Reprinted from the **BRITISH MEDICAL JOURNAL,** *27th January 1990, 300, 230-235*

Analysis of serial measurements in medical research

J N S Matthews, Douglas G Altman, M J Campbell, Patrick Royston

Abstract

In medical research data are often collected serially on subjects. The statistical analysis of such data is often inadequate in two ways: it may fail to settle clinically relevant questions and it may be statistically invalid. A commonly used method which compares groups at a series of time points, possibly with *t* tests, is flawed on both counts. There may, however, be a remedy, which takes the form of a two stage method that uses summary measures. In the first stage a suitable summary of the response in an individual, such as a rate of change or an area under a curve, is identified and calculated for each subject. In the second stage these summary measures are analysed by simple statistical techniques as though they were raw data. The method is statistically valid and likely to be more relevant to the study questions. If this method is borne in mind when the experiment is being planned it should promote studies with enough subjects and sufficient observations at critical times to enable useful conclusions to be drawn.

Use of summary measures to analyse serial measurements, though not new, is potentially a useful and simple tool in medical research.

Division of Medical Statistics, University of Newcastle upon Tyne, The Medical School, Newcastle upon Tyne NE2 4HH
J N S Matthews, PHD, *lecturer*

Medical Statistics Laboratory, ICRF, Lincoln's Inn Fields, London WC2A 3PX
Douglas G Altman, BSC, *director*

Medical Statistics and Computing, University of Southampton, Southampton General Hospital, Southampton SO9 4XY
M J Campbell, PHD, *senior lecturer*

Department of Medical Physics, Royal Postgraduate Medical School, London W12 0NN
Patrick Royston, MSC, *senior lecturer*

Correspondence to:
Dr Matthews.

Br Med J 1990;**300**:230-5

Introduction

A common study design in medical research is to give patients some intervention and then observe what happens to them over time. For example, blood glucose concentrations may be measured several times after a glucose drink. In many cases there may be more than one group of patients, possibly randomised to different treatments. Despite its apparent simplicity the analysis of this form of study presents statistical problems which, judged from published work, are not widely appreciated. The purpose of this paper is to propose a general simple method for a clinically useful and statistically valid analysis. We consider only studies in which each patient receives a single treatment or intervention, so excluding escalating dose studies or crossover trials which require more complicated analysis. Though we also restrict attention to outcome variables that are quantitative because these occur most commonly, the methods can also be applied to ordered data such as pain scores.

Types of time dependency

It is helpful to distinguish two main ways in which the outcome variable may change with time.

Peaked—In many studies the outcome variable starts from a baseline (sometimes zero), rises to a peak, and then returns to baseline. This is displayed as a peaked curve (fig 1). For example, in a study of post prandial energy expenditure during pregnancy the metabolic rate was measured in women after a 12 hour fast and then at 30 minute intervals for two hours in response to a test meal.[1] This was done during pregnancy and again once lactation had stopped. The metabolic rate rose to a peak after about 60 minutes and then fell steadily. Women were found to have a reduced energy expenditure during pregnancy. In that study the interest lay in both the total response and the time to reach the maximum value.

Growth—Sometimes the outcome variable steadily increases or decreases with time and does not start to return to its initial value over the period of study. This is displayed as a growth curve (fig 1). A recent study investigated the role of peripheral vascular tone in

hypotension induced by dialysis.[2] Each patient had sessions of dialysis with acetate fluid and with bicarbonate fluid. Blood pressure was measured every 15 minutes during the four hours of dialysis. The changes in the variables were shown to be roughly linear with time. Other examples of this type of time dependency might be irreversible changes in lung function in people exposed to toxic fumes, or growth of children in different social settings. In this type of study the important feature is the rate at which the variable changes.

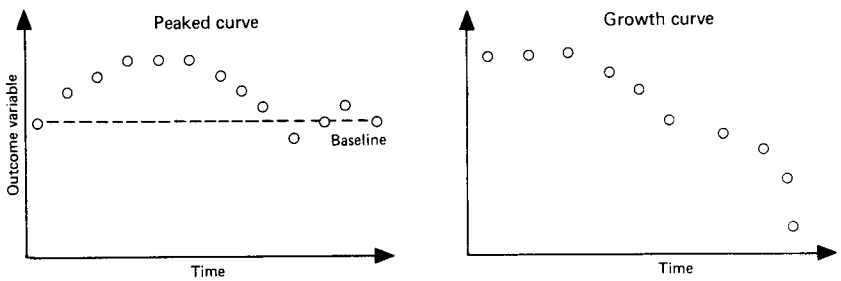

FIG 1—*Examples of peaked curve and growth curve*

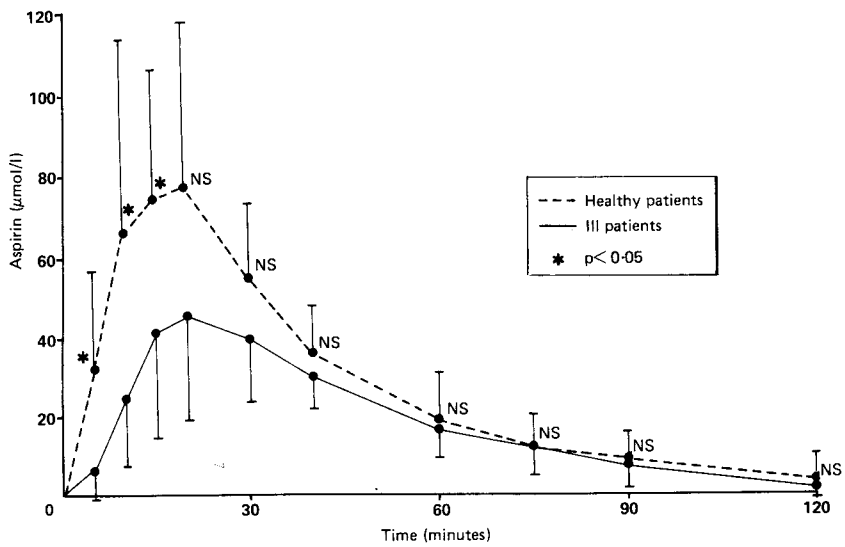

FIG 2—*Mean and standard deviation of aspirin concentrations in nine healthy and nine ill patients over time. ("Usual" method of display)*

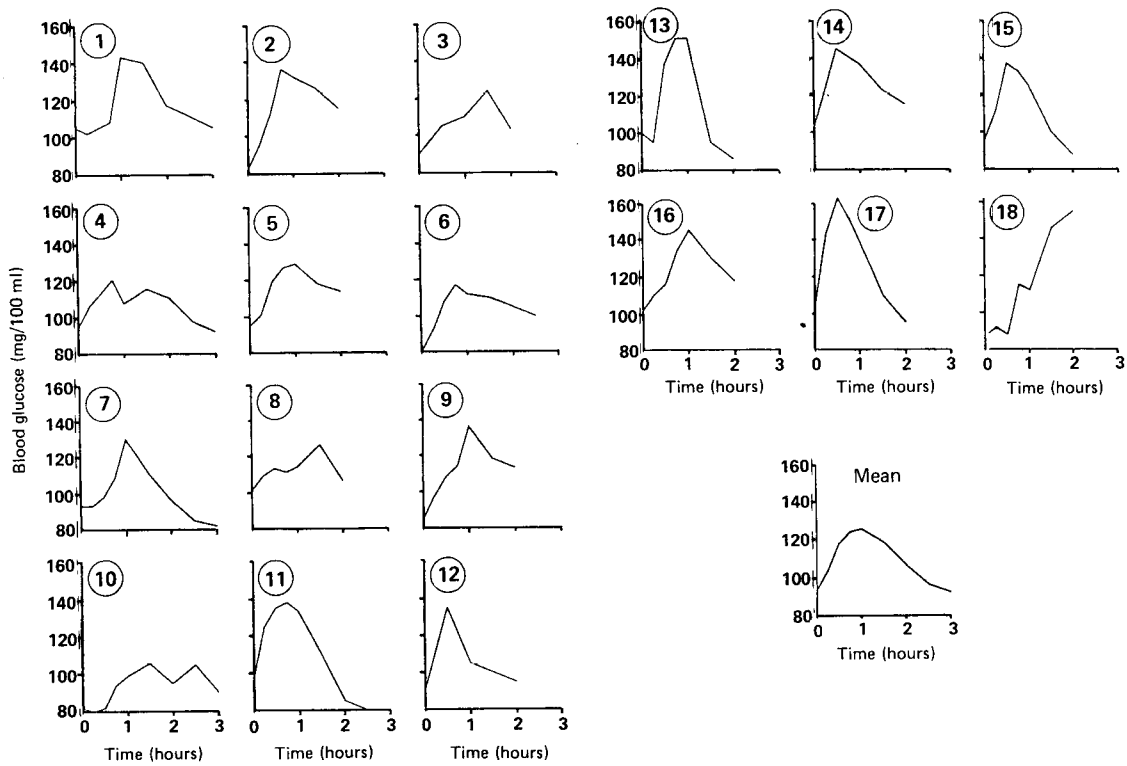

FIG 3—*Individual plots of data given by Fikri and Ghalioungui[3] of blood glucose concentrations against time in 18 patients with ancylostoma anaemia*

Usual method of analysis

When two groups are being compared a common but inappropriate analysis is to apply separate two sample tests at each time point—for example, the *t* test or Mann-Whitney U test. To illustrate this approach we compare aspirin absorption in patients who are either healthy or ill. Each patient was given the same dose of aspirin per kg body weight and had his or her blood aspirin concentration measured at time zero and after 5, 10, 15, 20, 30, 40, 60, 75, 90, and 120 minutes. Analyses of mean concentrations at each time point would result in the graph shown in fig 2, which is typical of graphs in many published papers. The important features of the analysis shown in fig 2 are: (*a*) the lines joining the means at each time point are drawn for each group; (*b*) "error" bars (often undefined) are attached at each time point; (*c*) an indicator of statistical significance is placed by each time point to summarise the results of the separate significance tests. What is wrong with this approach? There are several criticisms that can be made.

(1) The curve joining the means may not be a good descriptor of a typical curve for an individual—Important variation in the shapes and locations of curves for different subjects may be hidden. We illustrate this by some data published over 50 years ago.[3]

Figure 3 shows individual glucose tolerance curves for 18 subjects with ancylostoma anaemia. From the mean curve (also shown) we might deduce that for a typical patient the venous blood sugar concentration rises to a maximum at about one hour after drinking a glucose solution and then falls back to normal (around 90 mg/100 ml (5·0 mmol/l)) after some three hours. From the individual curves, however, it is evident that this summary hides a wide variety of curves, including multiple peaks (cases 4 and 10) and a steady rise (case 18). The maximum rise above the fasting value (time zero) occurs at times ranging from 30 minutes to two hours after the glucose drink. The summary of the results was that 12 out of 18 patients with ancylostoma anaemia showed abnormal glucose tolerance, information which could not be gleaned from the mean curve.

(2) No account in the analysis is taken of the fact that measurements at different time points are from the same subjects—It is inherent in the design that the main

interest lies in the way individual subjects respond over time, yet this is ignored when each time point is analysed separately.

(3) Successive observations on a given subject are likely to be correlated—The value at one time point is likely to influence successive time points, so that the significance tests will not be independent. If a test at one point in time gives a significant result, then it is likely that tests performed at points close in time will also give significant results.

The usual approach to the analysis of serial data may thus be criticised on both clinical and statistical grounds. Points (1) and (2) together indicate that the usual analysis may give a misleading impression of the way that individual subjects typically respond over time and give no information about variation among subjects in their response over time. The error bars often shown (fig 2) relate only to between subject variation at each time point. From point (3) it is evident that, in addition, there are convincing statistical arguments against multiple tests. Separate significance tests at different time points are often interpreted as if they gave independent information about the relative location of the groups. That this is untrue may be illustrated by considering what would happen if the blood aspirin concentrations had been measured every minute. The number of significant p values would increase enormously, even though the increase in the amount of information about the separation of the groups would be minimal. The tests are clearly not independent and so their interpretation is much more difficult.

Further, dividing the results into "significant" and "not significant" introduces an artificial dichotomy into serial data. Most biological variables change over time in a smooth and continuous manner, and so the idea that at one point in time the difference between two variables is not significant whereas at the next point in time it is significant is artificial. Thus separate significance tests are not a sensible way to assess the difference between sets of repeated measurements.

Recommended method of analysis: use of summary measures

One method that is clinically relevant, simple to use, and widely applicable is what we refer to as the method of summary measures. Though not new (it has been described several times since 1938[4-9]), it does not seem to be widely known among medical research workers. The method considers the individual as the basic unit and uses the responses for each individual subject to construct a single number which summarises some aspect of that subject's response curve. This approach avoids all the difficulties outlined above. Appendix I lists examples of some summary measures.

When the outcome measure is the concentration of a substance that has been given—for example, a drug or glucose—the response will often be peaked. The total uptake of the substance may be of interest and can be measured by the area under the response curve in an individual subject. Appendix II gives a method for calculating the area under the curve and shows that it may be interpreted as a type of weighted average of the responses. Thus the area under the response curve may be useful even in cases when a direct interpretation such as the amount of substance absorbed is not possible. An alternative measure of the overall value is the simple mean of the observations. If the intervals between successive observations are the same this will be closely related to the area under the curve. Another feature of peaked data that is frequently of interest is the maximum value, often denoted by C_{max}, which may be interpreted as a measure related to the maximum effect of the agent given. The time taken to reach this maximum may also be a clinically important variable. There may be instances where it is the minimum, not the maximum, level that is relevant.

For data in the "growth" category the rate at which the variable is changing over either the whole or part of the experiment is an important feature. A good measure of this rate will often be the slope of a line fitted to the data, which is most easily measured by the regression coefficient estimated by least squares. In some circumstances—for example, after giving different drugs to different groups of subjects—the final outcome, possibly expressed as a difference from baseline, may usefully represent some "achievable" value. This may be more appropriate for data that tend to level off over the last few values—for example, diastolic blood pressure in the treatment of hypertension. Once the appropriate summary measure has been calculated for each subject its values can be treated as raw data for an appropriate statistical analysis. For example, if there are two groups, then means or medians of the summary measures could be compared.

More than one summary measure can be constructed so that different aspects of the response may be investigated. It is not easy to give any general rule about how many measures can be constructed, though clearly each measure should aim at summarising a different aspect of the response. There are seldom more than two or three interesting aspects of a response curve, and so two or at most three different measures should be enough for a complete analysis of the data. As the principle is to reduce a large number of dependent observations to a smaller number of summaries, the method would be vitiated if too many summaries were used.

The summary measures should have some clear clinical or biological relevance and ideally they should be chosen before the data are collected. This avoids any temptation to choose a particular summary measure because it shows a maximal difference between groups. Indeed, an advantage of the method is that by thinking in terms of relevant summary measures the researcher is compelled to decide on specific questions that the data are required to answer. By thinking in these terms in advance it may be possible to improve the design of a study. For example, if the important summary measure is the time to maximum response, then to get a precise estimate frequent measurements would be needed around the time that the maximum response is expected. It would be dogmatic, however, to insist that a summary measure could not be chosen after the data have been examined, especially when the shape of the time-response curve is uncertain.

There are few computer packages that allow computation of summary measures in a routine manner. Nevertheless, studies of this nature are often fairly small, and calculating summary measures by hand or using an ad hoc computer program is quite feasible.

Graphic display

Graphs are an effective way to display data and illustrate conclusions. Nevertheless, it is more difficult to create informative and truthful graphs of repeated measures than might be supposed. The most informative approach and one which is strongly recommended when the data are first analysed is to produce separate graphs of the responses against time for each subject. In order to aid visual comparisons the same axis scaling should be used on all graphs. The plots may be arranged into a panel or grid with separate panels for each group. It may help to order the plots in some way, such as by increasing mean or maximum value. Depending on the sample size it may be feasible to include plots of the raw data in a paper—for example,

as was done in the blood glucose example—and we recommend this. If raw data cannot be shown it may be possible to classify the curves and plot representative examples. How the representative examples were chosen (for example, a one in five random sample) should be described in any publication. Fikri and Ghalioungui considered that there were seven classes of curve in their data (fig 3).[3]

The question whether the mean of a given set of responses represents a "typical" subject's curve is not simple to answer. The setting in which the mean curve is most likely to be useful occurs when all the peak responses occur at the same time—for example, when subjects all respond quickly to a stimulus.

Summary measures have considerable advantages for plotting because the usual graphical methods, such as histograms, scatter plots, and so on, may be applied to them. Insight may be gained from a scatter plot of any two summary measures. In particular, a useful exercise is to plot the maximum (or minimum) value for each subject against the time that the maximum (or minimum) occurred. This enables large quantities of data to be plotted succinctly and may disclose a relation between the two variables that could not be discerned in the plot of the raw data.

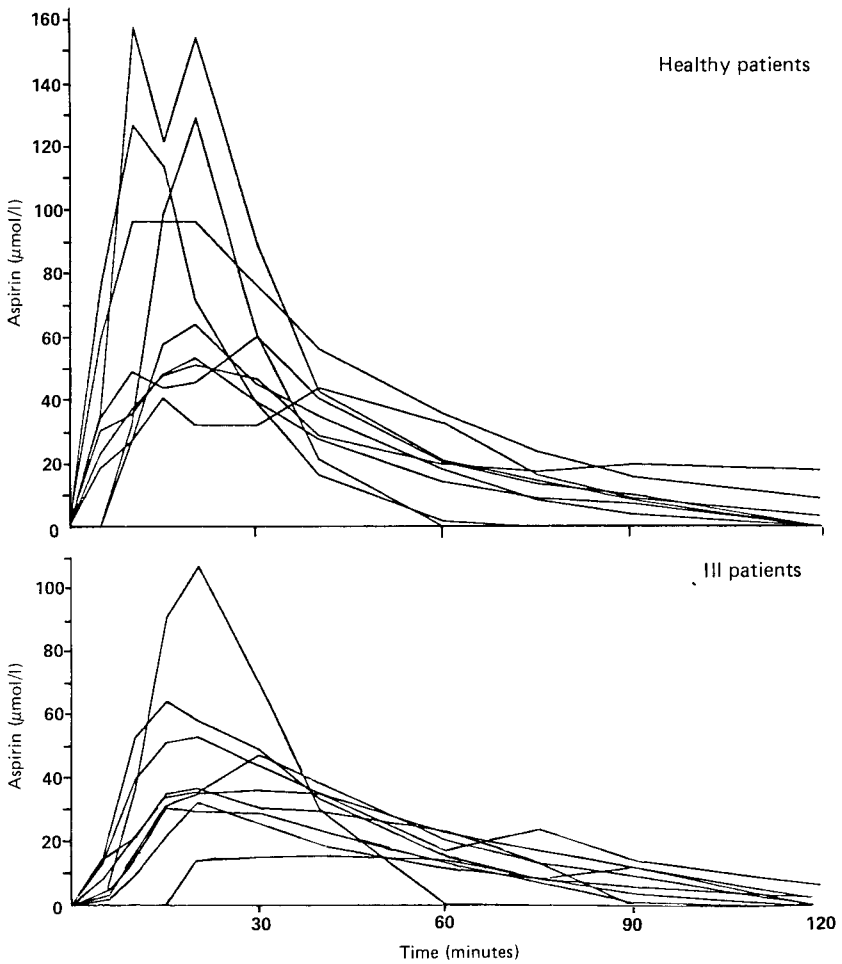

FIG 4—*Individual plots of aspirin concentrations against time in healthy patients and ill patients*

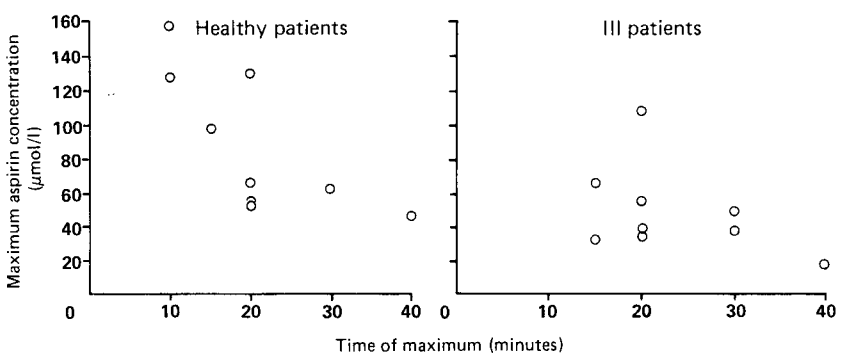

FIG 5—*Scatter plot of maximum aspirin concentrations by time of maximum in healthy patients and ill patients*

Examples

We illustrate the points discussed above with two examples relating respectively to peaked data and to growth data.

PEAKED DATA

We first reconsider the study of aspirin absorption. Figure 4 shows the individual responses in the healthy and ill patients, which clearly belong to the "peaked" category. It is clear that the mean curves in fig 2 hide considerable variability. The basic question posed by the researcher was, "Do ill patients have reduced absorption of aspirin?" We could answer this by using two of the summary measures given in appendix I. We could calculate the area under the curve for each patient and we could also look at the maximum value. Both these measures are meaningful and are familiar to pharmacologists. Table I gives statistics of these two summary measures for comparing the two groups. The distributions of both summary measures (not shown) are skewed, indicating that the data should be transformed or analysed by a non-parametric method.[10]

TABLE I—*Analysis of data from aspirin study*

	Area under curve	Maximum concentration
Healthy patients (n = 9)		
Arithmetic mean (SD) (μmol/l)	26·5 (8·8)	86·0 (41·5)
Geometric mean (μmol/l)	25·4	77·8
Ill patients (n = 9)		
Arithmetic mean (SD) (μmol/l)	17·5 (5·0)	46·7 (26·3)
Geometric mean (μmol/l)	16·8	41·2
Ratio of geometric means	1·52	1·89
95% Confidence interval	1·11 to 2·08	1·14 to 3·13
t Test	2·83 (df = 16)	2·66 (df = 16)
p Value	0·01	0·02

Also the standard deviations increase with the mean concentration. This suggests that the data should be transformed by a logarithmic transformation.[10] Table I gives the results of an analysis where the summary measures have been log transformed and where the difference between the groups is expressed as the ratio of the geometric means together with the 95% confidence interval.[11] There is strong evidence that there is both a greater peak value and a higher overall level for blood aspirin concentrations in healthy subjects than in ill subjects.

Figure 5 gives the plot of peak aspirin concentrations in the healthy and ill patients by the time the maximum occurs as derived from fig 4. This shows clearly that peak values tend to be lower and occur later in the ill patients, which can also be deduced less easily from fig 4. Nevertheless, there also seems to be a negative relation between the size of the peak and its time of occurrence—that is, the higher peaks occur earlier, particularly in healthy patients—which is not easily seen in fig 4.

GROWTH DATA

A second example describes changes in tumour volume in three groups of 10 rats at 11 time points, after different injections. The injections were tissue culture medium (group 1), tissue culture medium and normal spleen cells (group 2), and normal spleen cells, immune RNA, and tumour antigen (group 3).[12] Figure 6 plots the results (original data), which clearly belong to the "growth" category. Plainly the tumours are growing in almost all animals but at a variable rate,

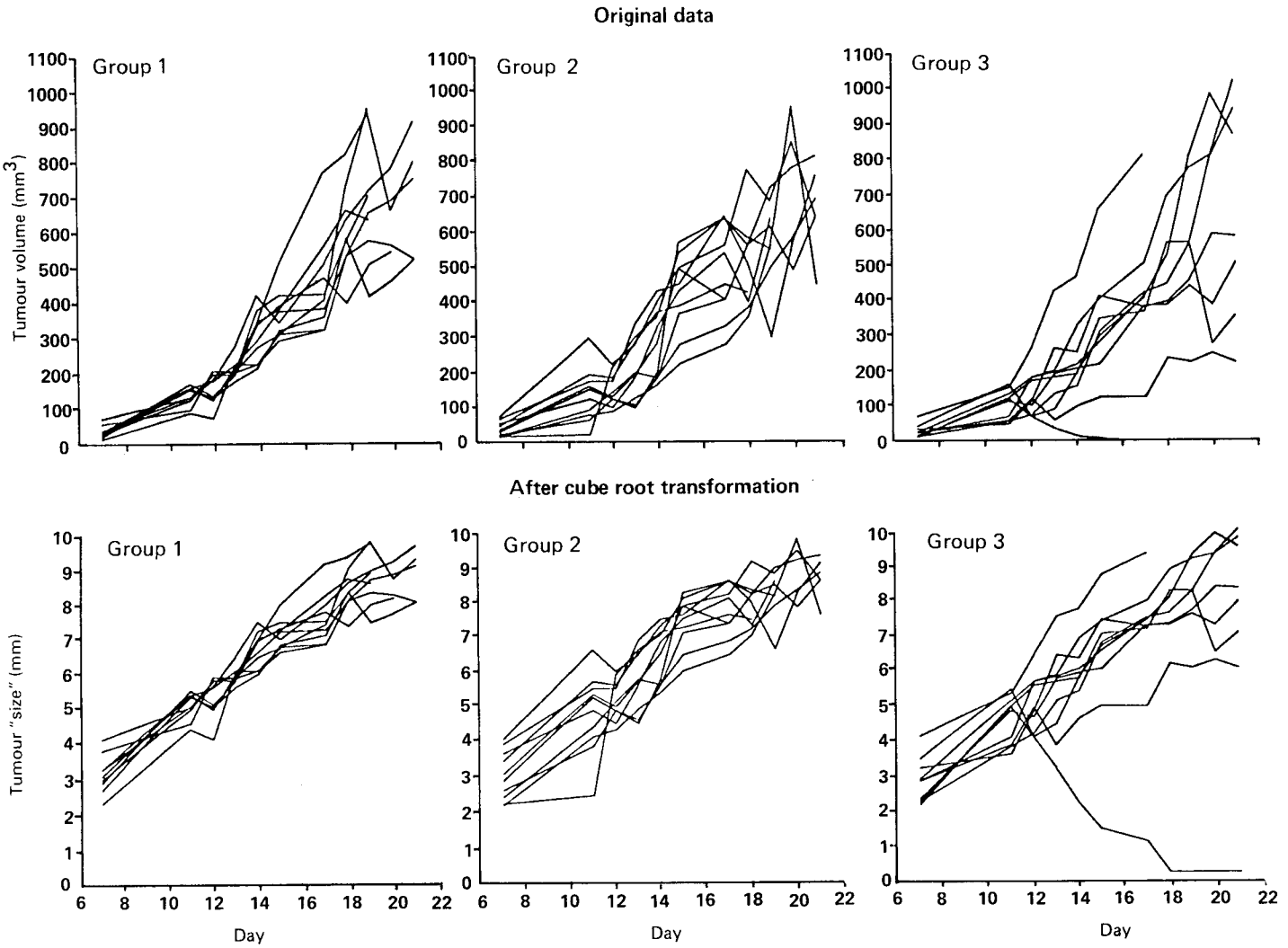

FIG 6—*Individual plots of tumour volume against time in three groups of 10 rats expressed as original data and cube root transformed data*

particularly in group 3. Figure 6 also shows the data after a cube root transformation, which converts a volume (mm^3) into a linear measure which might be termed tumour size in mm. The growth rates of the transformed data seem to be more linear than before; the one disparate animal in group 3 (with apparent tumour regression) stands out with clarity. The transformed plots suggest that the tumour growth rate for each animal, determined, for example, by linear regression of tumour size on day, summarises the interesting information in the data. This illustrates the general point that it may be necessary to transform the data on to a different scale before a simple summary analysis is possible. Table II shows the transformation of the data on two rats together with the slope obtained from linear regression. A one way analysis of variance (excluding the outlier) shows that there is no significant difference in the rate of growth of tumours in the three groups (mean tumour growth rates = 0·438, 0·438, and 0·435 mm/day for groups 1, 2, and 3 respectively; $F_{2,26} \approx 0{\cdot}0$, $p \approx 1{\cdot}0$).

TABLE II—*Full data on two rats in tumour growth study*[13] *with transformed values and corresponding summary measures*

	Tumour volume			
	Rat 1		Rat 2	
Day	mm^3	Cube root mm	mm^3	Cube root mm
7	55·0	3·803	70·0	4·121
11	95·0	4·563	129·7	5·062
12	205·9	5·905	196·0	5·809
13	205·9	5·905	205·8	5·904
14	270·0	6·463	375·7	7·216
15	307·3	6·748	419·1	7·484
17	405·1	7·398	421·2	7·496
18	726·0	8·988	573·4	8·308
19	950·4	9·832	701·8	8·887
20	661·5	8·713	–	–
21	798·6	9·278	–	–

Summary measures (slope of cube root of tumour volume on day): rat 1, 0·444 mm/day; rat 2, 0·404 mm/day.

Discussion

It is common in medical research for methods of statistical analysis to become standard for particular types of data. Their use becomes widely accepted and little thought is given to whether they are truly appropriate to the clinical question being posed. We have shown that there are serious problems associated with the common use of comparisons at each time point when analysing serial measurements on patients. In particular, the method is inappropriate because it does not provide clear answers to clinically relevant questions. In addition, there are important statistical deficiencies.

We have described the method of summary measures, which avoids all of the problems identified with the more usual analysis. Its main disadvantage is that it may be difficult to specify in advance an appropriate summary measure. The use of this type of analysis should encourage researchers to think about the features of the data that will be of most interest to them when designing the study. Posing the question simply as "How do the groups differ" is too vague to indicate the correct statistical analysis.

There are other methods of analysis of serial data, such as repeated measures analysis of variance,[10] multivariate analysis of variance,[11] Kenward's method,[14] and hierarchical models.[15] Methods such as the split plot analysis of variance which ignore the correlations within subjects are generally not valid.

None of these methods provides results that are as easy to understand as the method of summary measures, nor are they as easy to use. If there are missing values, as is common with this type of study, the method of summary measures will usually still allow the summary to be calculated, but other methods may be difficult to apply. If the times of observations differ among subjects, then all the methods, other than summary measures, will fail.

There are problems with all methods of analysis if the data collection period varies greatly among subjects or if data collection is stopped in some subjects for reasons which may relate to the outcome variable. For example, if the time dependence was really peaked but some subjects were withdrawn from the study early while their response was still rising it would seem that the time dependence was one of growth. If a large proportion of subjects do not complete the study, then in common with other types of study this will make sensible analysis of the results very difficult if not impossible.

We have described just two common types of serial data, but other forms of response over time are sometimes of interest. For example, in studies of basal body temperature around the time of ovulation the feature of interest is whether there has been a change in the temperature level (and when this occurred). In studies assessing the effectiveness of antihypertensive treatment the outcome of interest may be the level at which the blood pressure stabilises rather than the change from baseline or the rate of change.

One consequence of replacing the measurements on a subject by a single summary measure is that what seemed to be a lot of data suddenly seems rather small. This will be the case when trying to make up for lack of subjects by measuring each subject many times. The strong dependency between measurements close in time on the same patient means that the original wealth of data was to some extent illusory; the summary measures will give a more honest indication of the amount of information that has been collected.

When a researcher is planning a study in which serial measurements on each subject will be collected the intended method of analysis should be considered. It is valuable to have a clear idea of the features of the data that will be of prime interest. The sample size of the study should be taken as the number of subjects, not the total number of observations. If it is known that the timing of a particular feature, such as a peak, is of interest, than extra observations should be made around the time when the feature is likely to occur. For the analysis of such data we recommend the method of summary measures for extracting the useful information from the data. This method is in common use in clinical pharmacology and should become more widespread in all branches of clinical research.

We are grateful to John Williams for producing the figures and Lindsey Izzard for typing the revised drafts. We thank David Appleton, Michael Healy, Niels Keiding, Judy Simpson, Lene Skovgaard, and particularly Martin Gardner for constructive criticism of previous drafts of this paper.

1 Illingworth PJ, Jung RT, Howie PW, Isles TE. Reduction in postprandial energy expenditure during pregnancy. *Br Med J* 1987;**294**:1573-6.
2 Bradley JR, Evans DB, Gore SM, Cowley AJ. Is dialysis hypotension caused by an abnormality of venous tone? *Br Med J* 1988;**296**:1634-7.
3 Fikri MM, Ghaloungui P. Ancylostoma anaemia. *Lancet* 1937;i:800-2.
4 Wishart J. Growth rate determination in nutrition studies with the bacon pig, and their analysis. *Biometrika* 1938;**30**:16-28.
5 Oldham PD. A note on the analysis of repeated measurements of the same subjects. *J Chronic Dis* 1962;**15**:969-77.
6 Rowell JG, Walters DE. Analysing data with repeated observations on each experimental unit. *Journal of Agricultural Science (Camb)* 1976;**87**:423-32.
7 Healy MJR. Some problems of repeated measurements. In: Bithell JF, Coppi R, eds. *Perspectives in medical statistics*. London: Academic Press, 1981: 155-7.
8 Yates F. Regression models for repeated measurements. *Biometrics*1982;**38**: 850-3.
9 De Klerk NH. Repeated warnings re repeated measures. *Aust N Z J Med* 1986;**16**:637-8.
10 Armitage P, Berry G. *Statistical methods in medical research*. Oxford: Blackwell Scientific, 1987:355,360,411.
11 Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. *Br Med J* 1986;**292**:746-50.
12 Tsai K-T, Koziol JA. VARCOV II: a computer program for the multivariate analysis of growth and response curves. *Comput Methods Programs Biomed* 1988;**27**:69-74.
13 Morrison DF. *Multivariate statistical methods*. New York: McGraw-Hill, 1976.
14 Kenward MG. A method for comparing profiles of repeated measurements. *Applied Statistics* 1987;**36**:296-308.
15 Goldstein H. *Multilevel models in educational and social research*. London: Griffin, 1987:51-60.

Appendix I

Some summary measures

Type of data	Question to be answered	Summary measure
Peaked	Is the overall value of the outcome variable the same in different groups?	Overall mean (equal time intervals) Area under curve (unequal time intervals)
Peaked	Is the maximum (minimum) response different between groups?	Maximum (minimum) value
Peaked	Is the time to maximum (minimum) response different between groups?	Time to maximum (minimum) response
Growth	Is the rate of change of the outcome variable different between groups?	Regression coefficient
Growth	Is the eventual value of the outcome variable the same between groups?	Final value of outcome measure or difference between last and first values, or percentage change between first and last
Growth	Is the response in one group delayed relative to the other?	Time to reach a particular value (for example, a fixed percentage of baseline)

Appendix II

Calculation of area under the curve

The area under the curve is calculated by adding the areas under the graph between each pair of consecutive observations. If we have measurements y_1 and y_2 at times t_1 and t_2, then the area under the curve between those two times is the product of the time difference and the average of the two measurements. Thus we get $(t_2-t_1)\ (y_1+y_2)/2$. This is known as the trapezium rule because of the shape of each segment of the area under the curve.

If we have n+1 measurements y_i at times t_i (i = 0, ..., n), then the area under the curve (AUC) is calculated as:

$$\text{AUC} = \frac{1}{2} \sum_{i=0}^{n-1} (t_{i+1}-t_i)(y_i+y_{i+1}).$$

Consider the data for one ill patient in the aspirin absorption study. At times zero, 5, 10, 15, 20, 30, 40, 60, 75, 90, and 120 minutes the aspirin concentrations are 0, 8·3, 21·6, 33·9, 35·5, 47·2, 38·3, 20·5, 13·3, 0, and 0 μmol/l respectively. Thus we have

$$\text{AUC} = 5 \times \frac{8{\cdot}3}{2} + (10\text{-}5) \times \frac{(8{\cdot}3 + 21{\cdot}6)}{2} +$$

$$(15\text{-}10) \times \frac{(33{\cdot}9 + 21{\cdot}6)}{2} + \ldots + (90\text{-}75) \times \frac{13{\cdot}3}{2} +$$

$$(120\text{-}90) \times \frac{0}{2} = 2191\ \mu\text{mol min/l}.$$

If we standardise by the length of the study, 120 minutes, we get 2191/120 = 18·3 μmol/l, which is close to the mean value of the observations of 19·9 μmol/l.

(Accepted 27 November 1989)