

Clinical Trials are Often False Positive: A Review of Simple Methods to Control This Problem

Ton J. Cleophas* and Aeilko H. Zwinderman

European College of Pharmaceutical Medicine, Lyon, France

Abstract: *Background:* Statistical hypothesis testing is much like gambling. If, with one statistical test, your chance of a significant result is 5%, then, after 20 tests, it will increase to 40%. This result is based on the play of chance. In current clinical trials, instead of a single efficacy-variable of one treatment, multiple efficacy-variables of more than one treatment are increasingly assessed.

Methods: The current paper reviews some methods for reducing the problem of falsely positive results due to multiple testing.

Results and Conclusions: These methods include (1) the Bonferroni test, (2) the least significant difference (LSD) test, (3) other less conservative, more rarely used methods like Tukey's honestly significant difference (HSD) test, Dunnett's test, Student-Neuman-Keuls test, Hochberg's adjustment, and the Hotelling Q-square test. Alternative approaches to the problem of multiple testing include (4) the construct of composite endpoints, (5) no adjustment at all, but a more philosophical approach to the interpretation of the p-values, and (6) the replacement of the traditional 5% rejection level with a 1% rejection level or less. Evidence-based medicine is under pressure, because trials do not adequately apply to their target populations. As long as the effects of multiple testing are not routinely assessed in the analysis of clinical trials, it can not be excluded as one of the mechanisms responsible.

Keywords: Clinical trials, Multiple testing, Multiple comparisons, Type I error, Evidence-based medicine.

INTRODUCTION

Statistical hypothesis testing is much like gambling. If, with gambling once, your chance of a prize is 5%, then, with gambling 20 times, this chance will be close to 40%. The same is true with statistical testing of clinical trials. If, with one statistical test, your chance of a significant result is 5%, then after 20 tests, it will increase to 40%. This result is, however, not based on a true treatment effect, but, rather, on the play of chance. In current clinical trials, instead of a single efficacy-variable of one treatment, multiple efficacy-variables of more than one treatment are increasingly assessed. E.g., in 16 randomized controlled trials with positive results, published in the British Medical Journal (BMJ) in 2004, the numbers of primary efficacy-variables varied from 4 to 13 (Table 1). This phenomenon introduces the statistical problem of multiple comparisons and multiple testing, which increases the risk of false positive results, otherwise called type I errors. There is no consensus within the statistical community on how to cope with this problem. Also, the issue has not been studied thoroughly for every type of variable. Clinical trials rarely adjust their data for multiple comparisons. E.g., none of the above BMJ papers did. The current paper reviews some methods for that purpose.

BONFERRONI TEST [1]

If more than two samples are compared in a clinical trial, multiple groups analysis of variance (ANOVA) is often

applied for the analysis. E.g., three groups of patients were treated with different hemoglobin improving compounds with the following results:

	Sample Size	Mean Hemoglobin mmol / l	Standard Deviation mmol / l
Group 1	16	8.725	0.8445
Group 2	10	10.6300	1.2841
Group 3	15	12.3000	0.9419

The F-test produces a p-value <0.01, indicating that a highly significant difference is observed between the three groups. This leads to the not-too-informative information that not all group means were equal. A question encountered is, which group did and which one did not differ from the others. This question involves the problem of multiple comparisons. As there are 3 different treatments, 3 different pairs of treatments can be compared: groups 1 vs 2, groups 1 vs 3, and groups 2 vs 3. The easiest approach is to calculate the Student's t-test for each comparison. It produces a highly significant difference at $p < 0.01$ between treatment 1 vs 3 with no significant differences between the other comparisons. This highly significant result is, however, unadjusted for multiple comparisons. If the chance of a falsely positive result is, e.g., α with one comparison, it should be 2α with two, and close to 3α with three comparisons. Bonferroni recommends to reject the nullhypothesis at a lower level of significance according to the formula

*Address correspondence to this author at the Department of Medicine, Albert Schweitzer Hospital, Box 44, 3300 AK Dordrecht, The Netherlands; Tel: 00312 184 434222; Fax: 0031 184 434340; E-mail: ajm.cleophas@wxs.nl

Table 1. Positive Randomized Controlled Trials Published in the BMJ in 2004

	Numbers of Primary Efficacy Variables	Smallest P-values	Positive Study after Bonferroni Adjustment
1. Schroter <i>et al.</i> 328: 742-3	5	0.001	yes
2. Laurant <i>et al.</i> 328: 927-30	12	0.006	no
3. Yudkin <i>et al.</i> 328: 989-90	10	0.001	yes
4. Craig <i>et al.</i> 328: 1067-70	6	0.030	no
5. Kalra <i>et al.</i> 328: 1099-101	7	0.001	yes
6. Hilten <i>et al.</i> 328: 1281-1	5	0.05	no
7. James <i>et al.</i> 328: 1237-9	10	0.003	yes
8. Logan <i>et al.</i> 328: 1372-4	6	0.01	no
9. Smith <i>et al.</i> 328: 1459-63	13	0.002	yes
10. Powell <i>et al.</i> 329: 89-91	10	0.001	yes
11. Henderson <i>et al.</i> 329: 136-9	6	0.03	no
12. Collins <i>et al.</i> 329: 193-6	4	0.03	no
13. Svendsen <i>et al.</i> 329: 253-8	7	0.02	no
14. McKendry, M 329: 258-61	9	0.001	yes
15. Van staaij <i>et al.</i> 329: 651-4	8	0.01	no
16. Norman <i>et al.</i> 329: 1259-62	10	0.02	yes

rejection p-value = $\alpha \times 2 / k (k-1)$

k = number of comparisons, α = agreed chance of falsely positive result (mostly 0.05)

In case of three comparisons, the rejection p-value will be $0.05 \times 2/3(3-1) = 0.0166$.

A p-value of 0.0166 is still larger than 0.01, and, so, the difference observed remains significant, but using a cut-off p-value of 0.0166, instead of 0.05, the difference is not highly significant anymore.

LEAST SIGNIFICANT DIFFERENCE TEST (LSD) TEST [2]

As an alternative to the Bonferroni test a refined t-test, the so-called least significant difference (LSD) test, can be applied. This refined t-statistic has n-k degrees of freedom, where n is the number of observations in the entire sample and k is the number of treatment groups. In the denominator of this refined t-test, the usual pooled standard error (SE) is replaced with the pooled-within-group variance from the above-mentioned F-test. For the application of the LSD procedure, it is essential to perform it sequentially to a significant F-test of the ANOVA procedure. So, if one chooses to perform the LSD procedure, one first calculates the ANOVA procedure and stops if it is not significant, and calculates the LSD test only if the F-test is statistically significant. The LSD test is largely similar to the Bonferroni-test, and yields with the above example a p-value close to

0.05. Like with Bonferroni, the difference is still significant, but not highly significant anymore.

OTHER TESTS FOR ADJUSTING THE P-VALUES

None of the 16 BMJ trials discussed in the introduction were adjusted for multiple testing. When we performed their Bonferroni adjustment of them, only 8 trials continued to be positive, while the remainder turned into negative studies (Table 1). This does not necessarily indicate that all of these studies were truly negative. Several of them had more than 5 efficacy-variables, and, in this situation, the Bonferroni test is somewhat conservative, meaning that power is lost, and the risk of falsely negative results is raised. This is particularly so, if variables are highly correlated. A somewhat less conservative variation of the Bonferroni correction was suggested by Hochberg: if there are k primary values, multiply the largest p-value with 1, the second-largest p-value with 2, the third largest with 3....., and the smallest p-value with k [3].

	Calculated p-values	Reject null-hypothesis at
(1)	largest p-value	$\alpha_1 = 0.05 \times 1 = 0.05$
(2)	second largest p-value	$\alpha_2 = 0.05 \times 2 = 0.10$
(3)	third largest p-value	$\alpha_3 = 0.05 \times 3 = 0.15$
(k)	kth largest p-value	$\alpha_k = 0.05 \times k = \dots$

The mathematical arguments of this procedure go beyond this paper. What happens is, that the lowest and highest p-values will be less different from one another. There are other less conservative methods, like Tukey's honestly significant difference (HSD) test [6], Dunnett's test [7], Student-Neuman-Keuls test [8], and the Hotelling Q-square test [9]. Most of them have in common that they produce their own test-statistics. Tables of significance levels are available in statistical software packages including SAS and SPSS.

COMPOSITE ENDPOINT PROCEDURES

A different solution for the multiple testing problem is to construct a composite endpoint of all of the efficacy-variables, and, subsequently, to perform a statistical analysis on the composite only. For example, it is reasonable to believe that statin treatment has a beneficial effect on total cholesterol (Tc), high density cholesterol (HDL), low density cholesterol (LDL), and triglycerides (Tg). We can perform a composite analysis of the four variables according to

$$\text{Composite variable} = (\text{Tc} + \text{HDL} + \text{LDL} + \text{Tg})/4$$

$$Tc = (\text{Tc} - \text{mean}(\text{Tc})) / \text{SD}_{Tc} \text{etc}$$

A simple t-test produces

$$\text{Placebo: mean result composite variable} = -0.23 \text{ (SD } 0.59)$$

$$\text{Statin: mean result composite variable} = 0.15 \text{ (SD } 0.56)$$

$$p = 0.006$$

This p-value is lower than that obtained by a Bonferroni or LSD procedure. This is probably so, because of the added power provided by the positive correlation between the repeated observations in one subject. If no strong correlation between the variables is to be expected, the composite endpoint procedure provides power similar to that of the Bonferroni or LSD procedure.

Largely similar to the composite endpoint procedure are the so-called index methods. If the efficacy-variables are highly correlated, because they more or less measure the same patient characteristic, then they be best replaced with their add-up sum. In this way, the number of primary variables is reduced, and an additional advantage is that the standardized add-up sum of the separate variables is more reliable than the separate variables. E.g., the Disease Activity Score (DAS) for the assessment of patients with rheumatoid arthritis, including the so-called Ritchie joint pain score, the number of swollen joints, and the erythrocyte sedimentation rate, is an example of this approach [10].

NO ADJUSTMENTS AT ALL, AND PRAGMATIC SOLUTIONS

A more philosophical approach to the problem of multiple comparisons is to informally integrate the data, and to look for trends without judging one or two low p-values among otherwise high p-values as proof of a significant difference in the data. However, both the medical community and the investigators may be unhappy with this

solution, because they want the hard data to provide unequivocal answers to their questions, rather than uncertainties. An alternative and more pragmatic solution could be the standard use of lower levels of significance to reject the null-hypothesis. For the statistical analysis of interim analyses, that suffer from the same risk of increased type I errors due to multiple testing, Pocock's recommendation to routinely use $p < 0.01$ instead of $p < 0.05$ has been widely adopted [11]. A similar rule could, of course, be applied to any multiple testing situation. The advantage would be that it does not damage the data, because the data remain undamaged. Moreover, any adjustments may produce new type I errors, particularly, if they are post-hoc, and not previously described in the study protocol.

CONCLUSION

Approaches to reducing the problem of multiple testing include (1) the Bonferroni, test, (2) the LSD method, (3) other less conservative, more rarely used methods like Tukey's honestly significant (HSD) method, Dunnett's test, Student-Neuman-Keuls test, Hochberg's adjustment, and the Hotelling Q-square test. Alternative approaches to the problems of multiple testing include (4) the construct of composite endpoints, (5) no adjustment at all, but a more philosophical approach to the interpretation of the p-values, and (6) the replacement of the traditional 5% rejection level with a 1% rejection level or less.

Evidence-based medicine is increasingly under pressure, because clinical trials do not adequately apply to their target populations [12-14]. Many causes are mentioned. As long as the issue of multiple testing is rarely assessed in the analysis of randomized controlled trials, it cannot be excluded as one of the mechanisms responsible. We recommend that the increased risk of falsely positive results should be taken into account in any future randomized clinical trial, which assesses more than one efficacy-variable and / or treatment modality. The current paper provides 6 possible methods for assessment.

REFERENCES

- [1] Lu Y, Fang JQ. Bonferroni adjustment. In: Advanced Statistics. World Scientific, River Edge, NJ, 2003, pp. 870.
- [2] Cleophas TJ, Zwiderman AH, Cleophas AF. Multiple statistical inferences. In : Statistics applied to clinical trials. Kluwer Academic Publishers, Boston, MA, 2002, pp. 73-81.
- [3] Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* 1988; 75: 800-2.
- [4] Brown BW, Hollander M. Analysis of k-sample problems. In: Statistics: a biomedical introduction. Wiley, New York, 1977, pp. 287.
- [5] Motulsky H. Dunnett's method for multiple comparison. In: Intuitive Statistics, Oxford University Press, Oxford, UK, 1995, pp. 259.
- [6] Glantz SA. Student Newman Keuls test and analysis of variance. In: Primer of statistics, McGraw-Hill, New York, 1992, pp. 311.
- [7] Cleophas TJ, Zwiderman AH, Cleophas AF. Hotelling's T-square. In: Statistics applied to clinical trials, self-assessment. Kluwer Academic Publishers, Boston, MA, 2002, pp. 142.
- [8] SAS statistical software. <http://www.sas.com>
- [9] SPSS statistical software: <http://www.spss.com>
- [10] Fuchs HA. The use of the disease activity score in the analysis of clinical trials in rheumatoid arthritis. *J Rheumatol* 1993; 20: 1863-6.
- [11] Pocock SJ. Clinical trials. A practical approach. New York, Wiley, 1988.

- [12] Furberg C. To whom do the research findings apply? *Heart* 2002; 87: 570-4.
- [13] Julius S. The ALHATT study: if you believe in evidence-based medicine. Stick to it. *Hypertens* 2003; 21: 453-4.
- [14] Cleophas GM, Cleophas TJ. Clinical trials in jeopardy. *Int J Clin Pharmacol Ther* 2003; 41: 51-6.

Received: 06 April, 2005

Revised: 27 April, 2005

Accepted: 19 May, 2005