ANALYSIS OF REPEATED MEASUREMENTS DESIGNS

by

Dennis B. Gillings, Michael J. Symons and Mary M. Donelan

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 1215

MARCH 1979

ANALYSIS OF REPEATED MEASUREMENTS DESIGNS

Dennis B. Gillings

Michael J. Symons

Mary M. Donelan

ABSTRACT

Multi-clinic trials in which patients are assigned at random to two or more treatments within each clinic are considered for analysis. A profile of repeated measurements initially and at each of one or more follow-up visits are recorded for each patient. The recorded data are assumed to be ordinal categorical in the illustration, but discussion for continuous and nominal data situations is also included. Univariate and multivariate parametric, non-parametric, and categorical data methods are considered as potential analysis strategies. Selected strategies of each type are compared and results presented for each of three similar clinical trials. Recommendations for analysts faced with similar problems are suggested with particular reference to missing data situations.

Key Words Multivariate and univariate methods; repeated measurements designs; ordinal categorical data; multi-clinic trials

AUTHORS FOOTNOTES

Dennis B. Gillings and Michael J. Symons are both Associate Professors, and Mary M. Donelan is a Systems Analyst, all in the Department of Biostatistics, School of Public Health, University of North Carolina, Chapel Hill, N. C. 27514. This work was supported in part by Hoechst-Roussel Pharmaceuticals Inc. and Burroughs Wellcome Company. Further, the authors would like to thank Hoechst-Roussel Pharmaceuticals, Inc. for providing the data which were used for the examples in this paper. Special thanks is due to Irwin Ho of Hoechst-Roussel Pharmaceuticals, Inc. who spent many hours discussing data analysis strategies with the authors. Particular thanks is extended to Professor Gary G. Koch of the Department of Biostatistics at Chapel Hill for several lengthy discussions which provided important insights into the practice of statistics. Finally, the authors appreciate Bea Parker for her careful typing of the original manuscript.

1. INTRODUCTION

A commonly occurring experimental design is considered in which patients (or experimental units) are allocated to two or more treatments. The allocation may be an unrestricted randomization scheme or one in which blocks of patients are assigned to different treatments at random so as to equalize the numbers of patients in each treatment group within relevant strata. Usually there are two treatments: an experimental and a control or standard for comparison purposes. The follow-up of the treated patients produces a profile of possibly correlated measurements for each patient observed at several points in time. In addition, a baseline observation provides a measurement on the study variables before the treatments are administered.

It is assumed that the general situation outlined here will be carried out at several clinics or settings each involving a group of patients or experimental units. A well defined protocol must be agreed upon by the investigators in all the clinics to assure good quality data.

This paper discusses several approaches to the analysis of data generated by experiments of the type described and highlights many of the issues involved. Hence, the main concern is statistical practice and so many of the points covered may be controversial. Numerical results are compared for selected methods and recommendations concerning analysis strategies suggested.

2. NOTATIONAL CONVENTION

For notational convenience denote the groups of patients by G_i, i=1,...,g. One group of patients will be under study at one particular clinic or setting. Denote the treatments by T_j, j=1,...,t, the profile of the measurements by P_k, k=1,...,p, and the visits (or times) at which measurements are taken by V_ℓ, ℓ=1,...,v. If the measurements are categorical and categories of the profile P_k are labeled m = $1,2,\ldots,q_k$, study data may be represented in an array as shown in Table 1. This array identifies five patient groups (g=5), two treatments (t=2), a profile of two measurements (p=2) taken initially and at two follow-up visits (v=3), and assumes that each variable is categorical. The response for the first variable (P_1) is recorded under three categories, and for the second variable (P_2) under two categories. The (marginal) frequency with which a response may be labelled i,j,k,l,m is denoted by r_{ijklm}. If some or all of the data were interval, grouping of data would be necessary before they could be presented as in Table 1. However, Table 1 does represent the scope of data collected and is convenient only for reference. In general the grouping of continuous variables for analytic purposes is not recommended, although in specific instances, for example where the quality of measurements is doubtful, categorization may be desirable.

TABLE 1 HERE

The notation in Table 1 is different from that presented by

Koch et al. (1977). From a technical point of view, Koch et al. (1977) give notation that is preferable but at the same time more complicated. For example, the frequencies of Table 1 would be more precisely referred to as marginal frequencies. The simplified notation used here is adequate for the purposes of this paper.

3. ANALYSIS STRATEGIES

The analysis of this design when data are ordinal is of considerable interest since it is possible to argue for the application of statistical methods that are concerned with

A. Continuous data, i.e., traditional parametric methods

B. Ordinal data, i.e., non-parametric methods

C. Nominal data, i.e., categorical data methods.

General methods are available for the parametric and categorical approaches, both with an underlying linear model structure. Smith, Gnanadesikan, and Hughes (1962) give a practical description of MANOVA and Grizzle, Starmer, and Koch (1969) present an analogous methodology for the analysis of categorical data. The non-parametric procedures are less flexible as, for example, in their ability to handle interactions although some procedures for interactions do exist in specific instances (Koch, 1969). However, in any given situation, none of the foregoing approaches may be entirely adequate. In the parametric case, there is concern because the underlying data are not continuous; in the non-

parametric case the investigator may worry about occurrence of many ties; in the categorical data case, the cell sizes may not be large enough to be confident that the required asymptotic results hold.

A further choice with cases A and B is between a multivariate or univariate approach. If missing data were not an issue, then a multivariate approach is likely to be preferable. If there are several missing data items (say 10% or more), then multivariate analysis may be questionable due to the relatively large number of data points that would need to be estimated in order to proceed straightforwardly.[1] However, salvation will not be in univariate methods alone, as there is also need to demonstrate that it is legitimate to ignore the missing data. One is concerned when the reason for the missing values is related to the treatment effects. If, however, the missing data occur at random and not according to some specific pattern, then standard univariate methods could be applied at each time point.

Let us now consider the application of different analytical procedures to data taken from three clinical trials that were performed to test the efficacy of a new drug for psoriasis or atopic dermatitis. In the first trial, the new drug was compared with a placebo in each of five clinics. Up to 40 patients were randomly assigned to drug or placebo within each clinic, i.e., nearly 20 patients per drug per clinic were included. For each patient, several specific signs and symptoms related to the progress of the skin condition were rated, each on a 1-5 scale, at the baseline visit

and at each of three follow-up visits. An assessment of the initial status of the disease based on the patients' reports of the condition was rated on an ordinal scale at baseline. Only patients whose condition was stable or worsening were included. In addition an overall evaluation score, again on a 1-5 scale, was recorded for each patient at each follow-up visit. The overall evaluation summarized the degree of improvement of the patient's condition over baseline, and is the outcome (dependent variable) that is chosen here for purposes of illustrating relevant analyses.

The other two trials were essentially similar except for differences in the control drug used, number of clinics involved, and number of follow-up visits. In summary, the three trials to be considered may be specified as shown in Table 2. The data for the overall evaluation scores collected during Trial I are listed in the Appendix. These data are representative of the data for all three trials except that Trial III had very little missing data. It can be seen that the trials are quite similar. Trial I has one more follow-up visit than Trials II and III, and Trial III has one fewer clinic.

TABLE 2 HERE

The questions of interest in the analyses of the three trials were as follows:

A. Were there any differences between the test drug and the control? Did the difference in drug effects change over time?

B. Were there differences between clinics as regards patient outcomes and did such differences change over time?

It was recognized that clinic differences might include differences between clinic settings, differences between physicians who practiced in the clinics and the way they rated patients, and differences in the types of patients in each clinic. Some of the possibilities may be excluded if a detailed protocol were followed closely. However, the precise nature of clinic differences was not a question that these clinical trials were designed to address.

C. Were there different degrees (or reversals) of drug effects for the different clinics and did such effects change over time?

In statistical terms, the questions of interest may be stated in terms of the significance of drug effects and drug by visit interactions, clinic effects and clinic by visit interactions, drug by clinic interactions and drug by clinic by visit interactions.

Consider now, some selected univariate and multivariate approaches for continuous, ordinal, and categorical data in turn. These approaches will explore the presence and significance of the effects of interest. Not all possible analysis strategies will be covered but the analyses illustrated will give a flavor for types of approaches which are relevant and available to the analyst.

Univariate

a. ANOVA or ANCOVA. The initial severity may be used as a covariate. Other covariates such as age, sex or other demographic or clinical variables may be appropriate in particular situations. A separate analysis must be carried out at each time point. Hence effects which interact with visit effects cannot be tested directly. The missing data at each visit should be explored to see if there was any tendency for missing visits to relate to one of the drugs. Such a tendency might invalidate the univariate analyses at each time point. The model used in each analysis might include drug, clinic, and drug by clinic effects. All these effects are assumed to be fixed. Although one might argue that clinic effects should be random, usually there are a small number of clinics (say three to ten) that are specifically selected. No random selection of clinics is attempted and by the same token generalization to all potential clinics is not attempted, at least using statistical arguments. In this situation fixed clinic effects are felt to be most appropriate.

Another approach is to consider an analysis of "last recorded visit minus baseline visit" scores. Patients who do not complete the final visit (drop-outs) would appear in the analysis if they attended one of the preceding visits. However, some justification of this analysis is needed to ensure no selective biases led to patients on one drug dropping out more frequently than patients

on the other drug. Patients may drop out because they do not perceive any drug effect or perhaps because they feel cured and do not bother to return. Other reasons unrelated to the drug can also cause missed visits or drop-outs.

Care should be taken to test assumptions associated with the model. The investigator would expect a significant overall model, significant covariate(s) and the equality of slopes of the covariate(s) across cells. Exploration of the normality of residuals can be undertaken, but may be unnecessary in the situation where competing analyses (i.e., non-parametric analogues) are included for comparative purposes.

b. Non-parametric. Wilcoxon-Mann-Whitney tests (with three or more treatments, Kruskal-Wallis tests) may be applied to the outcome scores, a separate analysis at each time point. In some situations, it may be argued that tests might be applied to difference scores (Follow-up visit score minus Baseline Score). This would provide an adjustment for baseline. However, it is not possible to adjust for covariates in the non-parametric case in the same way as for the parametric analysis.[2] The tests may be applied to each clinic separately and clinics pooled to test for drug effects. Clinic effects may be explored using the Kruskal-Wallis test, by comparing mean ranks for each clinic.

Unfortunately, the pooling of clinics may not be a straightforward matter. In the absence of clinic by drug interactions,

pooling may be defended since the patients from the different clinics are similar statistically as regards drug effects. If clinic by drug interactions are present, it is not always easy to defend pooling across clinics. Subgroups of clinics with similar drug effects are good candidates for pooling. In addition, if the net drug effect is in the same direction for all clinics, pooling is still a reasonable strategy in the presence of clinic by drug interactions, representing significant differences in the magnitude of the similar drug effects. However, if one or more clinics show a net drug effect in the opposite direction to the remaining clinics, the possibility exists that there are different subpopulations for which different inferences can be made as regards relative drug efficacy. In this latter situation, a reasonable strategy might be to pool those clinics in which the direction of the net drug effect is the same.

Univariate (quasi-multivariate) Analysis

ANOVA for repeated measurements designs is discussed in Winer (1971), chapter 7. The technique could be applied here but is not as preferable as MANOVA due to restrictive assumptions on the form of the variance-covariance matrix. However, this procedure would allow for visit interactions to be tested directly. Alternatively, approximate tests could be used as described by Greenhouse and Geisser (1959), provided that a plausible basis for their validity can be postulated.

Multivariate Analysis

In the parametric and non-parametric cases, there are options concerning the choice of multivariate vector, namely

A. one variable of the profile at the successive time points--in this case the analysis is repeated for each variable in the profile;

B. the entire profile of measurements at a given time point--the analysis is repeated at each time point;

C. one "super" vector over both the profile of measurements and time.

For the purposes here, the first of these options was selected. Parametric and non-parametric strategies are multivariate analogues of the univariate procedures outlined in A and B.

a. MANOVA. The initial severity may be used as a covariate. If the outcome vector is one variable measured at successive follow-up visits, then drug effects, clinic effects, drug by clinic interactions, drug by visit interactions, clinic by visit interactions, and drug by clinic interactions can all be tested using standard methods as described in Smith, Gnanadesikan, and Hughes (1962), and Cole and Grizzle (1966) using a two-way model with drug and clinic main effects and drug by clinic interactions. However, some problems can arise in the interpretation of the multivariate test statistics.

Significant differences may occur according to the criterion of a multivariate test statistic that do not correspond to

differences observed in univariate situations. For example, consider the following hypothetical data in Figure 1 where the mean improvement scores for each follow-up visit over baseline is plotted at three time points for two drugs. The two profiles look somewhat erratic and may well lead to a significant drug difference based on a multivariate test of mean improvement scores simultaneously over the three follow-up visits. However, nearly all of the significant drug difference may be explained by the presence of drug by visit interactions. Hence, interactions with time (or any other variable on which the profile of multivariate observations is based) may be a factor as regards the presence of (multivariate) main effects. Further, these main effects are not always interpretable as differences in the same way as univariate main effects since a significant difference may not correspond to one drug being better than the other. Univariate test statistics at each time point will help clarify the nature of the differences observed. However, a more definitive assessment may be provided by a univariate test that averages the drug effects over time or measures the trend of the effects over time. Such univariate tests are really quasi-multivariate in the sense that they summarize differences over several dimensions by a one-dimensional quantity. In turn, this is analogous to any multivariate test statistic which selects a one-dimensional representation (such as largest root) of a multi-dimensional

entity (a matrix of sums of squares and products).

FIGURE 1 HERE

b. Non-parametric. Multivariate extensions of the Kruskal-Wallis test may be applied to test for drug effects, clinic effects, drug by visit interactions, and clinic by visit interactions as described by Koch (1969). Similar comments apply to the interpretation of the multivariate test statistics as in the parametric case.

c. Categorical Data Analysis. The methods described in Grizzle et al. (1969) may be applied to the mean scores at each time point for each of the drug by clinic subpopulations to produce categorical data analogues of the MANOVA tests outlined in (a). However, one difference here between the methods of Grizzle et al. (1969) and MANOVA is that covariates (e.g., initial score) cannot be included in the model in the same way as in MANOVA. The initial score may be incorporated into the Grizzle et al. (1969) analysis by analyzing difference scores, or by increasing the number of subpopulations according to initial visit scores.[3] The latter is often impractical as the cell sizes may become too small. The former may or may not be appropriate, depending on the scale of measurement as is the case with the non-parametric approach if difference scores are selected for analysis. It should also be emphasized that the analysis of categorical data using maximum likelihood rather than weighted least squares may be equally appropriate. The maximum likelihood approach is

described fully by Bishop, Fienberg, and Holland (1975).

An alternative categorical procedure that is relevant but not presented here is the Mantel-Haenszel (1959) test which allows the results from different investigators to be combined to generate a χ^2 test of treatment differences. Mantel and Haenszel also described a one degree of freedom test for the q×2 case. This would apply to the trials considered here as there are two drugs and q=5 ordered response categories for each clinic. This test was extended by Mantel (1963) to the q×t case, (i.e., q ordered categories, t treatments for each clinic). Landis et al. (1977) discuss general procedures for the q×t case as an extension of a partition strategy used by Koch and Reinfurt (1974). Thus, both (q-1)×(t-1) degrees of freedom test statistics as well as one degree of freedom tests like those of Mantel are considered.

Parametric, non-parametric, and categorical multivariate procedures were applied to data from Trials I, II and III carried out on human subjects. The measurement profile considered here comprises just one variable, an overall assessment of improvement over baseline, rated subjectively on a scale 1-5 (1 = recovery, through 5 = condition worsened; see Appendix for definitions). In the parametric case, a MANOVA was performed using a fixed effects model that included drug effects, clinic effects, drug by clinic interactions, and initial severity as a covariate. The multivariate response was the outcome at the successive follow-up

visits. For the non-parametric analysis, the methods of Koch (1969) were used to test for drug effects, clinic effects, and their interactions over follow-up visits. The categorical data analysis used a saturated model that included drug, clinic, and visit main effects and all possible second and third order interactions. The increases (Q_c) in the goodness of fit χ^2 statistics (Q) as the different effects were eliminated from the model were used to test the significance of each of the effects in turn [see Appendix I in Koch et al. (1977)].

Initial severity was a suitable covariate and for that reason was included in the MANOVA model. However, the covariate did not really contribute to the model from a statistical viewpoint and so the p-values presented would hardly change if the covariate was excluded. No comparable adjustments were carried out in the non-parametric and categorical data cases partly because of the MANOVA results for testing the model. In the categorical data case, further subpopulations could have been used but this would have reduced the cell sizes considerably. An entirely satisfactory non-parametric adjustment procedure is not available.

Trials I and II had about 10% of the data points missing for the follow-up visits but Trial III had very few missing data points. All missing values were estimated using mean values of all patients, regardless of drug, in the same clinic, and who

initially were of comparable severity. This procedure tended to be conservative as regards exploring drug effects since the estimated values substituted for missing data were averages of responses for both drugs. Comparable univariate tests excluding missing data were carried out in most parametric and non-parametric cases for Trials I and II, and the results for these tests are indicated in parentheses.

4. DISCUSSION

The results of the tests performed are shown in Table 3. Overall, the results from the different approaches are reasonably similar. There seems to be a tendency for the analysis of Grizzle et al. (1969) to give smaller p-values, sometimes considerably, than those for the corresponding parametric and non-parametric analyses. One reason for such small p-values may be the relatively small sample sizes which were involved for each clinic by treatment combination as they relate to asymptotic results. The situations in which all three tests reached significance, say at the 0.05 level, would appear to be definitive. More equivocal situations are represented by the cases for which one or two of the tests reached significance and the remainder did not. For example, the drug by visit interactions are significant for Trial II at the 0.05 level ($p=0.035$) for the non-parametric test, but are not significant in the parametric and categorical analyses. However, some of the equivocal situations may get resolved if there are corresponding

univariate and multivariate analyses. For example, in Trial I the non-parametric test for clinic effects at Visit 2 is not significant at 0.05, in contrast to the parametric and categorical data tests. However, all three of the Visit 1 tests are significant as well as the corresponding multivariate tests. Such overall consistency argues for the presence of clinic effects. Specific clinic differences would then warrant further investigation.

TABLE 3 HERE

For Trial II, the categorical data tests point to significant clinic by drug interactions. The MANOVA analysis appears to support this although the multivariate test ($p=0.065$) does not quite reach significance if the 0.05 criterion is used. The interactions serve to caution the analyst as regards pooling. The corresponding univariate tests are significant, however. Also, plots of the drug effect for each clinic show the interactions to be due to the varying degrees of drug effect in the different clinics rather than reversals of effect. Hence, the pooled non-parametric results are felt to be valid in this case.

In a given situation, the user would probably feel more comfortable with an actual recommendation as to which method to use. However, the problem of missing data must be discussed before recommendations are given. Problems which can be deferred to the statistician are often the responsibility of study management rather than mathematical wizardry. Missing data fall into

this category and should not be a statistical problem, rather one of good management and careful implementation of protocol. There are several statistical techniques available to deal with missing data but they either make assumptions which might be difficult to defend (e.g., using a mean value to substitute for the missing datum or perhaps using a more sophisticated model to estimate missing values) or they are so conservative that it is doubtful that significance could ever be obtained if they were applied (e.g., score treatment values which are missing as "worst possible" situation, and control values that are missing as the "best possible").

The problem of missing data gets more formidable, of course, in the multivariate situation where if one or more components of the vector are missing, then the whole vector must be ignored or the missing component(s) need(s) to be estimated. An ad hoc guideline that may be reasonable is to aim for less than five percent of the total number of data points missing. In this situation study management can be taken to be adequate as far as data collection is concerned. Ten percent or more missing data may be inadequate and missing data not estimated in this situation. Between 5% and 10% could be regarded as a no-man's land where the analyst can proceed with missing data techniques but caution should be observed.

If sample sizes are moderately large (five or more per cell and 20 or more per degree of freedom) and there are less than five percent missing data, then MANOVA, multivariate non-parametric and the methods of Grizzle et al. (1969) are about equally desirable. Familiarity or ease of access to particular computer programs may well be the deciding factor. Other factors to be included in the analysis may also be important. For example, if covariates were critical to the analysis, MANOVA may be the procedure of choice as the adjustments could be handled more satisfactorily.

With moderately large cell sizes but more missing data (10% or more), the ratio estimation procedure of Stanish et al. (1978) is likely to be preferable provided the assumption that missing data occur at random can be justified. This recommendation is made because of the ease with which the analyses may be conducted with one computer program MISCAT, described in Stanish et al. (1977). However, if covariates were essential to the analysis, cell sizes may become small due to too many subpopulations. In this case the ratio estimation procedure is not to be preferred and one may have to resort to an elaborate model to estimate missing values or perhaps be content with supporting a MANOVA analysis with univariate analyses at each time point ignoring missing data. With intermediate amounts of missing data and moderately large cell sizes, the analyst may select either the MANOVA, multivariate non-parametric or categorical analysis,

depending on arguments that are relevant in any particular situation.

If cell sizes are small, which is usually the case, then the missing data must be limited to five percent in order to be able to justify the analysis with some confidence. The choice now is between MANOVA and non-parametric analogues. The non-parametric procedures are usually preferred because they need less assumptions. It is common to derive non-parametric tests for location under the assumption of variance homogeneity (see Hollander and Wolfe, 1973) which is a more stringent requirement in practice than normality due to the convenient relevance of central limit theory. However, non-parametric procedures which test for differences in location as well as more general differences between distributions may also be derived from the underlying randomization distributions and, of course, the only assumption required in this case is randomization itself. In practice, the randomization distributions are not worked out and instead, normality of these distributions is assumed. In turn this is justified by randomization central limit theory which require no assumptions of variance homogeneity. However, a lack of homogeneity will reduce the power of the tests to discriminate against corresponding alternatives. It should also be noted that the speed of convergence of the randomization distributions to normality when there are many ties may be a problem as regards the

non-parametric procedures. Unfortunately, computationally practical non-parametric techniques are not available to test for drug by clinic interactions, and so a parametric procedure such as MANOVA needs to be used. In practice, the MANOVA and multivariate Kruskal-Wallis tests can be expected to give similar results in practical situations, as was demonstrated in this section, and so both tests may be used to confirm borderline cases of significance.

The final case to consider is smaller cell sizes and a good deal of missing data (10% or more). In this situation one can turn to univariate analyses ignoring missing data at each time point and for each variable in the profile. Other relevant approaches are more complex and require assumptions of multivariate normality[1]. However, the univariate approach ignoring missing data must be defended by testing the equality of missing data proportions in treatment subpopulations. If reasons for missing data are available, then additional comparisons of the relative frequencies of reasons for treatment subpopulations should enable the researcher to conclude whether or not it is justifiable to assume missing data occurred at random. However, small cell sizes will be a problem in any analysis of missing data proportions and so it may not be possible, in some cases, to carry out an analysis which is entirely defensible on statistical grounds.

5. CONCLUSION

Multiple analyses involving two or more distinct approaches

to the analysis as presented in Table 3 has much appeal, and we recommend a dual analysis approach for the following reasons. First, it safeguards against errors, particularly major ones, as the analyst looks for comparable results. Second, there is usually a greater feeling of comfort on the part of the analyst as regards degree of belief in the conclusions suggested. The second analysis provides a confirmation. Of course, it may be argued that the dual analysis can lead to ambiguity, since one test may be significant and the other not so. However, ambiguity only arises if both analyses are treated as compelling. If instead, one of the approaches is treated as the procedure of choice and the other a supportive analysis, no such ambiguity arises. Further, if the conclusions are based partly on p-values rather than significance levels alone, there is even less potential for ambiguity. In this situation there are three possible conclusions:

1) Yes there is a difference
2) No difference has been demonstrated
3) The results are suggestive of a difference but not definitive.

However, it must be remembered that p-values implied by statistical tests are, in practice, approximations to the actual p-values. Provided the distributions are not excessively degenerate, the approximations are usually adequate for moderate sample sizes except for the cases of small p-values. Large samples may be

needed to accurately reflect p-values as low as 0.001.

Finally, it is stressed that other considerations may be relevant and should be incorporated before a conclusion is reached. These might include data from other clinical trials, other types of medical evidence, data quality, and the implications of an erroneous conclusion. Unfortunately, it is unlikely that an algorithm for statistical analysis can be devised which will reflect all of the necessary features. The analysis will be based on a model that approximates the truth and so the results should always be used in concert with good judgment.

FOOTNOTES

1. Specialized analysis approaches have been devised to deal with missing data for multivariate problems. See, for example, Hocking and Smith (1968) or Orchard and Woodbury (1972). More recently, Stanish, Gillings, and Koch (1978) have devised an approach based on linear models for categorical data.
2. Indirect tests are sometimes possible. These are undertaken by subtracting the (multivariate) Kruskal-Wallis statistic for comparing treatments with respect to the covariables alone from the joint (multivariate) Kruskal-Wallis statistic for comparing treatments with respect to both the dependent variables and covariables simultaneously.
3. Alternatively, the initial score may be taken into account by subtracting test statistics pertaining to it alone from multivariate test statistics involving both it and specific visit scores. This strategy is essentially the same as that outlined in footnote 2. Its application is illustrated in Grizzle, Koch, and Sen (1977).

REFERENCES

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975). *Discrete Multivariate Analysis*, Cambridge: The MIT Press.

Cole, J. W. L., and Grizzle, James E. (1966). The Application of Multivariate Analysis of Variance to Repeated Measurements Experiments, *Biometrics* 22, 810-828.

Greenhouse, S. W., and Geisser, S. (1959). On Methods in the Analysis of Profile Data, *Psychometrika* 24, 95-112.

Koch, G. G., Grizzle, James E., Semenya, Kofi, and Sen, P.K. (1978). Statistical Methods for Evaluation of Mastitis Treatment Data, *Journal of Animal Science* 61, 830-847.

_____, Starmer, C. Frank, and Koch, Gary G. (1969). Analysis of Categorical Data by Linear Models, *Biometrics* 25, 489-504.

Hocking, R. R. and Smith, N. B. (1968). Estimation of Parameters in the Multivariate Normal Distribution with Missing Observations, *Journal of The American Statistical Association* 63, 159-173.

Hollander, M., and Wolfe, D. A. (1973). *Nonparametric Statistical Methods*, New York: John Wiley and Sons.

Koch, Gary G. (1969). Some Aspects of the Statistical Analysis of 'Split Plot' Experiments in Completely Randomized Layouts, *Journal of the American Statistical Association* 64, 485-505.

____, Landis, J. Richard, Freeman, Jean L., Freeman, Daniel H., and Lehnan, R. G. (1977). A General Methodology for the Analysis of Experiments with Repeated Measurement of Categorical Data, *Biometrics* 33, 133-158.

____, and Reinfurt, Donald W. (1974). An Analysis of the Relationship Between Driver Injury and Vehicle Age for Automobiles Involved in North Carolina Accidents During 1966-1970, *Accident Analysis and Prevention* 6, 1-18.

Landis, J. Richard, Stanish, William M., Freeman, Jean L., and Koch, Gary G., A Computer Program for the Generalized Chi-Square Analysis of Categorical Data Using Weighted Least Squares (GENCAT), *Technical Report No. 8, Department of Biostatistics, University of North Carolina at Chapel Hill,* Revised April 1976.

Mantel, N. (1963). Chi-Square Tests With One Degree of Freedom; Extensions of the Mantel-Haenszel Procedure, *Journal of the American Statistical Association* 58, 690-700.

____, and Haenszel, W. (1959). Statistical Aspects of Data From Retrospective Studies of Disease, *Journal of the National Cancer Institute* 22, 719-49.

Orchard, T. and Woodbury, M. A. (1972). A Missing Information Principle: Theory and Applications, *Proceedings of the 6th Berkely Synposium* 1, 697-715.

Smith, H., Gnanadesikan, R., and Hughes, J. B. (1962). Multivariate Analysis of Variance (MANOVA), *Biometrics* 18, 22-41.

Stanish, William M., Gillings, Dennis B., and Koch, Gary G. (1978). An application of multivariate ratio methods for the analysis of a longitudinal clinical trial with missing data, *Biometrics* 34, 305-317.

____, Koch, Gary G., and Landis, J. Richard (1977). A Computer Program for Multivariate Ratio Analysis (MISCAT), to appear in *Proceedings of the American Statistical Association.*

Winer, B. J. (1971). *Statistical Principles in Experimental Design, Second Edition*, New York: McGraw-Hill Book Co.

TABLE 1

Data Array for a Class of Multicenter Clinical Trials with a Profile of Repeated Measurements - Categorical Data Case

Group/ Treatment	Visit/ Profile/ Category	V_1					V_2					V_3				
		P_1			P_2		P_1			P_2		P_1			P_2	
		1	2	3	1	2	1	2	3	1	2	1	2	3	1	2
G_1	T_1	r_{11111}														r_{11232}
	T_2															
G_2	T_1															
	T_2							$r_{ijk\ell m}$								
G_3	T_1															
	T_2															
G_4	T_1															
	T_2															
G_5	T_1															
	T_2	r_{52111}														r_{52232}

i denotes groups, i=1,...,g (=5)

j denotes treatment, j=1,...,t (=2)

k denotes measurement from among profile, k=1,...,p (=2)

ℓ denotes visit, ℓ=1,...,v (=3)

m denotes category, m=1,...,q_k; (q_1=3 and q_2=2)

$r_{ijk\ell m}$ denotes (marginal) frequency of response

TABLE 2

Summary of Three Clinical Trials

Trial Specifications	Trial I	Trial II	Trial III
g = number of clinics	5	5	4
t = number of drugs	2	2	2
v = number of visits[a]	4	3	3
p = number of outcomes	1	1	1
q = number of ordinal categories for the outcome measurement	5	5	5

[a]The number of visits is the number of follow-up visits plus an initial visit.

TABLE 3

Comparison of P-Values for Parametric, Non-Parametric, and Categorical Data Approaches to the Analysis of Three Clinical Trials With Ordered Categorical Data

		VISIT 1			VISIT 2			VISIT 3			ALL VISITS			VISIT INTERACTIONS		
		Drug	Clinic	Drug × Clinic	Drug	Clinic	Drug × Clinic	Drug	Clinic	Drug × Clinic	Drug	Clinic	Drug × Clinic	Drug × Visit	Clinic × Visit	Drug × Clinic × Visit
	Parametric	.000	.021	.878	.000	.044	.776	.000	.005	.883	.001	.001	.394	.001	.003	.145
	MANOVA	(.001)[a]	(.027)	(.999)	(.001)	(.135)	(.318)	(.001)	(.022)	(.374)						
TRIAL I	Non-Parametric	.000	.022	[b]	.000	.113		.000	.023		.000	.003		.002	.002	
2 Drugs	Multi Kruskal-Wallis	(.001)			(.000)			(.000)								
5 Clinics																
3 Follow-up Visits	Categorical GSK	.000	.004	.835	.000	.022	.811	.000	.001	.907	.000	.000	.441	.000	.000	.186
	Parametric	.000	.532	.045	.000	.952	.046				.000	.730	.065	.101	.565	.320
	MANOVA	(.001)	(.999)	(.032)	(.001)	(.999)	(.205)									
TRIAL II	Non-Parametric	.000	.666		.000	.972					.000	.730		.035	.240	
2 Drugs	Multi Kruskal-Wallis	(.000)			(.000)											
5 Clinics																
2 Follow-up Visits	Categorical GSK	.000	.640	.007	.000	.966	.012				.000	.300	.005	.141	.131	.244
	Parametric	.010	.046	.587	.007	.319	.666				.018	.157	.738	.687	.574	.795
	MANOVA															
TRIAL III	Non-Parametric	.005	.038		.005	.155					.010	.068		.690	.358	
2 Drugs	Multi Kruskal-Wallis															
4 Clinics																
2 Follow-up Visits	Categorical GSK	.007	.026	.467	.005	.305	.665				.013	.059	.693	.662	.237	.654

[a] Corresponding univariate test with missing data excluded.

[b] No test conducted--data not available or no practical test available.

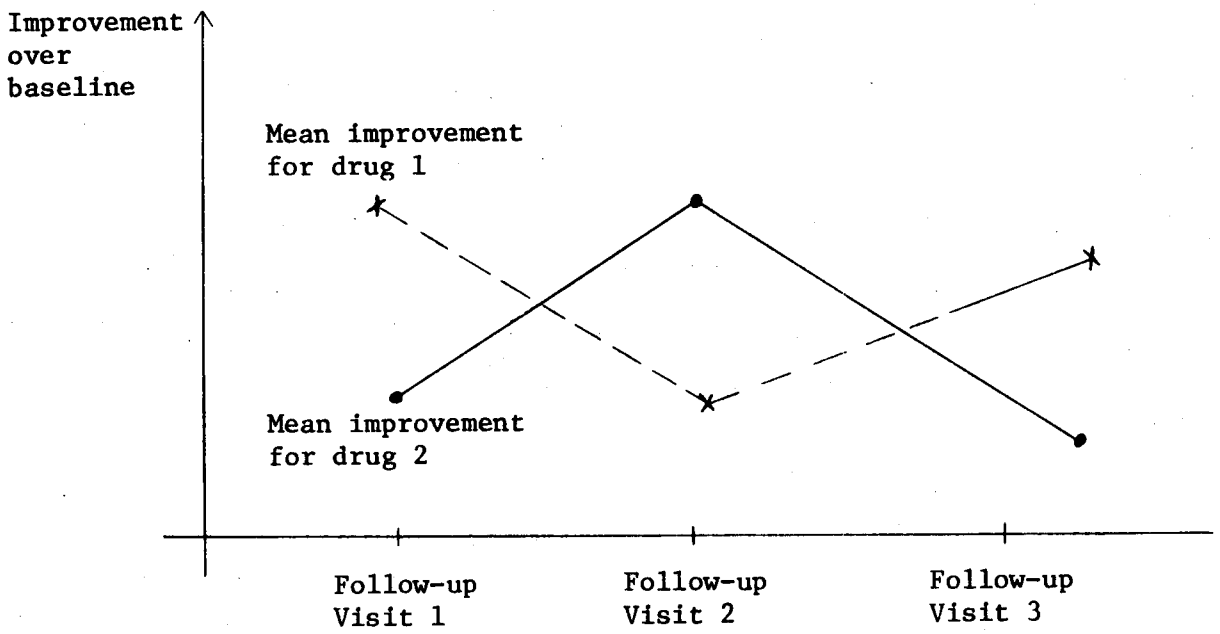

Fig. 1. Hypothetical data of mean improvements for two drugs at three follow-up visits

DATA LISTING FOR TRIAL I

		DRUG 1				DRUG 2			
		Initial[a]	Visit 1[b]	Visit 2[b]	Visit 3[b]	Initial[a]	Visit 1[b]	Visit 2[b]	Visit 3[b]
Clinic 1		3	3	0(3.0)[c]	3	3	4	3	3
		3	3	2	2	3	4	4	4
		4	3	2	2	4	4	5	4
		3	2	2	1	3	4	4	5
		3	3	2	2	3	4	4	4
		4	2	1	3	4	4	4	4
		4	1	1	1	4	4	0(4.0)	0(4.0)
		4	1	1	1	3	4	4	0(4.0)
		5	5	0(5.0)	0(5.0)	3	2	2	0(3.1)
		3	1	1	1	5	3	3	4
		4	4	4	4	3	4	4	4
		4	3	1	1	3	4	4	0(3.1)
		4	1	1	1	4	4	4	0(4.0)
		4	3	3	3	4	4	5	0(5.0)
		4	1	1	1	4	4	4	0(4.0)
		3	1	1	0(3.1)	3	4	0(3.0)	0(3.1)
		3	4	4	4	4	1	1	0(2.3)
		3	3	0(3.0)	0(3.1)	3	4	4	4
		4	0(2.8)	1	0(2.3)				
	n	19	18 (19)	16 (19)	15 (19)	18	18	16 (18)	9 (18)
	$\bar{x}$	3.63	2.44 (2.46)	1.75 (2.05)	2.00 (2.29)	3.50	3.67	3.69 (3.67)	4.00 (3.81)
	s	0.60	1.25 (1.22)	1.07 (1.27)	1.13 (1.24)	0.62	0.84	1.01 (0.97)	0.50 (0.67)

		DRUG 1				DRUG 2			
		Initial[a]	Visit 1[b]	Visit 2[b]	Visit 3[b]	Initial[a]	Visit 1[b]	Visit 2[b]	Visit 3[b]
Clinic 2		3	3	3	3	4	3	3	3
		4	2	2	2	4	4	4	4
		4	3	2	2	4	2	2	2
		4	4	0(4.0)	0(4.0)	4	4	4	0(4.0)
		4	2	2	2	4	2	2	2
		4	2	2	1	3	3	3	0(2.3)
		4	3	3	3	4	4	4	0(4.0)
		3	1	1	1	4	4	3	3
		4	3	1	1	4	5	0(5.0)	0(5.0)
		4	2	2	1	3	1	0(2.6)	1
		3	2	0(2.6)	1	3	4	2	4
		3	3	4	4	4	5	0(5.0)	0(5.0)
		5	2	2	2	5	4	5	0(5.0)
		4	2	1	1	4	4	4	3
		4	3	4	4	5	3	4	4
		4	1	1	1	4	4	3	3
		4	1	1	1				
	n	17	17	15 (17)	16 (17)	16	16	13 (16)	10 (16)
	$\bar{x}$	3.82	2.29	2.07 (2.21)	1.88 (2.00)	3.94	3.50	3.31 (3.48)	2.90 (3.39)
	s	0.53	0.85	1.03 (1.08)	1.09 (1.17)	0.57	1.10	0.95 (1.05)	0.99 (1.18)

APPENDIX (Continued)

		DRUG 1				DRUG 2			
		Initial[a]	Visit 1[b]	Visit 2[b]	Visit 3[b]	Initial[a]	Visit 1[b]	Visit 2[b]	Visit 3[b]
Clinic 3		4	0(2.4)	4	4	3	1	1	2
		4	3	2	1	4	2	2	3
		5	1	0(3.0)	1	3	2	2	3
		4	1	1	1	3	3	5	5
		3	2	1	0(2.5)	3	2	2	2
		4	2	1	1	4	3	3	3
		3	1	1	1	3	3	3	3
		4	2	2	2	5	4	3	3
		3	1	1	1	4	4	4	5
		4	3	3	4	5	4	0(4.0)	0(4.0)
		3	2	2	1	3	3	0(2.0)	5
		3	2	1	1	5	4	3	4
		4	2	1	1	3	2	3	3
		4	2	2	2				
		4	3	2	1				
		4	2	1	1				
		4	2	2	1				
	n	17	16 (17)	16 (17)	16 (17)	13	13	11 (13)	12 (13)
	$\bar{x}$	3.77	1.94 (1.97)	1.69 (1.77)	1.50 (1.56)	3.69	2.85	2.82 (2.85)	3.42 (3.46)
	s	0.56	0.68 (0.67)	0.87 (0.90)	1.03 (1.03)	0.86	0.99	1.08 (1.07)	1.08 (1.05)

		DRUG 1				DRUG 2			
		Initial[a]	Visit 1[b]	Visit 2[b]	Visit 3[b]	Initial[a]	Visit 1[b]	Visit 2[b]	Visit 3[b]
Clinic 4		3	1	1	1	3	3	3	3
		3	1	1	1	3	4	4	4
		3	2	2	1	3	1	1	1
		3	2	2	1	3	2	2	0(2.0)
		3	1	1	1	3	2	2	2
		3	3	2	1	3	4	4	0(4.0)
		3	2	2	2	3	1	1	2
		3	1	1	1	3	2	3	3
		3	3	1	1	3	4	3	3
		3	2	2	2	3	3	3	3
		3	3	2	2	4	3	3	4
		3	3	3	2	3	3	3	4
		3	1	1	1	3	3	3	3
		3	1	1	1	3	5	0(5.0)	0(5.0)
		3	3	3	3	3	2	2	1
		3	1	1	1	3	4	4	4
		3	2	2	2	3	4	3	3
		3	2	2	1				
	n	18	18	18	18	17	17	16 (17)	14 (17)
	$\bar{x}$	3.00	1.89	1.67	1.39	3.06	2.94	2.75 (2.88)	2.86 (3.00)
	s	0.00	0.83	0.69	0.61	0.24	1.14	0.93 (1.05)	1.03 (1.12)

APPENDIX (Continued)

		DRUG 1				DRUG 2			
		Initial[a]	Visit 1[b]	Visit 2[b]	Visit 3[b]	Initial[a]	Visit 1[b]	Visit 2[b]	Visit 3[b]
Clinic 5		4	2	1	1	4	2	2	1
		3	4	3	3	4	4	4	4
		5	3	0(1.8)	0(1.5)	4	4	4	4
		3	2	1	1	4	4	3	4
		4	0(2.9)	3	2	3	4	4	0(4.0)
		4	3	0(2.6)	0(2.4)	4	4	3	3
		4	2	2	2	4	2	2	2
		4	2	2	2	3	4	4	0(4.0)
		4	2	2	1	5	4	3	3
		5	2	1	1	4	4	3	3
		3	1	1	0(1.0)	4	3	3	3
		3	2	1	1	4	2	2	1
		3	3	2	2	3	4	3	3
		5	2	2	1	4	4	4	4
		5	1	1	1	3	4	4	3
		4	2	1	1	4	4	3	3
						3	4	3	3
	n	16	15 (16)	14 (16)	13 (16)	17	17	17	15 (17)
	$\bar{x}$	3.94	2.20 (2.24)	1.64 (1.71)	1.46 (1.49)	3.77	3.59	3.18	2.93 (3.06)
	s	0.77	0.78 (0.77)	0.75 (0.73)	0.66 (0.65)	0.56	0.80	0.73	0.96 (0.97)

[a]For the initial visit: 3 = stable; 4 = slowly worsening; 5 = rapidly worsening.

[b]For each of the follow-up visits:

1 = excellent; better than 75% clinical improvement over baseline

2 = good; better than 50% but less than 75% clinical improvement over baseline

3 = fair; better than 25% but less than 50% clinical improvement over baseline

4 = poor; less than 25% improvement or no improvement over baseline

5 = exacerbation; worsening of the condition.

[c]For the body of the table, the values in parentheses are estimates of the missing values, denoted by zeroes. These were estimated or specified in a manner conservative to the comparison of the drugs. For the bottom margin of the table the statistics in parentheses are obtained including the estimated missing values; the other statistics exclude the missing values.