Evaluation of test statistics in split-mouth clinical trials

P. P. Hujoel and L. H. Moulton
Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI U.S.A.

Hujoel PP, Moulton LH.: Evaluation of test statistics in split-mouth clinical trials. J Periodont Res 1988; 23: 378–380.

This study was undertaken to examine the appropriateness of test statistics used for split-mouth clinical trials. Twenty-two published trials were reviewed and the primary test statistics were labeled as appropriate, inappropriate, or absent. Only 5 of the 22 trials reported an appropriate test statistics. Of the remaining 17 trials, 12 used an inappropriate test statistics, and 5 did not report or did not use statistical tests. A serious problem is that more than half of the reviewed trials have reported the use of a one-way analysis of variance or a two-sample *t* test to investigate the data of the split-mouth trial. This mistake may have led to a large increase of the Type II errors due to the correlated nature of split-mouth data. Failure to detect genuine therapeutic differences because of inadequate attention to the data analysis can occur with inappropriate use of test statistics. It is concluded that there should be more concern for the use of appropriate test statistics.

Accepted for publication September 27, 1988

Introduction

Split-mouth experimental designs have been used extensively since their introduction in 1968 (1). Originally, the split-mouth was defined as a division of the mouth in the midsaggital plane between the two central incisors. Later, the definition was enlarged to include other types of divisions of the dental arches, such as quadrants or sextants. In the experimental design of clinical trials these divisions of the mouth constitute the experimental units which are randomly assigned to treatment modalities.

The use of the split-mouth as the experimental unit for clinical investigation has two principal advantages: 1) the elimination of the subject factor from the experimental error, and 2) the economical use of patients. The realization of the benefits of the split-mouth design depends, however, on the application of the appropriate statistical techniques in the analysis. The essential statistical characteristic of this design is that comparisons are made on a within-patient basis, not on a between-patient basis. The distinction between the two types of comparisons is of considerable importance for the choice of the appropriate test-statistics. Using the one-way ANOVA or the two-sample *t* test for the purpose of making within-patient comparisons is incorrect, provides incorrect Type I and Type II error rates, and can lead to erroneous conclusions.

These errors occur because the two different sizes of experimental units are ignored: the patient and the subdivision within a patient. This distinction is important for the separation of the error terms. The within-patient variance excludes all extraneous sources of variation commonly referred to as the host factor, such as oral hygiene, immune system, etc. When only two treatments are studied in a split-mouth design, the F statistic of the one-way ANOVA is equivalent to the two-sample *t* test, both of which are inappropriate; whereas the F statistic of the two-way ANOVA in that situation is equivalent to the one-sample (paired) *t* test, both of which do make appropriate treatment comparisons within patients.

The purpose of this paper is to review 22 published trials in the periodontal literature and to determine whether appropriate test statistics for the split-mouth design were used. An example is given to illustrate that inappropriate test statistics may affect the conclusions of a study.

Study of published trials

Materials and methods

Twenty-two split-mouth clinical trials were examined to determine if an appropriate analysis was conducted (1–22). We started with a set of 21 papers selected by Antczak et al. (23) (1986). They identified the papers by reviewing the Medical Subjects Heading Keywords in Medline (1980–1984) and by inspecting the bibliographies of original and review articles. Two papers which did not use a split-mouth design were excluded. Three recent publications were added using selection criteria similar to Antczak's.

The statistical analysis was considered appropriate if a paired *t* test, a two-factor ANOVA, or any other appropriate test statistic was used. The analysis was categorized as inappropriate if a one-way ANOVA or a two-sample *t* test was used to make within-patient comparisons. It was also classified as inappropriate if the paper referred to "previous methodology" where an inappropriate analysis was performed. The statistical test was called absent (a) if p-values were reported without mention of the test statistic used or (b) if the paper only reported descriptive statistics.

Results

The results show that only 5 of the 22 papers reported an appropriate statistical analysis. Twelve papers used an inappropriate test and 5 papers did not report the test statistics used. This leads to a total of 17 of the comparative clinical trials with no or inappropriate stat-

Table 1. Detailed presentation of the criticism of the 22 trials. Quotation marks indicate a citation from the publication. Of the 22 split-mouth clinical trials, 23% appropriately used a paired *t* test, 23% inappropriately used the two-sample *t* test, 32% inappropriately used the one-way ANOVA and 23% did not report the test statistic used. Note the absence of the two-factor ANOVA which could be used when more than 2 treatment modalities are applied within a patient

Overview of 22 papers using a split-mouth design	
Use of inappropriate one-way ANOVA	
1975	"Analysis of variance ... advantage of the paired design was not fully utilized."
1977	Reference to 1975.
1979	"One-way analysis of variance"
1980	Reference to 1975.
1980	Reference to 1979.
1981	"One-way analysis of variance"
1987	Reference to 1981.
Use of inappropriate two-sample t test	
1973	"student *t* test"
1982	"Conventional *t* test and Kolmogorow-Smirnow two sample test"
1982	"*t* test"
1984	"student *t* test"
1985	"student *t* test"
No test statistics published	
1968	
1981	
1984	
1984	
1984	
Use of appropriate one-sample t test	
1976	"paired *t* test"
1980	"paired *t* test"
1981	"paired *t* test"
1983	"paired *t* test"
1983	"paired *t* test"

istical tests. Specifics are summarized in Table 1.

Discussion

The application of inappropriate statistical tests in the set of 22 papers may have led to some misleading conclusions. A correct hypothesis test removes intuition and bias from decisions and provides conclusions with a known probabilistic distribution. Absent or inappropriate statistical tests provide no measure of confidence for the conclusions. Since the topic of this paper is concerned with statistical tests and analysis, the issues of clinical versus statistical significance and appropriateness of means as a summary value for a quadrant will not be addressed.

Only on rare occasions can conclusions be based on a comparison of descriptive statistics. Sometimes, differences between treatments are so large that no statistical tests are needed. Berkson has called this the traumatic intraocular test: the results hit one between the eyes. Unfortunately, such dramatic differences between periodontal therapies are not present. Even surgical versus non-surgical therapy comparisons seem to result in minimal differences of the descriptive statistics. Therefore, an absence of statistical tests for clinical trials in periodontics is hard to defend.

The use of inappropriate test statistics is a more serious problem since it may lead a reader to an unwarranted confidence in the common conclusion of 'no difference between periodontal treatment modalities'. Antczak reported that the majority of trials (81%) show no statistically significant differences between treatments for at least some classification of disease severity (23). These conclusions of no difference may be due to 1) a 'true' no difference status, 2) the possibility of a 'real' type II error, which may be due to an inadequate sample size, or 3) inappropriately applied test statistics with largely inflated probabilities of type II errors. This last problem occurred in 12 out of the 17 trials which reported statistical tests. For these 12 trials the 'no significant difference conclusion' should seriously be questioned. The fact that these trials used inappropriate test-statistics does not necessarily imply that erroneous conclusions were reached. It only indicates that the analysis had an unnecessarily low level of power, resulting in an increased probability of making a 'no significant difference' conclusion. Whether these changes in probability levels actually resulted in erroneous conclusions can not be investigated unless the data are reanalyzed.

To illustrate the possible impact of incorrect statistical analysis on the conclusions of a study we provide the following example based on published data of mean attachment level changes (2) and an estimate of 0.7 for the between quadrant correlation (24) (Table 2). An appropriate and an inappropriate analysis is performed to compare the mean attachment level changes of the Modified Widman Flap (mean = −0.24, S.D. = 0.85) and osseous surgery (mean = −0.41, S.D. = 0.70) at 2 yr post-operatively (pockets 4 to 6 mm). An inappropriate one-way analysis of variance (or 2-sample *t* test) leads to the conclusion that the treatments are not significantly different ($p = 0.169$). However, the appropriate two-factor analysis of variance (or paired *t* test) shows significant differences between the treatment modalities ($p = 0.016$). This example illustrates the occurrence of a Type

Table 2. Comparison of the inappropriately used one-way ANOVA to the appropriate two-factor ANOVA. The two methods lead in this example to opposite conclusions

One-way ANOVA				Two-way ANOVA			
Source	DF	SS	MS	Source	DF	SS	MS
Treatment	1	1.16	1.16	Treatment	1	1.16	1.16
Error	158	95.79	0.61	Error (between patients)	79	80.80	
				Error (within patient)	79	14.99	0.19
$F^* = 1.90$				$F^* = 6.10$			
Since $F^* = 1.90 < F(.95;1,158) = 3.90$ We conclude: No significant differences between the two treatment modalities have been demonstrated.				Since $F^* = 6.10 > F(.95;1,79) = 3.96$ We conclude: There are significant differences between the two treatment modalities.			

II error: a conclusion of no significant differences while in fact there are differences present. To assess the probability of this Type II error, the power of both the appropriate and the inappropriate test statistic was calculated with the type I error rate fixed at 5% (25). The one-way analysis of variance test statistic has a power of approximately 16%. In other words, there are about 16 chances in 100 that the decision rule will lead to a detection of a true treatment difference of .17 mm between two periodontal treatment modalities. With a two-factor analysis of variance (or a paired t test) the power of the decision rule is 86%. This large difference in power of the two test statistics makes it obvious that the use of an incorrect statistical analysis may alter the interpretation of the data. In addition to a large increase of the Type II error, the Type I error of the test is not 5% as presumed, but is somewhat less.

Of course, in actual clinical settings the data often will be more complex than the above example. More advanced techniques to handle repeated measures, unbalanced data, and multiple comparisons will be required. To make our point about accounting for within-patient correlation, we are focusing on the basic aspects of the experimental design and the subsequent choice of the primary test statistic, since these topics are so critical for the conclusions of comparative periodontal studies.

It is unfortunate that some investigators will put themselves to great trouble in designing and carrying out a clinical study and yet deny themselves the benefits of correct statistical techniques. The marginal labor and cost of a statistical analysis is small when compared to the total expenditure of a trial. Yet without proper attention to data analysis the conclusions can be doubtful.

Acknowledgment

We would like to thank Dr. Graham Kalton, Chairman of the Department of Biostatistics, for many helpful comments and suggestions.

References

1. Ramfjord SP, Nissle RR, Shick RA, et al. Subgingival curettage versus surgical elimination of periodontal pockets. *J Periodontol* 1968; **39**: 167.
2. Hill RW, Ramfjord SP, Morrison EC, et al. Four types of periodontal treatment compared over 2 years. *J Periodontal* 1981; **52**: 655.
3. Ramfjord SP, Knowles JW, Morrison EC, et al. Results of periodontal therapy related to tooth type. *J Periodontol* 1980; **51**: 270.
4. Ramfjord SP, Knowles JW, Nissle RR, et al. Results following three modalities of periodontal therapy. *J Periodontol* 1975; **46**: 552.
5. Ramfjord SP, Knowles JW, Nissle RR, et al. Longitudinal study of periodontal therapy. *J Periodontol* 1973; **44**: 66.
6. Knowles J, Burgett F, Nissle RR, et al. Results of periodontal treatment related to pocket depth and attachment level. Eight years. *J Periodontol* 1979; **50**: 225.
7. Knowles J, Burgett F, Morrison EC, et al. Comparison of results following three modalities of periodontal therapy related to tooth type and initial pocket depth. *J Clin Periodontol* 1980; **7**: 32.
8. Burgett FG, Knowles JW, Nissle RR, et al. Short term results of three modalities of periodontal treatment. *J Periodont* 1977; **46**: 131.
9. Ramfjord SP, Caffesse RG, Morrison EC, et al. Four modalities of periodontal treatment compared over 5 years. *J Clin Periodontol* 1987; **14**: 445.
10. Isidor F, Karring T, Attstrom R. The effect of root planing as compared to that of surgical treatment. *J Clin Periodontol* 1984; **11**: 669.
11. Olsen CT, Ammons WF, van Belle G. A longitudinal study comparing apically repositioned flaps, with and without osseous surgery. *Int J Periodontol and Res Dent* 1985; **4**: 11.
12. Pihlstrom BL, McHugh RB, Oliphant TH, et al. Comparison of surgical and non-surgical treatment of periodontal disease. *J Clin Periodontol* 1983; **10**: 524.
13. Pihlstrom BL, Ortiz-Campos C, McHugh RB. A randomized four year study of periodontal therapy. *J Periodontol* 1981; **52**: 227.
14. Smith DH, Ammons WF, Van Belle G. A longitudinal study of periodontal status comparing osseous recontouring with flap curettage. *J Periodontol* 1980; **51**: 367.
15. Waite IM. A comparison between conventional gingivectomy and a non-surgical regime in the treatment of periodontitis. *J Clin Periodontol* 1976; **3**: 173.
16. Lindhe J, Socransky SS, Nyman S, et al. "Critical probing depth" in periodontal therapy. *J Clin Periodontol* 1982; **9**: 332.
17. Lindhe J, Westfelt E, Nyman S. Healing following surgical/non-surgical treatment of periodontal disease. *J Clin Periodontol* 1982; **9**: 115.
18. Lindhe J, Westfelt E, Nyman S. Long-term effect of surgical/non-surgical treatment of periodontal disease. *J Clin Periodontol* 1984; **11**: 448.
19. Echeverria JJ, Cafesse RG. Effect of gingival curettage when performed one month after root instrumentation. *J Clin Periodontol* 1983; **10**: 277.
20. Badersten A, Nilveus R, Egelberg J. Effect of non-surgical therapy. I. Moderately advanced periodontitis. *J Clin Periodontol* 1981; **8**: 57.
21. Badersten A, Nilveus R, Egelberg J. Effect of non-surgical therapy. II. Severely advanced periodontitis. *J Clin Periodontol* 1984; **11**: 63.
22. Badersten A, Nilveus R, Egelberg J. Effect of non-surgical therapy. III. Single versus repeated instrumentation. *J Clin Periodontol* 1984; **11**: 114.
23. Antczak AA, Tang J, Chalmers TC. Quality assessment of randomized control trials in dental research. II. Results: Periodontal research. *J Periodont Res* 1986; **21**: 315.
24. Chilton N, Fleiss J, for the Forsyth Group. A Re-examination of correlations associated with attachment levels. Philadelphia: Meeting of the Dental Task Force, 1986.
25. Lindgren BW. Univariate Normal Inference. In: Lindgren BW, ed. *Statistical Theory*. New York: Macmillan Publishing Co., 1976: 354.

Address:

Dr. Philippe Hujoel
Department of Biostatistics
University of Michigan
School of Public Health
109 South Observatory Street
Ann Arbor, MI 48109
U.S.A.