Session II: Specific Design Considerations in Clinical Caries Trials—Neal W. Chilton, Chairman

The Use of Misclassification Models in the Evaluation of Caries Clinical Trials

G. F. REED

Epidemiology and Biometry Research Program, National Institute of Child Health and Human Development, Bethesda, Maryland 20205

J Dent Res 63(Spec Iss):727-730, May, 1984

Introduction.

Comparisons of dental caries incidence rates from clinical trial data are complicated by two conditions: that there is always some unknown amount of error in the diagnosis of caries, and that teeth are necessarily observed in clusters rather than as statistically independent individuals.

In response to the problem of diagnostic error, Carlos and Senning (1968), Lu (1968), and Poole *et al.* (1973) developed three different models of caries incidence with misclassification. These ingeniously conceived models, each based on its own set of assumptions about the mechanism of diagnostic error, produce incidence estimators that are accordingly adjusted for error. Estimators from all the models are functions of the observed frequencies of teeth, in the following categories:

- S_1 = CC, teeth diagnosed as decayed, missing, or filled (DMF) at pre-intervention and post-intervention examinations;
- S_2 = NN, teeth diagnosed as normal at both examinations;
- S_3 = NC, teeth diagnosed normal before intervention, DMF afterward (the size of this category is the observed incidence of caries during the trial); and
- S_4 = CN, teeth diagnosed DMF before intervention and normal afterward (*i.e.*, "reversals").

By the existence of three competing approaches, there naturally arise questions of the models' validity, fit, and relative strengths. Moreover, in addition to these concerns, there is the issue of how to describe the statistical behavior of the estimators resulting from the models, for, even if no adjustment for misclassification were necessary, the non-independence that characterizes frequency counts of teeth, by virtue of their clustering by mouth, precludes the application of standard techniques (see, *e.g.*, Fleiss, 1980) for comparing incidence rates from independent observations. Thus, at the time of publication of the models, the task of determining means, variances, and statistical distributions of the model-induced estimators was not satisfactorily addressed.

The models and foregoing issues surrounding their use were the subject of some work (Reed, 1978; Reed and McHugh, 1979) in which was suggested a methodology, applicable with any misclassification model, for making statistical inferences on caries incidence rates. The rationale of the approach is as follows: Measurement error and experimental error may be regarded as two distinct random components of an experiment; in a caries clinical trial, the measurement component, *i.e.*, diagnostic error, is deterministically removed by whatever misclassification model is employed, so that the execution of the research design remains as the source of the random element of the data. The specific design involves a finite population of teeth, clustered by mouth and subject to a single-stage cluster sampling procedure (see, *e.g.*, Sukhatme and Sukhatme, 1970) for drawing two samples as control and intervention groups. This well-known sampling design accommodates the clustering problem and provides for the distribution theory of the observed frequencies of the diagnostic categories. With means and variances of the frequencies as input, the multivariate delta method for obtaining the distribution of the incidence estimator completes the technique.

Unfortunately, this solution is not a universal one. Statistics from finite population sampling require either knowledge of the population size or assurance that sample sizes are negligible relative to that of the population. Such requirement is not met in research situations in which, for example, the population is neither well-defined nor well-enumerated. Next in this presentation, then, will be detailed a closely related but more generally applicable alternative: randomization testing.

The randomization approach.

A simple design for a caries clinical trial may start with the identification of N eligible trial participants, each of whom possesses a cluster of teeth whose number varies from cluster to cluster. At random, participants or, equivalently, clusters are assigned to a control group or to a group which is to receive some cariostatic intervention. If the control group is to contain n clusters, then under proper randomization there are $({}_N C_n)$ equally likely ways to assign the N clusters to the two groups. The trial is then conducted, and upon its conclusion diagnostic category frequencies for each group are counted, a choice of misclassification model is made, and resulting estimates of incidence by group are obtained.

The randomization testing technique for evaluating the observed difference in incidence rates of the groups proceeds under the hypothesis that there is no effect of intervention on incidence and that, consequently, differences observed between groups are due only to the groups' random composition, determined by the investigators before the trial. Each such composition has probability $1/({}_N C_n)$ of being chosen for the trial, and since each composition generates its own category frequencies and incidence estimators for both groups, the null distribution of these statistics is thereby induced. Since the randomization is of clusters rather than of individual teeth, cluster effects are explicitly and fully comprehended by the technique. For a more general exposition of the randomization approach, due to Fisher (1935), see Kempthorne (1955).

In symbols, the random assignment to groups is described by the indicator random variables c_m $(m = 1, \ldots, N)$, each of which assumes the value 1 with probability n/N and the value 0 with probability (N-n)/N; further, the c_m must sum to n. So, if $c_m = 1$, then the m^{th} cluster is assigned to

control, otherwise it receives the intervention. The c_m are the trial's elemental random quantities, to which all statistical behavior is traced. Their moments are easily derived from their distribution:

$$E(c_m) = n/N,$$

$$\mathrm{Var}(c_m) = \frac{n}{N}\left(1 - \frac{n}{N}\right), \text{ and}$$

$$\mathrm{Cov}(c_m, c_{m'}) = \frac{-n}{N}\left[\frac{N-n}{N(N-1)}\right].$$

Let X_{im} denote the number of teeth in category S_i from cluster m, k_i the number of category S_i teeth in the control group, and K_i the number of S_i teeth in both groups. Then, because of the relationship

$$k_i = \sum_{m=1}^{N} c_m X_{im},$$

The moments of the k_i are readily achieved:

$$\left.\begin{aligned} E(k_i) &= \frac{n}{N}K_i, \\ \mathrm{Var}(k_i) &= \frac{n(N-n)}{N}\sigma_i^2, \text{ and} \\ \mathrm{Cov}(k_i, k_j) &= \frac{n(N-n)}{N}\sigma_{ij}, \end{aligned}\right\} \quad \text{(Eq. 1)}$$

where

$$\sigma_i^2 = \sum_{m=1}^{N} (X_{im} - \bar{K}_i)^2/(N-1), \text{ and}$$

$$\sigma_{ij} = \sum_{m=1}^{N} (X_{im} - \bar{K}_i)(X_{jm} - \bar{K}_j)/(N-1); \text{ with } \bar{K}_i = K_i/N.$$

Because of the prodigious effort of enumeration usually required to know the exact distribution of the k_i, the alternative of large sample approximation is preferred. The k_i are sums (over m=1, . . . , N) of the variates $c_m X_{im}$, which, though not independent of each other, are assumed to satisfy conditions on dependent sequences that permit application of the Central Limit Theorem (see, *e.g.*, Puri and Sen, 1971). So, for sufficiently large n and N, the k_i are approximately normally distributed with parameters (1).

For the control group, the caries incidence rate, which is the true number of teeth in the group to become DMF during the trial per number of clusters in the group, is estimated according to Carlos and Senning by

$$\frac{(k_1 + k_2 + k_3 + k_4)(k_3 - k_4)}{n(k_3 - k_4) + n\sqrt{(k_1 + k_2)^2 - 4k_3k_4}}, \quad \text{(Eq. 2)}$$

and according to Lu by

$$\frac{(k_1 + k_2 + k_3 + k_4)(k_3 - k_4)}{n(k_1 + k_2 + k_3 - 3k_4)}. \quad \text{(Eq. 3)}$$

(For simplicity's sake, the model of Poole *et al.* has been omitted. Nevertheless, the present techniques are applicable to it.) Given a model, let $r(k_1, k_2, k_3, k_4; n)$ denote its estimator for the control group. The estimator for the intervention group, then, is similar, with $K_i - k_i$ and N-n replacing k_i and n, respectively, so that the difference in rates between groups is a function of the k_i:

$$d = r(k_1, k_2, k_3, k_4; n) - r(K_1 - k_1, K_2 - k_2, K_3 - k_3, K_4 - k_4; N-n).$$

Since (k_1, k_2, k_3, k_4) is approximately multivariate normal, by the multivariate delta method (see Bishop *et al.*, 1975), d has an approximately normal distribution with mean zero and variance

$$\sum_{i=1}^{4} D_i^2\sigma_i^2 + \sum_{\substack{i=1 \\ i\neq j}}^{4}\sum_{j=1}^{4} D_iD_j\sigma_{ij},$$

where D_i is the partial derivative of d with respect to k_i, evaluated at $k_1 = E(k_1)$, $k_2 = E(k_2)$, $k_3 = E(k_3)$, and at $k_4 = E(k_4)$. Expressions for the D_i are derived in the Appendices.

An example. — The randomization technique was applied to some data provided through the courtesy of Professor Lawrence Meskin of the University of Minnesota. Table 1 displays the observed frequencies of the categories S_i, and Table 2 gives the variances and covariances of the frequencies from (1). Table 3 lists estimates of the incidence rates from expression (2); their difference; calculated partial derivatives, D_i; and the resulting approximation of Var(d). The same quantities for the Lu model, with estimator (3), are shown in Table 4.

The statistic $d/\sqrt{\mathrm{Var}(d)}$ is approximately standard normal and has p-values of 0.512 (C—S) and 0.511 (Lu) for the hypothesis of no true difference in rates.

TABLE 1
DIAGNOSTIC CATEGORY FREQUENCIES

Diagnostic Category	Control (n=300) k_i	Intervention (N−n=277) $K_i - k_i$	Total (N=577) K_i
S_1	934	864	1798
S_2	4368	4107	8475
S_3	201	174	375
S_4	78	78	156

TABLE 2
VARIANCE-COVARIANCE MATRIX OF THE k_i

	k_1	k_2	k_3	k_4
k_1	377.8584	−359.5767	−103.1477	−47.7861
k_2		1855.7368	−4.0038	9.4190
k_3			112.3362	−11.5937
k_4				60.4599

TABLE 3
CARLOS-SENNING ESTIMATES AND THE APPROXIMATE VARIANCE OF THE DIFFERENCE

Control Incidence Rate	0.4223
Intervention Incidence Rate	0.3576
Difference, d	0.0647
D_1	−0.00000464
D_2	−0.00000464
D_3	0.00715776
D_4	−0.00685211
Approximate Var(d)	0.009736

TABLE 4
LU ESTIMATES AND THE APPROXIMATE VARIANCE OF THE DIFFERENCE

Control Incidence Rate	0.4343
Intervention Incidence Rate	0.3686
Difference, d	0.0657
D_1	−0.00000916
D_2	−0.00000916
D_3	0.00736000
D_4	−0.00674409
Approximate Var(d)	0.009996

Concluding remarks.

The randomization approach has been proposed here as more generally applicable than the finite population sampling strategy for testing hypotheses of no intervention effect, since randomization testing is conditional on the sample and requires no knowledge or assumptions about some possibly ill-defined population. Randomization is less forthcoming, however, with confidence limits and power functions. These require the specification of underlying randomization distributions that correspond to alternative hypotheses about the behavior of d; this is a difficult and unsolved problem.

Even more important than the matter of how to use the misclassification models is the question of whether to use them. Besides the three models discussed here, many others are conceivable, but the paucity of empirical research on the nature of diagnostic error provides little foundation for validating any of them. Advances in accurately measuring caries incidence clearly await finer scrutiny of diagnostic error.

REFERENCES

BISHOP, Y.M.M.; FINEBERG, S.E.; and HOLLAND, P.W. (1975): **Discrete Multivariate Analysis.** Cambridge, MA:MIT Press.

CARLOS, J.P. and SENNING, R.S. (1968): Error and Bias in Dental Clinical Trials, *J Dent Res* 47:142-148.

FISHER, R.A. (1935): **The Design of Experiments.** Edinburgh: Oliver and Boyd.

FLEISS, J.L. (1981): **Statistical Methods for Rates and Proportions,** 2nd ed. New York: Wiley.

KEMPTHORNE, O. (1955): The Randomization Theory of Experimental Inference, *J Am Statist Assoc* 50:946-967.

LU, K.H. (1968): A Critical Evaluation of Diagnostic Errors, True Increment and Examiner's Accuracy in Caries Experience Assessment by a Probabilistic Model, *Arch Oral Biol* 13:1133-1147.

POOLE, W.K.; CLAYTON, C.A.; and SHAH, B.V. (1973): The Estimation of Examiner Error and the True Transition Probabilities for Teeth or Surfaces in Dental Clinical Trials, *Arch Oral Biol* 18:1291-1302.

PURI, M.L. and SEN, P.K. (1971): **Nonparametric Methods in Multivariate Analysis.** New York: Wiley.

REED, G.F. (1978): Measurement of Dental Caries Incidence in the Presence of Diagnostic Error. Ph.D. thesis, University of Minnesota, Minneapolis, MN.

REED, G.F. and McHUGH, R.B. (1979): The Estimation of Dental Caries Incidence in the Presence of Diagnostic Error, *Biometrics* 35:473-478.

SUKHATME, P.V. and SUKHATME, B.V. (1970): **Sampling Theory of Surveys with Applications.** Ames, Iowa: Iowa State University Press.

Appendix A: Derivation of the D_i

Let $\underset{\sim}{K}$ and $\underset{\sim}{k}$ denote the vectors (K_1, K_2, K_3, K_4) and (k_1, k_2, k_3, k_4), respectively. Given a model, let $r(\underset{\sim}{k}; n)$ be its estimator of incidence rate for the control group, and let $f_i(\underset{\sim}{k}; n)$ be its partial derivative with respect to k_i. The partial derivative with respect to k_i of the estimator $r(\underset{\sim}{K}-\underset{\sim}{k}; N-n)$ for the intervention group is, by the chain rule,

$$-f_i(\underset{\sim}{K}-\underset{\sim}{k}; N-n).$$

The derivative of the difference between the estimators is therefore

$$f_i(\underset{\sim}{k}; n) + f_i(\underset{\sim}{K}-\underset{\sim}{k}; N-n).$$

Evaluate this at $\underset{\sim}{k}=E(\underset{\sim}{k}) = n\underset{\sim}{K}/N$ to get

$$D_i = f_i\left(\frac{n}{N}\underset{\sim}{K}; n\right) + f_i\left[\left(1-\frac{n}{N}\right)\underset{\sim}{K}; N-n\right].$$

From Appendices B and C, it is clear that, for either model, common multipliers of the k_i factor out of the f_i, so that

$$D_i = f_i(\underset{\sim}{K}; n) + f_i(\underset{\sim}{K}; N-n),$$
$$= \left(\frac{1}{n}+\frac{1}{N-n}\right) f_i(\underset{\sim}{K}; 1).$$

The f_i specific to each model are given in Appendices B and C.

Appendix B: Partial Derivatives of the Carlos-Senning Estimator of the Control Incidence Rate

From (2), the Carlos-Senning estimator of the rate is

$$r = \frac{1}{n}(k_1+k_2+k_3+k_4)(k_3-k_4)\left\{k_3-k_4+[(k_1+k_2)^2-4k_3k_4]^{1/2}\right\}^{-1},$$

so that the partial derivatives of r with respect to $\underset{\sim}{k}$ are as follows:

$$f_1 = \frac{1}{n}(k_3-k_4)\left\{k_3-k_4+[(k_1+k_2)^2-4k_3k_4]^{1/2}\right\}^{-1}$$
$$-\frac{1}{n}(k_1+k_2+k_3+k_4)(k_3-k_4)\left\{k_3-k_4+[(k_1+k_2)^2-4k_3k_4]^{1/2}\right\}^{-2}$$
$$\circ\left\{[(k_1+k_2)^2-4k_3k_4]^{-1/2}(k_1+k_2)\right\}.$$

$f_2=f_1$, since r is symmetric in k_1 and k_2.

$$f_3 = \frac{1}{n}(k_3-k_4)\left\{k_3-k_4+[(k_1+k_2)^2-4k_3k_4]^{1/2}\right\}^{-1}$$
$$+\frac{1}{n}(k_1+k_2+k_3+k_4)\left\{k_3-k_4+[(k_1+k_2)^2-4k_3k_4]^{1/2}\right\}^{-1}$$
$$-\frac{1}{n}(k_1+k_2+k_3+k_4)(k_3-k_4)\left\{k_3-k_4+[(k_1+k_2)^2-4k_3k_4]^{1/2}\right\}^{-2}$$
$$\cdot\left\{1-2[(k_1+k_2)^2-4k_3k_4]^{-1/2}\cdot k_4\right\}.$$

$$f_4 = \frac{1}{n}(k_3-k_4)\left\{k_3-k_4+[(k_1+k_2)^2-4k_3k_4]^{1/2}\right\}^{-1}$$
$$-\frac{1}{n}(k_1+k_2+k_3+k_4)\left\{k_3-k_4+[(k_1+k_2)^2-4k_3k_4]^{1/2}\right\}^{-1}$$
$$+\frac{1}{n}(k_1+k_2+k_3+k_4)(k_3-k_4)\left\{k_3-k_4+[(k_1+k_2)^2-4k_3k_4]^{1/2}\right\}^{-2}$$
$$\cdot\left\{1+2k_3[(k_1+k_2)^2-4k_3k_4]^{-1/2}\right\}.$$

Appendix C: Partial Derivatives of the Lu Estimator of the Control Incidence Rate

From (3), the Lu estimator of the incidence rate is

$$r = \frac{1}{n}(k_1+k_2+k_3+k_4)(k_3-k_4)(k_1+k_2+k_3-3k_4)^{-1},$$

so that the partial derivatives of r with respect to $\underset{\sim}{k}$ are as follows:

$$f_1 = \frac{1}{n}(k_3-k_4)(k_1+k_2+k_3-3k_4)^{-1} - \frac{1}{n}(k_1+k_2+k_3+k_4) \cdot (k_1+k_2+k_3-3k_4)^{-2}\,(k_3-k_4),$$

$$= \frac{4k_4(k_4-k_3)}{n(k_1+k_2+k_3-3k_4)^2}.$$

$f_2 = f_1$, since r is symmetric in k_1 and k_2.

$$f_3 = \frac{1}{n}(k_3-k_4)(k_1+k_2+k_3-3k_4)^{-1} + \frac{1}{n}(k_1+k_2+k_3+k_4) \cdot (k_1+k_2+k_3-3k_4)^{-1} - \frac{1}{n}(k_1+k_2+k_3+k_4)(k_3-k_4) \cdot (k_1+k_2+k_3-3k_4)^{-2},$$

$$= \frac{k_1+k_2+2k_3}{n(k_1+k_2+k_3-3k_4)} - \frac{(k_1+k_2+k_3+k_4)(k_3-k_4)}{n(k_1+k_2+k_3-3k_4)^2}.$$

$$f_4 = \frac{1}{n}(k_3-k_4)(k_1+k_2+k_3-3k_4)^{-1} - \frac{1}{n}(k_1+k_2+k_3+k_4) \cdot (k_1+k_2+k_3-3k_4)^{-1} + \frac{3}{n}(k_1+k_2+k_3+k_4)(k_3-k_4) \cdot (k_1+k_2+k_3-3k_4)^{-2},$$

$$= \frac{-(k_1+k_2+2k_4)}{n(k_1+k_2+k_3-3k_4)} + \frac{3(k_1+k_2+k_3+k_4)(k_3-k_4)}{n(k_1+k_2+k_3-3k_4)^2}.$$

The Use of Misclassification Models in the Evaluation of Caries Clinical Trials: Discussion of Dr. Reed's Presentation

J. D. GOLDBERG

Department of Biomathematical Sciences, Mount Sinai School of Medicine, New York, New York 10029

J Dent Res 63(Spec Iss):730, May, 1984

I want to thank Dr. Reed for his excellent paper dealing with the use of randomization procedures to compare the incidence of caries in two groups. In the context of the dental clinical trial, Dr. Reed raises many important and interesting aspects of the effects of misclassification on data analysis and inference.

It is interesting to note that the inferences under the two models of caries misclassification studied are similar, as was the case in the earlier work of Reed and McHugh, who used a finite sampling approach for the same problem. Further, the randomization approach and the finite sampling approach yield comparable inferences (see Table). Can some practical guidelines be given for the choice of method of analysis?

I also have several questions which result from my perspective and interest in problems of misclassification. Since the effects of the two error models under study on inference are the same, can Dr. Reed add any perspective on how to choose an error model? As Dr. Reed has pointed out, the lack of "empirical research on the nature of diagnostic error provides little foundation for validating" any of the conceivable misclassification models. The example presented here suggests that both the randomization and the finite sampling procedures are robust with respect to the class of error models considered. Are there ranges of incidence rates or possible error structures which would suggest one model over another?

The example presented raises some additional issues. Each of the error models under study requires several assumptions about the nature of the diagnostic errors. These assumptions can include the requirements that the probability of a misdiagnosis is independent of the true state, and that these probabilities are equal at each time point. From the data presented, no assessment of the appropriateness of assumptions such as these is possible.

While the results of the analysis seem independent of the error model, I am puzzled by the lack of data regarding the error structure. The "false negatives" and "false positives" result from changes in classification from pre-intervention to post-intervention. Evaluation at each of the two time points is required. It is quite possible that the probabilities of false negative and false positive results prior to intervention differ from these probabilities subsequent to intervention. With some standard for evaluation and some attention to study design, error rates can be estimated prior to the study and estimated again at the conclusion of the study. The two types of errors have differing effects on inference, with the false positive rates, in general, producing more serious effects (for prevalence or incidences < 0.5). These errors may also differ among subclasses of individuals based on susceptibility to caries or with observers. Randomization to treatment and control should permit balance here; however, if the errors are different in the comparison groups as well, additional confounding could result. What effects on the randomization procedure would be expected in such circumstances?

In summary, I would reiterate Dr. Reed's conclusion that much remains to be done in the area of assessing misclassification itself in dental caries trials.

TABLE
COMPARISON OF METHODS AND MISCLASSIFICATION MODELS

Estimated Quantity	Method: Randomization		Finite Sampling*	
Model:	C-S	Lu	C-S	Lu
Treatment Incidence	0.36	0.37	0.36	0.37
Control Incidence	0.42	0.43	0.42	0.43
Difference (C-T)	0.06	0.07+	0.06	0.06
S.E. of Difference	0.099	0.100	0.099	0.092

*From Reed & McHugh, 1979.
+Rounding.

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.

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.