IN BRIEF

- A description of some different types of repeated measures data
- An indication of the difficulties associated with the analysis of such data
- The use of a relevant summary measure to analyse repeated measures data
- An outline of more complex methods of analysis and a comparison of some different approaches

Further statistics in dentistry Part 7: Repeated measures

A. Petrie[1] J. S. Bulman[2] and J. F. Osborn[3]

Consider the situation in which there is a single quantitative variable of interest that is measured on each individual on several different occasions. Typically, these occasions are defined time points (usually including pre-treatment as well as various post-treatment times), so that each individual contributes a series of readings. The main objective of the study may be to compare the responses on this variable when each individual has been assigned to one of two or more treatments groups.

[1]Senior Lecturer in Statistics, Eastman Dental Institute for Oral Health Care Sciences, University College London; [2]Honorary Reader in Dental Public Health, Eastman Dental Institute for Oral Health Care Sciences, University College London; [3]Professor of Epidemiological Methods, University of Rome, La Sapienza
Correspondence to: Aviva Petrie, Senior Lecturer in Statistics, Biostatistics Unit, Eastman Dental Institute for Oral Health Care Sciences, University College London, 256 Gray's Inn Road, London WC1X 8LD
E-mail: a.petrie@eastman.ucl.ac.uk

Refereed Paper

Consider, secondly, another situation that involves the periodontal treatment of a particular condition occurring at several sites within each patient's mouth. Suppose a trial of two or more treatments is undertaken and patients are randomly allocated to the treatment groups. Although observations, for example of pocket depth, are made before and after treatment at each site, the same treatment is given to all the sites in a given patient. The response at each site is then recorded as the change in pocket depth, so that each patient has one response from each of his pockets.

These two examples are similar in the sense that each patient produces several observations, but in the first example, the observations are ordered in time whereas in the second, the observations represent different sites measured at the same time and are thus not ordered.

Difficulties in interpreting this type of data arise because there are two sources of variability, both within- and between-patients. Within a patient, the observations differ because they are taken at different times or are obtained from different pockets, while if the mean value of the responses is calculated for each patient, these means will differ between patients.

In the analysis of such data, various strategies could be adopted. These are discussed in, for example, Everitt (1995)[1] and Matthews *et al.* (1990).[2] Some methods are relatively simple, but not 100% efficient; others may be more efficient but are much more complex and may require the use of very specialised computer programs; there are others which are simple but which are invalid and thus may yield totally misleading results.

Thus three strategies are:

1. Calculate the mean value (or some other single summary measure of the response, for example a regression coefficient) for each patient. The data set is now reduced to one observation from each of the patients who have been randomly allocated treatments. The methods of analysis are the conventional two-sample *t*-test or its non-parametric equivalent if there are two treatments, or one-way analysis of variance (ANOVA) or the Kruskal-Wallis test if there are three or more treatments. This method loses information about the variability of the observations within patients.
2. Perform an analysis which takes account of both sources of variation. The first example in which the observations are taken serially in time will have a more complex analysis than the second where the observations are merely repeated observations on the same individual.
3. Ignore the variability between patients, and pretend that all the observations are independent; that is, analyse 50 observations on one patient as though they were the same as one observation on each of 50 patients. This strategy may produce a significant result but is totally *invalid*.

Repeated measures data

- Arise when each individual has more than one observation on the variable of interest
- May be analysed by replacing these multiple observations per individual by a summary measure (eg their mean)

STRATEGY 1. USE SUMMARY MEASURES

This first approach is the one which is recommended for its simplicity and the fact that it is safe in the sense that the significance tests will tend to be conservative. This implies that if a result is statistically significant using this method, it will almost certainly be significant using even the most complicated statistical techniques. Details of this method are given in this paper in the situation in which each patient receives a single treatment and the outcome variable is quantitative (although the method can be used for ordered qualitative data). All the information on each individual is reduced to just one single measure which is believed, in the context of the particular experiment, to be a useful summary of the responses. For example, it may be decided that for an individual patient, the mean of the responses at all the time points after the start of the treatment, or of all the sites within his or her mouth, is a sensible representation of the individual's overall response to treatment. The analysis is then restricted to this summary measure, and the summary measures are compared in the individuals in one treatment group with those in another group using a simple two-sample comparison, such as the two-sample *t*-test or the Mann-Whitney or Wilcoxon rank sum tests. If there are more than two treatment groups, it is possible to use the one-way ANOVA or the non-parametric Kruskal-Wallis test. Easy to understand, simple to execute and, fortunately, scientifically justifiable!

Often the choice of the summary measure is straightforward. However, on some occasions it may be difficult to know which single measure best describes the set of responses for an individual. The chosen measure must focus on the important issues and describe what is relevant in the particular investigation. In fact, if it is necessary to investigate different aspects of the response, the analysis may be repeated for two or more different summary measures.

In the context of the example of a clinical trial where observations are made at different points in time, two questions may be of interest: Are the responses on average higher (or lower) for treatment A than for treatment B? Alternatively, if the treatment were expected to change (consistently increase or decrease) the value of the response over the whole of the treatment period, it might be better to calculate the regression coefficient of the response on time, that is, the rate at which the response changes for each patient. It is important to realise that these two summary measures answer two different questions. In the first case, the analysis concentrates on the comparison of the mean level of response and totally ignores the rate at which the patients improve. In the second, the rate at which the patients improve is compared between the two treatments, totally ignoring the level of the responses. It is essential to choose the summary measure(s) *before* the study is conducted, thus ensuring that the choice cannot be influenced by the results, but this should be clear from the objectives and protocol.

In the periodontal disease example investigating change in pocket depth, there is a further complication in that the number of observations is unlikely to be the same for each patient; not all patients will have exactly the same number of gingival pockets. If the number of observations per patient varies very widely it may be more efficient to use a **weighted analysis** of the summary measure in order to take more account of the mean for a patient who has say 30 pockets than for a patient who has just one pocket. However, because the variability of the responses within patients will almost certainly be very much less than the variability of the means of the patients, there may not be a very great gain in efficiency as compared with the unweighted analysis.

Typical summary measures

The same summary measure is determined for every individual in the study. There are various summary measures that can be used. Some of the more usual ones in the context of the first time-related example are:

- The overall post-treatment mean of the responses for an individual.
- The difference between the initial and final responses. (This would correspond to the patient's mean change in pocket depth in the periodontal disease example)
- The percentage change between the initial and final responses.
- The maximum (or minimum) response.
- The time to reach the maximum (or minimum) value.
- The time to reach a particular value (eg some fixed percentage of the baseline value).
- The estimated slope of the linear regression line (provided a straight line relationship is appropriate). In this situation, it may be sensible to make adjustments in the analysis for the fact that some slope estimates may be more precise (ie have smaller standard errors) than others.
- The area under the curve.

Refinements on the basic procedure

The summary measure for the first time-related example may comprise one or more of the post-treatment responses, ignoring the pre-treatment value(s). However, the pre-treatment reading for each individual (or, if there is more than one pre-treatment reading for an individual, the mean of these pre-treatment readings) could be incorporated into the summary measure. For example, the summary measure might be the difference between the mean of the post-treatment responses and the pre-treatment value (ie the reading or the mean of the readings, as appropriate). Alternatively, instead of incorporating the pre-treatment value(s) into the measure itself, the power of the comparisons can be increased by using the **analysis of**

covariance, with the pre-treatment value as the covariate, instead of the simple comparison of groups which makes no adjustment for the pre-treatment value.

Plotting the data

It is always helpful, if possible, to plot the data to give a broad indication of what is happening. In the first example, the raw data could be plotted against time. All the information could be retained in one diagram, so that every response for all individuals is shown. However, if there are a large number of individuals, this can produce a confusing diagram which fails to achieve its aim of demonstrating trends and relationships. In such instances, a separate diagram can be produced for each treatment group (Fig. 1a and Fig. 1b). Alternatively, separate graphs of the responses against time for each individual can be drawn; each should be drawn on the same scale, perhaps grouping the graphs for each treatment in a grid. Often there are so many graphs that it becomes unwieldy to include them all in a paper. Then, just representative examples which are believed to illustrate particular types of response structures may be included. The subjectivity of this approach may be open to criticism, so the choices should be justified.

It is tempting to produce average curves for each treatment group by plotting the mean (with confidence interval or standard error bars) of the responses at each time point against time. Although this is a simple approach to incorporate into one diagram the information from many individuals, there is always the danger that the mean curve for a treatment group is not representative of any single individual's curve. Sometimes, depending on the quantity of the data, it is possible to show both the mean curve(s) and those for the individuals in a single graph, distinguishing the two by, for example, using a solid black line for the mean curve, and different pastel colours for the individuals.

If summary measures are used to analyse the data, it is often helpful to plot them. The distribution of the chosen summary measure should certainly be investigated by producing, separately for each treatment group, a diagram, such as a histogram or box-plot, of its values. The distribution will influence the choice of statistical test to compare values in the different treatment groups or may suggest that the data should be transformed. If the summary measures are not Normally distributed with approximately constant variance, a non-parametric method, such as the Wilcoxon rank sum test may be appropriate to compare these measures in the two treatment groups or the Kruskal-Wallis one-way ANOVA when there are more than two groups. Alternatively, a parametric analysis, such as the two-sample *t*-test or the one-way ANOVA, would be preferable for Normally distributed measures (or their transform). Sometimes, if more than one summary measure is investigated, it is useful to plot one summary measure against another. For example, the two summary measures might be minimum response and the time to reach that minimum response. The resulting scatter plot may provide insights which would not be available from the raw data alone.

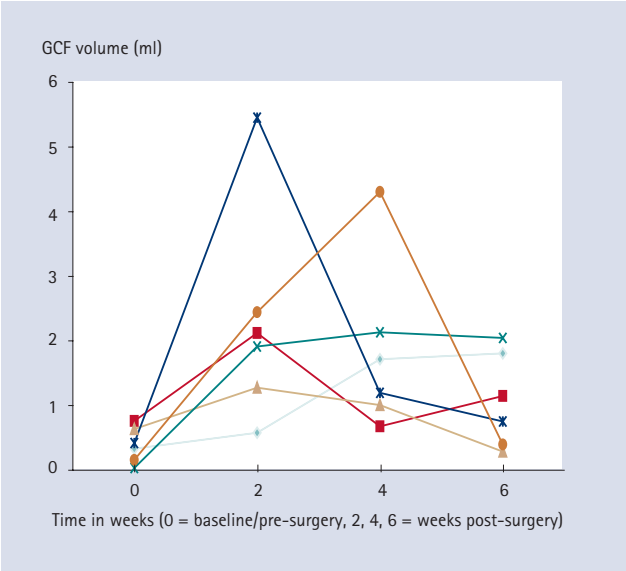

Fig.1a Mean GCF volume per site collected at four time points in each of six patients undergoing GTR surgery

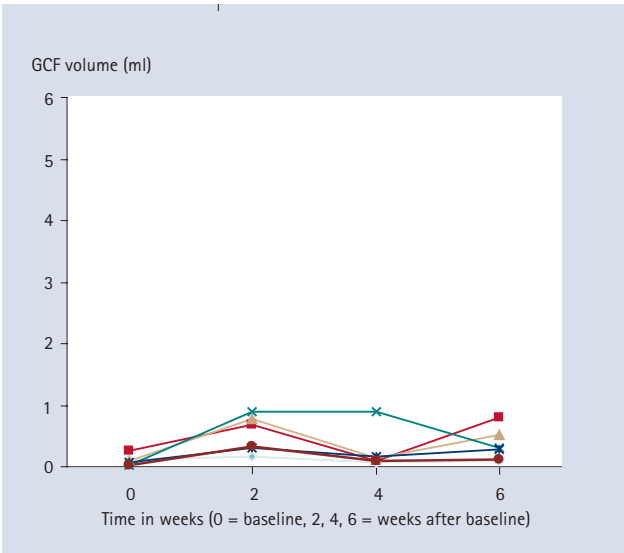

Fig.1(b) Mean GCF volume per site collected at four time points in each of six control patients

Example

As an example, consider a small study (based on data kindly provided by Dr Gareth Griffiths and Dr Leyla Kuru of the Eastman Dental Institute for Oral Health Care Sciences, University College London) which was concerned with investigating the cellular activities involved in periodontal wound healing. Gingival crevicular fluid (GCF) accumulates in the gingival crevice and is considered to reflect ongoing cellular activities in the surrounding periodontal tissue. For each patient undergoing guided tissue regeneration (GTR) surgery, the volume of GCF was collected from 3–4 sites adjacent to the intrabony defects on molar teeth. The volume of GCF was also collected from 2–4 randomly chosen healthy molar sites in each patient in a control group not undergoing surgery. Samples were collected at baseline (pre-surgery) and then at 2, 4 and 6

weeks after baseline. The mean volume of GCF per site was determined from the total volume collected from each patient at each time point. The preliminary results from 12 patients are shown in Fig. 1a and Fig. 1b. The maximum of the mean GCF volume for each patient at a given time point in the 6-week period was used as a summary measure. These maxima were 1.81, 1.28, 2.14, 5.46, 4.31 and 2.12 µl in the surgery group, and 0.16, 0.81, 0.79, 0.91, 0.31 and 0.32 µl in the control group. The median of these maxima was 2.13 µl (range 1.28 to 5.46 µl) in the surgery group and 0.56 µl (range 0.16 to 0.91 µl) in the control group: the difference in medians was 1.57 µl. Since the sample size was very small (there were only six patients in each group) and it was therefore difficult to establish Normality, the non-parametric Wilcoxon rank sum test was used to compare the two groups of maxima; this gave $P = 0.002$, indicating that, on average, the maximum volume of GCF produced was greater in the surgical group.

STRATEGY 2. MORE COMPLEX ANALYSES TAKING ACCOUNT OF WITHIN- AND BETWEEN-PATIENT VARIABILITY

The data can be analysed using more complex **analysis of variance** (ANOVA) techniques. The analysis of variance covers a wide range of experimental designs. It is important to ensure that all the appropriate considerations relating to the analysis of variance model are taken into account in any particular design. In this situation, **repeated measures ANOVA,** available in the more sophisticated statistical software packages, can be used.

Sources of variation

The variability of the observations *within* an individual is usually less than that *between* different individuals and both are sources of variation in repeated measures data

The use of repeated measures analysis of variance relies on being able to specify the model appropriately and understand its underlying assumptions, both of which present some difficulty to the novice statistician. Should fixed or random effects be specified in the model? Are the data Normally distributed? Is there sphericity or circularity of the covariance matrix? How should Mauchly's test be interpreted? What is the correction factor correcting? Should a multivariate approach be used? Do we need orthogonal polynomials to transform the variable. Are we able to interpret the ANOVA tables in the output? These are some of the questions to be answered if the repeated measures ANOVA approach to analysing this data is to be used. Clearly, it is not for everyone!

An added problem is that the design for a repeated measures ANOVA should be balanced. In the first time-related example, this requires that each individual should have the same number of measurements at equal time intervals. This may be reasonable in a well controlled experiment; a completely balanced design can be specified in which the same number of responses are measured at particular pre-determined time points for all individuals, recognising that there may be just a few missing observations because of factors which are beyond control (a transport strike, a bereavement in the family, etc). However, in some longitudinal studies, this may not be possible, and the data may be a series of readings on individuals at essentially random points in time. A repeated measures ANOVA on these data would have to ensure that there are complete data at specified time points, requiring many missing values at these times to be estimated and some existing values at other times to be ignored. This is an extremely inefficient approach to the analysis and is not to be recommended.

Even more **complex models** can be used, which take into account the fact that the time points at which measurements are made for different individuals may vary, and which allow for missing observations. For example, a suitable regression-type model might be specified and maximum likelihood used, instead of ordinary least squares, to estimate its parameters and standard errors. This approach is not simple, and it relies on the assumptions underlying the model being satisfied, and is best left to the experienced statistician.

Similarly, the *multilevel modelling* approach (Leyland and Goldstein, 2001)[3] is appropriate but requires considerable expertise and dedicated software (eg MLwiN: information on www.ioe.ac.uk/MLwiN) to execute. In multilevel modelling, the hierarchical (or nested) structure of the data is taken into account. When there are repeated measures, such as in the examples quoted, the hierarchy consists of times or sites within individuals. A regression model is specified with two 'levels', the lower level representing times/ sites and the upper level representing individuals. A covariate is included in the model to represent the treatment effect and additional covariates can also be included to represent the effects of other factors such as gender and age. The multilevel model incorporates random residuals which vary between the units at each level, and uses generalised least squares (rather than ordinary least squares as in straightforward regression) to estimate the parameters of interest.

STRATEGY 3. HOW NOT TO PROCEED

One method that should not be attempted in the time-related example is to compare the groups at each time point, using, for two groups, a two-sample *t*-test or a non-parametric equivalent, such as the Wilcoxon rank sum test. There are a number of reasons why this is inadvisable:

- The within-patient changes over time are ignored.
- The successive tests are not independent.
- The whole process may involve many significance tests, thereby increasing the probability of the Type I error.
- It may be difficult to come to some overall conclusion about the difference between groups, and impossible to obtain a single estimate of this difference.

Table 1 Results of three methods of analysis of a trial of the effect of three mouth washing solutions on gingival pockets (Abstracted from Osborn, 1987)[5]

	Metronidazole	Quinine sulphate	Saline
Number of patients	9	5	5
Total number of pockets per treatment group	263	145	139
Number of pockets per patient (range)	21 to 49	24 to 41	19 to 40
Mean reduction in pocket depth per patient (mm)	1.0288	0.7235	0.5905
SE(mean) (mm) simple summary measure method	0.2066	0.2772	0.2772
SE(mean) (mm) weighted summary measure method	0.2061	0.2762	0.2774
SE(mean) (mm) ignoring patients (invalid method)	0.0705	0.0949	0.0970

A second, even worse method is to classify all the observations only according to the treatments and pretend that the observations in each treatment group are independent; that is to take no account of the patients and the variability within patients. This may well lead to an apparent significant difference between the treatments but, because it ignores an important source of variability, will lead to a totally invalid result.

DISCUSSION

This paper has reviewed some methods of analysis to be used when there are repeated observations on the same study unit; in the case of the examples described, the study unit is the individual patient. However the problem is much more general and data are often analysed without taking account of the structure of the raw data. A very common error is to design a multi-centre study, perhaps because any one centre cannot generate sufficient patients, but to do the analysis pretending that the results are from a single centre, thus ignoring the big differences that almost always exist between centres. The error can be even worse in a large international multi-centre study, in which say, Great Britain is represented by a single clinic in North Wales. It is exceedingly important that multi-centre studies are analysed as multi-centred and not as though the data are generated by a random sample of patients from, say, Europe.

This problem has been recognised in periodontal research for some time and earlier papers on the subject are Blomqvist (1985)[4] and Osborn (1987).[5] Both papers investigate the multiple site per patient example. Blomqvist suggests that the simple summary method be used, the summary measure being the mean change in pocket depth for each patient. On the other hand, Osborn compares Blomqvist's method with a weighted summary measure method (that is, taking account of the number of observations on each patient and the variability of the observations within each patient), and also the invalid method which assumes that all the observations in each treatment group are independent. The data were from a small trial to compare three mouth washing solutions (containing metronidazole, quinine sulphate or saline solution) for the treatment of gingival pockets. Some of the results of the comparison of the methods of analysis are shown in Table 1. It can be seen that the weighted analysis, which is much more complicated than the simple summary method, produces practically identical values of the standard errors of the estimated treatment effects. There are clearly no statistically significant differences between the treatments if either of the two valid methods of analysis is done. In contrast, the differences are highly significant according to the mistaken analysis. The analysis of a repeated measures study must be based on the number of subjects, not only on the total number of observations.

1 Everitt B S. The analysis of repeated measures: a practical review with examples The Statistician 1995; **44:** 113-135.
2 Matthews J N S, Altman D G, Campbell M J, Royston P. Analysis of serial measurements in medical research. *Br Med J* 1990; **300:** 230-235.
3 Leyland A H, Goldstein H. (eds) *Multilevel Modelling of Health Statistics.* Chichester: John Wiley and Sons, 2001.
4 Blomqvist N. On the choice of computational unit in statistical analysis. *J Clin Periodontol* 1985; **12:** 873-876.
5 Osborn J. The choice of computational unit in the statistical analysis of unbalanced clinical trials. *J Clin Periodontol* 1987; **14:** 519-523.