# THE EVALUATION OF SOME CLINICAL DATA*

JAMES PARKHOUSE, M.A., M.D., F.F.A.R.C.S., D.A.†

DURING THE PAST FEW YEARS I have been involved in a number of studies relating to the clinical evaluation of analgesics. These studies have been of various types, and have involved a number of different experimental designs; this has given us the opportunity, indeed it has forced upon us the necessity, of thinking about a variety of methods of statistical analysis. The latest investigation was a series of aspirin dose response studies; much of what I propose to say relates to these studies, and our earlier investigations, but rather than discussing specific findings in detail I wish to use the present occasion as an excuse for putting forward some idle thoughts of an amateur about clinical statistics. The word amateur is important, and I beg that it be remembered. I am no statistician. If I were I should not have spent so many hours worrying myself about the mysteries which form the basis of this paper. Mysteries, to me, they certainly are; I have learned some lessons through trying to understand them and it is because of this, and because others of my acquaintance often seem to be equally mystified, that I dare to raise my voice.

## A VERSUS B

The simplest situation that can arise in a clinical drug study is where one group of patients receives drug A, another group receives drug B, and the response of each patient is represented by some kind of numerical score. The "mean score" of each treatment group can then be computed and the difference between these two means, taking account of the spread of the individual scores (the standard error of the mean) can be used to give an indication of how often the observed result is likely to occur by chance. This is the principle of the *t*-test, which for half a century has held pride of place in the statistical handling of this type of situation. Its theoretical validity rests on certain assumptions, notably about normality of distribution, but it is known that some departure of the data from the form assumed by the theory of the test is unlikely to matter a great deal—as the statisticians say, it is a "robust" test. A more serious objection arises in analgesic studies, where the scores are not "real numbers": they denote arbitrary grades of pain or relief and there is some danger in assuming that these can be manipulated in the familiar manner of geometrical proportions or trigonometrical ratios. Who knows, to quote a familiar criticism, that the size of the step from "mild" to "moderate" is the same as the size of the step from "moderate" to "severe"? Even this objection has not deterred many investigators

*Presented at the Annual Meeting, Canadian Anaesthetists' Society, June 25–30, 1967.
†Professor of Anaesthetics, University of Manitoba.

from using the *t*-test in pain studies. It did not deter us; we knew of no proven alternative.

There is no lack of possible alternatives. Firstly, there is the chi square test. Here the grading problem does not exist: there is only relief or not-relief. Much the same may be said of the usual form of sequential trial, in which results are paired arbitrarily, each pair then providing a possible preference for one treatment or the other. Although some kind of grading system may be used for the initial scoring, as is true even when the chi square test is to be used, the magnitude of the difference between scores for any one pair does not appear in the final diagram. This might be described as an all-or-none result derived from a pseudo-quantal response which, in a sense, is getting the worst of both worlds. There are sequential *t*-testing procedures, but with these we are back to the same assumptions as with the standard *t*-test. I do not decry the humble chi square test; it often turns out to be as sensitive as any, and it is certainly free from the potential viciousness of most. Nevertheless, one feels instinctively that there must be some loss of information inherent in a simple yes/no classification of results. If I have a patient whose response to drug A is graded 15, a patient whose response to drug B is graded 6, and another whose response to drug C is graded 2, my common sense tells me that A probably has a much greater superiority over C than does B; if the same kind of result is reflected in a group of cases treated with each drug I feel that, although I may not know exactly what this "15" means, I should like to be able to do something more with it than just call it a chi squared "yes" or a plain sequential preference, which is no different from what I should do with "6."

The use of "non-parametric" tests might seem to offer some hope. These tests allow more than an all-or-none categorization and yet do not depend on an assumption of normality. The Wilcoxon rank sum test, for example, is based on the ranking of scores in order of magnitude without taking account of the absolute value of any score; it is to some extent open to the criticism voiced at the end of the last paragraph, but theoretically it compares favourably with the *t*-test in "power." In practice, the test is difficult to apply when the range of possible scores is limited, as in pain studies. It is common to find several identical scores ("ties") and these considerably weaken the ranking procedure. A correction factor is available for "ties" but the test is said to be unreliable when ties are very frequent and little theoretical work has been done on its use in this situation.

Another attempt at the handling of non-parametric data is the use of Ridits. Without going into details, the idea of Ridit analysis is to refer the results in a given treatment group to a kind of reference standard, or arbitrarily chosen control group which is called the Identified Distribution. A potential advantage of the method is that it might enable the results of several trials to be compared despite obvious differences in the mean scores given by different investigators, rather in the manner of a sophisticated and statistically justifiable "scaling up" of all marks to 100 per cent. In my own attempts to apply Ridit analysis, however, I have been unable to discover what the practical advantages are and I think my feeling is shared by others.

## More Than A Versus B

In our aspirin dose-response studies we compared a placebo, 5 gr., 10 gr., and 20 gr. of aspirin, so that we did not have the simple situation of a group A and a group B. We had four treatment groups: instead of a single *t*-test there were six possible *t*-tests, since we could ask six different questions—is A better than B, is A better than C, is A better than D, is B better than C, is B better than D and is C better than D? We actually continued to use the *t*-test for this investigation, because it was simple to do so, having a computer programme already written, and because we knew it would offer us a useful guide. But we were not unaware of the realities of the situation. Cochran and Cox write: "If there are no differences between the true treatment means, the probability of obtaining an apparently significant result in one (5%) *t*-test is 0.05. It has long been recognized, however, that if several *t*-tests have been performed, the probability that at least one of these is apparently significant is greater than 0.05. If the *t*-tests are independent this probability is 0.23 for 5 tests, 0.40 for 10 tests, and 0.64 for 20 tests."[1] This is a very important truth which is often overlooked. If five *t*-tests are applied to a set of results there is a one-in-four chance of one comparison showing a "statistically significant" difference even if the treatments are identical. Sometimes an investigator tries to avoid this weakness by first inspecting his results and then making only one *t*-test, to compare the highest with the lowest-scoring group. But there is no assurance at this stage that the difference in scores is real, rather than a chance occurrence, and the investigator is in fact confining his statistics to the case in which experimental errors are likely to have combined to produce an apparently large difference in results. Again, Cochran and Cox estimate that even when there are no real differences between drugs a *t*-test applied simply to the highest and lowest of six treatment scores will appear significant at the 5 per cent level on 40 per cent of occasions.

If distinctions between individual pairs of treatments are needed in a multiple drug study, a test is required which will maintain its meaning in terms of statistical significance for all comparisons. Several such "multiple range" tests are available, including the Newman-Keuls method and the test described by Duncan,[2] although here the statisticians themselves are not in complete agreement since Scheffé,[3] who studied this problem extensively, expressed difficulty in understanding Duncan's test!

## Slopes and Scatters

In analysing a study of several doses of the same drug, like our aspirin investigation, an obvious alternative is the dose-response regression. This is an application of the classical bioassay technique of the pharmacologist to the clinical situation. Responses are plotted, graphically, to a series of doses which increase in logarithmic progression. The resulting points will usually lie roughly in a straight line and the technique of regression analysis can then be applied to test whether the *slope* of the line differs significantly from zero (i.e., whether the response to the drug does genuinely rise as the dose increases), and whether

the scatter of the points differs significantly from a straight line (i.e., whether there is a direct log-dose/response relationship).

This is a method of analysis which we used in our studies, and we learned much from it. When two different studies are made of the same drug a useful picture of the range of clinically estimated effectiveness can be obtained; when two different drugs are studied, their relative potencies can be assessed. But it must be admitted that worthwhile log-dose/response lines are harder to obtain in the hospital than in the laboratory. At least three points are needed for each drug in order to make any statements about linearity, or parallelism of response, and a placebo does not count as a dose. The simplest possible study, therefore, to establish the relative potencies of two drugs and to include a placebo requires seven treatment groups. I know that some conclusions have been drawn from "two point" and even "one point" clinical assays, but these must be dubious.

The usual questions about "statistically significant" differences between doses are not answered by a regression line. This is often no bad thing, since attention is focused on the more important question of what is clinically significant. There may be occasions, however, when a specific question has to be answered such as whether or not there is anything to be gained by using a certain big dose rather than a certain smaller dose. For this question a different type of study is more appropriate.

## The Time Factor

All the statistical methods discussed so far come into a class which might be called static. They give a result based on a single score, or on a single sum of scores, which is taken to represent the action of the drug. None of them gives any idea of the *progress* of the drug's action on a time scale, or, to look at the situation more humanistically, they paint no sort of dynamic picture of the patient's progress under the influence of his medication.

It is possible, of course, to repeat tests, and this we have done in our studies: pain and its relief are assessed at hourly intervals, so that a mean score for each drug and a *t*-test comparison can be produced at each hour after medication. The total number of *t*-tests generated by an investigation now becomes enormous and one has at least the dubious comfort of knowing that one or two "statistically significant" differences are almost bound to turn up. There are some other interesting problems about hourly mean scores. Some patients do not last more than an hour or two before needing another drug, especially after a placebo, and the question then arises whether these patients should be given an arbitrary score and included in the mean scores for hours 3, 4, and so on. If they are so included, an agreeable-looking curve is produced for the time-course of action of the drug, but at the later hours it is impossible to know from inspecting this curve whether all patients are still responding slightly or only a few patients are still responding excellently. I have discussed this in connection with a previous study[4] and pointed out that it may be important to make this distinction in order to exclude difference in response within the population, due to genetic or other abnormalities.

Inherent in this is an entirely different concept of drug assessment, based

merely on duration of action. It was obvious to us in looking at some early results that the proportion of patients who needed a second dose of a postoperative analgesic was much higher when the first dose was a placebo than when it was an active drug. The reason is also, I hope, fairly obvious. It led me to suggest on one occasion that when we had exhausted our repertoire of complex statistical tests we should begin assessing the efficacy of our analgesics simply by counting the people who asked for something else. I have not enough data to know what loss of information would result from applying only this one test, but if we turn our attention to our later studies I can illustrate how it appears in practice. Table I shows the percentage of patients, in five separate aspirin dose-response

TABLE I

PER CENT OF PATIENTS NOT HAVING RECEIVED SUBSEQUENT MEDICATION AT STATED TIME*

| Study | Dose | Hour | | | | | Per cent of patients |
|---|---|---|---|---|---|---|---|
| | | 2 | 3 | 4 | 5 | 6 | |
| 1419A | 0 | 96 | 76 | 64 | 60 | 56 | 25 |
| | 300 | 92 | 83 | 75 | 71 | 67 | 24 |
| | 600 | 87 | 78 | 65 | 65 | 52 | 23 |
| | 1200 | 85 | 81 | 73 | 69 | 65 | 26 |
| 1419B | 0 | 40 | 36 | 32 | 24 | 20 | 25 |
| | 300 | 62 | 58 | 46 | 31 | 23 | 26 |
| | 600 | 77 | 69 | 69 | 62 | 35 | 26 |
| | 1200 | 87 | 82 | 70 | 70 | 48 | 23 |
| 1425A | 0 | 58 | 23 | 23 | 15 | 15 | 26 |
| | 300 | 77 | 54 | 46 | 38 | 35 | 26 |
| | 600 | 76 | 68 | 60 | 52 | 48 | 25 |
| | 1200 | 87 | 83 | 65 | 65 | 61 | 23 |
| 1422A | 0 | 91 | 83 | 52 | 40 | 22 | 23 |
| | 300 | 100 | 96 | 89 | 74 | 60 | 27 |
| | 600 | 100 | 100 | 91 | 91 | 86 | 22 |
| | 1200 | 93 | 93 | 89 | 67 | 52 | 27 |
| 1422B | 0 | 100 | 93 | 82 | 82 | 70 | 27 |
| | 300 | 100 | 91 | 77 | 68 | 60 | 22 |
| | 600 | 100 | 96 | 96 | 82 | 82 | 27 |
| | 1200 | 100 | 100 | 100 | 96 | 96 | 23 |

*The first three studies were carried out consecutively by a single nurse-investigator in an orthopaedic hospital. The last two studies were carried out consecutively by two different nurse-investigators in a general hospital. Doses are milligrams of aspirin. The number of patients treated, and not having received subsequent medication after one hour, is taken as 100 per cent. For details see text.

studies, who had not yet required further medication for pain at various times. It is interesting to see the wide variation in "duration of action" of the placebo; in one study 77 per cent of the patients had been given a further drug by the third hour while in another study only 30 per cent had been given a further drug after six hours. It is also interesting to note that studies 1419A, 1419B, and 1425A were all carried out, successively, by the same nurse-investigator who apparently became more self-confident in her "weeding out" as she gained experience. But perhaps most interesting of all is the fact that in study 1419A, where the number of patients remaining after each hour was almost identical for placebo and for 600 mg. aspirin, the *t*-test, based on total relief scores over

the six hours, showed a significant difference ($p$ less than 0.01), while in study 1419B, where the numbers of patients "surviving" after these medications were markedly different, a $t$-test showed no significant difference.

## More Than One Dose

The time factor leads naturally to thoughts of studies involving two or more doses. Here the possibilities for statistical analysis are even more numerous and the opportunities for dissecting information are also, in theory, much greater. This can be illustrated by a fairly simple model.

Suppose we have two drugs, A and B, and we treat each patient with both drugs in succession so that half the patients get A followed by B and half get B followed by A. Suppose also that we assign an arbitrary neutral score of 5 to each patient assuming an equal effect from each drug. Suppose, finally, that we add and subtract the scores for A and B. We can now construct a table of possible events (Table II). If the underlying pain remains constant and A and B have the same effect (row 1), the scores for the two doses will be the same, and for both the AB patient and the BA patient the sum and difference, A + B and A − B, will work out to be 10 and 0. If the pain gets less with time, both patient AB and patient BA will score less (for pain) with the second dose (row 2); for patient AB, A + B will equal 9 and A − B will equal 1, while for patient BA, A + B will equal 9 and A − B will equal −1. If one patient has more pain than

TABLE II*

| | Dose 1 | Dose 2 | A + B | A − B |
|---|---|---|---|---|
| 1. Neutral | | | | |
| AB | 5 | 5 | 10 | 0 |
| BA | 5 | 5 | 10 | 0 |
| 2. Decline in pain | | | | |
| AB | 5 | 4 | 9 | 1 |
| BA | 5 | 4 | 9 | −1 |
| 3. Difference between patients | | | | |
| AB | 6 | 6 | 12 | 0 |
| BA | 4 | 4 | 8 | 0 |
| 4. Difference between drugs | | | | |
| AB | 6 | 4 | 10 | 2 |
| BA | 4 | 6 | 10 | 2 |
| 5. Persistent effect of B | | | | |
| AB | 6 | 4 | 10 | 2 |
| BA | 4 | 5 | 9 | 1 |
| 6. (3 + 4) | | | | |
| AB | 7 | 5 | 12 | 2 |
| BA | 3 | 5 | 8 | 2 |
| 7. (2 + 3 + 4) | | | | |
| AB | 7 | 4 | 11 | 3 |
| BA | 3 | 4 | 7 | 1 |

*Theoretical model illustrating effects of various factors on the result of administering two different drugs in succession, in both possible orders. The "neutral" situation is envisaged as that in which the pain remains constant throughout the time-course of both drug administrations, the treated patients respond identically, the drugs have exactly the same effect, and there is no carry-over of action from one drug period to the next. The last two columns are the sum and difference of the scores achieved from the two drugs. For details see text.

the other (row 3), A + B for one will be 12 and A + B for the other will be 8, A − B being 0 for both. If B is a better drug than A (row 4) the pattern of A + B and A − B will again be different; likewise if the effect of B carries over into the second dose period while the effect of A does not (row 5). All kinds of combinations of factors can be analysed in the same way (e.g., rows 6 and 7).

What we have now described is the principle of analysis of variance, which has been called the most powerful weapon in the armamentarium of the statistician. Its elegance is self-evident; its appeal enormous. But when applied to the study of postoperative pain it proves to have limitations, one of the most important of which is the loss of data from patients who need no second dose, or the adulteration of the study through inclusion of patients who are given a second dose whether they need it or not. At best it is, once again, a "static" assessment.

## Time Analysis Again

Finally, I want to look back, as I did with the need-for-a-subsequent-drug question, from the two-dose situation to the study of a single dose, this time trying to apply the analysis of variance type of dissection in a *dynamic* way. What I am thinking of is something like what statisticians call the study of stochastic processes; as far as I know clinical drug effects have not been looked at in this way.

A stochastic process is one which involves conjecture, the Greek origin of the word suggesting a guess. Mathematically, the term has been applied to many problems which involve a chain or sequence of events. One of these is the gambling or Monte Carlo problem—and it should be remembered that the analysis of games of chance was one of the earliest incentives to the study of probability. There is a large class of situations in which the outcome of each event, like the toss of a coin or the spin of a roulette wheel, is completely independent of the outcome of preceding events: the "Gambler's Ruin" is a classic problem of this type. There are many classes of other situations, called Markov chains after the famous Russian mathematician who studied them, in which the future behaviour of the system is dependent on its present state; a complex and extensive mathematical literature has grown out of the study and analysis of such cases.

The kind of chain that we can imagine ourselves to be concerned with is illustrated by the progress of the individual patient as the hours go by. We can begin to build a stochastic model by imagining all the possible "natural histories" of the condition we are interested in, for instance postoperative pain. It may remain constant with time; it may gradually diminish; it may get worse, as when the wound is infected or the bandage too tight; or it may fluctuate, in which case we may by chance begin our study while it is waxing or while it is waning, these being equally likely contingencies (Fig. 1). We can now assign probabilities to these trends either arbitrarily or as a result of studying large numbers of cases: we may say, for instance, that half the patients will have steady pain, one in five will have diminishing pain, and so forth.

If we start to study a single patient with no prior knowledge of the trend in his case, the probability that his pain will be less after one hour, according to the

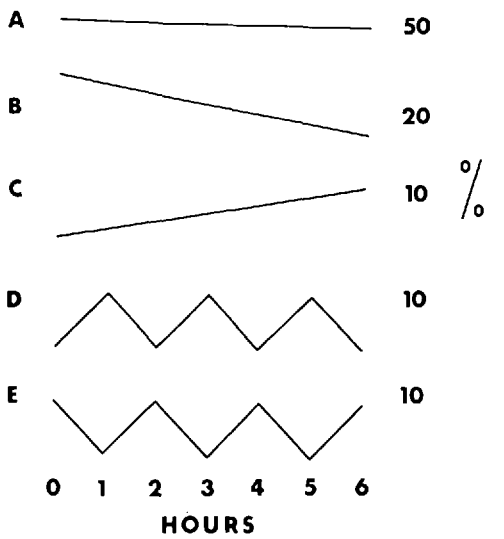


FIGURE 1. Possible "natural histories" of untreated postoperative pain, the ordinate representing intensity and the abscissa representing time. The figures at the right-hand side indicate the percentage of patients assumed to fall into each group for the purpose of this model (see text).

scheme of Figure 1, is 0.3 (30%). However, if we wait two hours before looking at him, the probability of finding spontaneous improvement is only 0.2 (20%) since he may have been in Group B or Group E and in the latter case his pain will have returned to its original intensity.

What happens if we look at the patient after one hour, and find improvement, and then try to estimate the probability of finding *further* improvement after a second hour? The situation is now quite different: at hour 0 we have no idea whereabouts in Figure 1 this particular patient should be placed; but when we see after an hour that his pain has become less we know that he is not in Group A, or in Group C, or in Group D. He must belong to Group B or Group E, with a 67 per cent chance of being in Group B. If he is in Group B, the probability of his pain becoming still less after a further hour is 100 per cent; if he is in Group E the probability is nil since his pain is now due to rise again; he therefore has a 67 per cent chance of further improvement. When we look at him after the second hour, and find that his pain has indeed continued to diminish, we know that further improvement is inevitable. This model is of course hopelessly oversimplified, especially with regard to Groups D and E, but it is by no means unreasonable in principle.

We can next superimpose the hypothetical effect of a drug, which will exert a certain relieving effect in proportion to the initial pain and which will reach a peak effect after a certain interval of time. Such a situation is represented in Figure 2, again in very simple form. Some intriguing speculations are invited by this model: the probability that the pain will be less after one hour is now 0.9 (90%); if we look at the patient only after two hours the probability of finding improvement is

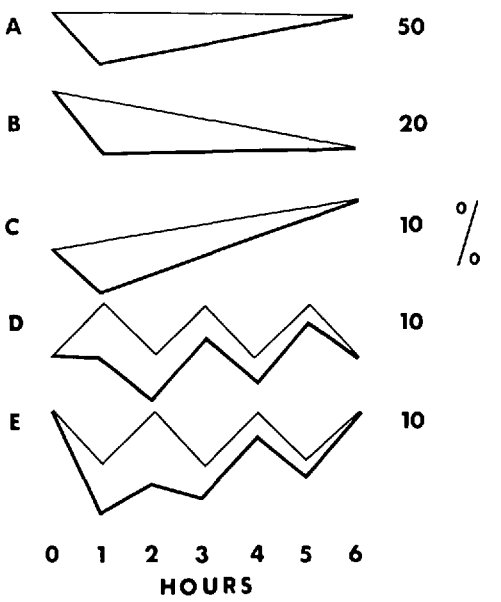


FIGURE 2. Theoretical effect of drug administration (heavy line) on various "natural histories" of postoperative pain (light lines are as in Fig. 1). The figures at the right-hand side indicate the percentage of patients assumed to fall into each group for the purpose of this model (see text).

100 per cent; if the assessment is deferred until three hours the probability is 80 per cent. But *if* we find improvement after one hour, the probability of finding *further* improvement after a second hour is nil. This tells us that the peak effect of the drug has passed in all patients after one hour.

We have now considered the spontaneous progress of the pain and the added effect of a drug; we could add the carry-over effect of previous drugs and other relevant factors. We do not want to consider differences between patients since we are now looking at individuals rather than groups. It would take a mathematician to make any kind of quantitative sense out of this Pilgrim's Progress, but it seems to me that the exercise might be well worthwhile. It could perhaps shed some light on the natural history of the condition we are seeking to influence with our drug, on the time-course of action of the drug, on the magnitude of the drug effect in relation to other factors and on the pattern of placebo responses. Most important of all, it might tell us something about individual people and their responses to drugs. This, after all, is what we need to know more about than the behaviour of those popular abstractions, "groups of cases"—particularly when the individual responses happen to be abnormal.

## REFERENCES

1. COCHRAN, W. A. & COX, G. M. Experimental Designs. 2nd ed., New York: Wiley (1957).
2. DUNCAN, D. B. Biometrics. *11*: 1 (1955).
3. SCHEFFÉ, H. Biometrika. *40*: 87 (1953).
4. PARKHOUSE, J. & HALLINON, P. A Comparison of Aspirin and Paracetamol. Brit. J. Anaesth. 39: 146 (1967).