J Clin Periodontol 2005; 32: 677–683 doi: 10.1111/j.1600-051X.2005.00735.x

Journal of Clinical Periodontology

Statistical methods for testing plaque removal efficacy in clinical trials

I. Heynderickx[1] & J. Engel[2]

[1]Philips Research Laboratories, Eindhoven, [2]CQM, Eindhoven, the Netherlands

Heynderickx I, Engel J. Statistical methods for testing plaque removal efficacy in clinical trials. J Clin Periodontol 2005; 32: 677–683. doi: 10.1111/j.1600-051X.2005.00735.x. © Blackwell Munksgaard, 2005.

***Abstract*:**

Objectives: To evaluate the ability of different statistical approaches in finding a statistically significant difference in plaque removal efficiency between brushes in clinical trials.

Materials and Methods: The approaches, which are evaluated, concern the scores after brushing only, the difference in scores before and after brushing and the relative difference scores (i.e. score before minus score after brushing divided by the score before brushing). In each case the scores before brushing may be included as a covariate. Except for the relative difference scores, the power of the test statistics of the approaches has been compared by assuming a simple statistical model. These theoretical results have been compared with the numerical results of two particular clinical trials – one with a between-subject design and one with a within-subject design.

Results: The numerical results of these clinical trials show that the calculated *p*-values support the conclusions drawn from the statistical model, i.e. the power of the *F*-test is highest when evaluating the data after brushing with the data before brushing included as a covariate. Using the differences in scores before and after brushing – again with the data before brushing as a covariate – does not add additional power to the test. Omitting the data before brushing as a covariate only gives satisfactory results when the variance over the subjects or the error variance is zero, which in general is not the case.

Conclusions: This investigation reveals that in general the approach of analysing the scores after brushing with the scores before brushing as a covariate yields the highest chance of finding a statistically significant difference between two brushes.

Keywords: brushing efficacy; clinical trial; plaque removal; statistical analysis

Accepted for publication 17 November 2004

With the introduction of various power-assisted toothbrushes, the literature on clinical trials to prove the efficacy of these brushes has rapidly increased. The short-term efficacy is often determined by measuring the removal of plaque as a consequence of brushing by means of various plaque indices. The long-term efficacy is evaluated mainly by following the health of the gingival tissue over longer periods of use of the brush (Heasman & McCracken 1999). But within this general tendency there is still substantial variation in the specific protocol used to evaluate the efficacy of a toothbrush (Heasman & McCracken 1999). And, apart from decisions relating to the experiment's protocol, there is yet no common standard on how to statistically analyse the experimental data. Therefore, we will in this paper focus on the statistical analysis of data evaluating the plaque removal efficacy of toothbrushes.

Clinical trials use protocols in which subjects are distributed over the various brushes to be tested or in which all subjects apply all the various brushes in the experiment. The former case is referred to as a parallel-group design, or more generally as a Between-Subjects Design (BSD). In the latter case a choice is made between a cross-over design, in which each brush is used in the whole mouth during subsequent periods of time, or a split-mouth design, in which the brushes are used in different quadrants of the mouth. More generally, both options are classified as a Within-Subjects Design (WSD). Independent of the design choice made, a clinical evaluation of the plaque removal efficacy of a brush generally proceeds as follows: after a professional instruction on the use of a specific brush, and possibly a learning or adaptation period, the plaque removal efficacy is determined by measuring the amount of plaque – gathered

over at least 24 h – before and after brushing. All plaque indices are designed such that the amount of plaque present in a well-defined sub-area of a tooth is expressed in terms of a number. Focussing on e.g. the Turesky modified Quigley and Hein (Q&H) index, six numbers between 0 and 5 are generated per tooth (Quigley & Hein 1962, Turesky et al. 1970). Measuring the plaque on 28 teeth in a mouth results in 168 values per evaluation per subject. These numbers are further processed to determine the statistical significance of the difference in plaque before and after brushing and among various brushes. Of course, depending on the expected cleaning behaviour of the brush relevant subsets of the full-mouth plaque scores may be evaluated in a similar way. Common examples are all inter-proximal areas only or all posterior teeth only.

Prerequisites for having confidence in statistical testing that is generally applied in clinical toothbrush studies are that the data are independent and normally distributed and that variances be equal. Moreover, parametric tests can only be used for data obtained on a linear scale. Strictly speaking, none of these requirements is fulfilled. In order to make the data independent it is common use to first average within one subject all the plaque values measured on the teeth brushed with one brush (De Rouen 1989, Addy 1995). In this way, a single measure is obtained that summarizes all the information of the brush applied to a subject. As a zero score is the lowest value one can get, floor effects may distort the distribution of data from normality. But it has been shown that the *F*-tests we shall apply are sufficiently robust with respect to deviations from normality (Chilton & Fleiss 1986, Sullivan & D'Agostino 1992). Finally, it should be mentioned that plaque indices attribute a number to a given amount of plaque on an arbitrary scale, which is not necessarily linear. So, theoretically speaking, the statistical analysis of these data should be based on non-parametric tests. In practice, however, parametric tests are preferred as they have more power to make small differences significant. It has been demonstrated that the robustness of the *F*-test allows it to be used with Silness & Loë index scores (Chilton & Fleiss 1986).

The simplest statistical analysis for evaluating the plaque removal efficacy of two brushes is to compare the mean plaque scores in the Fisher's *F*-test. Actually, most often the mean increments or decrements in plaque score because of brushing are compared among brushes (Chilton & Fleiss 1986, Sharma et al. 1992, van der Weijden et al. 1994, Heasman et al. 1999). In simulation studies based on gingivitis scores, little difference has been found between the use of differences in scores as compared with the ratio of scores for demonstrating either equivalence or superiority of one product over another (Kingman 1992). One step more complicated is the analysis of covariance using the initial plaque level as a covariate. It has been shown that this analysis may make the comparison between two brushes sharper when the treatments are successfully balanced by randomization (Chilton & Fleiss 1986, Samuels 1986).

In this paper, we will give additional evidence on the preferred choice concerning the approach in the statistical analysis to compare the plaque removal efficacy of two brushes. We will evaluate different approaches, and for each of them compute the power of the statistical test. Indeed, as most clinical trials pursue to prove that one brush performs statistically significantly better than the other, the statistical test that differentiates best between the two brushes is preferable. But one should keep in mind that a statistically significant difference in plaque removal efficacy between two brushes does not necessarily mean that there is also a clinically relevant difference between the brushes. Although it has not yet been proven what difference in plaque removal efficacy is clinically relevant, a relative difference of 25% is suggested as a guideline (Heasman & McCracken 1999).

Statistical Model and Approaches

Introduction

In this section, we will formulate the statistical model and testing procedures that were used to compare the plaque removal efficacy of toothbrushes by means of plaque scores measured before and after brushing. Here, the model and analysis is detailed for the particular case of the comparison of two toothbrushes, but the ideas can be generalized for more brushes. As mentioned above, two experimental designs will be considered separately: a BSD and a WSD.

The standard statistical means for testing the effect of a toothbrush on plaque removal is the *F*-test from the analysis of variance (ANOVA), or analysis of covariance (ANCOVA). A large power of this *F*-test results in a high probability of rejecting the hypothesis of an equal brushing efficacy of the brushes if these efficacies are in fact unequal. Remember that in all cases the probability of rejecting the hypothesis when it is true will be limited to say 5%. The power of a test is a function of the sample size, the residual variance and the expected difference in plaque removal between the two brushes. In the reasoning used here, an analysis approach is stated to be better than another approach if the corresponding test statistic is more powerful. Or, in other words, by comparing the power of various statistical approaches we will optimize the chance of finding a statistically significant difference in plaque removal between two brushes.

With respect to the experimental data, the following approaches can be formulated for the statistical testing:

Approach 1: Analysis of the full-mouth plaque scores after brushing only.
Approach 2: Analysis of the full-mouth plaque scores before brushing minus after brushing.
Approach 3: Analysis of the plaque scores after brushing, where the scores before brushing are included as a co-variable.
Approach 4: Analysis of the plaque scores before minus after brushing with scores before brushing being included as a co-variable.
Approach 5: Analysis of the relative plaque scores (before minus after brushing, divided by the scores before brushing).
Approach 6: Analysis of the relative plaque scores with the scores before brushing being included as a co-variable.

In the literature, Approaches 1, 2 and 5 are usually followed (Chilton & Fleiss 1986, Sharma et al. 1992, van der Weijden et al. 1994, Heasman et al. 1999), where the relative plaque scores of Approach 5 are usually expressed as a percentage plaque reduction. The argument for including the scores before brushing in the analysis is that these scores contain extra information on the subjects's plaque level. Therefore, the use of these scores may improve the power of the statistical test by reducing the residual variability.

We shall here consider only the first four approaches for the two design types separately. Approaches 5 and 6 will be considered only in the numerical analysis, as the computation of the power of the tests for the relative differences is rather complicated and can no longer be deduced from simple analytical expressions. However, we shall make some remarks on these approaches in the next two sub-sections.

BSD

The statistical model

The score before brushing is modelled as

$$Y_{i1k} = \mu_1 + S_i + e_{i1k},$$

with $i = 1, \ldots, n_k$ and n_k being the number of subjects for brush k. The index 1 in Y_{i1k} stands for the first measurement point (i.e. before brushing). μ_1 is the general mean, S_i is the effect for subject i and e_{i1k} is the error term. It is assumed that S_i and e_{i1k} are independent random variables, which are normally distributed around 0 with a variance of σ_S^2 for S_i and of σ_e^2 for e_{i1k}. As a consequence, the expected value of Y_{i1k} equals $E(Y_{i1k}) = \mu_1$, and the variance is var $(Y_{i1k}) = \sigma_S^2 + \sigma_e^2$.

The score after brushing is modelled as

$$Y_{i2k} = \mu_2 + S_i + e_{i2k} + B_k.$$

The index 2 in Y_{i2k} denotes the second point of measurement (i.e. after brushing). Now, B_k is introduced to present the effect of the kth toothbrush with k being 1 or 2 when two toothbrushes are compared. The assumptions for S_i and e_{i2k} are the same as for the scores before brushing and we further assume that $\Sigma B_k = 0$ for standardization. As a consequence, the expected value for the scores after brushing is $E(Y_{i2k}) = \mu_2 + B_k$ and the variance equals var $(Y_{i2k}) = \sigma_S^2 + \sigma_e^2$.

Test statistic and results for variances

The standard F-test statistic is used to test the null hypothesis that the two brushes have an identical efficacy in plaque removal, that is H_0: $B_1 = B_2$. Under this null hypothesis, the F-test has an F-distribution with 1 and n_1+n_2-2 degrees of freedom, where n_1+n_2 is the total number of observations, if no co-variables are included. If the scores before brushing are included as a co-variable, the denominator number of degrees of freedom equals n_1+n_2-3. Apparently, there is a loss of one degree of freedom that has, however, no large practical consequence if n_1+n_2 is not small. We shall ignore this in further comparisons. If the hypothesis is not true, and the removal efficacy is different for the two brushes, the F-test has a non-central F-distribution with a non-centrality parameter δ^2 that is given by the expression $\delta^2\sigma^2 = \sum_k n_k B_k^2$. A large value of δ^2 gives a high power to the test. For fixed B_k's this means that σ^2 should be small. We can then simply compare the four test statistics by means of the above variances σ^2. These variances are summarized in Table 1.

Discussion

The expressions summarized in Table 1 illustrate that the variances for Approaches 3 and 4 are exactly the same. In other words, if the scores before brushing are taken into account as a co-variable, then the power of the F-test is the same for the scores after brushing only as for the differences in scores before minus after brushing. In the particular case of $\sigma_e \approx 0$, Approaches 2, 3 and 4 yield the smallest variance and, hence, the highest power. If $\sigma_S \approx 0$, Approaches 1, 3 and 4 give the highest power. In practice, however, σ_e and σ_S are different from zero, and in that case Approach 3 or 4 is the most effective, both resulting in the same power. As the use of Approach 4 is more complicated, we advise the use of Approach 3 in all cases.

In certain cases, the models as proposed in this section do not hold for the original data but for the data obtained after transformation by the logarithm. Then the same conclusions are true, but for the transformed data. An interesting observation can be made with respect to Approach 5 (see subsection ''Introduction''). This approach uses the variable in test $(Y_{i1k}-Y_{i2k})/(Y_{i1k})$. This variable can be seen as a linear approximation of the variable log (Y_{i1k})–log (Y_{i2k}) that is used in Approach 2 when we work with the logarithm of the data. This means that Approach 2 applied to the logarithm of the data is expected to give approximately similar results as Approach 5 applied to the original data.

In the next main section on ''numerical analysis'', the preferred Approach 3 is also applied to the logarithm of the data, and the results are compared.

WSD

The statistical model

In the case of a WSD, the model for the score before brushing contains an additional factor that refers to the part in the mouth that is brushed with brush k. This results in

$$Y_{i1k} = \mu_1 + S_i + d_{ik} + e_{i1k}.$$

Here, $i = 1, \ldots, N$ with N being the total number of subjects. μ_1 is again the general mean, S_i the effect for subject i, d_{ik} the effect of the part of the mouth where brush k is applied by subject i and e_{i1k} the error term. It is again assumed that the variables S_i, d_{ik} and e_{i1k} are independent random variables, which are normally distributed around 0 with variances equal to σ_S^2, σ_d^2 and σ_e^2 for S_i, d_{ik} and e_{i1k}, respectively. As a consequence, the expected value of Y_{i1k} equals $E(Y_{i1k}) = \mu_1$ and the variance is var$(Y_{i1k}) = \sigma_S^2 + \sigma_d^2 + \sigma_e^2$.

Equivalently, the model for the scores after brushing is given by

$$Y_{i2k} = \mu_2 + S_i + d_{ik} + e_{i2k} + B_k,$$

with B_k being the effect of the kth toothbrush ($k = 1$, 2 for the two brushes). Again, we assume that

Table 1. Variances σ^2 for a BSD under four approaches

Approach	Variable in test	Variance σ^2
1. Score after brushing	Y_{i2k}	$(\sigma_S^2 + \sigma_e^2)$
2. Score before minus score after brushing	$Y_{i1k}-Y_{i2k}$	$(\sigma_S^2 + \sigma_e^2)$
3. Score after brushing with centred score before brushing as a co-variable	$Y_{i2k}\|(Y_{i1k}-Y_{.1k})$*	$(\sigma_S^2 + \sigma_e^2)(1-\rho^2)$ with $\rho^2 = \sigma_S^4/(\sigma_S^2+\sigma_e^2)^2$; equivalently:
4. Score before minus score after with centred score before as a co-variable	$(Y_{i1k}-Y_{i2k})\|(Y_{i1k}-Y_{.1k})$*	$(\sigma_e^2 + \sigma_e^2)[1 - \sigma_e^4/\{(\sigma_S^2+\sigma_e^2)(\sigma_e^2+\sigma_e^2)\}]$

*The co-variable has been positioned to the right of the |-symbol. Centred scores have been used for the co-variable in order to simplify the calculations.
BSD, Between-Subjects Design.

Table 2. Variances σ^2 for the WSD

Approach	Variable in test	Variance σ^2
1. Scores after brushing only	$Y_{i22} - Y_{i21}$	$2\sigma_d^2 + 2\sigma_e^2$
2. Scores before minus scores after brushing	$(Y_{i12} - Y_{i22}) - (Y_{i11} - Y_{i21})$	$4\sigma_e^2$
3. Scores after brushing with centred scores before brushing as a co-variable	$(Y_{i22} - Y_{i21})\mid (Y_{i11} - Y_{\cdot 11}, Y_{i12} - Y_{\cdot 12})$	$2\sigma_e^2 \times f(\sigma_S, \sigma_d, \sigma_e)^*$
4. Scores before minus scores after brushing with centred scores before as a co-variable	$(Y_{i12} - Y_{i22}) - (Y_{i11} - Y_{i21})\mid (Y_{i11} - Y_{\cdot 11}, Y_{i12} - Y_{\cdot 12})$	$2\sigma_e^2 \times f(\sigma_S, \sigma_d, \sigma_e)^*$

*With $f(\sigma_S, \sigma_d, \sigma_e) = [\sigma_e^4 + 2\sigma_e^2\sigma_S^2 + 2\sigma_d^4 + 4\sigma_d^2\sigma_S^2 + 3\sigma_d^2\sigma_e^2]/[\sigma_e^4 + 2\sigma_e^2\sigma_S^2 + \sigma_d^4 + 2\sigma_d^2\sigma_S^2 + 2\sigma_d^2\sigma_e^2]$. WSD, Within-Subjects Design

$\Sigma B_k = 0$. Under similar assumptions as mentioned above, the expected value is $E(Y_{i2k}) = \mu_2 + B_k$ and the variance $\text{var}(Y_{i2k}) = \sigma_S^2 + \sigma_d^2 + \sigma_e^2$.

Test statistic and results for variances

The ANOVA *F*-test for the brushes in the WSD has a standard expression. With this design, too, the power of the test statistic depends on the parameter δ^2 in the expression $\delta^2\sigma^2 = 1/2\ N\ (B_2 - B_1)^2$. Again, we can compare the tests by the values of σ^2 (see Table 2).

Discussion

As in the case of a BSD, the variances of Approaches 3 and 4 are the same. The expression of $f(\sigma_S, \sigma_d, \sigma_e)$ varies between 1 for small values of σ_d and 2 for small values of σ_e. In the particular case that $\sigma_d \approx 0$, Approaches 1 and 3 yield the same variance and, as Approach 1 is simpler to use, this is the one to be recommended. It can be proved that Approaches 3 and 4 have the lowest variance; so, in general, the scores before brushing can profitably be included in the analysis as a co-variable. But, as Approach 3 is simpler, we recommend this one. The variance of this approach varies between $2\sigma_e^2$ for small σ_d and $4\sigma_e^2$ for small σ_e. Finally, similar remarks can be made as in the Discussion of the previous sub-section with respect to the transformation of the data by the logarithm. Also, the similarity between Approach 5 for the original data and Approach 2 for the transformed data holds in the WSD case.

Numerical Analysis

The preference order in statistical approaches that follows from the calculation of the power of the test statistics based on a simple model is compared in this chapter to the numerical results of the statistical analysis for two particular clinical trials. For this evaluation we selected one clinical trial with a BSD and one with a WSD.

Table 3. Overview of the results of the first clinical trial expressed as the mean value with the standard error of mean between brackets for both brushes separately

Brush	Mean score before brushing	Mean score after brushing	Δ
TB1	2.65 (0.51)	1.50 (0.34)	1.15 (0.40)
TB2	2.65 (0.49)	1.29 (0.45)	1.36 (0.49)

Δ represents the mean difference in score before minus after brushing.
TB, power-assisted toothbrush.

Description clinical trials used

BSD (Trial 1)

In this trial, the plaque removal efficacy of two power-assisted toothbrushes – hereafter referred to as TB1 and TB2 – was evaluated by recording the Q&H plaque index before and after brushing. The trial was based on a single-blinded, randomized, parallel-group design with 31 subjects in each group. After providing written consent, the subjects were enrolled for a baseline visit (visit 0). During this visit a baseline plaque level was determined and the subjects were randomly allocated to one of the two treatment groups (TB1 or TB2). The subjects were then requested to abstain from any form of oral hygiene for the next 48 h. Two days after visit 0 (i.e. at visit 1) the plaque level was determined before and after 3 min. brushing with the assigned brush. Brushing was preceded by a professional instruction on the use of the brush and subsequently supervised by a dental hygienist. The subjects were then requested to use the distributed brush for their daily oral hygiene for the next 2 weeks. After this habituation period, the same sequence of measurements was repeated (visit 2 and visit 3). The data used to evaluate the statistical approaches are the plaque indices before and after brushing at visit 3. Their mean values are summarized in Table 3.

WSD (Trial 2)

In this trial, the plaque removal efficacy of two toothbrushes – hereafter referred to as TB3 and TB4 – was compared in a cohort of 16 adults. The test protocol was based on a single-blinded, split-mouth design, in which each subject used each of the brushes in two opposed quadrants of the mouth. The plaque removal efficacy was determined by recording the Q&H plaque indices at six surfaces of all the teeth in the mouth. The baseline visit (visit 0) was preceded with a 2-weeks habituation period, in which subjects got detailed brushing instructions and were requested to brush daily with both brushes. At the baseline visit, the subjects were given a prophylaxis. They were then requested to abstain from all oral hygiene procedures for the next 24 h. On the next day (visit 1) plaque levels were determined on all the teeth before and after brushing. The 3 min. brushing was supervised by a dental hygienist, so that it was assured that each quadrant got 45 s cleaning time. The data of visit 1 are used for the analysis and their mean values are summarized in Table 4.

Analysis

As mentioned above the individual plaque scores measured on the six surfaces of all the teeth of a subject are first averaged to obtain full-mouth scores. In the case of the WSD, only those scores that correspond to one treatment (i.e. one brush) are averaged, yielding, in this particular case, two ''full-mouth''

scores per subject. The data are further statistically processed with the aid of the program SPSS, version 8.0 (SPSS Inc., Chicago, IL, USA).

For each of the approaches described above the hypothesized equality of the mean plaque removal efficacy of the two brushes is tested with a ''General Linear Model'' (GLM). For the BSD the factor brush is the only dominant effect taken into account in Approaches 1, 2 and 5. This implies that in these cases the *F*-test is equivalent to the independent-samples t-test. We confirmed that both tests yielded the same value for the significance level *p*. In the analysis of Approaches 3, 4 and 6 the scores before brushing are added as a co-variable. It was checked that the interaction between the scores before brushing and the brush itself has no statistically significant contribution in all these cases.

For the WSD the factor subject and the factor brush are taken into account in the GLM for Approaches 1, 2 and 5, whereas the factor score before brushing is added as a co-variable for Approaches 3, 4 and 6. Again, in the latter cases, the interaction between the scores before brushing and the brush itself has no significant contribution. In all cases the factor subject is highly statistically significant, meaning that the plaques scores vary strongly between the subjects.

In general, a *p*-value smaller than or equal to 0.05 is accepted as the limit to allow the conclusion that the two brushes differ in plaque removal efficacy. It means that there is a chance of 5% of this conclusion being abusively drawn if, in reality, the two brushes do not really differ. We will compare the statistical approaches on the basis of the *p*-values obtained in the *F*-tests. In the subsequent discussion, we will evaluate whether these results correspond to the guidelines resulting from the model calculations given in the previous section.

Table 4. Overview of the results of the second clinical trial expressed as the mean value with the standard deviation between brackets for both brushes separately

Brush	Mean score before brushing	Mean score after brushing	Δ
TB3	2.91 (0.56)	2.03 (0.48)	0.88 (0.37)
TB4	2.88 (0.57)	1.85 (0.52)	1.03 (0.33)

Δ represents the mean difference in score before minus after brushing.
TB, power-assisted toothbrush.

Table 5. *F*-values and *p*-values calculated with SPSS for the six different approaches defined for a BSD

Approach	*F*-value	*p*-value
1. Score after brushing	4.302	0.042
2. Score before minus score after brushing	3.518	0.066
3. Score after brushing with score before brushing as a co-variable	5.886	0.018
3b. Logarithmic score after brushing with logarithmic score before brushing as a co-variable	7.325	0.009
4. Score before minus score after with score before as a co-variable	5.886	0.018
5. Relative scores (before minus after brushing divided by scores before brushing)	6.621	0.013
6. Relative scores with scores before brushing as a co-variable	6.714	0.012

BSD, Between-Subjects Design.

Table 6. *F*-values and *p*-values calculated with SPSS for the six different approaches defined for the WSD

Approach	*F*-value	*p*-value
1. Score after brushing	7.475	0.015
2. Score before minus score after brushing	6.670	0.021
3. Score after brushing with score before brushing as a co-variable	9.064	0.009
3b. Logarithmic score after brushing with logarithmic score before brushing as a co-variable	8.453	0.011
4. Score before minus score after with score before as a co-variable	9.064	0.009
5. Relative scores (before minus after brushing divided by scores before brushing)	9.663	0.007
6. Relative scores with scores before brushing as a co-variable	8.673	0.011

WSD, Within-Subjects Design.

BSD

For each of the approaches introduced above, the *p*-values of the BSD study are summarized in Table 5 together with the *F*-values resulting from the GLM analysis. From the analysis result of Approach 2 we conclude that the two brushes do not really differ in plaque removal performance ($p > 0.05$), whereas the results of the other approaches lead to the conclusion that the two brushes do differ in their plaque removal efficacy. The analysis of the differences in scores before and after brushing apparently results in a higher *p*-value than the analysis of the scores after brushing only. Adding the scores before brushing as a co-variable further reduces the *p*-values. Indeed, Table 5 shows that the *p*-values resulting from Approaches 3, 4 and 6 are lower than those resulting from Approaches 1, 2 and 5, respectively. Performing the analysis with relative scores (Approaches 5 and 6) results in the lowest *p*-values.

In ''Statistical model and approaches'', we advised the use of Approach 3 when basing on power considerations. We also performed this approach on the data transformed by the logarithm. These results are presented in Table 5 as Approach 3b. In fact, when testing at the 5% level the conclusions are the same as those from Approach 3 applied to the untransformed data. Graphical inspection of the residuals with respect to normality does not lead to the conclusion that any of the methods is better in this respect. All in all, we conclude that the difference between the brushes is significant.

WSD

The *F*-values and *p*-values resulting from the analysis with each of the approaches introduced above are given in Table 6 for the WSD study. Here, all approaches lead to the same conclusion: a statistically significant difference in plaque removal efficiency exists between the two brushes. As in the BSD study, the *p*-values obtained with

Approaches 3 and 4 are the same. They both are lower than the *p*-values obtained with Approaches 1 and 2. This demonstrates that, here too, the *p*-values are decreased by extending the analysis with the scores before brushing as a covariate. The *p*-values are further reduced by performing the analysis of the relative difference scores, i.e. Approach 5. Including here the data before brushing as a covariate (i.e. Approach 6) yields a higher *p*-value.

Also here, Approach 3 is applied to the data transformed by the logarithm and the results are included in Table 6 as Approach 3b. The conclusion remains the same.

Discussion

The importance of making the right choice in a statistical approach is well illustrated by the particular example of the clinical trial with the BSD. The numerical results obtained in this experiment show that the conclusions drawn from the experiment depend on the choice of statistical approach, as statistical significance is not obtained with all six approaches. The efficacy of the brushes in the WSD is significantly different whichever approach is used.

The results of the numerical analysis fully support the conclusions drawn from the theoretical evaluation in that the *F*-value for Approach 3 equals that of Approach 4 for both the BSD and the WSD. In addition, considering for a moment the first four approaches only, Approach 3 yields the lowest *p*-value in both studies, being in agreement with the theoretical predictions. The confirmation of the theoretical results found for these two particular clinical trials does not necessarily imply that the analysis of any trial will be in agreement with the theoretical predictions. A disagreement between a trial's analysis and the theoretical results may have various causes:

- The guidelines for a choice between the different statistical approaches are based on the chance of finding an *F*-value that is large when the null hypothesis is false. This chance should be as high as possible, which is equivalent to saying that we optimize the chance of finding a statistically significant difference between the two brushes if it is there. This, however, does not necessarily imply that the *p*-value resulting from the *F*-test for the advised statistical approach is always lowest in an actual experiment. We are just optimizing the chance of finding a lowest *p*-value, which does not necessarily mean that we actually have to find it in any experimental verification.
- Centred scores before brushing are used in the model calculations for the different approaches, whereas the numerical analysis is based on non-centred scores. However, especially in the case of the WSD, in which both brushes are used by all the subjects, one would not expect a big difference between centred and non-centred scores.
- The statistical model is oversimplified for the description of a particular test.

We opted for a simple statistical model because that allows the power of the corresponding *F*-test to be calculated in an analytical way. At the choice of the model the following assumptions were made:

- the same variance for the error term for the two brushes used in the test;
- the same variance for the error term before and after brushing;
- the same variance for the subjects before and after brushing.

Basically, the statistical model can be expanded to account for unequal variances in one of these aspects. In practice, it is then no longer possible to calculate the power of the corresponding *F*-test in a simple way. By analysing the data of the two particular clinical trials used in this investigation, however, we have found that there is no statistically significant difference between the error variances for the two brushes and between the error variances before and after brushing. So, at least in these particular cases, this supports the validity of our assumptions.

The two studies analysed here show that the *p*-values can be further decreased by using relative instead of absolute difference scores. The difference in contribution to the brush effect for different initial plaque levels may be an explanation for this observation. When analysing the relative difference scores the values based on a low initial plaque level will have a higher weight (divided by a lower number) than the values based on a high initial plaque level (divided by a higher number). When a positive relation between the absolute difference scores and the scores before brushing exists, this difference in weighting is minimized. Nevertheless, if the difference in plaque removal efficacy between the two brushes is most prominent at low initial plaque levels, analysing the relative difference scores might increase the chance of finding a statistically significant difference between the two brushes. In none of the studies analysed here did we find a statistically significant relation between the relative difference scores and the scores before brushing. This might explain why adding the scores before brushing as a covariate in the analysis of the relative difference scores hardly changes the *p*-values. Eventually, a model different from the one introduced in ''Statistical model and approaches'' may apply here.

It was remarked by one of the referees that, with respect to the analysis of covariance, the scores before brushing can be hampered by measurement error. Measurement error is in fact already included in the residual error term in our model in ''Statistical model and approaches''. It may have an effect on the estimated value of the regression coefficient for the score before brushing. However, we apply the analysis of covariance just to obtain a reduction of the residual variation and not to obtain a proper estimate of this coefficient. Therefore, measurement error does not lead to any bias in our analysis.

The issue of (how) including the scores before a treatment in the statistical analysis of the scores after a treatment, as addressed in this investigation, is also relevant in completely different contexts. The results obtained here can be straightforwardly generalized to the case where three or more brushes are included in the experiment, with likely the same choice of testing approach. The conclusions can even be further generalized to medical trials that can be described by numerical scale data obtained before and after a given treatment on a cohort of subjects. Another generalization concerns different types of (dental) clinical experiments, in which e.g. longitudinal treatment effects may play a role. Here, too, a best testing approach may be found, but a few more terms like time effects may have to be included in the statistical models we used.

The covariance analysis that has to be performed in Approaches 3, 4 and 6 can easily be done with the statistical package SPSS, version 8.0, that we used; other generally available statistical packages such as BMDP, STATA and SAS will do it too.

Conclusions

We have investigated different statistical approaches to compare the chance of finding a statistically significant difference between two treatments in clinical trials in which data before and after a treatment were collected. This investigation has been worked out for the particular example of determining the difference in plaque removal efficacy of two different toothbrushes.

On the basis of a simple statistical model we have shown that in the case of a BSD as well as in the case of a WSD the *F*-test has the highest power with respect to an evaluation of the data after brushing when the data before brushing are included as a co-variable (i.e. Approach 3). Subtracting the data after brushing from the data before brushing does not add additional power to the test. If some variances would be equal to zero, the more simple Approaches 1 and 2 may also be satisfactory. However, this is generally not the case, and Approach 3 is best in all circumstances.

The numerical results of the clinical trials discussed in this paper – with a BSD as well as with a WSD – show that the calculated *p*-values support the conclusions drawn from the statistical model. They also illustrate that for these two particular studies the *p*-value can be decreased by analysing the relative plaque scores.

Acknowledgements

The authors gratefully thank Dr. Ir. M. de Jager, Dr. Ir. F. Starmans, Dr. D. Sturm and Dr. T. Denteneer for their contributions via fruitful discussions.

References

Addy, M. (1995) Evaluation of clinical trials of agents and procedures to prevent caries and periodontal disease: choosing products and recommending procedures. *International Dental Journal* **45**, 185–196.

Chilton, N. W. & Fleiss, J. L. (1986) Design and analysis of plaque and gingivitis clinical trials. *Journal of Clinical Periodontology* **13**, 400–406.

De Rouen, T. A. (1989) Biostatistical and methodological issues in demonstrating efficacy of therapeutic agents for periodontal disease. *Journal of Dental Research* **68**, 1661–1666.

Heasman, P. A. & McCracken, G. I. (1999) Powered toothbrushes – A review of clinical trials. *Journal of Clinical Periodontology* **26**, 407–420.

Heasman, P. A., Stacey, F., Heasman, L., Sellers, P., MacGregor, I. D. M. & Kelly, P. (1999) A comparative study of the Philips HP735, the Braun/Oral B D7 and the Oral B35 Advantage. *Journal of Clinical Periodontology* **26**, 85–90.

Kingman, A. (1992) Specific statistical considerations relevant to the design and analysis of gingivitis trials demonstrating product superiority or equivalence. *Journal of Periodontal Research* **27**, 378–389.

Quigley, G. A. & Hein, J. W. (1962) Comparative cleaning efficacy of manual and power brushing. *The Journal of the American Dental Association* **65**, 26–29.

Samuels, M. L. (1986) Use of analysis of covariance in clinical trials: a clarification. *Controlled Clinical Trials* **7**, 325–329.

Sharma, N. C., Galustians, J., Rustogi, K. N., McCool, J. J., Petrone, M., Volpe, A. R., Korn, L. R. & Petrone, D. (1992) Comparative plaque removal efficacy of three tooth brushes in two independent clinical studies. *The Journal of Clinical Dentistry Supplements C* **3**, C13–C20.

Sullivan, L. M. & D'Agostino, R. B. (1992) Robustness of the t-test applied to data distorted from normality by floor effects. *Journal of Dental Research* **71**, 1938–1943.

Turesky, S., Gilmore, N. D. & Glickman, I. (1970) Reduced plaque formation by the chloromethyl analogue of vitamin c. *Journal of Periodontology* **41**, 41–43.

van der Weijden, G. A., Timmerman, M. F., Reijerse, E., Danser, M. M., Mantel, M. S., Nijboer, A. & van der Velden, H. (1994) The long-term effect of an oscillating/rotating toothbrush on gingivitis. An 8-month clinical study. *Journal of Clinical Periodontology* **21**, 139–145.

Address:
J. Engel
CQM Building HCZ-3
P.O. Box 414
5600 AK Eindhoven
The Netherlands
E-mail: engelj@natlab.research.philips.com