APPLICATION OF ADAPTIVE INTERIM ANALYSIS IN PHARMACOLOGY

LUDWIG A. HOTHORN, PhD

University of Hannover, Hannover, Germany

ULLRICH MARTIN, PhD

Boehringer Mannheim GmbH, Penzberg, Germany

A Drug Information Association Workshop on "Statistical Methodology on Non-Clinical and Toxicological Studies" was held March 25–27, 1996. The purpose of this meeting was to discuss the appropriateness of current and new biostatistical methods in this field of drug development. This paper describes the application of adaptive interim analysis in pharmacological studies. Bauer and Köhne (1) published a two-step approach in the case of unknown a priori information. This approach is now widely used for clinical trails. Here, the advantages of use in some pharmacological studies will be discussed.

Key Words: Interim analysis; Pharmacological studies

THE PROBLEM

IN PHARMACOLOGY animal studies can be categorized as screening studies for a series of substances on a well-defined animal model in which the outcome is frequently dichotomous (eg, responder/nonresponder), standardized bioassay for estimation of biological activity, and studies to demonstrate a clinically relevant effect for a selected substance in a specific—sometimes not established—animal model, usually based on a design "treatment or dose groups versus negative control." The last type of study will be considered here. The paradigm of these studies is similar to clinical trials: randomized, placebo (negative control C–) controlled, based on experimental design with sample size estimation, definition of an (clin-ical) endpoint, and using a standardized protocol and analysis.

Frequently for these studies, a stepwise testing scheme can be observed: starting with a small sample size pilot study, continuing with a larger sample size extended pilot study, and decision making by a final per-protocol study. The reasons for this stepwise procedure are:

- The *clinically relevant endpoint* is unknown a priori. This primary endpoint should be selected from some candidates using this study,
- The *variance* σ is unknown a priori. The same problem occurs in clinical trials as described by Gould (2),
- The *relevant effect difference* δ (either to negative control or to a competitor's drug) is unknown a priori,
- The *effective dose* is unknown a priori in dose-finding studies, and
- The *relevant time* is unknown a priori within the usually repeated measurements.

In the real data situation two approaches seem to be possible:

1. Learning from pilot studies and designing the per-protocol study using this information, however, rejecting the pilot studies from the direct final decision making, and
2. Definition of an "internal" pilot study.

Here, the concept of "adaptive interim analysis" according to Bauer and Köhne (1) will be used considering the information from the pilot study directly into the test decision.

ADAPTIVE INTERIM ANALYSIS

The principle of the adaptive interim analysis according to Bauer and Köhne (1) is as follows:

- Two independent randomized studies are to be performed: pilot and second ("main") study,
- Use all suitable information from the small sample size (undesigned) pilot study to adapt the second (main) study for new design and sample size: relevant endpoint, relevant dose group, relevant time point, relevant effect difference δ, variance σ, p-value of the first pilot study p_1, and so forth,
- Combine both studies to a global result by Fisher's product criterion based on both p-values: $p_1 p_2 \le c_\alpha$ (with p_i . . . p-value of the ith study with $i \in \{1,2\}$ [An extension to $i = 3$ is given in Bauer and Köhne (1) too]). That means that the null-hypothesis of no effect will be rejected, if the product of both p-values is smaller than a predefined critical value c_α. The critical value c_α was determined by Bauer and Köhne (1) as a modification of Fisher's product criterion taking the early stopping rules due to ineffectiveness or effectiveness in the first study into account, and
- Use early stopping rules: due to ineffectiveness, if $p_1 > \alpha_{upper}$, then H_0 is accepted; respective due to effectiveness, if $p_1 < \alpha_{lower}$, then H_0 is rejected, otherwise adapt the sec-

ond study. For an experimentwise type I error rate $\alpha_{exp} = 0.05$ possible values for α_{upper} and α_{lower} are given in Table 1 according to Bauer and Köhne (1).

A practical choice for pharmacological studies could be $\alpha_{upper} = 0.5$ (decision for early stopping due to ineffectiveness if the effect tends in the wrong direction) and based on $c_\alpha = 0.0087$ a level for $\alpha_{lower} = 0.0233$ results. An extension to three studies is given in Bauer and Köhne (1) but should not be discussed here.

Two questions arise for this approach; how large is the loss of power of this two-step approach in comparison with a one-step uniformly most powerful test and which price must be paid for this approach? The first question was answered by Banik et al. (3): the loss of power is practically negligible down to $N_{global} = n_1 + n_2 = 12$. Therefore, an application in pharmacological studies is not limited. The price to be paid is:

1. A combination of p-values instead of pooling the data, that is, two independent studies still exist, also for descriptive statistics, but not a unique study with pooled data,
2. More than double time is necessary, and
3. Only the intersection of both hypotheses $H_{01} \cap H_{02}$ can be rejected and, therefore, the interpretation is possible only for this intersection according to Bauer and Röhmel (4).

Several adaptations exist for the second study:

TABLE 1
Selected Values for
Adaptive Interim Analysis

$c_\alpha = 0.0087$	
α_{upper}	α_{lower}
0.3	0.0299
0.4	0.0263
0.5	0.0233
0.6	0.0207
0.7	0.0183

- Sample size n_2 (for sample size adaptations other approaches exist too, eg, according to Proschan and Hunsberger [5]),
- The primary endpoint with the objective to reduce the number of multiple endpoints, for example, using multiple endpoint tests of different dimensions in the pilot and main study,
- Dose groups with the objective to select the "best" and reduce the number of dose groups (see [4]),
- Dose steps: increase or decrease the distances between dose levels,
- Time points with the objective to select the "best" and reduce the number of time points, and
- The type of test statistics, for example, t-test in the primary and u-test in the secondary study after a heavily-skewed distribution was observed in the pilot study.

This list indicates that this approach is very flexible in adaptation because it is based on two independent randomized studies with their p-values only. The adaptations are primarily restricted by the interpretation of the outcome of Fisher s product criterion consisting of different tests in the first and secondary studies.

APPLICATION IN PHARMACOLOGY

In comparison with clinical trials, adaptive interim analysis can be easily used in pharmacology because of: less serious problems with unblinding; short-term studies, where the doubling of time is not too critical; and more missing a priori information about treatment effects, endpoints, dose levels, and so forth. With two examples the simple application of adaptive interim analysis in pharmacology is demonstrated.

Sample Size Adaptation

In a pharmacological study the novum Protease should be compared with the verum Alteplase based on the endpoint blood loss in a rabbit model of jugular vein thrombolysis (6). In an early stage of research the exper-

imental question was two-sided, that is, for a difference of blood loss. The decision on an appropriate two-sample test for this data is not simple. Because of small sample sizes and unknown distribution, a permutative U-test was used, although heterogeneity of variance exists. Table 2 presents the raw data for the undesigned internal pilot study.

Analyzing this data by StatXact (7), a p-value $p_1 = 0.132$ for the permutative U-test results. The following conditions for adaptive interim analysis were used (see Table 1): $\alpha_0 = 0.5$, $\alpha_{lower} = 0.0233$ and $c_\alpha = 0.0087$. Because $p_1 = 0.132$ is smaller than 0.5 and larger than 0.0233 no early stopping can be performed based on this pilot study. Therefore, sample size should be estimated for the second study. Sample size estimation was approximately performed for the asymptotic U-test using Noether's (8) approach. From this pilot study a variance estimation of $\sigma^2 \approx 2250000$ was assumed. Based on the predefined type II error $\beta = 0.20$ and the relevant effect difference defined a-priori based on medical information of $\delta = 2800$, sample sizes of $n_{2i} = 5$ were estimated for the local size of the second study $p_2 = 0.066$, based on $p_2 = c\alpha/p_1$.

From the raw data of the second study a p-value of $p_2 = 0.0434$ was estimated. Because of $p_1 * p_2 = 0.0057 < c_\alpha$ the global decision is that Protease reveals a significantly different blood loss in comparison to Alteplase. Moreover, two independent randomized studies demonstrate an effect in the same direction.

Multiple Endpoint Analysis

In a placebo-controlled pharmacological study the effect of Anti-L-Selectin should be analyzed on the endpoint's survival time and

TABLE 2
Blood Loss Raw Data for the Pilot Study

Substance	Blood loss
Alteplase	148, 4080, 8857, 1582, 6201, 4301
Protease	847, 1658, 299, 1983, 873, 424

TABLE 3
Survival Time and Lung Weight in the Pilot Study

Endpoint	Anti-L-Selectin	Placebo
Survival time	72, 72, 20, 68, 72, 72, 72, 72	38, 41, 72, 72, 72, 34, 48.5, 24
Lung weight	8.78, 8.97, 8.0, 4.78, 9.72, 11.18, 8.91, 8.47	16.49, 15.61, 11.14, 9.3, 7.56, 12.81, 8.19, 7.44

lung weight, using a subacute primate model of hypovolemic-traumatic shock with a three-day observation period (9). One-sided tests (increase of survival time and decrease of lung weight) should be performed in a multiple endpoint analysis. Table 3 presents the raw data for the undesigned internal pilot study.

Using rank transformed multiple endpoint analysis according to O'Brien (10) a p-value $p_1 = 0.145$ was estimated. The univariate p-values: survival time $p_{logrank\ test} = 0.086$ and lung weight $p_{u-test} = 0.399$ indicated that the more sensitive endpoint is survival time. Therefore, the test statistics were changed: from the multiple endpoint analysis containing both endpoints in the pilot study to the univariate test on survival time only (using logrank test) in the second study.

The same conditions for adaptive interim analysis as in example 1 were used (see Table 1): $\alpha_0 = 0.5$, $\alpha_{lower} = 0.0233$ and $c_\alpha = 0.0087$. No early stopping is possible. Therefore, sample size should be estimated for the second study, resulting in $n_{2i} = 5$. The second study revealed a p-value: $p_{2,survival\ time} = 0.0271$. Because of $p_1 * p_2 = 0.0039 < c_\alpha = 0.0087$ Anti-L-Selectin reveals a significant effect over placebo. Moreover, two independent randomized studies demonstrate an effect in the same direction.

CONCLUSIONS

Adaptive interim analysis represents a new technique in the case of uncertainty to make a clear decision: significant/nonsignificant effects. It is adapted to the experimental paradigm of pharmacologists: stepwise. Using this principle, one methodological dilemma can be overcome by answering the question in case of a nonsignificant study: "Is the substance ineffective or was the study design invalid, for example, was the sample size too small?" The application of adaptive interim analysis seems to be simpler in preclinical studies than in clinical trials because of fewer problems with study protocols and unblinding.

A price must be paid, however, for using this approach: combination of p-values instead of pooling the data, increase of necessary time, and interpretation of the intersection of both hypotheses only. To avoid abuse, in the study protocol the kind of adaptations to be intended should be fixed a priori.

Summarizing, adaptive interim analysis is a simple approach, frequently easy for pharmacologists to perform; only one table is needed (eg, Table 1), as well as p-values from usual tests, that is, t- or U-tests.

REFERENCES

1. Bauer P, Köhne K. Evaluation of experiments with adaptive interim analyses. *Biometrics.* 1994;50: 1029–1041.
2. Gould LA. Planning and revising the sample size for a trial. *Stat Med.* 1995;14:1039–1051.
3. Banik N, et al. On the power of Fisher's combination test for two stage sampling in the presence of nuisance parameters. *Biometrical J.* 1996;38:25–38.
4. Bauer P, Röhmel J. An adaptive method for establishing a dose-response relationship. *Stat Med.* 1995; 14:1595–1607.
5. Proschan MA, Hunsberger SA. Designed extension of studies based on conditional power. *Biometrics.* 1995;51:1315–1324.
6. Martin U, Kohnert U, et al. Effective thrombolysis by a recombinant Eschericha Coli produced protease domain of tissue-type plasminogen activator in the rabbit model of jugular vein thrombolysis. *Fibrinolysis.* 1996;10:87–92.
7. Mehta C, Patel N. StatXact 3 for Windows, Cytel Software Corp, Cambridge; 1995.

8. Noether GE. Sample size determination for some common nonparametric tests. *J Am Stat Assoc.* 1987; 82:645–647.

9. Schlag G, Redl H, Khakpour Z, Davies J, Pretorius J. Hypovolemic-traumatic shock models in baboons.

I: Schlag, G. and Redl. H. (eds.). *Pathophysiology of shock, sepsis and organ failure.* Berlin: Springer Verlag Berlin; 1993.

10. O'Brien PC. Procedures for comparing samples with multiple endpoints. *Biometrics.* 1984;40:1079–1087.