Evaluation of test statistics in split-mouth clinical trials


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.

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 sta-
demonstrated. Two treatment modalities have been compared. The two methods lead in this example to opposite conclusions. A correct hypothesis test reduces type I errors, in the sense of means as a summary value for the eyes. Unfortunately, such dramatic differences between treatment modalities 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 I error in the analysis of clinical trials.

### Table 1. Detailed presentation of the criticism of the 22 trials. Quotation marks indicate a citation from the publication.

<table>
<thead>
<tr>
<th>Year</th>
<th>Citation</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1979</td>
<td>Reference to 1979.</td>
<td></td>
</tr>
<tr>
<td>1983</td>
<td>Reference to 1983.</td>
<td></td>
</tr>
</tbody>
</table>

Use of inappropriate one-way ANOVA

- 1975: “Analysis of variance was not fully utilized.”
- 1979: “One-way analysis of variance.”

Use of inappropriate two-sample t test

- 1973: “Student t test.”

No test statistics published

- 1968
- 1981
- 1984
- 1984
- 1984
- 1985

Use of appropriate one-sample t test

- 1976: “Paired t test.”
- 1980: “Paired t test.”
- 1981: “Paired t test.”
- 1983: “Paired t test.”

### 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.

<table>
<thead>
<tr>
<th>Source</th>
<th>One-way ANOVA</th>
<th>Two-way ANOVA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DF</td>
<td>SS</td>
</tr>
<tr>
<td>Treatment</td>
<td>1</td>
<td>1.16</td>
</tr>
<tr>
<td>Error</td>
<td>158</td>
<td>95.79</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

F* = 1.90

Since F* = 1.90 < F(95; 1, 158) = 3.90

We conclude:

No significant differences between the two treatment modalities have been demonstrated.

F* = 6.10

Since F* = 6.10 > F(95; 1, 79) = 3.96

We conclude:

There are significant differences between the two treatment modalities.

### 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 intracocular 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.

Statistics in split-mouth studies

739
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


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.
This document is a scanned copy of a printed document. No warranty is given about the accuracy of the copy. Users should refer to the original published version of the material.