Management and Evaluation of the Effects of Misclassification in a Controlled Clinical Trial: Discussion of Dr. Bell's Presentation

M. R. SMITH

Department of Mathematical and Computer Sciences, Michigan Technological University, Houghton, Michigan 49931

J Dent Res 63(Spec Iss):735, May, 1984

I would like to thank Dr. Bell for a well-prepared and important paper on the effects of misclassification in a controlled clinical trial. I agree with most aspects of the paper and certainly on his overall conclusion that we really do not take "reliability" scores very seriously. I would like to comment on certain aspects of the paper which may not generalize to all caries clinical trials or with which I felt uncomfortable. Please take these comments in the context of – I hope – generating some discussion.

In discussing problems caused by systematic bias with multiple examiners, Dr. Bell implies that it doesn't appear to be important to maintain examiners over two years. It has been my experience that this is critical to reducing the variability in a study. The pairing of examinations by examiner has the additional advantage of being able to block in the analysis model, therefore testing the assumption that examiner differences are important. In two specific studies, I can recall that this blocking has paid off in terms of reducing the estimate of error variance.

Reliability vs. *Reproducibility*. – Reliability as Dr. Bell has defined it is in terms of being reproducible. What is the criterion of a good examiner? The bottom line is that he should be able to detect changes in the study populations, if indeed they do exist. In many cases, this means he is forced into calling caries at the earliest possible moment. This is particularly important if the regimen being tested gives a protective effect on a sound surface only. In this case, it is imperative that the examiner call the early lesion, and consequently he puts his reproducibility results in jeopardy. It is extremely easy to be a consistent examiner, and placing an inordinate amount of emphasis on that aspect will tend to make for a very poor and conservative examiner. Further, in cases where the examiner is rewarded for consistency, I fear he will give a low power performance. Power is used here in terms of statistical power – the ability to detect real differences between populations.

Should calibration sessions be held throughout the study? I think that the necessity for doing this must be carefully weighed with not only the dollar cost but also the human cost in terms of the examiner. It is essential that the examiner remain confident and consistent in his criteria for calling caries. Too many reviews of reliability make for a very poor and (justifiably) slightly paranoid individual. It is always true that an examiner will have a bad day or a poor run of exams. It is not beneficial to remind him of this continually. The situation reminds me of some of W. Edward Deming's work in industrial quality control. Adjusting your process too often can result in setting up a situation of huge swings in your output and can even result in oscillating the process out of control. I believe that once a study begins you have to put your trust in the examiner and assume his objectivity will prevail, and even if it doesn't, you have blindness to protect the integrity of the study. Too much calibration and continual update can increase variability.

I am not aware that there are any studies that prove that calibration at any level will reduce the variability of the data. Has anyone ever run a trial in which one examiner is monitored and the others are not? One would have to randomize the examiners into at least three groups: one that is not monitored, one that is monitored during the trial, and one that is simply *told* it is being monitored. If the latter group wins, then some of the cost issue, at least, can be put to rest. Do we have any hard scientific evidence that periodic monitoring has any beneficial effect on the study?

I agree wholeheartedly with Dr. Bell's section on the issues of reporting reliability results. The results do depend heavily on the study population. They also depend on the methods of carrying out the reliability study. They also depend on the statistics one chooses to report the reliability, on the criteria that are important to the regimen in question, and on the prevalence incidence levels of the population. There are other issues that need to be identified as well. For example, what other parameters affect the reliability of the examiner? I can think of two easy ones: the time of day that the examination is done, and how the examiner feels during that day. I don't believe we have reached a consensus on the "best" statistical methods to measure reliability or reproducibility in a caries clinical trial.

Dr. Bell and I are in complete agreement in terms of training and calibration, particularly new examiners, before the trial begins. Philosophically, we do not agree on the monitoring aspect of reproducibility, especially what is to be done with those data. I personally believe the data should not be used to choose between two good examiners. Their usefulness lies in:

(1) eliminating the examiner who has gone completely off (This is a very, very rare case and can usually be detected by other methods.); and

(2) discovering the parameters that are important to the individual examiner's reliability, *i.e.*, is there a daily time trend effect or weekly or even longer? (This could also be school or place effect, etc. If these are identified, they can be taken into account in the analysis phase of the project.) These methods need to be developed, and certainly this Conference is a big step in that direction.

In conclusion, I am not convinced that the increased monitoring of the examiner is beneficial. I am, for once – a very rare event – in complete sympathy with the examiner on this issue. It is my personal opinion that increasing the pressure on the examiner will decrease his reliability. if not in percentage terms, then certainly in terms of slowing him down. Most steps an examiner takes to increase his reliability (I mean a good examiner, of course) will tend to slow him down. This amounts to a reduction in sample size. The reduction in sample size counteracts the gain in reliability. The trade-off point will be different for each examiner, and perhaps monitoring will help each examiner find that point. This may be the beneficial aspect of reliability monitoring. It must, however, be viewed by all concerned as being beneficial to the examiner. All parties are interested in an efficient clinical trial.

I again thank Dr. Bell for his most informative paper. I hope a consensus can be reached on some of the issues raised. We are in complete agreement that the problem has been under the rug far too long.

Interim Clinical Evaluations in Caries Clinical Trials

S. POULSEN and E. KIRKEGAARD

Department of Pedodontics and Preventive Dentistry, Royal Dental College, Aarhus, Denmark

J Dent Res 63(Spec Iss):736-738, May, 1984

Introduction.

In the design of caries clinical trials, specific decisions have to be made before the commencement of the trial. Such decisions are: the minimum caries reduction which is considered of interest, choice of method of allocating subjects to different groups, and the duration of the trial. Furthermore, the minimum number of subjects needed to test the null hypothesis for given values of α and β must be determined from an estimate of the variability in caries increment during the trial and the minimum treatment effect which is considered to be of clinical relevance.

These decisions are crucial, because once they have been made and the trial started, they cannot be changed.

The length of caries clinical trials is still under discussion. Most trials are conducted for two years, and many investigators seem to consider three years as an ideal. If the time needed to plan a caries clinical trial and the time needed to analyze the data are added to the trial period, four to five years may well elapse from the time a trial is conceived until the results are available and can be applied. This may explain why interim evaluations have been included in so many clinical trials, and why results are published before the final examinations have been made.

However, it could be discussed whether interim clinical evaluations should be carried out at all, once the duration of the trial has been determined. So far, little consideration has been given to the rationale for including interim clinical evaluations as a standard component in clinical trials. Actually, such evaluations may result in erroneous conclusions, unless proper statistical methods are applied. Thus, we need to reconsider the importance of interim clinical evaluations in caries clinical trials.

The Table shows a summary of statements made by different authors, including the working groups under FDI, on the subject of re-examinations in caries clinical trials.

The purpose of interim clinical evaluations.

As can be seen from the Table, only a few papers have dealt with the problem, and scant information is found in the literature as to the stated purpose of interim clinical evaluations. Because statistical evaluation of differences between groups is carried out at interim clinical evaluations, it is possible that the purpose has often been to determine the time necessary to obtain a statistically significant difference. This may cause problems, however, concerning the values of α and β. The problem can be illustrated by the example of drawing white and black chips from a hat to test the null hypothesis that the number of black chips is equal to the number of white chips. Assuming that six chips are drawn and all are either black or white, the null hypothesis can be rejected, because the probability of this event under the null hypothesis is less than 0.05 ($2 \times \frac{1}{2}^6 = 0.031$). If just one of the chips has a color different from the others, a new set of three chips is drawn. If these three are of the same color as the first five, we have to reject the null hypothesis, since we now have eight chips out of nine of the same color. The probability of this under the null hypothesis is also less than 0.05 ($12 \times \frac{1}{2}^9 = 0.023$). If one of the last three drawn chips is different in color from the other two, the experiment is continued.

The problem illustrated by this example can easily be applied to caries clinical trials. If a trial continues for a sufficiently long period of time, and the data are tested at each interval, the probability of committing a type one error increases. If it is planned to analyze interim clinical evaluations as the trial progresses, then the investigator is actually performing a sequential trial, and the methods available for analysis of this type of trial should be applied. It would, however, seem advisable to clarify the advantages and disadvantages inherent in this design before recommending its general adoption.

Interval between interim clinical evaluations.

Most authors recommend intervals of one year between re-examinations. Two main reasons for such a recommendation are:

(1) the lack of reproducibility of the clinical caries examination (Grainger, 1972), or
(2) lack of difference in caries increment between the various groups in the trial (Horowitz *et al.*, 1973; FDI, 1982).

Lack of reproducibility.

It has often been stated that, due to the problems of lack of repeatability inherent in the clinical diagnosis of dental caries, small differences in increments cannot be identified. Lack of repeatability can be due to two different types of error: *systematic error* or *random error.*

In the caries clinical trial, *systematic error* would result in all increments being either too big or too small. This type of error is caused by a shift in diagnostic criteria over the period of the trial and results in biased increment counts and affects the mean. The problem of biased increments is, however, not *a priori* dependent on the interval between examinations, because a shift in diagnostic criteria may occur at any time during the trial. Maintaining constant diagnostic criteria still seems to be an unsolved problem, and we do not know if systematic error is related to intervals of a certain length. Repeatability in clinical caries diagnoses has been determined by repeat examinations and by computing different types of consistency ratios (Shaw and Murray, 1975) or percentage deviations on mean DMFT or DMFS scores (Davies and Cadell, 1963). These methods of quantifying repeatability have been used when training examiners and for reporting examiner consistency in clinical and epidemiological studies. However, it is not possible, on the basis of these methods, to determine to what extent lack of repeatability affects the results in a clinical trial.

To obtain a higher repeatability, a detailed description of diagnostic criteria, standardization of examination conditions, and careful training of examiners seem to be the only answers at the present time. It should be added that the extent to which these measures reduce bias is largely unknown.

When lack of repeatability is due to *random error* in the

TABLE
THE INTERVAL BETWEEN INTERIM CLINICAL EVALUATIONS, THE RATIONALE FOR SUCH INTERVALS, AND THE PURPOSE OF PERFORMING INTERIM CLINICAL EVALUATIONS

Author	Interval	Rationale	Purpose
Baume (1961)	"The second examination should be carried out after an interval of exactly one year . . . . . ."	–	–
Finn (1962)	"A 1-year interval is adequate for clinical trials."	–	". . . when the first indication of a trend is evidenced . . . . . ."
Grainger (1972)	"The frequency of re-examination of subjects should be at yearly intervals . . . . . ."	"Observations at an interval of less than one year are difficult to interpret and analyse because reproducibility of examinations is generally less than ideal . . ."	–
Horowitz *et al.* (1973)	Cites Baume (1961)	"Intervals . . . . shorter than one year . . . do not permit enough . . . caries lesions . . . to be detected at a differential rate between test- and control-groups."	–
FDI (1982)	"The usual time between examinations is one year."		

diagnosis of caries, it results in an unbiased increment, because those diagnostic decisions which are subjected to lack of repeatability tend to equal each other. According to this, the mean increment of an individual (DMFS) can be expressed as the sum of the true increment ($DMFS_x$) and the error component ($DMFS_e$):

$$DMFS = DMFS_x + DMFS_e$$

Since the mean of the error component due to random error is assumed to be zero, it does not affect the mean of the increments, while it does increase the variance.

Methods of determining the reliability of caries data have been described (Rugg-Gunn and Holloway, 1974). For incremental data, the "sum of prevalence error variances" method can be used. According to this method, the incremental error variance is equal to the sum of the prevalence error variance at the baseline examination and the follow-up examination – in this case, the interim evaluation. This allows the total variance to be partitioned into both the true variance and the error variance. If it is correct that a test of differences between the mean incremental counts in the different groups in the trial can be made after elimination of the error variance component from the total variance, then lack of repeatability should not influence the choice of interval between re-examinations.

Lack of difference in caries increment between groups.

The second argument for having intervals of no less than one year between re-examinations is based on the assumption that a shorter period would not permit enough caries lesions to be detected at a differential rate between test and control groups.

While the first argument concerned itself with the denominator in the formula for a *t* test, this second argument relates to the numerator. Since dental caries, except in the initial process, is irreversible, the difference in mean caries increment between a control group and an experimental group would increase with time. However, as the mean caries increment increases, so also does the standard deviation increase. This is illustrated in the Fig., which is based on data from a large number of published caries clinical trials compiled recently (Kirkegaard and Poulsen, 1980). At least for values of increments of six or less, the relationship seems to be linear, with a slope close to one. For higher values of mean increments, the number of

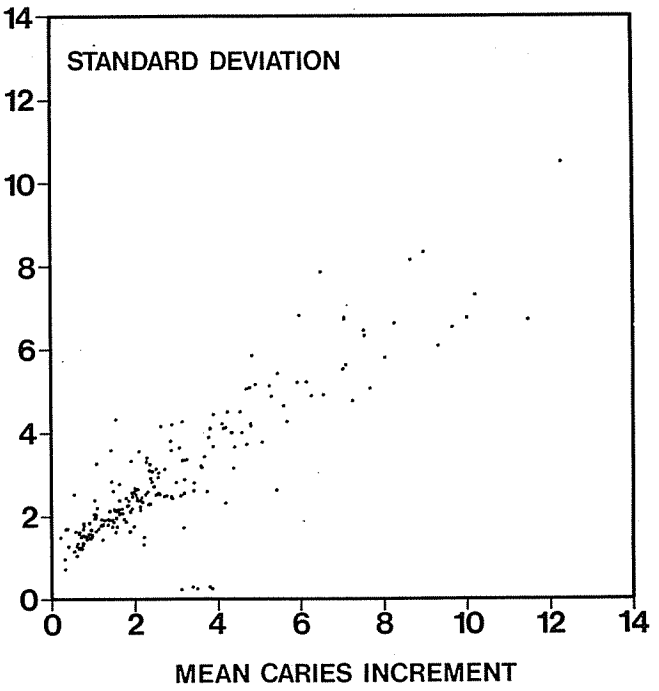

Fig. – Scatter-diagram of mean caries increments and the respective standard deviations reported in a large number of clinical trials compiled recently (Kirkegaard and Poulsen, 1980).

observations is too small for the type of relationship to be determined. In light of the fact that studies conducted on different fluoride compounds and by different investigators are pooled in this graph, the relationship is surprisingly clear.

Since larger increments obtained over larger intervals of time are accompanied by larger standard deviations, the efficiency of a clinical trial may not necessarily be enhanced by designing the trial in such a way that large increments are obtained (O'Mullane, 1976).

Role of interim clinical evaluations.

With this background, a re-definition of the value of interim clinical evaluations is attempted by listing a number of examples where such examinations may or may not be indicated.

In order to determine the presence of the preventive measure in the oral cavity, interim clinical evaluations may be of value. Trials on fissure sealants and trials on possible future methods for slow intra-oral release of fluorides are examples of this.

In trials where the agent is professionally applied, the need for separate interim clinical examinations seems less obvious, since the application of the agent can be combined with a screening of the oral conditions of the study subjects.

Obviously, interim evaluations are indicated in order to determine the occurrence of side-effects. Several clinical trials on chemical agents such as chlorhexidine and stannous fluoride have been associated with discoloration of the teeth and oral mucosa. Other examples of side-effects are soreness of the oral mucosa and staining of the dorsum of the tongue. If previous laboratory or short-term clinical testing of the agent in question has indicated a risk of side-effects, interim clinical evaluations should be included in the protocol for the trial.

Drop-out of experimental subjects is one of the major concerns of the investigator conducting a clinical trial. Good knowledge of the extent of the problem as well as a good understanding of the reasons for withdrawal of the subjects from the trial can only be obtained by constant supervision of the trial. Furthermore, a number of unforeseen complications can occur during the two- to three-year trial period. It has been our experience, from several trials, that these complications may not get to the knowledge of the investigator, unless close contact with the trial is maintained. Pre-scheduled interim clinical evaluations during a long-term clinical trial may be one way of securing this.

The last reason to be mentioned for performing interim clinical evaluations is the control of developing caries lesions. It could be argued that subjects participating in a clinical trial conducted by a professional team may expect to be made aware, during the trial, of developing treatment needs. In areas where regular dental treatment is provided to all children included in the trial, this reason for performing interim clinical evaluations is eliminated. However, in these areas interim clinical evaluations may be needed for some of the other reasons mentioned previously.

Conclusions.

In summary, the reason for performing interim clinical evaluations has never been clearly or sufficiently stated. If such evaluations are performed, they can be used in a sequential analysis of the trial. In other cases, interim clinical evaluations are conducted for reasons other than the testing of the null hypothesis. Once the length of the trial has been determined, the investigator should decide whether interim clinical evaluations should be performed, at what intervals, and with what methods. Thus, it follows that if interim examinations are performed for reasons other than statistical testing, they need not be as exhaustive and detailed as the baseline and the final follow-up examinations.

REFERENCES

BACKER-DIRKS, O.; BAUME, L.J.; DAVIES, G.N.; and SLACK, G.L.: Principal Requirements for Controlled Clinical Trials, *Int Dent J* 17:93-103, 1967.

BAUME, L.J.: ORCA Team I on Caries Statistics. Directives for Collecting and Recording Data on Dental Caries Increments by Means of Serial Examinations, *Arch Oral Biol* 4:217-223, 1969.

DAVIES, G.N. and CADELL, P.B.: Four Investigations to Determine the Reliability of Caries Recording Methods, *Arch Oral Biol* 8:331-348, 1963.

Federation Dentaire Internationale: Principal Requirements for Controlled Clinical Trials of Caries Preventive Agents and Procedures, *Int Dent J* 32:292-310, 1982.

FINN, S.B.: Conduct of Clinical Assays of Caries-controlling Agents. In: **Art and Science of Dental Caries Research,** R.S. Harris, Ed., New York and London: Academic Press, 1968, pp. 163-175.

GRAINGER, R.M.: Committee Consensus on Design and Analysis. In: **Conference on Clinical Testing of Cariostatic Agents,** Chicago: American Dental Association, 1972.

HOROWITZ, H.S.; BAUME, L.J.; BACKER-DIRKS, O.; DAVIES, G.N.; and SLACK, G.L.: Principal Requirements for Controlled Clinical Trials of Caries Preventive Agents and Procedures, *Int Dent J* 23:506-516, 1973.

KIRKEGAARD, E. and POULSEN, S.: A Review of Studies on the Effect of Topical Fluoride Applications and Fluoride Mouthrinses. Marburg: **ORCA Meeting,** 1980, Abstract No. 19.

O'MULLANE, D.M.: Efficiency in Clinical Trials of Caries Preventive Agents and Methods, *Community Dent Oral Epidemiol* 4: 190-194, 1976.

RUGG-GUNN, A.J. and HOLLOWAY, P.J.: Methods of Measuring the Reliability of Caries Prevalence and Incremental Data, *Community Dent Oral Epidemiol* 2:287-294, 1974.

SHAW, L. and MURRAY, J.J.: Inter-examiner and Intra-examiner Reproducibility in Clinical and Radiographic Diagnosis, *Int Dent J* 25:50-53 and 280-288, 1975.