Drug Information Journal, Vol. 33, pp. 471–477, 1999
Printed in the USA.

0092-8615/99

THE EVALUATION OF MULTIPLE CLINICAL ENDPOINTS, WITH APPLICATION TO ASTHMA

MARKUS NEUHÄUSER AND VOLKER W. STEINIJANS
Byk Gulden Pharmaceuticals, Department of Biometry, Konstanz, Germany
FRANK BRETZ
University of Hannover, Department of Bioinformatics, Hannover, Germany

The situation where four primary endpoints are classified into two groups is considered, with the clinical objective to show at least one positive effect in each group. This multiple endpoint scenario may be appropriate in clinical asthma trials, with the two groups being pulmonary function variables and patient recorded outcomes. The statistical motivation for the classification is that the within-group correlation is usually higher than the between-group correlation. Three methods for the evaluation of this multiple test problem are proposed. A combination of the intersection-union principle with the Simes method leads to a powerful level-α test. Two clinical studies are used to illustrate the methods. Strategies are discussed for the incorporation of one further endpoint, that is, the premature drop-out rate due to lack of efficacy.

Key Words: Clinical trials; Multiple endpoints; Intersection-union test; α-adjustment; Asthma

INTRODUCTION

STATISTICAL ISSUES FOR multiplicity have gained increasing importance in clinical trials since it is often difficult to focus on only one primary variable. For example, anti-inflammatory treatment of chronic asthma should improve both the pulmonary function and the subjective outcomes reported by the patients.

There are some global assessment measures, for example, methods based on ordinary and generalized least squares which were proposed by O'Brien (1). Unfortunately, these global measures do not provide specific information on the variables contributing to the possibly significant difference. A closed testing procedure, however, can be applied to O'Brien's method (2,3), but the tests concerning single endpoints may not be significant although there is a global significance. Moreover, according to Zhang et al. (4), O'Brien's approach is powerful when all variables have similar treatment effect sizes, but even the Bonferroni adjustment is more powerful if one variable has a very small effect while another has a much larger effect. In the case of asthma trials there are differences in the effect sizes between different endpoints; this is described in the section called "Examples."

The authors consider the situation where the endpoints can be classified into different groups and that the clinical objetive is to

Reprint address: Dr. Markus Neuhäuser, Byk Gulden Pharmaceuticals, Department of Biometry, Byk-Gulden-Str. 2, D–78467 Konstanz, Germany.

show at least one positive effect in each group. A suggestion for the handling of multiple endpoints in the evaluation of asthma trials was given by Capizzi and Zhang (5), whose approach is presented next. A modified approach is introduced and compared with a Bonferroni-type and a Simes-type adjustment. The methods are applied to two clinical trials. Then the results are discussed and a further extension presented.

EVALUATION OF MULTIPLE ENDPOINTS

Decision Rules According to Capizzi and Zhang (5)

In the following, the suggestions of Capizzi and Zhang (5) for the evaluation of placebo-controlled asthma trials with multiple endpoints are described. There are four primary endpoints: the forced expiratory volume in one second (FEV_1), the peak expiratory flow rate (PEF), asthma-specific symptom scores, and the use of rescue medication. The endpoints can be classified into two groups: Pulmonary function: FEV_1 and PEF, and Patient recorded outcomes: symptom scores and use of rescue medication. All four endpoints are considered to be sensitive measures of the treatment effect.

Capizzi and Zhang (5, p. 954) provided two decision rules:

- *Capizzi and Zhang's rule 1:* In each group, one endpoint must be significant at the 0.05 level and the other endpoint has to trend in the right direction, that is, the other endpoint must be significant at the 0.2 level, and
- *Capizzi and Zhang's rule 2:* In each group, one endpoint must be significant at the 0.05 level and the other endpoint has to trend in the right direction, that is, the other endpoint must be significant at the 0.1 level.

According to Capizzi and Zhang (5), the objective is to show a positive effect in each of these four endpoints, that is, the statistical alternative is that there are treatment differences concerning all four endpoints.

Nevertheless, Capizzi and Zhang (5) considered the null hypothesis "no treatment effect"; therefore, they do not cover the entire parameter space. In Capizzi and Zhang's opinion, this is reasonable in placebo-controlled Phase IIb/III clinical trials. The authors cannot agree with this point of view.

The complete complement of the alternative "treatment differences concerning all four endpoints" is "no effect concerning at least one endpoint." For example, when three endpoints show very strong effects and one endpoint shows no effect, this constellation is an element of the complement, that is, the null hypothesis. However, in this constellation both rules would lead to significant results with more than 5% probability. Consequently, both rules do not control the experimentwise type I error rate when the complete complement of the alternative is considered as a null hypothesis. An appropriate level-α test would be the intersection-union test (6). An intersection-union test leads to a significant result, if every single endpoint is significant with regard to the unadjusted α-level.

Proposed Methods

As mentioned in the introduction, the anti-inflammatory treatment of chronic asthma should improve both the pulmonary function and the subjective outcomes reported by patients. Hence, the clinical objective is to show at least one positive effect in each group, that is, one considers the alternative that there is a treatment effect concerning at least one endpoint in each group. The null hypothesis is that there is no effect in at least one group.

According to the intersection-union principle, one needs a level-α test for each group. Three methods are considered. The first one is a combination of the intersection-union principle with a within-group Bonferroni adjustment: *Method 1:* In each group, at least one endpoint must be significant at the 0.025 level.

The endpoints, however, are correlated,

especially within the two groups, and, consequently, a Bonferroni-type adjustment can be very conservative. Simes (7) suggested an adjustment which is more powerful than a Bonferroni adjustment. Let $P_{(1)} \leq P_{(2)} \leq \ldots \leq P_{(r)}$ be the ordered P-values of r tests. According to Simes (7), the null hypothesis that there is no effect can be rejected if $P_{(j)} \leq j\alpha/r$ for at least one $j = 1, 2, \ldots, r$. This method controls the type I error rate when the r test statistics are independent (7). Recently, Sarkar and Chang (8) proved that the Simes method controls the type I error rate for multivariate distributions exhibiting a type of positive dependence. For some other multivariate distributions which are positively dependent but do not satisfy a special condition, such as the multivariate Student's t distribution, a simulation study of Sarkar and Chang (8) indicates that the Simes method can still be used. Furthermore, in the bivariate case, the Simes procedure controls the type I error rate for two-sided tests based on normally distributed test statistics regardless of the correlation (9, 10). It seems plausible that a similar result "should hold also for t-tests, at least for sufficiently large degrees of freedom" (10).

Method 2 is a combination of the intersection-union principle with a within-group Simes adjustment: *Method 2*: In each group, at least one endpoint must be significant at the 0.025 level *or* both endpoints must be significant at the 0.05 level. In addition, another method, which is similar to the rule 2 of Capizzi and Zhang (5), is investigated. In order to control the type I error rate for the null hypothesis that there is no effect in at least one group, one must use a lower bound than 0.05:

Method 3 states that: In each group, one endpoint must be significant at the 0.04 level and the other endpoint has to trend in the right direction, that is, the other endpoint must be significant at the 0.1 level. These three methods are investigated with respect to size and power in the following. It should be noted that instead of the Bonferroni and Simes adjustment, respectively, the Westfall and Young (11) methods could be used. This approach, however, is not considered in this paper.

The Size and Power of the Proposed Methods

According to the intersection-union principle, the methods guarantee the experimentwise type I error rate if a level-α test is used within each group. Hence, the three methods are investigated within the groups, that is, for the case of two endpoints in one group. In Method 1, one of the two endpoints must be significant at the 0.025 level. Concerning Method 2, one of the two endpoints must be significant at the 0.025 level or both endpoints must be significant at the 0.05 level. In Method 3, one of the two endpoints must be significant at the 0.04 level and the other endpoint must be significant at the 0.1 level. For the results presented here it is assumed that there are 50 patients in the placebo arm and 50 patients in the treatment arm, and that an univariate two-sided Student's t test is applied for each endpoint. Details about the bivariate t distribution which was used for the authors' computations can be found in reference 12.

Figure 1 presents the size of the three methods; the actual size within one group of endpoints is displayed. All methods lead to level-α tests. Method 2 is less conservative than Method 1 since the rejection region of the Simes method contains that of the Bonferroni method. Method 3 is more conservative than the other methods in cases of low correlations between the endpoints and utilizes the α-level in a more complete way only in case of very high correlations. Other sample sizes lead to similar curves.

The power of the methods is investigated for the following correlation structure:

$$\begin{pmatrix} 1 & \rho_1 & \rho_2 & \rho_2 \\ \rho_1 & 1 & \rho_2 & \rho_2 \\ \rho_2 & \rho_2 & 1 & \rho_1 \\ \rho_2 & \rho_2 & \rho_1 & 1 \end{pmatrix}.$$

That is, the correlation between the endpoints within a group is ρ_1 and the correlation be-

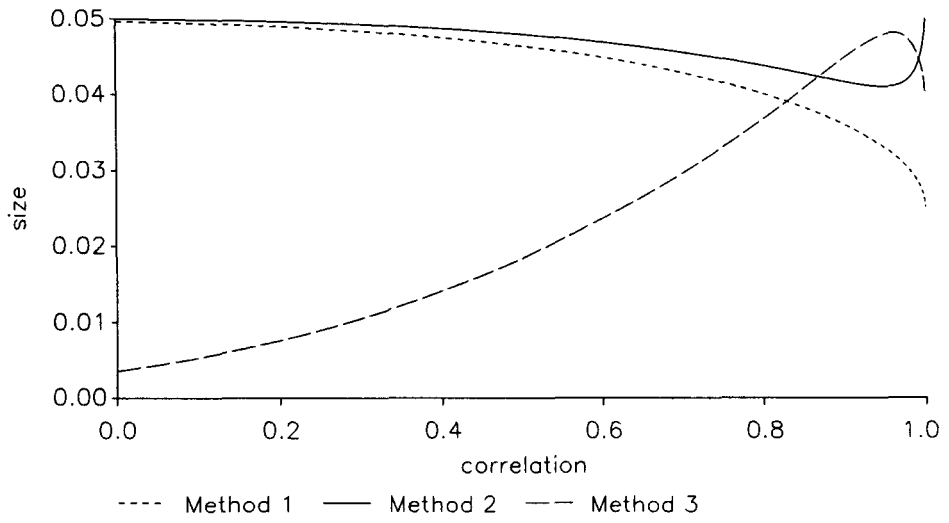

FIGURE 1. Actual within-group size of the three methods in dependence on the correlation (sample size per treatment group: 50). Method 1: 0.025 for one endpoint (Bonferroni). Method 2: 0.025 for one endpoint or 0.05 for both endpoints (Simes). Method 3: 0.04 for one endpoint and 0.1 for the other endpoint.

tween any two endpoints of different groups is ρ_2. In a simulation study, which has been performed with SAS version 6.09 (13), a vector of correlated standard normal random variables was created according to Bratley et al. (14). Again, a parallel-group design with a placebo and a treatment arm with 50 patients each is considered. In the treatment arm, the location of the simulated standard normally distributed variables were shifted. For all four endpoints, univariate two-sided Student's t tests were performed.

Table 1 contains representative simulation results. For each combination of parameters, 10,000 simulation runs were generated. According to these results Method 2 (in each group using the Simes adjustment) should be preferred to the other methods. Method 2 is always at least as powerful as Method 1 (in each group using the Bonferroni adjustment). Method 2 is much more powerful than Method 3 (in each group 0.04 for one and 0.1 for the other endpoint) if there is no effect concerning at least one out of the four endpoints. According to further simulations with sample sizes of 25 and 100 per group, respectively, these conclusions also hold for other sample sizes (results not shown).

EXAMPLES

Zhang et al. (4) gave an example of a randomized, multicenter, double-blind, parallel design clinical trial with 34 asthmatic patients in the test drug group and 35 asthmatic patients in the placebo group. The following two-sided P-values were obtained: 0.0037 (FEV_1), 0.0077 (PEF), 0.0274 (symptom scores), and 0.0369 (use of rescue medication). All P-values are smaller than 5%. Nevertheless, the requirements of Method 1 are not fulfilled, since both P-values in Group 2 are larger than 0.025. However, Method 2 as well as Method 3 lead to a significant result in this example. According to Zhang et al. (4), O'Brien's and some other methods also yield significant results.

As a further example a randomized, multicenter, double-blind, parallel design clinical trial which was designed to examine the efficacy of budesonide 400 μg twice daily with added theophylline compared to bude-

TABLE 1
Simulated Power of the Three Proposed Methods (Sample Size per Treatment Group: 50)

Differences* Concerning the Four Endpoints	Power					
	Method 1	Method 2	Method 3	Method 1	Method 2	Method 3
	$\rho_1 = 0.5$, $\rho_2 = 0.2$			$\rho_1 = 0.8$, $\rho_2 = 0.2$		
0.5, 0.5, 0.5, 0.5	0.60	0.62	0.48	0.51	0.54	0.51
0.7, 0.7, 0.7, 0.7	0.93	0.94	0.88	0.88	0.89	0.89
1.0, 1.0, 1.0, 0.5	1.00	1.00	0.80	0.99	1.00	0.81
1.0, 1.0, 0.5, 0.5	0.76	0.77	0.67	0.70	0.72	0.70
0.7, 0.5, 0.7, 0.5	0.84	0.85	0.63	0.81	0.81	0.65
1.0, 0.5, 1.0, 0.5	0.99	0.99	0.66	0.99	0.99	0.66
0.7, 0.7, 0.0, 0.7	0.87	0.87	0.08	0.85	0.85	0.07

*Standardized differences between the population means of the two treatment arms.

sonide 800 μg twice daily in asthmatic patients (15) is considered. In each group 31 patients completed the study. The correlation between the endpoints is displayed in Table 2.

The within-group correlation is much higher than the correlation among endpoints from different groups. Therefore, the motivation for the classification of the endpoints is satisfied.

Although this is an active-controlled study, there was a benefit among the theophylline treated group for FEV_1 ($P_{\text{two-sided}} = 0.0275$). For PEF a two-sided P-value of 0.1629 was obtained. As for the above-mentioned placebo-controlled trial, however, the P-values of the patient-recorded outcomes are much larger and failed to show a significant difference in this active-controlled study. Hence, none of the methods leads to a significant result. Active-controlled studies, however, are often conducted in order to show noninferiority or equivalence between treatments as was the case in this second example, and other statistical methods, for example, one-sided tests for noninferiority, are appropriate. In this active-controlled study, the noninferiority of the theophylline-treated group was statistically proven; these results and methodical issues for equivalence tests are presented by Steinijans et al. (16).

DISCUSSION

As demonstrated in the last section, the proposed methods may not be applicable in active-controlled studies. This was also mentioned by Capizzi and Zhang (5) for their decision rules. In placebo-controlled asthma trials, however, there is a further important

TABLE 2
Pearson Correlation Coefficients Between the Four Endpoints Observed in the Clinical Study Reported by Evans et al. (15)

	FEV_1	PEF	Symptom Scores	Use of Rescue Medication
FEV_1	1	0.439	0.146	0.033
PEF	0.439	1	−0.126	−0.084
Symptom scores	0.146	−0.126	1	0.678
Use of rescue medication	0.033	−0.084	0.678	1

variable which is not considered above, that is, the drop-out rate due to lack of efficacy.

Steroid-naive asthma patients are rarely available; therefore, in placebo-controlled trials a randomized subset of patients is switched from the previous therapy to placebo, and premature drop-out due to lack of efficacy is to be expected. Wolfe et al. (17) and Chervinsky et al. (18) reported drop-out rates due to lack of efficacy of 72% and 63%, respectively, in the placebo arm, but lower drop-out rates ranging from 4% to 23% in the different dose groups. Therefore, the drop-out rate due to predefined lack of efficacy criteria and the time-course of drop-outs can provide a discriminative endpoint.

The drop-out rate cannot be included in one of the two groups of endpoints and, hence, represents a new variable group. Consequently, according to the intersection-union principle one can additionally demand that the drop-out rate due to lack of efficacy is significant at the unadjusted level α.

In comparison with lung function variables such as FEV_1 and PEF, however, the drop-out rates due to lack of efficacy may have greater power to discriminate between treatments. Therefore, the drop-out rate is considered to be the most important endpoint by some investigators and regulatory bodies. Consequently, one can use the principle of a priori ordered hypotheses (19,20,21) in the following way. The first test is a logrank test concerning the drop-out times at the full α-level, for example, 5%. If and only if this test is significant, the other endpoints will be tested according to the methods described above. It should be noted, however, that high drop-out rates may make other endpoints difficult to evaluate and may lead to lack of comparability of the treatment groups for the other endpoints.

In the placebo-controlled fluticasone studies reported by Wolfe et al. (17) and Chervinsky et al. (18) the drop-out rates due to lack of efficacy were highly significant ($P \leq 0.001$ in both studies). The use of ordered hypotheses has the advantage that a significance concerning the drop-out rate can be proven without the need to demonstrate significant effects of the other variables.

REFERENCES

1. O'Brien PC. Procedures for comparing samples with multiple endpoints. *Biometrics.* 1984;40:1079–1087.
2. Lehmacher W, Wassmer G, Reitmeir P. Procedures for two-sample comparisons with multiple endpoints controlling the experimentwise error rate. *Biometrics.* 1991;47:511–521.
3. Kieser M, Reitmeir P, Wassmer G. Test procedures for clinical trials with multiple endpoints. Vollmar J, ed. In: *Testing Principles in Clinical and Preclinical Trials.* Stuttgart: G. Fischer Verlag; 1995:41–60.
4. Zhang J, Quan H, Ng J, Stepanavage ME. Some statistical methods for multiple endpoints in clinical trials. *Control Clin Trials.* 1997;18:204–221.
5. Capizzi T, Zhang J. Testing the hypothesis that matters for multiple primary endpoints. *Drug Inf J.* 1996; 30:949–956.
6. Berger RL. Multiparameter hypothesis testing and acceptance sampling. *Technometrics.* 1982;24:295–300.
7. Simes RJ. An improved Bonferroni procedure for multiple tests of significance. *Biometrika.* 1986;73: 751–754.
8. Sarkar SK, Chang C-K. The Simes method for multiple hypothesis testing with positively dependent test statistics. *J Am Stat Assoc.* 1997;92:1601–1608.
9. Hochberg Y, Rom D. Extensions of multiple testing procedures based on Simes' test. *J Statist Plan Inf.* 1995;48:141–152.
10. Samuel-Cahn E. Is the Simes improved Bonferroni procedure conservative? *Biometrika.* 1996;83:928–933.
11. Westfall PH, Young SS. *Resampling-based Multiple Testing.* New York: Wiley; 1993.
12. Dunnett CW, Sobel M. A bivariate generalization of Student's *t* distribution, with tables for certain special cases. *Biometrika.* 1954;41:153–169.
13. SAS Institute Inc. Statistical Analysis System (SAS), Version 6.09. SAS Institute Inc.: Cary; 1992.
14. Bratley P, Fox BL, Schrage LE. *A Guide to Simulation.* New York: Springer; 1987.
15. Evans DJ, Taylor DA, Zetterström O, Chung KF, O'Connor BJ, Barnes PJ. 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.
16. Steinijans VW, Neuhäuser M, Hummel T, Leichtl S, Rathgeb F, Keller A. Asthma management: the challenge of equivalence. *Int J Clin Pharmacol Ther.* 1998;36:117–125.
17. Wolfe JD, Selner JC, Mendelson LM, Hampel F, Schaberg A. Effectiveness of fluticasone propionate in patients with moderate asthma: A dose-ranging study. *Clinical Therapeutics.* 1996;18:635–646.

18. Chervinsky P, van As A, Bronsky EA, Dockhorn R, Noonan M, LaForce C, Pleskow W. Fluticasone propionate aerosol for the treatment of adults with mild to moderate asthma. *J Allergy Clin Immunol.* 1994;94:676–683.
19. Bauer P. Multiple testing in clinical trials. *Stat Med.* 1991;10:871–890.
20. Maurer W, Hothorn LA, Lehmacher W. Multiple comparisons in drug clinical trials and preclinical assays: a-priori ordered hypotheses. Vollmar J, ed. In: *Testing Principles in Clinical and Preclinical Trials.* Stuttgart: G. Fischer Verlag; 1995:3–18.
21. Koch GG, Gansky SA. Statistical considerations for multiplicity in confirmatory protocols. *Drug Inf J.* 1996;30:523–534.