Interim Clinical Evaluations in Caries Clinical Trials: Discussion of Dr. Poulsen's Presentation

B. G. GREENBERG

Kenan Professor of Biostatistics, University of North Carolina at Chapel Hill, 27514

J Dent Res 63(Spec Iss):739-740, May, 1984

Preamble.

It is a privilege to be asked to discuss this paper, which deals with several complex issues encountered in statistical theory and principles. As a biostatistician, however, I must point out various statistical inaccuracies. It is my wish to be sympathetic to the authors' approach to questions raised, but I would like to discuss several problems from a statistical point of view.

Two main points.

(1) The authors raise the question of whether interim clinical evaluations should be carried out in a clinical trial of caries development or prevention. They point out that there seems to be almost universal agreement on selecting intervals which are not less than one year apart, because a shorter time period between examinations does not permit enough caries lesions to develop, or, at least, not enough differences between the experimental and control groups. They correctly argue that lack of reproducibility should not influence the *interval* between examinations. They then argue for interim clinical evaluations, probably at annual intervals, for several reasons.

First, let me say that this topic is a general one in statistics called the "repeated-measurements" problem. It has been studied for well over 100 years with particular reference to growth studies. In 1876, Pagliani[1] studied institutionalized children for three years to determine the effects of changes in living conditions on the growth of children. The large numbers of growth studies on stature, weight, and bone development, both in this country and abroad, produced dozens of manuscripts in the first half of this century on how to design these longitudinal studies and how to analyze them.[2-4] Most of the statistical papers were concerned with the question as to whether it was more profitable to use a greater number of subjects by having only baseline and final measurements, or to have fewer subjects with more interim measurements.[5-8] These papers were based upon the desideratum of reducing the variance of the estimates, and basically the conclusion is that the answer depends upon the objective.

The decision on how to allocate one's resources in these types of clinical trials will depend upon the question being asked. Assuming that there are fixed resources to conduct a total of, say, 12 n dental observations, if the sole objective is to determine if there is a difference in the caries increment from baseline to final observation, one might use 6 n subjects for 6 n initial evaluations and 6 n final evaluations three years later. Of course, this assumes no attrition. If the objective is to study a difference between the treatments in the growth rate of lesions, and even if at the end of the three years both groups are alike, one can use 4 n subjects examined at baseline, mid-term, and final periods. This will allow the observer to see whether both groups acted alike in the first and second halves of the clinical trial. Finally, if the objective is to study the decay process itself, it might be best to have 3 n subjects studied at baseline, year 1, year 2, and the final year. The latter permits the fitting of growth curves to individuals as well as to groups. The fitting of growth curves is the most powerful statistical tool, and one should recognize that examination of the increment score using only baseline to final evaluation is equivalent to fitting a straight line to the two points. The curve may well be curvilinear, and intermediate observations are needed to detect the departure from linearity.

The authors provide other substantial but less probabilistic reasons for interim evaluation – such as less attrition, observation of side-effects, better prevention (especially if sealants need to be re-applied), and, perhaps, early termination of a trial if statistical significance develops sooner than expected.

(2) The question of early termination of a trial deserves special comment. The authors correctly point out that there are clinical trials with a fixed number of subjects as well as sequential clinical trials which usually require fewer subjects. Unfortunately, sequential clinical trials are less adaptable in dental studies where it takes a long period of time for the result to be observed. For instance, a decision region may be crossed by the path line in a sequential clinical trial, and if one then allows those in midstream to continue on the study, the path may cross back into the region of uncertainty or no difference.

There is a slight error in how the authors describe the fallibility of examining data in a clinical trial more than once. They claim that if one continues the experiment long enough, the null hypothesis will inevitably have to be rejected unless use is made of sequential trial methods. This is not true. One does need to adjust the nominal level of significance, however, if the data in a fixed trial are to be examined more than once in order to assure that the Type I error is held constant. For example, in the table provided by McPherson[9], if the investigator looks at the data ten times using the test statistic for a 5% level of significance, the true Type I error is not 5% but 19.3%. Examining the trial data early is to be encouraged even in fixed trials, not only for statistical significance but also to ascertain whether there is likely to be sufficient power to detect differences. For example, in the National Preventive Dentistry Demonstration Program conducted by the American Fund for Dental Health, experience during the first two or three years of the study showed that caries incidence even in the control group was considerably less than it had been when the study was designed. As a result, the trial was extended an extra year to obtain four-year increment DMFS scores.

Other points worth mentioning.

(1) As indicated above, all caries incidence clinical trials need not be limited to three-year increment scores. This may be especially true in the case of sealants being tested in a population with fluoridated water and a reasonable degree of oral hygiene.

(2) The authors report that the lack of reproducibility of

the caries examination is a good reason for selecting one-year intervals. If the lack of reproducibility is truly a systematic error, at each examination the systematic error should be the same in both the experimental and control groups and, therefore, the *difference between the groups* remains unbiased. If the systematic error has a synergistic relationship with the amount of caries, then either a suitable transformation of the data would correct this, or it may not be a systematic error. Thus, in the Fig., where the authors show that the standard deviation varies directly with the mean caries increment, a logarithmic transformation of the data will stabilize the variance so that standard statistical procedures apply.

(3) Where the authors use the illustration of the black and white chips, I am afraid their probabilities are in error. For example, after drawing six chips out of the hat where five are one color and the sixth the opposite, the authors consider the probability of drawing a new set of three chips where all three are of the same color as the original five. This is a conditional probability, and the probability of the three being the same color as the five is only $(1/2)^3$, or 1/8. However, even if one ignores the conditional probability aspect and asks what is the probability of having eight out of nine chips of the same color, the answer is $(2 \times 9)/2^9 = 18/512$. If one wants to reject the null hypothesis that the color distribution of the chips is equal, then one must also consider the odds of drawing all nine of the same color. This adds the value of 2/512, for a total of 20/512.

(4) Finally, in discussing reproducibility of examination data, the authors mention systematic error and random error. They state that random error does not bias the increment scores, because random errors tend to "equal each other". Unfortunately, not all random errors are sampling errors with a symmetrical distribution, and, even if they were, small samples could still distort the comparisons badly. Moreover, the variance of the mean difference may or may not be increased, depending upon the correlation between the baseline and final examinations. This comes about because:

$$\text{Variance (Increment)} = \text{Variance (Initial)} + \text{Variance (Final)} - 2\,\text{Covariance (Initial, Final)}$$

The last term on the right is usually much smaller than the sum of the variances, but it need not always be so.

A recent paper by Pocock[10] has considered the question of interim analyses in clinical trials. He bases the conclusions on sequential or multi-stage designs which enable the investigator to be statistically more rigorous in conducting periodic analyses in a clinical trial involving two treatments. By this approach, he is able to estimate the approximate time intervals for interim analyses and concludes that there is no advantage in analyzing trial data on more than five occasions.

REFERENCES

1. PAGLIANI, L.: **Sopra Alcuni Fatteri dello Suilippo Umano.** Torino, 1876; Milano, 1879.
2. SHUTTLEWORTH, F.K.: Sexual Maturation and the Physical Growth of Girls Age Six to Thirteen, Monograph, *Soc Res Child Dev* 2(5), 1937.
3. KODLIN, D. and THOMPSON, D.J.: An Appraisal of the Longitudinal Approach to Studies of Growth and Development, Monograph, *Soc Res Child Dev* 67(1), 1958.
4. TANNER, J.M.: Some Notes on the Reporting of Growth Data, *Hum Biol* 23:93-159, 1951.
5. PATTERSON, H.D.: Sampling on Successive Occasions with Partial Replacement of Units, *J R Stat Soc* Ser B 12:241-253, 1950.
6. WILLIAMS, J.S.: The Variance of Weighted Regression Estimators, *J Amer Stat Assoc* 62:1290-1301, 1967.
7. GRIZZLE, J.E. and ALLEN, D.M.: Analysis of Growth and Dose Response Curves, *Biometrics* 25:357-381, 1969.
8. TIMM, N.H.: **Multivariate Analysis with Applications in Education and Psychology.** Monterey, CA: Brooks/Cole Publishing Co., 1975, Sects. 5-12.
9. McPHERSON, K.: Statistics: The Problem of Examining Accumulating Data More Than Once, *N Engl J Med* 290:501-502, 1974.
10. POCOCK, S.J.: Interim Analyses and Stopping Rules for Clinical Trials. In: **Perspectives in Medical Statistics,** J.F. Bithell and R. Coppi, Eds., London: Academic Press, 1981, pp. 191-214.