Drug Information Journal, Vol. 34, pp. 591–597, 2000
Printed in the USA.

0092-8615/2000

SEQUENTIAL AND MULTIPLE TESTING FOR DOSE-RESPONSE ANALYSIS

WALTER LEHMACHER, PHD
Professor, Institute for Medical Statistics, Informatics and Epidemiology, University of Cologne, Germany
MEINHARD KIESER, PHD
Head, Department of Biometry, Dr. Willmar Schwabe Pharmaceuticals, Karlsruhe, Germany
LUDWIG HOTHORN, PHD
Professor, Department of Bioinformatics, University of Hannover, Germany

For the analysis of dose-response relationship under the assumption of ordered alternatives several global trend tests are available. Furthermore, there are multiple test procedures which can identify doses as effective or even as minimally effective. In this paper it is shown that the principles of multiple comparisons and interim analyses can be combined in flexible and adaptive strategies for dose-response analyses; these procedures control the experimentwise error rate.

Key Words: Dose-response; Interim analysis; Multiple testing

INTRODUCTION

THE FIRST QUESTION in dose-response analysis is whether a monotone trend across dose groups exists. This leads to global trend tests, where the global level (familywise error rate in a weak sense) is controlled. In clinical research, one is additionally interested in which doses are effective, which minimal dose is effective, or which dose steps are effective, in order to substantiate the choice of a dose. Dose differences are considered effective if they are clinically relevant and statistically significant. Therefore, multiple comparisons which control the experimentwise level (familywise type-I error rate in a strong sense) are necessary. The global and multiplicity aspects are also reflected by the actual International Conference for Harmonization Guidelines (1) for dose-response analysis. Appropriate order restricted tests can be confirmatory for demonstrating a global trend and suitable multiple procedures can identify a recommended dose.

For the global hypothesis, there are several suitable trend tests. For multiple comparisons, there are several approaches for multiple test procedures, some of which are described in the following. It is easy to perform global tests in group sequential or adaptive interim versions, but it seems that there are only a few methods which combine multiple test procedures with interim analyses. In this paper, some new approaches are proposed for the combination of these multiple and sequential dose-response analyses, which can link proof of global trend with

Reprint address: Prof. Dr. Walter Lehmacher, Institut für Medizinische Statistik, Informatik und Epidemiologie der Universität zu Köln, Joseph-Stelzmann-Str. 9, D-50931, Köln, Germany. E-mail: Walter.Lehmacher@medizin.uni-koeln.de.

proof of efficacy of a chosen dose. Sequential procedures can be especially useful in this area because the multiplicity of this problem makes a reasonable sample size calculation difficult.

TESTS OF TREND

K groups with increasing dose levels are considered, where the smallest dose is often the null-dose (placebo) as a control. A one-way layout with n_k experimental units tested at the k-th dose level, $k = 1, \ldots, K$, is considered. All x_{ki} observations are assumed to be mutually independent with $x_{ki} \sim N(\mu_k, \sigma^2)$, $k = 1, \ldots, K$, and $i = 1, \ldots, n_k$. In the following, a monotone dose-response relationship is assumed: $\mu_1 \leq \ldots \leq \mu_K$. The global hypothesis is H_0: $\mu_1 = \ldots = \mu_K$, and the trend alternative is $|H|$: $\mu_1 \leq \ldots \leq \mu_K$ with at least $\mu_1 < \mu_K$. Related trend tests are reviewed in Tamhane et al. (2). Most are based on special contrasts (eg, Helmert contrasts or simple pairwise contrasts between the maximal and minimal dose). Phillips (3) described methods for sample size estimation in the randomized k-sample design including single-step dose-response studies. Recently, a review of the performance of several tests of trend for dose-response studies was given by Phillips (4). Here a detailed discussion of the performance of these trend tests is not repeated, because in the framework of this paper all such test statistics can be used providing one considers their well-known pros and cons.

MULTIPLE TEST PROCEDURES

Three multiple test procedures which can be extended to designs with interim analyses are mentioned here.

A-priori Ordered Hypotheses

The family of the $K - 1$ elementary hypotheses which contain the smallest dose group is considered, that is, H_{1k}: $\mu_1 = \ldots = \mu_k$, $k = 2, \ldots, K$. Then for $K = 3, 4, 5$ one obtains the following families of hypotheses:

$K = 3$: $H_0 = H_{123} = H_{13}$: $\mu_1 = \mu_2 = \mu_3$
H_{12} : $\mu_1 = \mu_2$

$K = 4$: $H_0 = H_{1234} = H_{14}$: $\mu_1 = \mu_2 = \mu_3 = \mu_4$
$H_{123} = H_{13}$
H_{12}

$K = 5$: $H_0 = H_{12345} = H_{15}$:
$\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$
$H_{1234} = H_{14}$
$H_{123} = H_{13}$
H_{12}

As usual, the definition $H_{1\ldots k}$: $\mu_1 = \mu_2 = \ldots = \mu_k$ is used; the monotonicity implies $H_{1\ldots k} = H_{1k}$. For each K, these families of $K - 1$ hypotheses are closed under intersection. Under the assumption of a monotone dose-response relationship all relevant hypotheses which are necessary to identify an (minimally) effective dose are included. In practical applications, the recommended dose will be chosen among the effective doses by carefully considering the magnitude of the observed benefits and side effects.

For this family of hypotheses the closed test procedure is identical to the method of a-priori ordered hypotheses, where the hypotheses are tested step by step, beginning with the global hypothesis $H_0 = H_{1K}$, followed by $H_{1,K-1}$, $H_{1,K-2}$, ..., H_{13} and ending with the "smallest" hypothesis H_{12}. All tests are performed at local level α, and all hypotheses are rejected as long as significance at level α is reached; if a test for a hypothesis H_{1k} is not significant at level α, H_{1k} is not rejected, and no further test (for $H_{1,k-1}$, $H_{1,k-2}$, etc.) is necessary because no further rejection is possible. For a more detailed description of this a-priori method see Hothorn and Lehmacher (5), Kieser and Lehmacher (6), or Maurer et al. (7). For local tests of these elementary hypotheses, all (local) trend tests can be chosen (eg, linear contrasts or comparisons of highest vs. lowest doses), as mentioned earlier. In view of the monitonicity, these tests should be 1-sided. The assumption of monotonicity can be omitted, because the a-priori method works generally with each predefined order of hypotheses. In this

case, however, pairwise t-tests should be preferred as local test statistics to trend tests. Some other multiple testing methods fail to control the experimentwise error rate if monotonicity is not given; see Bauer (8).

The Closed Test Procedure

The $K-1$ elementary hypotheses $H_{k,k+1}$ stating the identity of two neighboured dose groups are considered, that is, $H_{12}: \mu_1 = \mu_2$, $H_{23}: \mu_2 = \mu_3, \ldots, H_{K-1,K}: \mu_{K-1} = \mu_K$. The family of hypotheses is chosen, which is generated by all intersections of these elementary hypotheses. As usual, the abbreviation $H_{k_1 k_2 \ldots k_m}: \mu_{k_1} = \ldots \mu_{k_m}$ is used. For example, for $K = 3, 4, 5$ the closed systems of hypotheses are given by the families:

$K = 3$: $H_0 = H_{123} = H_{13}$
$H_{12}\ H_{23}$

$K = 4$: $H_0 = H_{1234} = H_{14}$
$H_{123} = H_{13}\ H_{234} = H_{24}$
$H_{12/34} := H_{12} \cap H_{34}$
$H_{12}\ H_{23}\ H_{34}$

$K = 5$: $H_0 = H_{12345} = H_{15}$
$H_{1234} = H_{14}\ H_{2345} = H_{25}$
$H_{123/45} := H_{123} \cap H_{45} = H_{13} \cap H_{45}$
$H_{12/345} := H_{12} \cap H_{345} = H_{12} \cap H_{35}$
$H_{123} = H_{12}\ H_{234} = H_{24}\ H_{345} = H_{35}\ H_{12/34}$
$H_{12/45}\ H_{23/45}$
$H_{12}\ H_{23}\ H_{34}\ H_{45}$

For each K, these families of hypotheses are closed under intersection by definition. They consist not only of simple pairwise comparisons such as $H_{123} = H_{13}$, but even of partition hypotheses such as $H_{12/34} := H_{12} \cap H_{34}$. Under the assumption of a monotone dose-response relationship, this family contains all pairwise comparisons, because under order restriction, for example, $H_{234} = H_{24}$ and so forth. Therefore, this family contains all hypotheses necessary to identify effective dose steps as well as the minimally effective dose, too. For this family of hypotheses, the application of the closed test principle was proposed by Marcus et al. (9).

Each hypothesis H of this closed family is tested at local level α; a hypothesis H is rejected if its local test and all local tests of hypotheses H' inducing H (ie, $H' \subset H$) are significant at level α; if a local test of H is not significant at level α, H is not rejected, and all tests for H'' induced by H (ie, $H \subset H''$) are not necessary because rejections of H'' are not possible.

Alternatively to the original approach of Marcus et al. (9), a partition hypothesis $H_{12/34} = H_{12} \cap H_{34}$ can be tested very simply by combining the two 1-sided p-values with the Bonferroni method or with the Fisher combination test; or the test statistics can be combined to $(z_{12} + z_{34})/\sqrt{2}$ or another weighted sum of these statistics. Rom et al. (10) proposed a modified closed test procedure, where the partition hypotheses are tested with a test for the "smallest" partial hypothesis, for example, $H_{12/34}$ is tested by a level-α-test for H_{12}, which is evidently a level-α-test for $H_{12/34}$ as well. As a consequence, H_{34} can be tested only after the rejection of H_{12}, and according to the general closed test principle after the rejection of all other inducing hypotheses H_0, H_{234}, and so forth. As tests for the simple pairwise comparison hypotheses all (local) trend tests (see above) can be chosen. The above mentioned Rom procedure is "uniformly better" than the a-priori method, because after the rejection of a simple pairwise comparison hypothesis H_{1k} of the family described in the above a-priori ordered hypotheses section, only additional partial hypotheses induced by H_{1k} can be tested without loss of power concerning these $K-1$ tests.

The Bonferroni-Holm Procedure

Here the $K-1$ elementary hypotheses of neighboured comparisons $H_{k,k+1}: \mu_k = \mu_{k+1}$ are tested according to the Bonferroni-Holm method; see Budde and Bauer (11). This procedure has the primary goal of indentifying effective dose steps. This method, however,

has bad power characteristics if the investigated doses are too close together.

SEQUENTIAL TEST PROCEDURES

Group Sequential Procedures

For many of the global trend tests the classical group sequential test procedures can be directly performed with 1-sided tests; see DeMets and Ware (12). Bauer and Budde (13) applied these 1-sided group sequential tests to multiple comparisons concerning neighboured dose groups in combination with the Bonferroni-Holm and the a-priori ordered methods.

Adaptive Interim Analyses

Bauer and Köhne (14) proposed procedures with one (or two) adaptive interim analyses, which, in contrast to group sequential designs, allow for a new estimation of the final sample size using the results of the interim analysis. The approach is based on the combination of the independent 1-sided p-values of the sequences of the study by Fisher's combination test. For $\alpha = 0.05$, and one interim analysis and one null hypothesis to be tested in confirmatory analysis, this procedure works as follows: Let p_1 denote the 1-sided p-value of the first sequence of the study. If $p_1 \geq \alpha_0$, the trial stops without rejection of the null hypothesis. If $p_1 \leq \alpha_1$, the null hypothesis can be rejected and the trial can stop. If $p_1 \in (\alpha_1; \alpha_0)$, no final decision can be made, and the second sequence has to be planned. For example, $\alpha_1 = 0.023$ for $\alpha_0 = 0.5$.

Let p_2 denote the 1-sided p-value of the test statistic generated by the data of the second sequence. If $p_1 \cdot p_2 \leq 0.0087$ (ie, Fisher's product criterion) then the null hypothesis can be rejected.

The procedure allows a free, that is, adaptive choice of sample sizes after an interim analysis. A similar adaptive approach was proposed by Proschan and Hunsberger (15). Recently, a modification of the classical group sequential procedures was proposed (Lehmacher, Wassmer [16]), which allows for data-driven sample size reestimation after each of the interim analyses; this approach is essentially based on the combination of the Φ^{-1}-transformation of the independent p-values. Such adaptive interim analyses are of great importance especially in dose-response analyses, because here in most cases no valid sample size estimation is possible, and therefore the idea of an internal pilot study looks attractive. Because the adaptive methods of Bauer-Köhne and Lehmacher-Wassmer are based only on the combination of independent p-values, they have the additional advantage that a change of the test statistic is possible, too. Even a change of the hypotheses to be tested is possible, say, H_1 in the first sequence and H_2 in the second sequence. In this case, however, a significant result in general enables only the rejection of the global intersection hypothesis $H_0 = H_1 \cap H_2$. Bauer and Röhmel (17) applied the Bauer-Köhne procedure to dose-response analyses, where a change of dose groups (ie, a change of hypotheses) is admitted. This approach has the primary goal of demonstrating a global dose-response relationship, and in their paper it is only mentioned that after the rejection of the global hypothesis H_0, a closed test procedure can follow.

For many trials which aim to demonstrate only a global trend, a fixed sample size design may be sufficient. But even in these simple cases, group sequential trials can reduce the average sample number (ASN). The multiplicity of the additional proof of effective doses makes a sample size calculation essentially more complicated, and sequential approaches are especially useful. Therefore, multiple comparisons within group sequential approaches are particularly proposed in the following section.

COMBINATION OF MULTIPLE AND SEQUENTIAL PROCEDURES

All types of *interim analyses* (*group sequential* or *adaptive*) can be combined with *multiple test procedures* (*a-priori ordered*, *closed test*, or *Bonferroni-Holm*) described above.

The general construction rule is: "Each hypothesis H_i of the family of hypotheses is tested with a sequential analysis at local level α, or the related p-value is calculated. It is locally rejected, if it is rejected in one of the interim or final analyses of the respective sequential test, taking into account the α-adjustments induced by the performance of interim analyses." An hypothesis H_i is rejected if the multiple test procedure allows for the rejection of H_i (possibly in dependence from other rejections).

In the a-priori and closed test procedures, each hypothesis H_i is tested at local level α; in the Bonferroni-Holm method the p-values are compared with the $K-1$ adjusted levels $\alpha/(K-1), \alpha/(K-2), \ldots, \alpha/2, \alpha$.

Evidently, these procedures control the experimentwise error rate α. Flow charts of the a-priori method with $K=3$ and of the closed test procedure with $K=2$ are given for designs with one interim analysis in Figures 1 and 2. For a detailed description of the decision rules see Kieser, Bauer, and Lehmacher (18).

In practical applications, relevant shortcuts in dependence of early rejections are possible: If an hypothesis is rejected at an interim analysis, certain dose groups can be eliminated in the following sequences. In adaptive interim analyses, the sample sizes can be recalculated for the second (or third) sequence; even unbalanced sample sizes can be considered.

The adaptive interim approaches enable further flexibilities of the study conduct: If after an interim analysis a certain dose group gives no hope of a relevant contribution, in the second sequence this dose group can be cancelled, even without acceptance or rejection of related hypotheses. For testing a local (intersection) hypothesis in these adaptive approaches, only independent p-values have to be combined: In the closed test procedure, in the second sequence the needed p-value of an intersection hypothesis can be "substituted" by a p-value of one of its inducing hypotheses (by an a-priori definition). This is again possible because a level-α-test for an inducing hypothesis H_i is also a level-α-test for an intersection hypothesis $H_i \cap H_j$.

For example: If with $K=3$ after the interim analysis the largest dose group has to be dropped (eg, for safety reasons), in the closed test procedure (see Figure 2) for the test of the global hypothesis H_0, the p-value for $H_0 = H_{123}$ from the first sequence has to be combined with the p-value of H_{12} from the second sequence. The same strategy can be used if the method of a-priori ordered hypotheses is chosen. Additionally in the adaptive interim approaches, the test statistic can be changed, depending upon the type of contrasts which seem most appropriate.

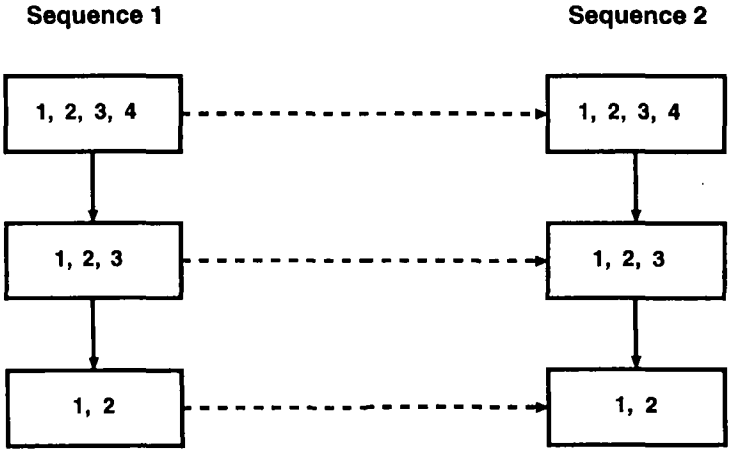

FIGURE 1. A-priori ordered hypotheses.

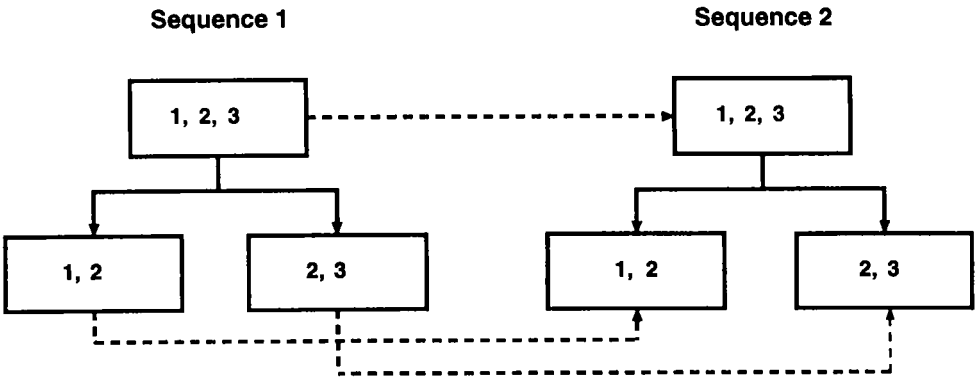

FIGURE 2. The closed test procedure.

A PRACTICAL EXAMPLE

In a three-armed randomized, double-blind multicenter study, the clinical efficacy of two different extracts of St. John's wort (Hypericum perforatum) was investigated against placebo ($k = 1$) in patients suffering from mild or moderate depression according to the criteria of the Diagnostic and Statistical Manual of Mental Disorders, fourth edition (DSM-IV) of the American Psychiatric Association (19). The Hypericum preparations differed only with respect to their content of hyperforin ($k = 2$: 0.5 % hyperforin; $k = 3$: 5 % hyperforin), and it was the aim of the trial to explore the relationship of the antidepressant efficacy of Hypericum extract from the hyperforin content. The study was planned according to the adaptive two-stage Bauer-Köhne design with 1-sided $\alpha = 0.05$, $\alpha_1 = 0.0299$ and $\alpha_0 = 0.3$. It was assumed that the treatment effect measured by the change in Hamilton Rating Scale for Depression (HAM-D) (20) increases with increasing hyperforin content of the Hypericum extract and the hypotheses were a-priori ordered accordingly. The prespecified nonparametric Jonckheere-test (21) revealed a 1-sided p-value of $p_{13} = 0.017 < \alpha_1$ for $H_0 = H_{13} = H_{123}$, leading to the rejection of H_0 and thus demonstrating the efficacy of the extract with the higher content of hyperforin. The U-test for H_{12} showed a 1-sided p-value of $p_{12} = 0.19 > \alpha_1$. Therefore, the null hypothesis concerning the comparison between placebo and the Hypericum extract with a content of only 0.5 % hyperforin could not be rejected in the interim analysis. Due to the small difference in HAM-D between placebo and the extract with the lower hyperforin content the study was stopped with this result. In principle, a second sequence could have been planned for testing H_{12}, where the 1-sided p-value must fall short of $0.0087/0.19 = 0.0458$.

DISCUSSION

In principle, all of the sequential and multiple procedures mentioned can be combined. Due to the complexity of the problem, however, an optimal procedure cannot be derived in general. The choice of the test statistic, and the multiple and sequential procedures must be determined by the well-known pros and cons of these procedures.

Because valid sample size estimations are rarely available, the combinations of the procedures described in this paper are suitable to join pilot and confirmative studies. For identifying primarily minimally effective doses, a simple but rather effective combination is the a-priori method (or the slightly more complicated but generally "better" Rom method) with the adaptive interim analyses based on simple pairwise contrasts. Further, the assumptions of variance homogeneity are not necessary, if only pairwise contrasts with the Welch modification as local test statistics are used. Related nonpara-

metric versions or trend tests for binary data are available; therefore, all the procedures proposed can be applied essentially for the nonnormal situation, too.

Acknowledgement—The authors thank two anonymous referees for critical comments and helpful suggestions.

REFERENCES

1. *Note for guidance on dose-response information to support drug registration.* International Conference for Harmonization. Topic E, CPMP/ICH/378/95.
2. Tamhane AC, Hochberg Y, Dunnett CW. Multiple test procedures for dose finding. *Biometrics.* 1996; 52:21–37.
3. Phillips A. Sample size estimation when comparing more than two treatment groups. *Drug Inf J.* 1998; 32:193–200.
4. Phillips A. A review of the performance of tests used to establish whether there is a drug effect in dose-response studies. *Drug Inf J.* 1998;32:683–692.
5. Hothorn L, Lehmacher W. A simple testing procedure "control versus k treatments" for one-sided ordered alternatives, with application in toxicology. *Biom J.* 1991;33:179–189.
6. Kieser M, Lehmacher W. Multiple testing in clinical trials with interim analyses and a-priori ordered hypotheses. (Multiples Testen bei klinischen Prüfungen mit Zwischenauswertungen und a priori geordneten Hypothesen.) In: Trampisch HJ, Lange S. (eds.): Medical Research—Medical Practice. (*Medizinische Forschung—Ärztliches Handeln.*) Proceedings, MMV Medizin Verlag München. 1995; 162–165.
7. Maurer W, Hothorn L, Lehmacher W. Multiple comparisons in drug clinical trials and preclinical assays: a priori ordered hypotheses. In: Vollmar J (ed.). Biometrics in the Pharmaceutical Industry. (*Biometrie in der chemisch-pharmazeutischen Industrie.*) Stuttgart: Fischer: 1995;6:3–21.
8. Bauer P. A note on multiple testing procedures in dose finding. *Biometrics.* 1997;53:1125–1128.
9. Marcus R, Peritz E, Gabriel KR. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika.* 1976;63:655–660.
10. Rom DM, Costello RJ, Connell LT. On closed test procedures for dose-response analysis. *Stat Med.* 1994;13:1583–1596.
11. Budde M, Bauer P. Multiple test procedures in clinical dose finding studies. *J Am Stat Assoc.* 1989;84: 792–796.
12. DeMets DL, Ware JH. Group sequential methods in clinical trials with a one-sided hypothesis. *Biometrika.* 1980;67:651–660.
13. Bauer P, Budde M. Multiple testing for detecting efficient dose steps. *Biom J.* 1994;36:3–15.
14. Bauer P, Köhne K. Evaluation of experiments with adaptive interim analyses. *Biometrics.* 1994;50: 1029–1041.
15. Proschan MA, Hunsberger SA. Designed extension of studies based on conditional power. *Biometrics.* 1995;51:1315–1324.
16. Lehmacher W, Wassmer G. Adaptive sample size calculation in group sequential trials. *Biometrics.* 1999;55:1286–1290.
17. Bauer P, Röhmel J. An adaptive method for establishing a dose response relationship. *Stat Med.* 1995; 14:1595–1607.
18. Kieser M, Bauer P, Lehmacher W. Inference on multiple endpoints in clinical trials with adaptive interim analyses. *Biom J.* 1999;41:261–277.
19. Laakmann D, Schüle C, Baghai T, Kieser M. St. John's Wort in mild to moderate depression: the relevance of hyperforin for the clinical efficacy. *Pharmacopsychiatry.* 1998;31:54–59.
20. Hamilton M. A rating scale for depression. *J Neurol Neuro-surgery Psychiatry.* 1960;23:56–62.
21. Jonckheere AR. A distribution-free k sample test against ordered alternatives. *Biometrika.* 1954;41: 133–145.