

This material may be protected by copyright law (Title 17 U.S. Code).

Measurement and Expression of Treatment Effects in Caries Clinical Trials

H. S. HOROWITZ

National Institute of Dental Research, Bethesda, Maryland 20205

J Dent Res 63(Spec Iss):709-712, May, 1984

This talk will cover a wide array of topics, including:

- (1) ways of expressing caries inhibition,
- (2) the use of confidence intervals,
- (3) clinical significance vs. statistical significance, and
- (4) cost-effectiveness.

Each of these topics merits a full presentation or might even justify an entire conference. However, within my allotted time, I shall attempt to give you some thoughts on each.

Ways of expressing caries inhibition.

The usual design for a clinical trial of a caries-preventive agent or method compares the DMF surface increment in a group of individuals who have received a preventive agent for various periods (usually years) with the increment in a similar, placebo-treated or untreated control group.

Most trials test the null hypothesis — that any observed differences between test and control groups result solely from chance errors of sampling, that is, that no real difference exists between the groups.^{1,2} Sometimes more than one treatment group is included in the design, and, particularly in recent years, ethical concerns have dictated omission of an untreated or placebo-treated control group, so that only one or more treatment regimens are compared with a standard regimen, or positive control. However, if a negative control group is not included in the study's design, measuring the absolute effectiveness of a test agent is impossible; only its relative effectiveness compared with the positive control can be ascertained.³

Studies with control groups require statistical analysis to determine whether the observed difference in DMFS increments between the test and control groups is real (a two-tailed test) or if an observed smaller increment in the test group is truly different from that in the control group (a one-tailed test). A decision to use one- or two-tailed tests should be made during the planning phase of a study, and not after the data are collected and tabulated.⁴ When only treated groups are compared, analysis is done to determine if statistically significant differences exist between or among the test groups, and a two-tailed test should be used for statistical analysis.

Most investigators present mean scores and a standard measure of the variation in observations, such as a standard deviation, for each group. Whether the differences in means are significant depends on the numbers of participants in each group, the magnitude of the differences in mean DMFS scores, and the variation in those scores among participants in each group. Generally, large differences in increments, small variation, and large samples will increase the likelihood of achieving statistical significance. Levels of significance may be set at 5% or 1% or some other value, depending on the confidence the investigators want that an observed difference is real.

Statements about significance in a clinical trial are usually based on the differences in mean increments and not on percentage differences. Yet, most investigators frequently

speak of statistically significant percentage reductions in decay. Moreover, most reviews compile results of studies expressed as percentage inhibitions.⁵ Using abstract percentage figures without reference to actual mean values, however, can be highly misleading. As shown in Table 1, a mean annual increment of 1.4 DMFS is 30% lower than a mean annual increment of 2.0, but so is a mean annual increment of 0.42 30% lower than one of 0.6. In the first instance, the average child in a test group has developed 0.6 fewer DMF surfaces in a year's time than one in the control group, whereas, in the second example, the average difference is only 0.18 surfaces. Both differences may be statistically significant, but the differences between the hypothetical findings in the two examples are great in clinical significance and, perhaps, in cost-effectiveness.

The strength of the differences in statements of statistical significance is unexpressed.⁶ For example, a significant difference in mean DMFS increments between a test and control group at a level of 0.05 merely means that a difference that large or larger is unlikely to occur by chance 5% of the time, or one time in 20, if the study were repeated an infinite number of times with random samples of the same size drawn from the same population. A highly significant difference denotes only that the probability that the observed difference will occur by chance is less than 1%, or one time in 100, if the study were repeated infinitely. Even when precise probability levels (P) of detecting the observed difference are presented, we know neither the strength of the effect nor the range of likely outcomes for the investigation (confidence interval).

Although a controlled prospective study, with concurrent test and control groups, is the best design for evaluating cariostatic measures, ethical considerations are causing more and more studies of known-to-be-effective measures to be done with no control group. In these studies, the caries prevalence of participants is compared with retrospectively established baseline data of children drawn from the same population before the preventive measure was started, as shown in Table 2. Although this type of design has been described as "... notoriously fallacious"⁷, it has been used for most evaluations of the caries-preventive effects of community water fluoridation, school water fluoridation, and for the 17 demonstration projects of weekly mouthrinsing with dilute fluoride solutions sup-

TABLE 1
TWO EXAMPLES OF 30% REDUCTIONS IN ANNUAL DMFS INCREMENTS

| Group | Study No. 1 | | Study No. 2 | |
|------------|----------------|-----------------------|----------------|-----------------------|
| | DMFS increment | Percentage difference | DMFS increment | Percentage difference |
| Control | 2.0 | — | 0.6 | — |
| Test | 1.4 | 30% | 0.42 | 30% |
| Difference | 0.6 | | 0.18 | |

ported by the National Caries Program. In fact, the value of many thoroughly accepted medical drugs — including morphine, digitalis, aspirin, insulin, penicillin, and the corticosteroids — was established in studies without concurrent comparisons.⁴ Although this design has been criticized for valid reasons and is particularly hazardous now, when the prevalence of dental caries may be changing, several checks for internal consistency of data can be made in such studies to ascertain the validity of their results. My colleague, Dr. Heifetz, will discuss these checks on validity later at this Conference. I must point out, however, that estimates of caries reductions made from studies with this design are compromised, and confidence limits cannot be estimated nor P-values derived from these types of data.

The way that caries inhibition is expressed in studies of pit-and-fissure sealants differs somewhat from either of the two methods discussed so far. In most studies of sealants, a particular sealant is placed over the pits and fissures of teeth in one-half of the mouth, and the pits and fissures of homologous teeth in the other half of the mouth are either left unsealed or are covered with another sealant. Because subjects serve as their own controls, this design has the advantage of avoiding variability among subjects. It has the disadvantage of reducing the number of observations of teeth that receive a particular preventive regimen in a subject. The number of sealed teeth that become carious after various yearly intervals in each half of the mouth is ascertained and reported in one of several ways, for example, as shown in Table 3. Note that, in the example, tooth pairs in which both the sealed tooth and its homologous pair remain sound do not enter the calculation of percentage differences or effectiveness, and that pairs in which both sealed and unsealed teeth become carious are entered in both halves of the numerator and in the denominator.

Use of confidence intervals.

Statistical tests tend to focus attention on whether the results of a study are considered statistically significant (P

value is less than 0.05) or highly significant (P less than 0.01). This emphasis on cut-off points serves well to decide between alternative actions based on the statistical results of a study, but it fails to characterize the extent of a difference between groups.⁶ Highly significant P values sometimes result from negligible differences if a study has many participants, and, conversely, P values that are far from significant cut-off points may occur when differences are great if a study has few participants. The tendency to characterize the results of a study as "yes (significant)" or "no (not significant)" may be misleading, if the study was poorly designed, with an improper number of participants. Even if it was well-designed, P values of, for example, 0.06 and 0.04 should not intellectually lead to definitively opposite conclusions but to closely similar ones.

An analysis that measures the possible range of an effect by calculating the confidence interval for the observed result may be more informative. Confidence intervals may be calculated for actual differences in means or for percentage differences.^{5,6,8} Discussion of findings usually occurs in terms of percentage reductions for reasons of convenience and ease of interpretation. Therefore, I shall discuss confidence intervals for percentage reductions.

Assume that results shown in Table 4 occurred in a clinical trial after one year. The findings show a difference of 0.32, or 20%, in the mean DMFS increment between the test and control group. The observed 20% difference, based upon the particular set of circumstances that existed in this specific trial, should be considered only as an approximation of the effectiveness of the test agent. The percentage reduction provides only a convenient point-estimate of the treatment effect, but a more informative statistic is a confidence interval or the range of results likely to include the true percentage reduction.⁹

In the example in Table 4, the observed difference is not significant, because one this large could have occurred by chance six times in 100 if the study were repeated an infinite number of times with samples drawn from the same population. The 95% confidence interval for the observed percentage reduction includes the zero percentage value, which confirms the lack of significance. A decision on whether to adopt the test procedure based on the P value must lead to its rejection as ineffective. However, the 95% confidence interval for the observed percentage reduction also extends to a 41% reduction in dental caries, or more than 0.65 fewer DMF surfaces in the test group than in the control group — a finding consistent with a conclusion of a fairly strong effect of the test procedure. Rothman contends that it is unlikely that a confidence interval including both zero and a percentage indicating a strong treatment effect within its boundaries represents the findings of a worthless preventive agent.⁶ However, his statement is valid only if the study is designed with sufficient power, and in that case, the confidence interval is not likely to

TABLE 2
AVERAGE PREVALENCE OF DMFS IN 12-TO-15-YEAR-OLDS
BEFORE AND AFTER INTRODUCTION OF A
PREVENTIVE REGIMEN

| Age | Average Baseline DMFS (1980) | Average Post-treatment DMFS (1983) | % Difference from baseline |
|-----|------------------------------------|--|-------------------------------|
| 12 | 8.5 | 5.8 | 31.8 |
| 13 | 11.0 | 8.6 | 21.8 |
| 14 | 13.8 | 10.2 | 26.1 |
| 15 | 17.3 | 12.4 | 28.3 |
| All | 12.65 | 9.25 | 26.9 |

TABLE 3
OCCLUSAL CARIES STATUS OF SEALED/UNSEALED PAIRS OF STUDY TEETH, BY ARCH, AFTER FIVE YEARS

| Arch | No. of Study Prs. | Occlusal Caries Status of Tooth Pairs | | | | % Diff. |
|------------|----------------------|---------------------------------------|-------------------------------|-----------------------------|---------------------------------|------------|
| | | Treated Sound, Control DMF | Treated DMF, Control Sound | Treated DMF, Control DMF | Treated Sound, Control Sound | |
| Maxillary | 260 | 50 | 20 | 40 | 150 | 33.3 |
| Mandibular | 240 | 60 | 10 | 30 | 140 | 55.6 |
| Both | 500 | 110 | 30 | 70 | 290 | 44.4 |

* No. of pairs with DMF control teeth — No. of pairs with DMF sealed teeth

No. of pairs with DMF control teeth

TABLE 4
FINDINGS AFTER ONE YEAR

| Group | Number of Subjects | Mean DMFS Increment | Percentage Reduction | Significance | P Value | 95% Confidence Interval |
|---------|--------------------|---------------------|----------------------|--------------|---------|--------------------------|
| | | | | | | for Percentage Reduction |
| Control | 150 | 1.60 | — | | | — |
| Test | 147 | 1.28 | 20.0 | N.S. | 0.06 | -1.1 to 41.1% |

contain both zero and a high percentage reduction. It must be stressed that when the confidence interval for an observed percentage reduction is broad, the breadth indicates a lack of assurance of the true reduction — in other words, the trial has provided only a vague determination of an agent's effectiveness.⁹ Moreover, the trial was probably designed with inadequate power — that is, the probability of missing a real treatment effect was high.¹⁰ This situation undoubtedly exists in the hypothetical example in Table 4.

The least informative way to present the outcome of a statistical analysis of data is to report merely that an observed difference is either significant or not significant. This approach focuses all interest on the lower boundary of a confidence interval. An actual P value is somewhat more informative; however, a notation that $P > 0.05$ says no more than that the findings are not significant.

The most informative way to present the findings is to give an unambiguous summary of the strength of the treatment effect in the form of a confidence interval.^{6,11,12} In the hypothetical example, an indication that the difference in means of 0.32 DMFS is not significant tells us little, only that the difference could occur by chance more than 5% of the time. The information that $P = 0.06$ shows us that this chance was only slightly greater than 5% of the time, but presentation of the full confidence interval indicates that the findings, although compatible with a conclusion of no preventive effect, may also be compatible with one of a potentially strong preventive effect. The breadth of the confidence interval also denotes the precision achieved by the study design. Therefore, giving confidence intervals in connection with tables that present findings of a clinical study of caries prevention will help readers ascertain the strength of an observed effect.⁶

Clinical significance vs. statistical significance.

Readers of scientific reports should be wary of poorly conceived or carelessly conducted studies that provide statistically significant findings. Statistical significance *per se* denotes nothing about the biological or clinical meaning of a difference in numbers or values. Readers who think that statistically significant results automatically authenticate the value or meaning of a study are deluded.

It is most important to distinguish between statistical significance and clinical significance. The former is a mathematical expression of the degree of confidence that an observed difference between groups is a real difference, that a zero-response would not occur often if the study were repeatedly done, and that the observed difference is not due merely to chance.¹³ In contrast, clinical significance is a judgment made by an investigator or reader that differences between groups in response to intervention are important for health.^{4,14} Clinical significance is a subjec-

TABLE 5
THREE-YEAR RESULTS BY STUDY GROUP

| Group | No. of Children | DMFS Increment | Percentage Reduction | 95% Confidence Interval |
|------------|-----------------|----------------|----------------------|-------------------------|
| Control | 597 | 2.50 | — | — |
| Test | 589 | 2.00 | 20.0% | 9.6 to 30.4% |
| Difference | | 0.50 | | |

tive determination based upon clinical experience and an understanding of the characteristics of the disease or condition being measured.

Consider the summarized hypothetical results of a study shown in Table 5. The children in the test group, living in a fluoridated community, received semi-annual applications of a professionally applied, topical fluoride gel for three years. The three-year increments of dental caries are low in both groups, undoubtedly because the children have consumed fluoridated water, many since birth. Children in the test group have developed 0.5 fewer DMF surfaces during the course of the study than their counterparts in the control group. The difference, representing a 20% reduction in dental caries, is statistically significant, probably because the groups are very large. However, it is doubtful whether the benefit to the test group children can be considered clinically significant, if one considers that it necessitated six professionally applied treatments, each of which may have required one-half hour to give. In other words, two children had to be treated six times during a three-year period in order to prevent one DMF surface from developing — hardly a clinically significant result.

If a difference in caries increments between test and control groups is judged clinically insignificant or lacking in practical importance, tests of the data for statistical significance are unlikely to contribute much to the interpretation of the findings. A trivial difference between study groups may be statistically significant when excessively large samples have been studied. If an observed difference is thought to be clinically significant, determining whether the difference is statistically significant or could easily have occurred by chance is important. But, a statistically significant result does not necessarily increase its clinical importance.¹⁵ Whenever possible, studies should be designed and sample sizes should be chosen so that outcomes of analyses of statistical and clinical significance will closely agree. Considerations of the practical value of the differences between caries increment in a test and control group, or clinical significance, are often related to considerations of cost-effectiveness.

Cost-effectiveness.

Cost-effectiveness analysis is a formal and systematic way to determine the least expensive of several alternative methods of achieving a stated objective.^{16,17} The term is frequently misused in biomedical literature, where one sees titles and passages of reports claiming that a particular treatment, diagnostic approach, or preventive method is or is not cost-effective, when what is really meant is that the treatment, approach, or method is or is not worth its cost. Because cost-effectiveness analysis, by definition, entails a comparison of the costs of two or more ways to achieve a desired goal¹⁶, one cannot compute the cost-effectiveness of an isolated regimen.

less than
to decide
al results
of a dif-
ues some-
has many
from sig-
are great
character-
"no (not
as poorly
s. Even if
0.06 and
opposite

an effect
observed
rvals may
r for per-
s usually
as of con-
shall dis-
red in a
difference
ween the
ce, based
xisted in
approxi-
percent-
timate of
tistic is a
o include

erence is
occurred
eated an
the same
observed
ge value,
ision on e
P value
the 95%
eduction
or more
an in the
ision of a
nan con-
cluding
reatment
ngs of a
ement is
t power,
likely to

| % Diff. |
|---------|
| 33.3 |
| 55.6 |
| 44.4 |

Although several investigators have discussed cost-effectiveness and suggested formulae for its calculation¹⁷⁻¹⁹, no one methodology is presently widely accepted. The costs of a caries-preventive program comprise the obvious expenses of providing the specific regimen, but determining these costs is not always free of difficulties. Although the costs for facilities, equipment, personnel, materials, and supplies are straightforward, it is debatable whether to include in the analysis the costs of such items as services of school personnel, overhead for school-based programs, and travel costs or work-time lost by parents who take children to an office for preventive treatments.¹⁶

Another problem in determining costs is that they are most often ascertained from studies funded by national institutions or industry. The methods, conditions, and personnel used in these studies often approach the ideal, which can inflate both the estimated costs for a procedure and the observed effectiveness compared with studies done under more realistic conditions and with public financing. O'Mullane has advocated distinguishing between studies that determine clinical effectiveness, or efficacy, and those in which community effectiveness is measured.²⁰ However, his approach focuses almost entirely on effectiveness, rather than on the costs of the two types of studies.

The effectiveness portion of cost-effectiveness analysis may be expressed in monetary terms (such as the most efficient way to reduce costs for restorative care for a community's schoolchildren by \$100,000) or in physical terms (such as the cheapest way to reduce the annual incidence of dental caries among schoolchildren in a community by one-half DMF surface).

Effectiveness will vary according to the thoroughness with which a preventive procedure is carried out, regardless of whether it is professionally applied or self-applied. It will also be affected directly by the caries susceptibility of the target population. Moreover, the frequency of applications will also influence effectiveness and costs. It should be evident that, in general, effective methods of caries prevention that are self-applied by large groups under general supervision are more cost-effective than those that are professionally and individually applied by professional personnel.^{16,17}

In common parlance, cost-effectiveness analysis permits program administrators to get "the biggest bang for the buck". Exercises in determining the efficiency of preventive programs can help administrators set aside pre-conceived notions on the value of various preventive regimens, view their activities objectively as they affect target populations, and measure critical monetary issues that affect decisions in these hard economic times.

REFERENCES

- OSBORN, J.: Statistical Principles in the Analysis of Dental Data. In: Dental Public Health - An Introduction to Community Dentistry, Slack, G.L. and Burt, B.A., Eds., Bristol, England: John Wright & Sons, 1974, pp. 221-244.
- HOROWITZ, H.S.; BAUME, L.J.; BACKER DIRKS, O.; DAVIES, G.N.; and SLACK, G.L., Eds.: Commission on Classification and Statistics for Oral Conditions, Federation Dentaire Internationale: Principal Requirements for Controlled Clinical Trials of Caries Preventive Agents and Procedures, *Int Dent J* 23:506-516, 1973.
- HEIFETZ, S.B. and KINGMAN, A.: The Impact of Ethical Considerations on Future Dental Caries Research: Clinical Field Trials, *J Dent Res* 59:1337-1340, 1980.
- HOROWITZ, H.S.: Evaluation of Scientific Information. In: Dentistry, Dental Practice, and the Community, 3rd ed., Striffler, D.F., Young, W.O., and Burt, B.A., Eds., Philadelphia, PA: W.B. Saunders, 1983, pp. 257-290.
- MARTHALER, T.: Statistical Treatment of Percentage Inhibitions of Human Dental Caries Incremental Data, *Helv Odont Acta* 15:15-29, 1971.
- ROTHMAN, K.J.: "A Show of Confidence" (editorial), *N Engl J Med* 299:1362, 1978.
- GARDNER, A.F.: Clinical Testing of Oral Pharmacological Agents, *J Can Dent Assoc* 34:29-31, 1968.
- WALLENSTEIN, S.; FLEISS, J.L.; and CHILTON, N.W.: Confidence Intervals for Percentage Reduction in Caries Increments, *J Dent Res* 61:828-830, 1982.
- DePAOLA, P.F.: The Interpretation of Findings in Clinical Trials, *J Dent Child* 41:11-18, 1974.
- Commission on Oral Health, Research and Epidemiology, Federation Dentaire Internationale: Principal Requirements for Controlled Clinical Trials of Caries Preventive Agents and Procedures, 3rd ed., *Int Dent J* 32:292, 1982.
- DUBEY, S.D.; LEHNHOFF, R.W.; and RADIKE, A.W.: A Statistical Confidence Interval for True Per Cent Reduction in Caries-incidence Studies, *J Dent Res* 44:921-923, 1965.
- ABRAMS, A.M.; McCLENDON, B.J.; and HOROWITZ, H.S.: Confidence Intervals for Percentage Reductions, *J Dent Res* 51:492-497, 1972.
- MILLER, S.L.: Introductory Statistics for Dentistry and Medicine. Reston, VA: Reston Publishing Co., 1981, pp. 145-166.
- SWANGO, P.A.: What is the Difference Between Statistical Significance and Clinical Significance? *Dent Hygiene* 54:12, 1980.
- SUOMI, J.D.: Evaluating the Effectiveness of Preventive Dental Programs, *J Public Health Dent* 34:56-59, 1974.
- HOROWITZ, H.S. and HEIFETZ, S.B.: Methods for Assessing the Cost-effectiveness of Caries Preventive Agents and Procedures, *Int Dent J* 29:106-117, 1979.
- HEIFETZ, S.B.: Cost-effectiveness of Topically Applied Fluorides. In: The Relative Efficiency of Methods of Caries Prevention in Dental Public Health, Burt, B.A., Ed., Ann Arbor, MI: University of Michigan, 1978, pp. 69-104.
- DAVIES, G.N.: Cost and Benefit of Fluoride in the Prevention of Dental Caries. Geneva: World Health Organization, Publication No. 9, 1974.
- DOHERTY, N.; POWELL, E.; and SMITH, M.: Cost-effectiveness Analysis of Alternative Preventive Treatments: First Year Results, *IADR Progr & Abst* 57:No. 327, 1978.
- O'MULLANE, D.M.: Efficiency in Clinical Trials of Caries Preventive Agents and Methods, *Community Dent Oral Epidemiol* 4:190-194, 1976.