Management and Evaluation of the Effects of Misclassification in a Controlled Clinical Trial

R. M. BELL and S. P. KLEIN

The Rand Corporation, Santa Monica, California 90406

J Dent Res 63(Spec Iss):731-734, May, 1984

Introduction.

Despite the well-accepted criteria outlined in the 1968 ADA Conference on the Clinical Testing of Cariostatic Agents, the diagnosis of caries is a subjective decision that exhibits substantial inconsistency in practice. In this paper, we try to make three main points about this process:

(1) Examiner error (inconsistency[1]) is a problem that affects caries clinical trials in numerous, and often unexpected, ways. Thus, this problem should receive the researcher's attention at all stages of a clinical trial.

(2) Clinical trials that use examiners should collect data on their reliability, and those data should meet certain minimum standards.

(3) These clinical trials should also report their reliability results in a form that will be useful to other researchers.

Unfortunately, the current practice falls far short of these standards on several points. Often, no reliability data are collected. And, when they are collected, the collection procedures often negate the value of the data. Some trials collect data but fail to report the results. Finally, the reporting methods are haphazard at best, often making meaningful comparisons among studies impossible.

The recommendations in this paper are based primarily on a comprehensive review of the dental examination reliability literature and our work on the National Preventive Dentistry Demonstration Program (NPDDP), a recent study of school-based preventive procedures offered in ten sites throughout the United States (Klein and Bohannan, 1984). This study provided dental examinations to over 30,000 children and included 9000 pairs of concurrent reliability examinations obtained with 31 trained examiners. Although certain points may be specific to the population we studied, most should hold more generally.[2]

Types of examiner errors. – It proves useful to distinguish two types of errors, systematic and random.

Systematic errors are attributable to factors which tend to recur under similar circumstances. The most familiar example is that some examiners systematically call more caries than do other examiners. But there are other potentially important sources of systematic errors. For example, one examiner or group of examiners may drift over the course of a trial in the use of the formal criteria. Added to the problem of known secular changes, this possibility makes the use of retrospective control groups very questionable. Also, there are likely to be differences in standards *between* clinical trials.

Most inconsistencies are non-systematic ones which we will refer to as *random errors.* These include "mental coin flips" that an examiner must make on close decisions, non-systematic misapplications of the criteria, and recording errors.

Types of reliability data. – The most important data for evaluating examiner reliability are the concurrent pairs of examinations. The subject receives two independent examinations on the same day, either both by the same examiner or one by each of two different examiners. These will be referred to as intra- and inter-examiner pairs, respectively.

Other data can aid in the evaluation of examiner reliability. For example, longitudinal data enable determination of the frequency of diagnostic *reversals,* where a surface is classified as carious on one examination and sound on an examination one or two years later. Comparison of mean DMFS or DMFT scores among examiners tests for systematic differences among the examiners. However, both of these data sources miss important types of errors, so that neither substitutes satisfactorily for concurrent reliability examinations (Klein *et al.*, 1984).

Why care about examiner errors?

Examiner error can potentially affect every stage of a caries clinical trial, from the design to the interpretation of results. Thus, researchers conducting such trials should consider the consequences of examiner error at each stage of their work.

Problems caused by systematic bias. – The possibility of systematic differences among examiners should be considered when subjects are assigned to examiners. Assignments should balance the combination of examiners and treatment groups. That is, if an examiner does 15% of all the examinations, he or she should see 15% of each treatment and control group. If balance can be achieved, then any systematic bias of a particular examiner would probably cancel when groups are compared.

A secondary consideration in the assignments of examiners is to maintain examiner/subject pairings over time. The justification is that examiner bias will cancel when increment scores are computed. However, reliance on this principle to remove examiner bias problems is naive. Although examiners agree much better with themselves than with each other on *concurrent* pairs of examinations, there is little evidence that this phenomenon persists over time. The first part of Table 1 shows the proportion of the time that a decayed call on one examination in the NPDDP was "reversed" by the other concurrent examination. Different examiners (row 2) disagreed almost twice as often as an examiner disagreed with him or herself (row 1). However, examiners were unable to maintain this high level of self-consistency over a period of two years. The rate of *longitudinal* reversals was only slightly higher when the examiner was switched as opposed to maintained over time.

Even if careful calibration eliminates systematic differences among examiners within a study, there is also the concern of differences across studies. This problem is most significant for determining trends in caries prevalence. For

[1] The term "error" is used synonymously with "inconsistency", without the intention of claiming that there is always an obvious correct call.

[2] Detailed evidence for many of the points made here appears in Klein *et al.* (1984).

example, the fact that two national dental health surveys found a surprisingly large drop in caries among children between the early and late 1970's has important consequences (Miller *et al.,* 1981). Unfortunately, there is no way to know how much differences in application of the same formal criteria may have contributed to that finding.

Reduction of precision. – Examiner error affects every clinical trial by increasing the variability of estimated treatment effects through the addition of random error to scores. Many studies report the reliability coefficient (intraclass correlation; see Fleiss *et al.,* 1979) for a particular examination. That number indicates how examiner error affects the precision of *prevalence* studies (*i.e.,* studies which look at the amount of decay or the relationship between decay and other characteristics *at a fixed point in time*). Table 2 shows the estimated impact of examiner error on the precision of prevalence studies, using reliability data from the National Preventive Dentistry Demonstration Program. To obtain the same level of precision that would be available from examining 100 children without any examiner error, we would have needed to examine from 105 to 108 children. The fact that these reliability coefficients are typical of those reported by other studies suggests that the price paid for examiner error in prevalence studies is fairly small. We use the word *suggests* because the Table does not account for the potential problems that systematic errors may cause.

Compared with the impact on prevalence studies, examiner error can substantially affect the precision of estimated treatment effects in clinical trials. Table 3 shows how the amount of examiner error observed in the NPDDP affected the information available about treatment effects through the analysis of two-year DMFS increments. About 20 to 25% more children were needed to obtain the same information that would have been available in the absence of any examiner error.[3] Considering the expense of conducting even a small clinical trial, increases of this sort are very significant.

There are two reasons for the greater impact on increment scores – the examination error occurs twice, and the true change during two years is quite small. Thus, examiner error has a greater impact on precision of estimates in a clinical trial than in a prevalence study, and the problem is greatest during a short study.

Clearly, the above comparisons are unrealistic. One could never eliminate all examiner error, and it might be too costly to reduce error much below that observed in the NPDDP. Still, the value of keeping down the amount of systematic and random examiner error in caries clinical trials should be apparent. Among the steps for doing so are:

- Carefully training the examiners to follow a rigid set of criteria. Review of the criteria should continue throughout the study.
- Holding calibration sessions on a population similar to that under study. Again, calibration sessions should continue throughout the study.
- Collecting reliability data as a regular part of the examination process. This encourages examiners to maintain good concentration and to adhere to the formal criteria. The incentive results both from competition among the examiners and from the desire of the examination team to compare favorably with teams from other trials. But this incentive fails if, as in many trials, the reliability data are collected during a separate calibration period, or the examiners somehow know which subjects compose the reliability sample. Also, to have full effect, reliability results should be fed back to the examiners at regular intervals.
- In selected studies, it may help to provide multiple examinations to each participant.[4] Providing two examinations to each participant and using the mean score from the two would substantially increase the reliability. In the example of Table 3, one would need to examine, and therefore treat, only 112 10-year-olds, as opposed to 123. Additional reduction might be achieved by following a suggestion of Kamen and Schmee (1974), "Two examiners diag-

TABLE 1
INTRA- AND INTER-EXAMINER REVERSAL RATES, USING CONCURRENT AND LONGITUDINAL DATA

	Reversal Rate	
	Age 6-9	Age 10-12
Concurrent exams		
Same examiner	0.15	0.12
Different examiners	0.29	0.22
Longitudinal exams		
Same examiner	0.20	0.17
Different examiners	0.22	0.22

Note: Only surfaces classified as decayed (not filled or missing) on the first exam are included. Concurrent results have been averaged across three years. Longitudinal reversals cover an elapsed period of two years.

TABLE 2
NUMBERS OF CHILDREN REQUIRED TO PROVIDE THE SAME INFORMATION ABOUT CARIES PREVALENCE WITHOUT AND WITH EXAMINER ERROR

		Sample Size	
Age of Children	Reliability Coefficient	No Error	Inter-examiner Error
6-9	0.93	100	108
10-12	0.95	100	105

TABLE 3
NUMBERS OF CHILDREN REQUIRED TO PROVIDE THE SAME INFORMATION ABOUT TWO-YEAR DMFS INCREMENTS WITHOUT AND WITH EXAMINER ERROR

		Sample Size	
Age at Start of Trial	Reliability Coefficient for Two-year Increment	No Error	Inter-examiner Error
6-7	0.82	100	122
10	0.81	100	123

[3] If, as in the NPDDP, analysis of covariance reduces the residual variance below that of raw increments, the relative importance of examiner error increases.

[4] If the reliability of one examination is r, then the reliability of the average of n examinations would be $nr/[1+(n-1)r]$. Essentially, providing two examinations *per* child would halve the number of excess examinations that need to be given to overcome the impact of examiner error.

nose each subject independently. Wherever there is a disagreement, a third examiner is consulted and the majority vote on the status of a surface would hold." Depending on the trade-off between marginal costs of providing more examinations *vs.* providing treatment to more participants, a multiple-examination design might be cost-effective.

Other procedures for limiting examiner error are discussed in two papers by Horowitz (1972). Particularly noteworthy is the admonition not to allow the examiners to determine the treatment group to which any child belongs.

Why collect examiner reliability data?

Besides the potential benefit of increasing examiner reliability, there are other important reasons why every clinical trial should collect reliability data.

The attainment and reporting of a high level of examiner reliability lend credibility to the examination phase of a study. Poor results or the absence of any reported results only raises questions about whether data collection problems may have affected the outcome of the trial.

When new training methods or examination procedures are employed, reliability data have special value in assessing the merits of the new approaches.

In large multi-year trials, attrition of subjects often reduces the number of examiners required to complete examinations on schedule. Clearly, examiner reliability should be a prime consideration in choosing which examiners to retain. This strategy has seldom been used in the past because of the admonition that examiner pairings should be retained throughout a trial. Although maintenance of examiner pairings *is* a desirable policy where possible, this goal should be secondary to using the most consistent examiners and assigning them to treatment groups in a balanced manner.

Unfortunately, reliability results provide little, if any, information about how the analysis of dental caries data should proceed.[5] For example, we second the recommendation of Radike (1972), that diagnostic reversals be taken at face value, even when the evidence indicates that a "correctable" error has been made.

Minimum standards for collection of reliability data.

The desirable scope and details of reliability data collection in a given clinical trial depend on many factors: the size of the trial, the specific uses planned for the data, characteristics of the subjects, the novelty of the examination procedures, the experience of the examiners, and the available budget. However, there are certain standards that every clinical trial should meet:

(1) At least a moderate amount of concurrent *inter*-examiner data should be collected. If limitations present a conflict between the amount of intra-examiner and inter-examiner data that can be obtained, precedence should go to inter-examiner data. It is worth reiterating that simple comparison of mean scores across examiners and analysis of reversals are insufficient.

(2) All reliability data should be collected on the main sample of subjects studied in the clinical trial, and every effort should be taken to ensure that the examiners have no way to distinguish reliability examinations from regular ones.

(3) Specification of a minimum sample size is arbitrary, but the following might serve as a useful rule-of-thumb: Until each examiner has diagnosed at least 50 lesions as part of reliability pairs, there is insufficient evidence to distinguish even the best examiner from a very poor one.[6] In general, more data would be desirable, especially in a large study. Thus, one may require a substantial reliability sample in a study of young children.

Issues in reporting of reliability results.

The ability to disseminate results about the quality of the examination process is a major reason, but not the only one, for collecting reliability data. Thus, it is important that the report of every clinical trial devote some space to presentation of examiner reliability results. Further, it is just as important that these results be reported in a precise and understandable manner. Unfortunately, current practice often fails to meet these goals. Even when good data are collected, the reporting of results in published papers often shrinks to the point where it is of no value to the reader.

Reliability measures are not always defined precisely. For example, a reversal rate might be defined as the mean number of reversals *per* subject, counting either all subjects, just those children with at least one erupted permanent tooth, or just those with a positive DMFS score at baseline. In a study of young children, the result would be very sensitive to the definition.

As this suggests, results for certain reliability measures are very sensitive to the characteristics of the subjects. As another example, the reliability coefficient is very sensitive to the homogeneity of the population being studied. A reliability coefficient of 0.95 might be quite good for a study of only 11-year-olds but hopelessly inadequate for a study of six- to 17-year-olds.[7]

Commonly reported reliability indices do not always convey the most useful information. For example, reliability of radiographic diagnoses are usually reported in terms of the *consistency index.* This index is the percentage of the time that two readers agreed, given that at least one of the readers called a surface carious. The first line of Table 4 shows that, for 10-year-olds in the NPDDP, two readers agreed that a surface was carious slightly more often (175 times) than they disagreed (165 times), for a consistency index of 51%. However, radiograph calls are usually combined with clinical calls to yield a combined score, using the rule that the radiograph call for a surface only affects the combined score if the clinical examiner classified the surface as sound. The last two lines of the Table indicate that the 51% was comprised of both very good agreement (85%) when the clinical examiner had rendered the radiograph call unnecessary and poor agreement (36%) when the radiograph call really mattered. Thus, the overall consistency of 51% is misleading.

[5] The one place where information about examiner reliability might profitably alter the analysis is by suggesting the need to adjust for systematic differences among examiners. Even so, the most accurate adjustments would probably come from including the examiner as a covariate in an analysis of covariance.

[6] For example, after 40 diagnoses by each examiner, there would only be about an even chance of detecting a statistically significant difference (at $p = 0.05$) between an examiner who agreed with his peers on 80% of caries calls and one who agreed 60% of the time.

[7] In the former case, the reliability coefficient is a quite meaningful index if the population is homogeneous. In the latter case, however, the fact that much of the variance in scores could be explained by using age as a covariate renders the reliability coefficient nearly meaningless (see footnote 3).

TABLE 4
INTER-READER CONSISTENCY INDICES FOR RADIOGRAPH READINGS ON 10-YEAR-OLD CHILDREN

	Frequency		Consistency Index for Radiograph Readings
	Readers Agreed Carious	Readers Disagreed Sound/Car.	
Full Sample	175	165	51
When clinical examiner called			
Carious	91	16	85
Sound	84	149	36

Minimum standards for reporting of reliability results.

Reporting reliability results from clinical trials should be an expected, standard practice. Also, these results should receive more space and care than is now typical, with the following minimum standards kept in mind:

(1) Authors should give precise details about the conditions under which reliability data were collected and the methods used to compute reliability indices.

(2) Authors should avoid reliability indices that are too sensitive to the particular population studied.

(3) Authors should anticipate how their examination data will actually be used when determining what reliability data to collect and how to report their reliability results.

Conclusions.

Many findings have been uncovered in our analysis of dental examination reliability data and our review of the associated literature. These led to the following main conclusions:

- Examiner errors are expensive. At the least they increase the sample size required to meet a certain objective. At the worst, they raise crippling questions about the validity of a clinical trial.
- Extensive training and calibration are essential, and they should continue throughout the course of a clinical trial.
- Every clinical trial should include collection of concurrent inter-examiner reliability data as part of the regular examination process. First, it can be one of the most effective ways to maintain examiner consistency. Second, it is the only way to ensure credibility in the face of a rightfully skeptical scientific community.
- Finally, much more attention needs to be devoted to careful reporting of reliability results. One problem, no doubt, is that journal space is tight. Not surprisingly, reliability results are often the first to be pared, or deleted completely. The only remedy is to convince editors and referees of the importance of this issue. We hope that this paper is a step in that direction.

REFERENCES

FLEISS, J.L.; SLAKTER, M.J.; FISCHMAN, S.L.; PARK, M.H.; and CHILTON, N.W.: Inter-examiner Reliability in Caries Trials, *J Dent Res* 58:604-609, 1979.

HOROWITZ, H.S.: Examiner Bias, **Proceedings of the Conference on the Clinical Testing of Cariostatic Agents.** Chicago: American Dental Association, 1972, pp. 95-96.

HOROWITZ, H.S.: Inter- and Intra-examiner Variability, **Proceedings of the Conference on the Clinical Testing of Cariostatic Agents.** Chicago: American Dental Association, 1972, pp. 97-98.

KAMEN, A. and SCHMEE, J.: Diagnostic Errors and Multiple Examiners in Anticariogenic Studies, *J Dent Res* 53:1500, 1974.

KLEIN, S.P.; BELL, R.M.; BOHANNAN, H.M.; DISNEY, J.A.; and WILSON, A.: **Reliability of Dental Examination Data in the National Preventive Dentistry Demonstration Program.** Santa Monica, CA: The Rand Corporation, R-3138-RWJ, 1984.

KLEIN, S.P. and BOHANNAN, H.M.: **Summary of the Major Findings in the National Preventive Dentistry Demonstration Program.** Santa Monica, CA: The Rand Corporation, 1984, in press.

MILLER, A.J.; BRUNELLE, J.A.; CARLOS, J.P.; and SCOTT, D.B.: **The Prevalence of Dental Caries in United States Children.** Bethesda, MD: USDHHS, NIH Publ. No. 82-2245, 1981.

RADIKE, A.W.: Examiner Error and Reversals in Diagnosis, **Proceedings of the Conference on Clinical Testing of Cariostatic Agents.** Chicago: American Dental Association, 1972, pp. 92-95.