doi: 10.1111/j.1472-8206.2006.00437.x

REVIEW ARTICLE

How to deal with multiple endpoints in clinical trials

Markus Neuhäuser

Department of Mathematics and Technique, RheinAhrCampus Remagen, Koblenz University of Applied Science, Südallee 2, 53424 Remagen, Germany

Keywords
biostatistics,
multiple primary endpoints,
multiplicity

Received 15 November 2005;
revised 19 January 2006;
accepted 15 June 2006

Correspondence and reprints:
neuhaeuser@rheinahrcampus.de

ABSTRACT

Multiple endpoints are common in clinical trials. This article discusses statistical methods that can be applied to control the rate of false positive conclusions at an acceptable level. The considered methods include the Bonferroni adjustment and related methods, the intersection-union test, ordered hypotheses and gatekeeper procedures, composite endpoints and global assessment measures, closed testing procedures, and combinations of different approaches.

1. INTRODUCTION

Multiplicity of inferences can arise in a variety of settings in clinical trials. Examples are multi-sample comparisons, interim and subgroup analyses, and, particularly, multiple endpoints. Therefore, statistical issues for multiplicity have gained increasing importance. In this article, methods appropriate for multiple endpoints in clinical trials are discussed. The main focus is on two-armed trials.

The primary endpoints are the most important endpoints from both the scientific and the practical point of view [1]. They should provide 'the most clinically relevant and convincing evidence directly related to the primary objective of the trial' [2]. In order to avoid or reduce multiplicity one should select and predefine one or a few primary variables and should declare all remaining secondary variables supportive. The ICH E9 guideline on biostatistics in clinical trials [2] recommends selection of one primary variable. However, investigators typically examine more than one primary endpoint [3]. A single primary endpoint is not sufficient in many disease areas. An example for such a multifaceted disease is asthma [4]: anti-inflammatory treatment of chronic asthma should improve both the pulmonary function and the subjective outcomes reported by the patients.

It is well known that the likelihood of obtaining significant results, just by chance, increases considerably with the overall number of statistical tests carried out [5]. Therefore, the appropriate threshold to declare a test's p-value significant becomes complex when more than one test is performed. When a significant difference in at least one endpoint is sufficient to claim a treatment effect, the classical Bonferroni scenario is given. This approach may be meaningful in case of a very serious disease with a paucity of treatments [1]. In this case it is clearly inappropriate to declare statistical tests as significant based on unadjusted individual (single-test) p-values.

An appropriate multiple testing procedure should control the rate of false positive conclusions at an acceptable level. Assume that a global null hypothesis is made up of several subhypotheses. A multiple testing procedure *weakly* controls the type I error rate (i.e. the rate of false positive decisions) when the probability to reject the global null hypothesis, although all subhypotheses are true, is controlled at level α. The type I error rate is *strongly* controlled when the probability of erroneously rejecting any configuration of subhypotheses is

controlled by α, although some subhypotheses may be true. To be precise, strong control means that the probability of rejecting at least one true null hypothesis is not larger than α, irrespective of which and how many individual null hypotheses are true [4]. Any procedure that strongly controls the type I error rate will also provide weak control, but the converse is not true [4]. The strong control seems to be the more appropriate option [6,7]. It is the best protection against wrong conclusions and leads to the strongest statistical inference [8].

The outline of the paper is as follows: section 2 discusses the Bonferroni adjustment and related methods. Sections 3 and 4 review the intersection-union test, ordered hypotheses and gatekeeper procedures. Section 5 gives combinations of the different approaches. An example is presented in section 6. Section 7 covers composite endpoints and global assessment measures, whereas closed testing procedures and correlated endpoints are discussed in sections 8 and 9. Section 10 is devoted to the false discovery rate. Section 11 is a discussion.

2. THE BONFERRONI ADJUSTMENT AND RELATED METHODS

We assume that a collection of k tests is simultaneously carried out. To perform some of the methods discussed in this section the p-values of the k tests have to be ordered from the smallest $p_{(1)}$ to the largest $p_{(k)}$, i.e. $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(k)}$.

According to the classical Bonferroni technique [9], each single test's p-value has to be compared with the adjusted level of α/k rather than α. For instance, in case of 10 tests and the usual $\alpha = 0.05$ the new threshold is $\alpha/10 = 0.005$. Thus, the Bonferroni procedure partitions α evenly among the tests. Instead to compare each p-value with α/k one can calculate an adjusted p-value, i.e. each observed p-value is multiplied by k. Obviously, the adjusted p-values have to be compared with α. As mentioned above the classical Bonferroni scenario is given when a significant difference in at least one endpoint is sufficient to claim a treatment effect. Thus, one p-value smaller than or equal to α/k is sufficient even when the majority of the tests give large p-values.

Holm [9,10] developed a sequentially rejective procedure that provides a more powerful test than the standard Bonferroni method. Note that procedures that find more differences are more powerful than those that find fewer differences. Both adjustments, Bonferroni and Bonferroni-Holm, provide strong control of the type I error rate. If the smallest p-value $p_{(1)} \leq \alpha/k$, then the null hypothesis of the corresponding test can be rejected. With regard to this single test there is no difference between the Bonferroni and the Bonferroni-Holm methods. If $p_{(1)} > \alpha/k$, the procedure stops and all tests are non-significant. However, in case of $p_{(1)} \leq \alpha/k$ the second-smallest p-value $p_{(2)}$ can be compared with $\alpha/(k-1)$. If $p_{(2)} > \alpha/(k-1)$ this test and all other test with larger p-values cannot reject their null hypotheses. If $p_{(2)} \leq \alpha/(k-1)$, then the corresponding test is significant and the procedure proceeds until $p_{(i)} > \alpha/(k-i+1)$ or until the last null hypothesis is rejected in case of $p_{(k)} \leq \alpha$.

All null hypotheses which can be rejected with the Bonferroni method can also be rejected with the Bonferroni-Holm technique. However, with the latter technique additional null hypotheses may be rejected. Therefore, Holm's method provides a substantial increase in power when some null hypotheses are rejected with a high probability, as the remaining p-values are compared with less stringent levels. It should be noted that, within the Bonferroni-Holm procedure, it is possible to assign different weights to the endpoints according to their importance [4].

Sometimes, there are logical implications among the hypotheses being tested. When using these implications the Bonferroni-Holm method can be improved to obtain a further increase in power [9,11]. Assume that $j-1$ hypotheses have been rejected. Then, one has to determine t_j, the maximum number of hypotheses that could be true, given that at least $j-1$ hypotheses are false. This maximum number t_j, which obviously is $\geq k-(j-1)$, can be used in the denominator of the α-level instead of $k-(j-1)$, which would be used in Holm's procedure as the denominator is reduced by 1 each time a null hypothesis is rejected. However, this procedure proposed by Shaffer [11] is rarely helpful when testing multiple endpoints because the different endpoints are usually not logically interrelated.

Another procedure was introduced by Hochberg [12]. That procedure contrasts the ordered p-values with the same set of critical values as the Bonferroni-Holm method. Hochberg's procedure rejects all null hypotheses with smaller or equal p-values to that of any found less than its critical value [12]. According to Hochberg's procedure all k null hypotheses can be rejected if $p_{(k)} \leq \alpha$. If $p_{(k)} > \alpha$, the null hypothesis corresponding to $p_{(k)}$ cannot be rejected and one has to compare $p_{(k-1)}$ with $\alpha/2$. If $p_{(k-1)} \leq \alpha/2$, the null hypotheses corresponding to $p_{(k-1)}, p_{(k-2)}, \ldots, p_{(1)}$ are rejected. If $p_{(k-1)} > \alpha/2$,

the corresponding null hypothesis cannot be rejected and the comparison $p_{(k-2)}$ with $\alpha/3$ follows, etc. [12].

Hommel [13] introduced a further procedure that is always more powerful than Hochberg's procedure (for $k = 2$ both procedures coincide). Hommel's procedure starts with the computation of

$$j = \max\left\{i \in 1,\ldots,k : p_{(k-i+m)} > \frac{m\alpha}{i} \text{ for } m = 1,\ldots,i\right\}.$$

All null hypotheses can be rejected if the maximum j does not exist. Otherwise, all null hypotheses corresponding to p-values with $p_i \leq \alpha/j$ can be rejected [13]. These procedures [12,13] are sharper than Bonferroni-Holm's one. However, the improved power of both procedures with respect to the Bonferroni-Holm procedure comes at the cost of having to make additional assumptions, e.g. independence between the p-values.

In the classical Bonferroni scenario it is sufficient that one out of k endpoints is significant. That means the experimental treatment must be superior for at least one endpoint. However, what if the experimental treatment is extremely negative on another endpoint? To deal with such a scenario, a more stringent requirement is that the experimental treatment is superior for at least one of the endpoints and not biological inferior for the remaining endpoints. This approach may be especially justified in certain studies in areas such as arthritis, rheumatism, or cancer, when treatment effects are measured by both efficacy and toxicity and both efficacy and toxicity may be measured by more than one endpoint [14]. Tests for this test problem were proposed by Bloch et al. [14] and Perlman and Wu [15]. In both papers an intersection-union test is used. The principle of an intersection-union test will be explained in the following section.

3. THE INTERSECTION-UNION TEST

When a significant difference in at least one endpoint is sufficient to claim a treatment effect, the global null hypothesis is the intersection of all individual null hypotheses. Thus, the global null hypothesis is the collection of all individual null hypotheses (i.e. the set of the elements that are in all individual null hypothesis). This global null hypothesis is tested vs. the global alternative which is the union of all individual alternative hypotheses. Thus, the global alternative is the set of all of the elements that are contained in at least one individual alternative hypothesis. Therefore, the classical Bonferroni adjustment is appropriate within a so-called union-intersection test [16].

However, if the decision rule is that a statistical significance is needed for all primary variables, the global null hypothesis can be expressed as the union of all single hypotheses regarding the individual endpoints. This union is tested vs. the global alternative which can be expressed as the intersection of all single alternative hypotheses. As all individual null hypotheses have to be rejected to claim significance, no multiplicity adjustment, i.e. no special method to consider the multiplicity of tests (such as e.g. the Bonferroni-Holm procedure), is needed. This procedure is called intersection-union test [16]. Within an intersection-union test all tests can be performed with the unadjusted α-level, but no single p-value is allowed to be larger than α. If one p-value is larger than α, one cannot reject any null hypothesis. This procedure inflates the type II error rate, i.e. there is an increased probability of not rejecting a false null hypothesis. This inflation must be taken into account for sample size determination [17].

A widespread application of an intersection-union test is the assessment of comparative bioavailabilities in a bioequivalence trial [18]. In such studies the issue is to reject the null hypothesis of non-equivalence with regard to two pharmacokinetic variables that characterize extent and rate of absorption, for instance area under the concentration/time curve (AUC) and maximum concentration (C_{max}).

Further examples are clinical trials in patients with chronic obstructive pulmonary disease (COPD) or Alzheimer's disease because the respective CPMP Points to Consider (19) and Note for Guidance (20) request that two primary variables are needed to describe clinically relevant treatment benefits. In COPD a significant benefit for both forced expiratory volume in 1 s (FEV_1) and a symptomatic benefit endpoint is requested.

4. ORDERED HYPOTHESES AND GATEKEEPER PROCEDURES

Similar to the case of an intersection-union test, an adjustment of the significance level is not necessary when multiple hypotheses are to be tested in a prespecified order according to their clinical relevance. Then each hypothesis can be tested in the preassigned order at full level α, as long as the prior null hypotheses have been rejected [21,22]. If a p-value is larger than α, the procedure has to stop. Such hierarchical testing procedures are also known as 'gatekeeping procedures' [23,24].

This method was applied, for example, in a placebo-controlled clinical trial to investigate the effect of roflumilast in the treatment of allergic rhinitis [25]. The primary variable was the rhinal airflow determined by rhinomanometry at seven consecutive days. First, the treatment difference was tested with the data of day 7. As there was a significant difference, a test with the data of day 6 could be performed, and so on backward to the day for which a significant difference could no longer be detected. In this example, the hypotheses were ordered in time. However, in other settings, the hierarchical order may be based on the seriousness of the considered endpoints or may result from the particular interests of the investigator [17]. In any case, no confirmatory claims can be based on variables that have a rank lower than or equal to that endpoint whose null hypothesis was the first that could not be rejected [17].

Another application of ordered hypotheses is a test for superiority in case of a proven non-inferiority according to Morikawa and Yoshida [26]. To be precise, in a study with an active control, non-inferiority of the new drug is tested first. If non-inferiority is demonstrated, a test for superiority of the new drug over the active control can follow without the need to adjust the α-level.

Usually, confirmatory claims are not based on secondary endpoints. When statistical tests are applied to secondary endpoints their analysis has an exploratory character and p-values are used as a descriptive measure for the difference between groups. However, a hierarchical order can be used in order to include secondary endpoints within the confirmatory strategy. Of course, within the hierarchical order the primary endpoints must have a higher rank than the secondary ones. Thus, secondary endpoints can be tested only after the primary objective of the clinical trial has been demonstrated.

Dmitrienko et al. [24] and Chen et al. [27] proposed procedures where one may proceed to the secondary endpoints when at least one, not necessarily all, of the primary endpoints exhibits significance. Of course, a multiplicity adjustment is necessary within such a parallel gatekeeping procedure in which a complete order of the primary hypotheses is not necessary. An example provided by Dmitrienko et al. [24] is a placebo-controlled clinical trial in patients with acute respiratory distress syndrome. Primary endpoints are the number of ventilator-free days and the 28-day all-cause mortality whereas the number of days the patients were out of the intensive care unit and quality of life are considered as secondary endpoints.

5. COMBINATIONS OF DIFFERENT APPROACHES

The usual application of the Bonferroni procedure is to test each of the k null hypotheses at the level $\alpha_i = \alpha/k$. However, any value between 0 and 1 is possible for an α_i as long as $\sum_{i=1}^{k} \alpha_i = \alpha$, i.e. the sum of all α_is is α. Such an unequal allocation is always subject to criticism because it is arbitrary. As well, it relies on the honesty of those reporting. Wiens [28] used such an unequal allocation to present the following approach that combines the Bonferroni adjustment with ordered hypotheses.

The a priori ordered null hypotheses are denoted by H_0^1, H_0^2,..., H_0^k, and α_i' is assigned to the ith hypothesis H_0^i, with $\sum_{i=1}^{k} \alpha_i' = \alpha$. Then, H_0^1 is tested at the level $\alpha_1 = \alpha_1'$. Subsequent hypotheses H_0^i are tested at the level $\alpha_i = \alpha_i'$ if $H_0^{(i-1)}$ was not rejected. If $H_0^{(i-1)}$ was rejected, H_0^i can be tested at the level $\alpha_i = \alpha_i' + \alpha_{i-1}$. The usual testing of ordered hypotheses discussed above is a special case with $\alpha_1 = \alpha$ and $\alpha_i = 0$ for all $i > 1$ [28,29].

The example discussed by Wiens [28] is a study in symptomatic heart failure with functional capacity as the primary endpoint. Mortality is a secondary endpoint and, as Wiens [28] pointed out, a strong impact on mortality would be important to know whether or not the primary endpoint is significant. Assume $\alpha_1' = 0.04$ for functional capacity and $\alpha_2' = 0.01$ for mortality. If the primary endpoint was significant at the $\alpha_1' = 0.04$ level, mortality can be tested using the $\alpha_2 = \alpha_2' + \alpha_1 = 0.05$ level. If the primary endpoint was not significant, mortality could not be tested in the usual fixed testing sequence procedure. However, in the procedure proposed by Wiens [28] mortality can be tested at the $\alpha_2 = \alpha_2' = 0.01$ level. The price to pay is a reduced power for the first test because functional capacity is tested at the 0.04 level instead at the full level 0.05.

Instead of ordering them, the different primary endpoints may sometimes be classified into two or more groups of endpoints. This may be justified when the clinical objective is to show at least one positive endpoint in each group of endpoints. Let us consider clinical asthma trials. As mentioned above, the anti-inflammatory treatment of chronic asthma should improve both the pulmonary function and the subjective outcomes reported by the patients [1,30]. Thus, the two endpoint groups are pulmonary function variables such as FEV_1 and the peak expiratory flow rate (PEF), and patient

recorded outcomes such as asthma-specific symptom scores and the use of rescue medication.

When the clinical objective is to show at least one positive effect in each group, a combination of the intersection-union principle with a within-group Bonferroni adjustment is suitable. To be precise, the null hypothesis that there is no effect in at least one group can be rejected if, in each group of endpoints, at least one endpoint is significant at the $\alpha/2$-level.

According to the intersection-union principle, the method guarantees strong control as a level-α test is used within each group. Neuhäuser et al. [30] presented the actual size in dependence on the correlation as well as a power simulation study. The power can be slightly increased when using the Simes [31] instead of the Bonferroni adjustment. In case of two tests there is a significance according to the Simes method [31] if the smallest p-value, i.e. $p_{(1)}$, is $\leq\alpha/2$ or if both p-values are $\leq\alpha$ (under additional assumptions, see also [32] and [33]). Thus, in each group of endpoints, at least one endpoint must be significant at the $\alpha/2$-level or both endpoints must be significant at the α-level.

6. EXAMPLE

The example discussed here was presented by Zhang et al. [4] and comes from a randomized, multicenter, double-blind, parallel design clinical trial in asthmatic patients. There were 34 patients in the test drug group and 35 patients in the placebo group. Four endpoints are considered: FEV_1, PEF, symptom scores, and use of rescue medication. The corresponding p-values of two-sided tests are 0.0037, 0.0077, 0.0274, and 0.0369.

Usually, investigators in medical research set α at 0.05 [3]. All four unadjusted p-values are smaller than this commonly applied significance level of 0.05. Therefore, if the intersection-union test was selected, one could claim a difference regarding all four endpoints. Similarly, there would be a difference regarding all endpoints if the fixed testing sequence procedure was selected, irrespective which ordering was prespecified. However, the Bonferroni adjustment could detect a difference only with regard to FEV_1 and PEF because their p-values are smaller than $\alpha/4 = 0.0125$. The other two p-values are larger than 0.0125. In this example, the Bonferroni-Holm method would not provide an advantage. A difference would be detected for FEV_1 and PEF because their p-values are smaller than $\alpha/4 = 0.0125$ and $\alpha/3 = 0.0167$, respectively. However, the third smallest p-value, 0.0274, is larger than $\alpha/2 = 0.025$, therefore, no additional significance is found by Bonferroni-Holm.

When the four endpoints are classified as described in section 5, one could detect a difference if the Simes method was selected as the within-group adjustment. Then, because all p-values are smaller than 0.05, the condition that both endpoints are significant at the 0.05 level is fulfilled for both groups of endpoints.

7. COMPOSITE ENDPOINTS AND GLOBAL ASSESSMENT MEASURES

Sometimes multiple measurements can be summarized in one summary variable. For example, when an endpoint is repeatedly measured over time the AUC can be calculated. The AUC is a composite, but univariate endpoint for which no multiplicity adjustment is necessary. AUCs are, for example, commonly applied in bioequivalence studies where the serum concentrations are repeatedly measured after drug intake [34]. An alternative is to simply average the measurements of different time points during the treatment period, as applied, for instance, by Evans et al. [35].

Several different endpoints may be combined into one composite endpoint. Of course, one has to define prospectively the components of a composite endpoint. Consider a clinical trial in patients with unstable angina: an event may be defined as occurrence of any one of, for example, death, myocardial infarction, recurrent angina, urgent intervention, etc. In such a case, the increased overall number of events can give an increased power [36]. Another example of a composite endpoint is a global disease score that is common, for example, in trials for inflammatory bowel disease and usually includes several components [36].

Composite endpoints are challenging for the interpretation of results [37]. Therefore, the individual components should be additionally analyzed in case of a significant difference in the composite endpoint. However, even when all-cause mortality is incorporated into the composite primary outcome the reporting of outcomes is generally inadequate as Freemantle et al. [37] found in their review. When the components are tested after a significance in the composite endpoint the question arises whether an adjustment is necessary. A reviewer suggested declaring individual components as secondary endpoints, and not adjust. However, results on secondary endpoints might need to be confirmed in further studies.

Another approach is to combine results from different endpoints in one test statistic rather than to define a composite variable that combines several endpoints. Hotelling's T^2 test is the generalization of Student's t-test for this multivariate scenario. This test assumes that the endpoints are normally distributed, but it is reasonably robust against departures from normality [38, p. 365]. In clinical trials, one can often assume that the different endpoints show a treatment difference, if any, of the same direction. Then, tests introduced by O'Brien [39] are more powerful than Hotelling's T^2 test. O'Brien's tests (see [4] and [40] for an overview) are powerful when all endpoints have similar treatment effect sizes, but even the Bonferroni adjustment provides more power if one variable has a very small effect while another one has a much larger effect [4]. Sankoh et al. [36] also noted that O'Brien's tests may not be optimal if a consistent treatment effect is not expected across a majority of the endpoints. However, it should be mentioned that improvements of the asymptotic O'Brien method exist: so-called stable tests proposed by Läuter et al. [41] can lead to exact level α tests.

The global measures do not provide specific information on the variables contributing to a possibly significant difference. Please note in this context that O'Brien [39] proposed his methods as a supplement, not an alternative, to univariate tests. However, when there are more outcome variables than patients it may be preferable to create some subgroups of variables rather than to test each individual endpoint. O'Brien [39] discussed as an example a randomized trial comparing two therapies for the treatment of diabetes. There were 34 variables, but only 11 patients in total. In that study seven subgroups of variables were formed (without viewing the data) in order to understand the nature of the treatment effect [39].

8. CLOSED TESTING PROCEDURES

When the 30-year-old concept of a closed testing procedure [42, see also 20] is applied, all tests can be performed at the level α, but additional tests may be necessary. To be precise, the set of null hypotheses to be tested must be closed under intersection, i.e. the intersection of any two of the null hypotheses must be an element of the set of null hypotheses. To be a closed set, often additional hypotheses must be included. A null hypothesis H_0^i is rejected if it was rejected in the corresponding level α test, and if all other null hypotheses that imply H_0^i were also rejected in their level α tests.

The closed test can be performed in a step-down procedure. It starts with testing the global null hypothesis, i.e. the intersection of all null hypotheses. In case the global null hypothesis is rejected at level α, the tests proceed to the intersection null hypotheses one stage lower. Detailed examples are given by Kropf [43]. Step-down procedures for determining which endpoints differ following a significant global test are also discussed by Geller [40] in good detail. A closed testing procedure can also be applied to O'Brien's methods [44].

The closed testing procedure strongly controls the type I error rate. It comprises many well-known procedures as special cases. However, a closed test can have some disadvantages. On the one hand, the tests concerning single endpoints may not be significant although there is a global significance, see, for instance, Kropf's [43] second example. On the other hand, a closed testing procedure may be cumbersome to handle, especially when there are a lot of endpoints, and in particular in multiple armed trials. The reason is that the number of additional hypotheses to be included in order to form the closure may increase exponentially [45]. Furthermore, corresponding confidence intervals are, in general, not available.

However, it should be noted that not all mentioned disadvantages hold for each closed testing procedure. For example, the Bonferroni-Holm method, a special case of a closed test, is not cumbersome to handle. Its advantages are its simplicity and the possibility to identify significant endpoints.

9. CORRELATED ENDPOINTS

As mentioned above multiple endpoints are usually correlated. Most of the methods presented so far are applicable whether or not the endpoints are independent. However, methods such as the Bonferroni adjustment are overly conservative in case of highly correlated endpoints. As a result, power is lost. This conservatism and the accompanied lack of power when the endpoints are correlated is the major drawback of the Bonferroni procedure [4].

Adjustment procedures for testing multiple correlated endpoints are less common. James [46] introduced a p-value adjustment approximation based on the standard multivariate normal distribution [see also 47,48]. The disadvantage of the James adjustment is that an equal pairwise correlation among the multiple endpoints is assumed, i.e. any two of the multiple endpoints have the same correlation coefficient. This assumption is rather

unrealistic as pointed out by James [46] and shown by Neuhäuser et al. [30] for an example. An alternative is to estimate the correlations from a previous study or a pilot trial [46]. However, sometimes neither data of a previous study nor of a pilot study are available. Then, it may be possible to estimate the correlations from the actual data, which leads at least to an asymptotic test. Please note that the numerical treatment of multivariate *t*-probabilities is possible using methods described by Genz and Bretz [49].

So-called resampling methods are based upon repeated sampling within the same sample, for instance, the data may be permuted (shuffled) to create multiple new pseudo data sets [see e.g. 50–52 for details]. These methods do not require any assumption or prior estimation of the correlation between the endpoints. In a resampling-based multiple test adjusted *p*-values are computed. The adjusted *p*-value for the *i*th test is a function of the *i*th test statistic and depends on the joint distribution of all test statistics [4]. This joint distribution heavily depends on the correlations between the endpoints which are typically unknown. Therefore, resampling techniques such as bootstrap or permutation approaches [50] are applied. Details about these computationally intensive methods, including applications and software, are given by Westfall and Young [52].

10. THE FALSE DISCOVERY RATE

The procedures to control the type I error rate when multiple statistical tests are performed come at a high cost: a reduction in power. The power for an individual test may become very low when the number of tests increases [53]. Therefore, in procedures that control the multiple significance level only relatively strong effects are likely to be recognized as significant when a lot of tests are carried out. That may be too stringent in purely exploratory studies, especially because the sample sizes are often not so large in these studies. Thus, analysis without multiplicity adjustment has been proposed for exploratory studies [8]. However, it is obvious that in that case results must be clearly labeled as exploratory ones. To confirm these results, the found effects have to be tested in confirmatory studies [8].

Here, we will briefly discuss an alternative approach: the control of the so-called false discovery rate. When many tests are performed in an exploratory study, keeping the proportion of type I errors (false discoveries) at a low level may be an alternative to controlling the chance of making even a single type I error. The control of the false discovery rate (FDR) means that the expected proportion 'number of type I errors/number of significant tests' is maintained at a desired level. This approach was developed by Benjamini and Hochberg [54] and is an active field of research [53]. Control of the FDR is also possible when the multiple endpoints are, as usually, correlated [55]. For instance, control of the false discovery rate is being widely adopted in genomic research [53,56]. However, in contrast to exploratory studies, methods developed for controlling the false discovery rate cannot be recommended for the analysis of clinical trials in regulatory settings.

11. DISCUSSION

Although there are methods to avoid or reduce multiplicity, for example the identification of one or two primary variables, multiple test problems often occur in clinical trials. Different procedures are available to consider multiplicity. The main approaches can be distinguished: single-step procedures and stepwise procedures. In a single-step procedure such as the Bonferroni technique each single test is tested without reference to any other. In stepwise procedures, such as testing in a hierarchical order, the decision on already tested hypotheses influences whether subsequent tests are conducted. In practice, the choice of a special procedure depends on the clinical objective and has to be identified in the study protocol. Note that several of the discussed approaches are implemented in the SAS procedures MULTTEST and GLIMMIX (SAS Institute Inc., Cary, NC, USA).

The CPMP Points to Consider on multiplicity issues in clinical trials [17] recommend considering whether a method allows a satisfactory clinical interpretation. Confidence intervals of effects have to be given in addition to the results of statistical tests. These confidence intervals must be consistent with the tests and may not be available for many of the more complex procedures [17]. Hence, simple methods such as the Bonferroni adjustment have an advantage in this context.

The primary endpoint(s) of a clinical study will usually be efficacy variable(s) (ICH E9). Safety variables may sometimes be part of the confirmatory strategy. In that case the primary safety endpoint(s) should be treated as efficacy endpoints [17]. However, safety and tolerability variables are more often used as a flagging device to signal potential risks caused by the investigated treatments. Then, no adjustment for multiplicity is justified [17]. A multiplicity adjustment would be counter-

productive in such a case because the power is decreased by that adjustment. A different approach very recently proposed is a multivariate test that compares the proportions of side effects of two treatments in a manner similar to Hotelling's T^2 test [57].

In general, there is no consensus in which cases multiple test procedures are appropriate. However, in confirmatory clinical trials with a prespecified objective represented by multiple endpoints, the use of multiple test procedures is mandatory [7,8]. If, in such a situation, no multiplicity adjustment was planned, an explanation of why adjustment is not thought to be necessary should be given [2]. Finally, note that not only the ICH E9 guideline [2], but also the CONSORT Statement [58] recommends a multiplicity adjustment in case of multiple testing.

ACKNOWLEDGEMENTS

The author thanks two anonymous referees for helpful comments and suggestions.

REFERENCES

1 Capizzi T., Zhang J. Testing the hypothesis that matters for multiple primary endpoints. Drug Inf. J. (1996) **30** 949–956.
2 ICH E9 Expert Working Group. ICH harmonised tripartite guideline: statistical principles for clinical trials. Statist. Med. (1999) **18**, 1905–1942.
3 Schulz K.F., Grimes D.A. Multiplicity in clinical trials I: endpoints and treatments. Lancet (2005) **365** 1591–1595.
4 Zhang J., Quan H., Ng J., Stepanavage M.E. Some statistical methods for multiple endpoints in clinical trials. Control. Clin. Trials (1997) **18** 204–221.
5 Beck-Bornholdt H.P., Dubben H.H. Potential pitfalls in the use of *p*-values and in interpretation of significance levels. Radiother. Oncol. (1994) **33** 171–176.
6 Tamhane A.C. Multiple comparisons, in: Ghosh S., Rao C.R. (Eds), Handbook of statistics, Vol. 13. Elsevier, Amsterdam, 1996, pp. 587–630.
7 Sankoh A.J., Huque M.F., Dubey S.D. Some comments on frequently used multiple endpoint adjustment methods in clinical trials. Statist. Med. (1997) **16** 2529–2542.
8 Bender R., Lange S. Multiple test procedures other than Bonferroni's deserve wider use. BMJ (1999) **318** 600.
9 Hochberg Y., Tamhane A.C. Multiple comparison procedures. Wiley, New York, USA, 1989.
10 Holm S. A simple sequentially rejective multiple test procedure. Scand. J. Statist. (1979) **6** 65–70.
11 Shaffer J.P. Modified sequentially rejective multiple test procedures. J. Am. Statist. Assoc. (1986) **81** 826–831.
12 Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. Biometrika (1988) **75** 800–802.
13 Hommel G. A stagewise rejective multiple test procedure based on a modified Bonferroni test. Biometrika (1988) **75** 383–386.
14 Bloch D.A., Lai T.L., Tubert-Bitter P. One-sided tests in clinical trials with multiple endpoints. Biometrics (2001) **57** 1039–1047.
15 Perlman M.D., Wu L. A note on one-sided tests with multiple endpoints. Biometrics (2004) **60** 276–280.
16 Casella G., Berger R.L. Statistical inference. Duxbury Press, Belmont, MA, USA, 1990.
17 CPMP. Points to consider on multiplicity issues in clinical trials. EMEA, London, UK. Biom. J. (2001) **43** 1039–1048.
18 Berger R.L., Hsu J. Bioequivalence trials, intersection-union tests, and equivalence confidence sets (with discussion). Statist. Sci. (1996) **11** 283–319.
19 CPMP. Points to consider on clinical investigation of medicinal products in the chronic treatment of patients with chronic obstructive pulmonary disease (COPD). EMEA, London, UK, 1999.
20 CPMP. Note for guidance on medicinal products in the treatment of Alzheimer's disease. EMEA, London, UK, 1997.
21 Bauer P. Multiple testing in clinical trials. Statist. Med. (1991) **10** 871–890.
22 Maurer W., Hothorn L.A., Lehmacher W., Multiple comparisons in drug clinical trials and preclinical assays: a-priori ordered hypotheses, in: Vollmar J. (Ed.), Testing principles in clinical and preclinical trials, G. Fischer Verlag, Stuttgart, Germany, 1995, pp. 3–18.
23 Westfall P.,Krishen A. Optimally weighted, fixed sequence and gatekeeper multiple testing procedures. J. Statist. Plan. Inf. (2001) **99** 25–41.
24 Dmitrienko A., Offen W.W., Westfall P.H. Gatekeeping strategies for clinical trials that do not require all primary effects to be significant. Statist. Med. (2003) **22** 2387–2400.
25 Schmidt B.M.W., Kusma M., Feuring M. et al. The phosphodiesterase 4 inhibitor roflumilast is effective in the treatment of allergic rhinitis. J. Allergy Clin. Immunol. (2001) **108** 530–536.
26 Morikawa T., Yoshida M. A useful testing strategy in phase III trials: combined test of superiority and test of equivalence. J. Biopharm. Statist. (1995) **5** 297–306.
27 Chen X., Luo X., Capizzi T. The application of enhanced parallel gatekeeping strategies. Statist. Med. (2004) **24** 1385–1397.
28 Wiens B.L. A fixed sequence Bonferroni procedure for testing multiple endpoints. Pharm. Statist. (2003) **2** 211–215.
29 Wiens B.L., Dmitrienko A. The fallback procedure for evaluating a single family of hypotheses. J. Biopharm. Statist. (2005) **15** 929–942.
30 Neuhäuser M., Steinijans V.W., Bretz F. The evaluation of multiple clinical endpoints, with application to asthma. Drug Inf. J. (1999) **33** 471–477.
31 Simes R.J. An improved Bonferroni procedure for multiple tests of significance. Biometrika (1986) **73** 751–754.
32 Hochberg Y., Rom D. Extensions of multiple testing procedures based on Simes' test. J. Statist. Plan. Inf. (1995) **48** 141–152.
33 Samuel-Cahn E. Is the Simes improved Bonferroni procedure conservative? Biometrika (1996) **83** 928–933.

34 Sauter R., Steinijans V.W., Diletti E., Böhm A., Schulz H.U. Presentation of results from bioequivalence studies. Int. J. Clin. Pharmacol. (1992) **30** (Suppl. 1), S7–S30.
35 Evans D.J., Taylor D.A., Zetterström O., Chung K.F., O'Connor B.J., Barnes P.J. A comparison of low-dose inhaled budesonide plus theophylline and high-dose inhaled budesonide for moderate asthma. N. Engl. J. Med. (1997) **337**, 1412–1418.
36 Sankoh A.J., D'Agostino R.B., Huque M.F. Efficacy endpoint selection and multiplicity adjustment methods in clinical trials with inherent multiple endpoints issues. Statist. Med. (2003) **22** 3133–3150.
37 Freemantle N., Calvert M., Wood J., Eastaugh J., Griffin C. Composite endpoints in randomized trials: greater precision but with greater uncertainty? JAMA (2003) **289** 2554–2559.
38 Armitage P., Berry G. Statistical methods in medical research. Blackwell Scientific Publications, Oxford, UK, 1994.
39 O'Brien P.C. Procedures for comparing samples with multiple endpoints. Biometrics (1984) **40** 1079–1087.
40 Geller N.L. Design and analysis of clinical trials with multiple endpoints. in: Geller N.L. (Ed.), Advances in clinical trial biostatistics, Marcel Dekker, New York, USA, 2004, pp. 101–119.
41 Läuter J. Exact *t* and *F* tests for analyzing studies with multiple endpoints. Biometrics (1996) **52** 964–970 [correction: *Biometrics* (2000) **56** 324].
42 Marcus R., Peritz E., Gabriel K.R. On closed testing procedures with special reference to ordered analysis of variance. Biometrika (1976) **63** 655–660.
43 Kropf S. Application of multiple test procedures to the combination of multivariate and univariate tests with varying variable sets. Biom. J. (1988) **30** 461–470.
44 Lehmacher W., Wassmer G., Reitmeir P. Procedures for two-sample comparisons with multiple endpoints controlling the experimentwise error rate. Biometrics (1991) **47** 511–521.
45 Pigeot I. Basic concepts of multiple tests – a survey. Statist. Papers (2000) **41** 3–36.
46 James S. Approximate multinormal probabilities applied to correlated multiple endpoints in clinical trials. Statist. Med. (1991) **10** 1123–1135.
47 Arani R.B., Chen J.J. A power study of a sequential method of *P*-value adjustment for correlated continuous endpoints. J. Biopharm. Statist. (1998) **8** 585–598.
48 Leon A.C., Heo M. A comparison of multiplicity adjustment strategies for correlated binary endpoints. J. Biopharm. Statist. (2005) **15** 839–855.
49 Genz A., Bretz F. Comparison of methods for the computation of multivariate *t* probabilities. J. Comput. Graph. Statist. (2002) **11** 950–971.
50 Good P.I. Permutation, parametric and bootstrap tests of hypotheses. Springer, New York, USA, 2005.
51 Good P.I. Resampling methods. Birkhäuser, Boston, USA, 2006.
52 Westfall P.H., Young S.S. Resampling-based multiple testing: examples and methods for *P*-value adjustment. Wiley, New York, USA, 1993.
53 Verhoeven K.J.F., Simonsen K.L., McIntyre L.M. Implementing false discovery rate control: increasing your power. Oikos (2005) **108**, 643–647.
54 Benjamini Y., Hochberg Y. Controlling the false discovery rate – a practical and powerful approach to multiple testing. J. Roy. Statist. Soc. B (1995) **57** 289–300.
55 Benjamini Y., Yekutieli D. The control of the false discovery rate in multiple testing under dependency. Ann. Statist. (2001) **29** 1165–1188.
56 Allison D.B., Cui X., Page G.P., Sabripour M. Microarray data analysis: from disarray to consolidation and consensus. Nat. Rev. Genet. (2006) **7** 55–65.
57 Agresti A., Klingenberg B. Multivariate tests for comparing binomial probabilities, with application to safety studies for drugs. Appl. Statist. (2005) **54** 691–706.
58 Altman D.G., Schulz K.F., Moher D. et al. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. Ann. Intern. Med. (2001) **134** 663–694.