

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

General Discussion

RINGELBERG: I have a couple of points, particularly on the Grainger/Lehnhoff paper. The first one is concerned with the correlation matrices. They stated that the correlation coefficients are all lower in the treatment groups, which seems to be true, and that if the treatment was absolute zero the r 's would also be equal to zero. I wonder if it might be of interest to discuss this and these coefficients. Looking at Kingman's paper, it appears to be true if there is correlation between MGSI and caries incidence where you have a control r of 0.36 and 0.32. I think there is only one thing I disagree with in Dr. Kingman's paper, and that was the last statement that investigators are more comfortable with post-randomization than with post-stratification analysis of covariance. Certainly the assumption is there when there is interaction present, and there usually is. But since pre-stratification seems to be preferred, I see no practical reason, as an investigator, why it shouldn't be done.

Second, although I am comfortable with covariance analysis, I don't think it is as universal a cure as Dr. Slakter suggested. Another point was the re-ordered dental age group where they put clusters zero to five, age six to 11, 12 to 19, and 20 to 28 - I think these clusters relate to a disease process and are derived from the data. Certainly the topical trials have shown a close association with smooth surface caries. With the MGSI of two above, which involves proximal lesions, I am not so sure that we are not looking at two different disease entities here. Certainly the organisms are different, suggesting a different etiology. What we are doing primarily is testing these products within pit and fissure populations from 6 to 11, and then extrapolating the results to smooth surface caries in populations of 12 to 19. So I have to disagree with some speakers this morning who suggested that we should be looking for at least topical fluoride studies for an older population. As an epidemiologist, I am aware of the need for having an attack rate which showed a difference for this. As an investigator, I am also fully aware of the constraints of trying to get that population over a period of time.

My final point: I haven't heard anyone compare comparisons. In these days when it costs upward to \$750,000 for a three-year trial, I doubt if there are going to be very many trials with just one treatment group. Nearly always it seems that we have two or more treatment groups. Certainly we are constraining two-tailed tests with these positive controls. There may be some reason for using one with a placebo, but not with the positive control. I'd like to see some discussion or attention to perhaps a more sensitive comparison of procedures.

RULE: Dr. Kingman pointed out that pre-randomization techniques are usually avoided, and this was due to prophylactic use of good prognostic variables as well as logistical problems in the running of the studies. As I was thinking over the nature of this conference and preparing for it, I took a look at almost all (about 25) of the clinical dentifrice studies and mouthrinse studies over the last 14 to 15 years. Actually, it turns out that almost all of them did have some pre-randomization done. Usually it was age and sex and grade and school, that sort of thing. There were very few that used other things like initial DMFS or DMFT. Unfortunately, what was surprising was the number of studies that showed some imbalances after two or three years in the study - over 20% of the papers showed these kinds of imbalances, usually in the numbers of surfaces. When imbalances did appear, the analysis of covariance was usually used and sought out the effect. As Grainger

mentioned, the analysis of covariance is largely unexploited. In fact, only two papers appeared in which analysis of covariance was used as part of the planning design to improve efficiency in which a large array of covariates was used.

CASH: Dr. Grainger indicated in his paper that, in all likelihood, some of the assumptions of the covariance analysis are likely to be violated to some degree. I wonder how one could statistically justify the use of 12 covariables. My concern is that by chance alone you could reduce the mean square error on the magnitude of 25%. Also, the likelihood is that, among 12 covariables, you will have some baseline imbalance among just the covariables which would likely result in some real adjustment in the final increments. In the example that was presented, the adjustment in the second set of data was almost 50%, that is, from 16% to 25%. I think that when you get into this size of adjustment, the assumptions of the covariance analysis become more important, and one would have to justify to some degree the validity of the covariance assumptions and analysis.

With regard to Dr. Kingman's paper, it mentioned that the model was good for several reasons, but one is that it allows for an adjustment due to differences among the prognostic factors, but I wasn't quite sure of the adjustment that was being made. I thought it was by choice of the weights. Maybe you can comment on that. With regard to Dr. Fleiss' paper, I would like to ask how he felt the power of his model was compared to parametric models, and also how he would go about testing for interaction.

With regard to Dr. Varma's paper, I have a strong feeling for non-parametric statistics. Some of Jay Conover's research indicates that, even when you have normal distributions but unequal variance, there is a real loss in power in the parametric test vs. the non-parametric test.

In my cursory review of published papers over the past year, I have often noticed that the equality of variance assumption can be easily violated. There have been several reasons presented here today to explain why that occurs. I think more research is needed in this area to investigate parametric vs. non-parametric in the light of unequal variances and the large sample sizes generally found in caries trials. There is no question of the robustness of these tests, but the comparison of power under a variety of conditions is of definite interest.

LASTER: Dr. Slakter made a case for putting to bed - I got the impression, forever - dealing with increments, basically because increments lack power in a certain sense but are continually used because of tradition. I am somewhat confused about that, because, in the body of his remarks, he also said, quite accurately, that covariance on increments is identical in a general sense with using a conventional approach in covariance, where the conventional approach implies covarying the final on the initial score. That's true. So a very simple way out of this is merely to covary on increments. This solves several problems: We have the more efficient model, and the *clinician* gets to see his data in the traditional way. That's why I couldn't understand the emphasis on not dealing with increments. I understood what he meant. He was opposed to the use of increments without covariance, obviously. Remember that the two procedures - that is, covariance on finals or increments - are directly convertible to each other to produce the appropriate point estimates. Very simply, adjusted increments plus baseline equals adjusted finals. An item you

should be aware of is that the test of the regression coefficient, *i.e.*, that the slope equals zero, is different. If you utilize the increment model, you may well appear to wipe out the impact of the covariate altogether. For example, if the pooled estimated Beta (the regression coefficient) turns out to be exactly one, then in the increment model the sum of squares for Beta will equal zero. Dr. Slakter was also concerned about parallel slopes when you use covariates. If you are able to pre-stratify or block on the covariable that you use in the analysis, you greatly enhance the chance that the test for parallel slopes will come out such that they are parallel (see Snedecor and Cochran).

In relation to Dr. Kingman's paper, I had spoken with him at length on several occasions. There is just so much you can do in 20 minutes! What he has done was well done and accurately done. I would like to compliment him on his work. I would like to add some comments about power relationships in this situation which are not generally discussed. Normally, when you hear the problem of unbalanced data discussed, at least as far as it has been considered here within the fixed model, it is about what to do in the case of interaction and how does one estimate marginal or average effects. Dr. Kingman handled this all very well.

I would like to discuss what problems can be encountered, especially in the computing area, when you have the fixed effects model and there is *no* interaction. That is to say, you have a statistical basis to assume that main effects are estimable. You should be aware, and I'm sure most of you probably already are, of the following items: Most of the package programs that are available today give you, by default, the weighted-squares-of-means solution when dealing with unbalanced data. In the BMDP2V program, for example, you get the weighted-squares-of-means algorithm when the data are unbalanced. When you use the SAS package, another commonly used program, you will get, by default, type 1 and type 4 sums of squares. With all cells filled, type 4 sums of the squares are equivalent to the method of weighted-squares-of-means. What's wrong with that? Well, if you were working as Dr. Kingman was, with the two-way model, you generally look at the test of interaction first. It is the last sum of squares after you have taken out the general mean and the two main effects. If the interaction is deemed to be non-significant, and you continue to evaluate the main effects in the model with the sum of the squares produced by weighted squares of means, the procedure is inefficient. It will lack power. In essence, you should really be using the method of fitting constants in the absence of interaction. The sums of squares for the method of fitting constants are readily available from SAS, type 2 sums of squares. Now, mind you, these two procedures do not generally give the same marginal mean estimates. They will usually be different. In the case of weighted squares of means, the procedure will be inefficient in the absence of population interaction.

Finally, a comment was made that post-stratification will not be preferred by the clinician. One of the important things presented in Dr. Kingman's paper was that this is not to be preferred as post-stratification, but the two-way model for dealing with these kinds of data is preferred to the analysis of covariance. I think that this should be seen as one of the major contributions of Dr. Kingman's paper. While available time does not allow for a full discussion, a major assumption of analysis of covariance is that of parallel slopes — say, in the fully randomized design — and in my experience it has been a rare occasion when this assumption has been met when analyzing unblocked data.

It is usually the case that they are not parallel or they created enough disturbance to mask the main effects that may be present. In this case, the two-way design is to be preferred. I think the major thrust of Dr. Kingman's work should be considered to be that the two-way model should not be discarded just because it is unbalanced, and an analysis of covariance performed instead on the stratifying variable, pre- or post-.

JOHNSON: I will be very general in my remarks. There is a nice cross-section of the various alternative types of analyses that can be used for dental caries data. The variety of methods, though, has to be acknowledged here, during the planning stage of the trial. I think to avoid problems of multiplicities, not only in interpretation of results by way of multiple subgroups or repeated tests over time, it is wise, at least from the FDA perspective, to consider the alternative types of analyses before you set out on a trial. You have to lay out in the protocol, if at all possible, the scoring indices that are planned and the general types of analyses that will be conducted. It sounds difficult and it surely is, but I think we see, at the FDA, new drug applications that come in that have been analyzed probably a zillion different ways. It is possible to come in with the most favorable results based on a specific technique. Many of these techniques vary in terms of their appropriateness for particular sets of data, and we need specifications for the ones that are used.

Another point I want to make is about losses. Even in the case where the dropout rates are similar across treatment groups, it makes me wonder why patients who fail to return for follow-up are distributed, in fact, evenly among the groups. Maybe this is a picky point, but in other cases we see similar rates for dropouts, but we may be concerned because side-effects or other treatment-related variables are the reason for people dropping out of the active drug group; whereas, in the placebo group, the reasons could be as difficult as compliance or something else. One can't rule out possible bias with respect to reasons for withdrawal. I wonder if anyone had looked into this, or whether it was a concern in dental caries trials.

Finally, I think it is great that there has been a search for powerful, sensitive approaches to these data. It is highly relevant in the case of active control trials, which I think many of us perceive as coming down the pike. I would like to echo some of the points made yesterday and earlier today about the importance of obtaining precise estimates of comparative treatment effects, and the difficulties created by low prevalence and very small differences in caries increments that are being sought in these positive control trials. But I think, further, that we need to pay some attention to the level of the effect of the positive control group to ensure that it is beyond that which could be attributed to a placebo.

As caries prevalence declines, and as we are looking at new populations to compare our positive control on new drugs, it is not that simple to interpret a non-significant difference between treatment groups with respect to caries increments. It may or may not mean that the new drug is efficacious. It depends on the positive effect of the control in the population under study. So you have to look back at information to ensure that the level of effects seen in current trials is similar to that in past trials which prove the positive control to be positive. I'm not sure this was ever brought out, even though it may be a simple point. I think we are going to have to deal with it as we see more and more controls.

GRAINGER: A lot of these questions were asked, such as, would you trust a covariance with 12 or 48 covariables

when you know so little of interactions and when you know that part of the mark doesn't have a linear relationship on variable? All we can say is that, if you keep trying, it is a useful thing. We are not wedded to it. But that sort of remark is quite sensible to make; you can't prove it or disprove it, really.

We do want to point out to Dr. Slakter that, really, our analysis was a covariance of increments in which the initial DMF was put in and then dropped out, in the sense that it wasn't quite as good as some of the other covariables. So I hope you weren't dissatisfied with what we did. Indeed, the initial DMF turns out to be a rather poor covariable, because, in a wide spectrum study, there are ranges (such as age six or seven) when the individual may have only the first molars in his mouth. If they are okay, then the correlation of the value of initial DMF is inverse, and you have a negative correlation. You will wait for another three or four years to get more teeth. So throughout those data, that relationship is pretty confused, and, as a result, if you end up with a little positive correlation, you are lucky. I feel you can argue for quite a while as to it being a little bit better, but I feel that the clinical interactive data make it easier for them to envisage the results and make it feel that you have to do it that way. Whether you want to do it another way in addition, that's all I can say.

KINGMAN: I will respond to Dr. Cash's concern first. The question was, were adjustments of any consequence made to the treatment groups as a result of the application, and how does the model actually make the adjustments? The observed marginal means are a function of the relative sample size within a treatment group in each of the four strata. So, unequal weighting of the sample means for the individual cells completely ignores the relative proportion of subjects within the specific cell means. Now, if you would use the relative marginal weights in the definition of, let's say, a generalized Yates procedure, in this specific example you would get essentially the observed marginal mean, because the balance among the three groups was very good within each stratum. However, with equal weights, the adjustment was of little consequence here, because essentially we are weighting each of the sample means. In terms of Dr. Rule's comment, I will have to change paragraph three, that started with "usually" and ended up being "often", to "sometimes". To one comment I can only say that if there are too many observations in a given cell, so that the data cluster along a significant diagonal, you are going to have essentially total confounding, in which case you won't be able to do anything. So I don't know the answer to that.

In the study I referred to, the placebo group was a *weekly* rinse with sodium chloride, so it is not really a double-blind study in the sense that one of the groups of subjects knew they were in an active treatment group, which was the *daily* group since they rinsed five days a week. One can argue for including a fourth treatment group in which there was a *daily* placebo rinse.

As far as drop-out as a function of treatment group (Dr. Johnson's comment), I don't think we have the information as to why drop-outs occurred, other than to say that I'm sure that the majority of it was because of transient moves on the part of the community. As far as Dr. Ringelberg's comment — suggesting that it is more difficult to explain to a clinician the results of a covariance analysis when you have strong interaction in the mouth than it is to explain stratification — for me to say that the level of treatment effect depends on whether you have

an initial DMFS of 8 or a DMFS of 14, I think is hard to interpret clinically. The reason I used the example I did is that I think the strata made clinical sense in terms of further specifying a treatment effect in a way clinicians can understand. For example, in stratum one, for kinds that only had pit and fissure decay, you talk about treatment effects for those kinds of subjects and make more sense, I believe, than to say for subjects who had initial DMFS of 4, as opposed to the third stratum. This third stratum had an MGSJ value of 2, which specified subjects that had not only pit and fissure caries but also poor proximals. So I think your interpretation of the stratum that was used in the example would further make it easier for me to explain to the investigators what interaction meant in this sense.

FLEISS: I have to join in the general attack on Slakter. Bob Grainger pointed out that, in the data set he analyzed, there were other more powerful predictors of increments than initial DMFS score. Three or four years ago, Al Kingman had a paper which showed that in several studies, almost without exception, the pre-treatment level on the Grainger severity index was a better predictor of the increment than were initial DMFS scores. So, theoretically, what happened the other way around, didn't. It may be paradoxical when you have to go by the data.

With respect to Ralph D'Agostino, and the question from the panel: If the numerals that were attached to the several levels of the Grainger severity index had been taken as honest-to-goodness numbers, zero, 1, 2, 3, 4, and if the *t* test were applied to those numbers, then there would have been a trivial difference between the chi-squares I got and the chi-squares that would have been obtained by a traditional parametric kind of analysis. Therefore, the powers would have been almost the same. The only time the parametric approach would have apparently greater power than the chi-square kind of analysis would be if there were enormous treatment effects, which in this business is rare. So there could have been a difference, but for most distributions the powers are almost identical. Peter Imrey pointed out something of which we have to keep reminding ourselves, and that is that the so-called distribution-free methods do not have assumptions. The next time some ivory-tower type says you have no right doing the *t* test because the variances in the groups are not the same and therefore you should have applied the non-parametric procedures, I think it is appropriate to turn the tables and say, prove to me that your non-parametric procedure is as robust for inequality as the *t* test is.

FERTIG: I must reply to Mickey Slakter. You can use increments and also use analysis of covariance, as you well know. Increments all by themselves are okay, but you can use the covariate as an increment. Furthermore, if you use a final value and give the table of adjusted final values, the investigator wants to see adjusted increments. I know you can get the adjusted increments from the adjusted final value, but he doesn't know how to do it, so please give him increments with the analysis of covariance.

D'AGOSTINO: I don't believe I suggested that categorical techniques shouldn't be used. I suggested a set of criteria about the properties before using them, then I went on to say that we know something about the parametric techniques. I have a sense that a lot of these categorical techniques will, in fact, be more powerful than the parametric techniques. I believe that one should consider all the options, and what is happening in the categorical field is that there are an awful lot of techniques, and it is very hard to sort one out. Going back to Mary Johnson's

comment, I also like to think that before you run the study you should have an idea of what procedures you are going to use. I think you should state, in the protocol, that this is the type of procedure you would like to use. You should know a lot about the procedure in terms of its power and in terms of what you can expect. This is what I was trying to drive home, not the idea that these techniques shouldn't be used. I think these techniques are quite fine, as a matter of fact, and have some contributions to make.

SLAKTER: Every time I lecture, I receive more reinforcement that I am not as good a teacher as I think I am. I didn't mean to imply that the analysis of covariance is the be-all and the end-all. All I am saying is, it should be used in place of increment analysis. I also agree that increment analysis of covariance will end up with the same results as analysis of covariance. I am not quite sure why you go through the increment analysis. I see nothing wrong with reporting the increment. In doing a test on the increments as well as the analysis of covariance, then it seems to me you are doing two tests on the same data.

LASTER: It is somewhat easier, given the increments that have been adjusted, to add them to the baseline to produce the finals. The reason I suggested increments is because — as you said, and many people agree — most clinicians are accustomed to and feel more comfortable with looking at the increments.

SLAKTER: We are pretty much agreed. I would just do the analysis of covariance, and I think we end up with the same thing. In regard to the Grainger and Fleiss discussions, I want to make clear that the reason why the initial DMFS is a poor predictor of the increments is that it should be zero if the increments were doing the job. The correlation between initial DMFS and the increment would be zero, so I don't think we can critique it as a barrier on the basis of using the criterion as increment. If you use the criterion as post-DMFS or final DMFS, I'm sure you will find that initial DMFS is the best predictor and therefore the best covariate. Let me say again, I would like to put to rest the increment analysis of variance. You are sacrificing precision at a time when you can ill afford to do that. A plain analysis of variance of increment scores is apt to be a very imprecise approach. Now, as I understand it, the slopes may even be getting to the point where increment analysis will be less precise than analysis of variance of just the post-treatment scores. When that slope gets to be less than half, it turns out that increment analysis will be less precise than just analysis of variance of the post-scores.

FERTIG: I think it is a matter of terminology. To say that DMFS is predicting for final score — nobody disputes that. But we are disputing that it is not much of a predictor for increments.

SLAKTER: I am suggesting we shouldn't be doing increments.

FERTIG: The investigators want to see the increment, adjusted or not. Ask Hersch Horowitz — he is always saying this. Adjusting the final value isn't good enough for him, and the investigator doesn't know how to get the adjusted increment. Please, if you want to do the analysis of covariance, using final value as a response variable, okay, but give him the adjusted increment somewhere. We weren't suggesting doing a test on increments. We were suggesting doing it as a response variable and the analysis with the covariable.

LU: I would like to offer a partial answer to the question that Dr. Johnson raised. I'm glad she brought this out, because this has been a problem that's plaguing all of us.

For example, if you run a test and your examiner is on the ball, then you get a higher rate of decay by someone. Then, when the x-rays get to the dentist, he goes in and puts in eight fillings and the parents get mad and then their child quits the study. So they quit, and it lowers the average of the effect on the control group. If one of these things happens in the test group, it diminishes the true effect of its efficacy. With informed consent, they can quit whenever they please, and I have no way of keeping the subject in the study, so this is a problem that we have to live with.

BELL: I want to point out that there aren't that many different ways to analyze a particular data set, although statisticians have a good way of making it look like there are. Dr. Fleiss' thought about using riddits is essentially the Wilcoxon Signed Rank test — I think it is evident from the Table in Dr. Imrey's presentation. What's more, it is essentially the same as doing a two-sample *t* test on the ranks. I bring this up not to imply that you should look at different approaches to going after a data set. Instead, I wish to indicate that all of these methods are really familiar ones. If we have a two-sample *t* test, then, if we feel like understanding it, we really understand these other methods as well. The only relevant area is the response ratio. One uses the ranks of the data rather than the raw scores. There are certainly pro's and con's to this approach. I can see each of those. I think there is something to be said from the point of view of doing a *t* test on the variable. You quickly generalize these methods back to analysis of covariance in an attempt to draw together all four papers. I hope that no one felt bombarded by different equations. There really isn't, as far as I can tell, that much difference in the procedures.

IMREY: There are differences in these procedures in spirit. I would like to mention a paper recently published in *Biometrics*.* It is an encyclopedic-type review of analysis of covariance, using both continuous and discrete covariates for different application schemes of both parametric and non-parametric content. The paper contains material that brings together these techniques presented here and many others in a fairly unified framework that makes it all understandable.

D'AGOSTINO: The *t* test and other parametric procedures are basically permutation tests, and one way of viewing them is that you may have a continuous underlying phenomenon that you can't measure directly but you can, in fact, categorize individuals into bins. Then, based on that, you assign the numbers zero, minus 1, and you generate a permutation test. These tests have been shown over the years to have extremely high power qualities. You will, in fact, have a lot of these techniques doing as well as the categorical techniques. I do think the techniques have some weaknesses in terms of the assumptions and the way one can generate a lot of very specific tests, and also get some estimates. After you perform the *t* test, all you have done is performed the tests of hypotheses, and you are stuck with trying to quantify estimates. You don't have these things, and you don't have the riddits available, and so forth. These are things which you gain from pursuing these other techniques. Again, I will emphasize that, before we want you to use them in these types of data, we should feel comfortable in what they are doing. But they certainly have something to offer beyond the *t* test. Let me say one other thing:

*KOCH, G.G.; AMARA, I.A.; DAVIS, G.W.; and GILLINGS, D.B. (1982): A Review of Some Statistical Methods for the Analysis of Covariance, *Biometrics* 38:563-596.

My interest in the *t* test and parametric procedures comes not so much in terms of why or how I should analyze the next study, but rather because a lot of people have analyzed previous studies using the *t* test, using the parametric procedures where they were wrong. Many of the investigations we went through were to see just how well those techniques actually perform, and to justify or not justify previous analyses.

BURCHELL: I have a couple of points. I think that,

with the small increments of difference, we still feel a lot more comfortable with the two-tailed test. On this question that Dr. Johnson brought up, about checking if the positive control is still working, I can't see any ethical way in which we can do that. All that we can do is to make sure that the positive control that we are going to use is a well-established positive control. Historically, it has been shown to work. I don't see any way that you can show that it is still working against a placebo.

on the
Then,
uts in
child
ge of
things
ect of
never
ect in
with.
many
ough
there
tially
from
, it is
n the
ok at
ead, I
miliar
d like
thods
. One
There
each
n the
ticky
ice in
at no
really
pro-

es in
ished
alysis
riates
and
that
many
it all

roce-
y of
nder-
but
Then,
us 1,
have
ower
iques
think
the
very
you
l the
g to
you
are
ques.
use
table
thing
hing:

NGS,
alysis