of The American Statistical Association* **50**, 1096 - 1121
- [6] Hothorn, L. A., Neuhauser, M., Koch, H.F. (1997). Analysis of randomized dose-finding-studies: Closure test modifications based on multiple contrast tests. *Biometrical Journal* **39**, 467 - 479
- [7] Hothorn, L. A. (2006). Multiple comparisons and multiple contrasts in randomized dose-response trials-confidence interval oriented approaches. *Journal Of Biopharmaceutical Statistics* **16**, 711 - 731
- [8] Hothorn, L.A.(2007). How to deal with multiple treatment or dose groups in randomized clinical trials? *Fundamental and Clinical Pharmacology* **21**, 137 - 154
- [9] Hothorn, L.A., Konietzschke, F. (2010). Evaluation of toxicological studies using a non-parametric Shirley-type trend test for comparing several dose levels with a control group *Submitted*
- [10] ICH (1998). Statistical Principles for Clinical Trials, 1998
- [11] Jonckheere, A. R. (1954). A Distribution-free k-sample Test Against Ordered Alternatives. *Biometrika* **41**, 133 - 145
- [12] Konietzschke, F. (2009). *Simultane Konfidenzintervalle für nichtparametrische relative Kontrasteffekte*, Dissertation, Georg-August University of Göttingen
- [13] Neuhauser, M. (2006). How to deal with multiple endpoints in clinical trials *Fundamental and clinical pharmacology* **20**, 515 - 523
- [14] Ryu, E., Agresti, A. (2008). Modeling and inference for an ordinal effect size measure. *Statistics in Medicine* **27**, 1703 - 1717
- [15] Shirley, E. (1977). Nonparametric Equivalent of Williams Test for Contrasting Increasing Dose Levels of a Treatment. *Biometrics* **33**, 386 - 389
- [16] P. H. Westfall, R. D. Tobias, D. Rom, R. D. Wolfinger, Y. Hochberg (1999) Multiple Comparisons and Multiple Tests Using the SAS System. *SAS Institute Inc; Cary, NC* , 1 -
- [17] Williams, D. A. (1971). Test For Differences Between Treatment Means When Several Dose Levels Are Compared With A Zero Dose Control. *Biometrics* **27**, 103 - 117