Research in anaesthesia ranges from work which is related to fundamental and theoretical problems at one extreme, to investigations of direct clinical application at the other. Although a good deal of such research may appear to be straightforward laboratory work, yet it is usually involved with all the problems of clinical research, e.g. great variability between subjects (individual patients) and the presence of many important factors which cannot be controlled. When it comes to clinical trials on anaesthetics, the problems that arise are very much those of clinical trials of treatments in other branches of medicine. Such differences as are usually found are mainly of emphasis rather than anything fundamental. For example, it is generally true that by the time a drug is applied to human subjects, comparison of effects with a dummy, i.e. attempts to demonstrate that the drug has some sort of effect as compared with nothing at all, is not often of interest. Most of the questions worth investigating, therefore, involve a comparison between different agents. This is less true for analgesics than it is for anaesthetics.

PRELIMINARY REQUIREMENTS

Selection of subjects.

The organizers of such trials must always consider carefully some of the fundamental problems and criteria of clinical trials. The first and most important is that of selection of subjects: who is to be included and why, and who is to be excluded. Furthermore, the characteristics of the group of patients must be so defined that it is easy to recognize whether a given individual can be said to be a member of that group. This is necessary not only for the proper evaluation of the results, e.g. whether they were obtained from young patients and not from old ones, or from chronic cases and not from acute, but also that the results may be capable of generalization to new cases. The subjects to be included should be suitable for the investigation, e.g. an analgesic should be evaluated in patients whose pain is expected to be stable over the experimental period. An important consideration is that patients selected should be such that they can be expected to come to no harm, neither from the investigation as a whole, nor if the new drug does not come up to expectations. This may limit considerably the range of cases available, but this need not lead to more than slight slowing of the investigation. If the first results show that the new drug is no worse than the standard with which it has been compared, it is then justifiable to go ahead on a wider range of clinical conditions.

Dosage and route of administration.

Next comes a consideration of the dose and mode of administration of the substance under investigation. This must be suitably defined though the definition need not be rigid. Different methods of administration can be used as “factors”, i.e. the trial is repeated as a complete experiment under different conditions of administration. If the different conditions show the same result in the comparisons, then all of them can be amalgamated into one large experiment, and nothing is lost. If they do show different results, then everything has been gained.

SOME FALLACIES

Clinical trials with anaesthetics are of particular interest in that they expose three fallacies about clinical trials:

1. that they must always be “double blind”;
2. that they can compare fixed doses of drug only;
3. that a direct criterion, i.e. the measurement of the action of the drug, is the best method of assessment.

Blind trials.

There is much confusion about the requirements of “blindness” in a clinical trial. Where
the patient's knowledge of the treatment given may affect the criterion for assessing the results, it is important that the patient should not know which of the various treatments he is receiving. This is the single blind trial. If such information, when available to the physician (and all those looking after the patient), may also affect the criterion for assessment then he too should not know which treatment the patient is receiving. This is the double blind trial. If the nature of the criterion is such that it will be affected neither by one nor the other, there is no need for the trial to be blind in this sense. Sometimes it is extremely difficult for a clinical trial to be blind, single or double, even where this would be advisable, though there is usually no difficulty in doing this when the various drugs under investigation have the same form and are administered in the same way, e.g. intravenously or by inhalation. If necessary, the treatments can be given in very elaborate, disguised forms. For example, if it is necessary to compare an intravenous anaesthetic with a gaseous one, the two can be compared by giving each one an appropriate dummy, i.e. the intravenous agent can be accompanied by inhalations of air or oxygen and the gaseous anaesthetic can be accompanied by intravenous injections of saline. The difficulty of this procedure is that it raises the question of the ethics of such a trial. It becomes necessary to weigh the importance of the findings and the need of maintaining double blind against the giving of unnecessary injections.

Comparison of unlike doses.

The notion that in a properly controlled trial all patients should receive the same or equivalent doses (as between drugs) is often held, and mistakenly, as being fundamental to the nature of the trial. Whether a fixed or variable dose of drug is given depends on the nature of the question being asked and vice versa. If a fixed dose is given, then the experiment answers the question of the relevant merits of the particular fixed doses used, but gives no information about others. If the dose is varied to suit the patient, then the investigation evaluates the different treatments under these conditions, and obviously gives no information about the value of fixed doses. In the case of anaesthetics, the dose given is not a fixed one, but that which is necessary for the required purpose. The amount of drug given may depend upon the depth of anaesthesia required and the length of time for which it is needed, in addition to other reasons for variation, e.g. size of patient, sensitivity, etc. This does not invalidate the trial, since the results indicate the differences between treatments according to the appropriate conditions.

It is customary to evaluate drugs by the intensity of the effects they produce with a given dose, but clearly this is not often applicable to anaesthetics which are usually used to produce an appropriate level of anaesthesia. In other words, comparison is not between the effect produced (in terms of depth of anaesthesia), and it will not be in terms of quantity of anaesthetic used, since this is not particularly relevant. In general, the criterion will have to be of some other kind.

CRITERIA OF CHANGE

The choice of a suitable criterion for evaluation of a drug can be very simple and obvious. If one wishes to investigate the changes in blood pH, oxygenation or sodium concentration, there is no difficulty in applying standard tests to this purpose. It is on the occasions when one is interested not so much in the biochemical changes produced by a drug as in the clinical effects, that difficulties may arise in the choice of a suitable criterion for measuring them, e.g. several anaesthetics may appear to be equally effective, so that one would be interested not so much in their effect as in the frequency of complications resulting from their use. It would first be necessary to define what could be included as a "complication" of use and then to count the number of such. Indeed it may be necessary first to carry out a pilot study to determine the nature and frequency of complications, and how best to assess these. It might be better to count patients who have had complications rather than individual complications, since these might be associated. Another useful criterion might be that which considers the ease of administration of an anaesthetic judged in terms of the stability of level of anaesthesia. In this case, it would be necessary to devise a suitable method of measuring the level of anaesthesia such that it could be used sufficiently often for the purpose required (and this would have to be decided
beforehand) and then some sort of index of variation would be used for statistical analysis. This could be the maximum divergence from the average (very unsatisfactory) or the standard deviation or the frequency of divergencies above a defined level, per minute or hour. It must be remembered that an excessive depth of anaesthesia is a very different problem from an insufficient level, so that the two would have to be considered separately.

Another kind of criterion would be to investigate the amount and kind of metabolic disturbance produced by a series of given anaesthetics. In this case it would be necessary to decide which changes to investigate and at what point in time. Even for the straightforward measurement of metabolic changes it might be more important to determine the time required for restitution to normal than to measure the actual degree of disturbance. In other words, the rate of recovery from the anaesthetic might be the important criterion for a particular investigation. Under this category one could also include the possibility of using the postoperative effect of an anaesthetic, e.g. the amount of sickness, vomiting, confusion, headache, etc. The rate could include not only the restoration of body chemistry but the return to consciousness as well. A rate of recovery is easier to measure than a time until complete recovery, because of fluctuations in the patient’s condition. This is particularly true for level of consciousness. For this reason, it might be better to devise psychological tests of continuity of attention rather than of level of performance.

Finally, an appropriate index of a series of measurements might be considered to be more informative than an observation on a particular function.

In these circumstances, many investigators often prefer to choose a criterion on the grounds that it is objective, as opposed to subjective. An example of this was reported recently by Parbrook (1966), who decided to measure the effectiveness of relief of pain by use of an index number, the Respiratory Restoration factor, as originally described by Bromage (1965). This index is an objective measure which is apparently related to the presence of postoperative pain. It might be thought that because it consisted of objective measures and a mathematical formula relating them, it avoided some of the errors of subjective assessments of pain, but this is mistaken. In the first place, the patient’s reaction to the pain plays an important part in determining the actual values of the index so that although the index itself is objective, it is very much under the influence of subjective factors. In the second place, although it can be proved that the index is related to the intensity of pain, such proof must have been carried out under specific conditions, which may not be applicable in other investigations.

Whenever an index like this is used, the investigator should provide evidence (or indicate where such evidence can be found) that it does measure what it purports to, i.e. it is valid. A low validity is sometimes the explanation for discrepant results between different investigators. It is also important to know the reliability of measurement of the criterion. In general, assessments of subjective state tend to have low reliabilities and it is for this reason that investigators try to use objective measurements; instruments give reliable measurements. If objective indexes have low validity, this means that they are very responsive to influences other than the criterion and therefore the great reliability may be bought at too great a price. This is a matter which must be very carefully considered, as usually there is no great difficulty in increasing reliability to an adequate level, e.g. by taking the mean of repeated observations.

**DESIGN OF TRIAL**

**Randomization.**

When it comes to the design of experiments, there are no special problems, but the nature of the subject makes for certain limitations. For example, in trials of anaesthetics, the type of design in which each subject is given all the treatments (as in Latin Square designs) is generally impracticable, although it might be possible for comparing two anaesthetics when two-stage operations are standard treatment. Once again, trials of analgesics are much easier in this respect. Generally, the trial will be made on two or more groups of patients, each group receiving one of the treatments. It is of fundamental importance that such groups should be formed so that they may legitimately be regarded as random samples.
from the same population. As soon as it is decided that a patient is suitable for inclusion in the trial he should then be allotted to one of the treatments by some method of randomization. It is quite surprising that so many investigators, having reached the point where they recognize the importance of control groups, yet hesitate at the next step of "random" allotment to groups, and try to devise some alternative scheme. Any scheme which involves a known order of allocation is liable to go astray for all sorts of reasons. The one exception is the use of "paired cases" in a sequential study. If the first of the two paired cases is randomly allocated treatment A, then the second one must be given B.

The allocation to treatments can be done very simply, in the case of two groups, by tossing a coin. When there are more than two, it is easier to use a table of random numbers. Any point on the table is picked out, and all the digits running downwards, upwards or to a side are taken. Each digit is allotted to a case number, and in the case of two groups the even digits are allotted to treatment A and the odd digits to B, zero counting as an even number. The list is kept by somebody other than the investigator, and when it is decided to include a particular patient he is given a case number and his treatment then ascertained. If there are three groups, the procedure is similar but the zeros are omitted and the digits are then divided by three and the remainder recorded. If the remainder is zero, he is given treatment A; if it is 1 he is given treatment B, and if 2, then C. It is obvious that the full agreement of all those concerned in the treatment of the patient must be obtained for the use of this procedure. Most apprehensions about random allotment are concerned with the risk to a seriously ill patient; this can be avoided by explicit rules concerning the exclusion of "bad risks". Once the decision has been made, it should not be altered, and the patient should not be taken out of the trial. This difficulty can be overcome by using a recognized scheme for grading patients according to their physical status (such as that devised by the ASA),* and agreeing to limit the trial to those who conform to grades 1 and 2. It may seem platitudinous to say that any of such changes destroy the random nature of allocation and produce groups which are not comparable, but unless everybody concerned has been really convinced of the importance of avoiding last-minute changes, there will always be a tendency for such changes to be made "just this time".

Size of groups.

When the comparisons are straightforward differences between several groups, there is no need for the numbers in the groups to be exactly equal. The experiment is most efficient when they are, i.e. its sensitivity is greatest for a given number of cases when the numbers in the group are equal, but small divergencies from equality do not make much difference. When the comparisons are in the form of a factorial design, then it is customary to emphasize the importance of having equal numbers or proportionate numbers in each of the cells. The reason for this is the tremendous increase in complexity of calculating the statistical tests when the numbers are unequal and disproportionate. To conform to this requirement, the usual procedure is for the investigator to decide on the required number in each of the cells and to go ahead until he has filled one cell, after which he does not bother with further cases for that cell. This makes for the minimum of work, since he is not coping with additional cases while he is waiting to fill the other cells. Nevertheless, when one compares the difficulty of obtaining suitable patients and making observations on them with the relative ease of doing statistical calculations, it seems a great pity that additional material should be rejected even when it is available. The investigator could make use of the additional material and the extra work involved would then entail a good deal of extra work in doing the necessary calculations. The benefit from this is that the results of the experiment would be derived from the maximum number of cases obtainable and this is very important from the point of view of the generalization and ultimate usefulness of the results.

Despite this advantage, the formidable task of doing the extra calculations has deterred investigators in the past from following this procedure; the advent of computers which can take over the job of doing the necessary elaborate calculations is therefore, a tremendous boon, but at the time of

* See Conroy, Cassels and Stodsky, 1948.
writing, the number of programmes available for
carrying out this task is minimal.*

Presumably this situation will alter one day.
Returning to the point concerning randomization,
in the case of the factorial design, the procedure
consists essentially of a random allotment to dif-
f erent treatments done independently for each
cell of the design.

PROBLEMS OF FACTORIAL DESIGNS

The factorial design of clinical experiments is the
most powerful method for obtaining the maxi-
mum information from the minimum number of
cases under the widest range of conditions. The
essence of the design can be described briefly as
consisting of a comparison between treatments
repeated under every possible combination of the
different "factors". It may be added here that
the different treatments themselves constitute one
factor. Thus the patients may fall into a set
of groups each characterized by a different disorder,
the types of treatment may vary, e.g. three modi-
fications of a surgical operation and the character-
istics of the patients may differ. This may be of
sufficient importance to warrant classifying, e.g.
the patients may be divided into so many age
groups.

Thus, if three different anaesthetic agents are
being compared with two disease conditions for
which three types of operations may be performed
and the patients are divided into three age groups
this will constitute a $3 \times 2 \times 3 \times 3$ factorial design
containing 54 cells. Assuming a minimum of 2
patients per cell this requires 108 patients, though
it would be much better if the numbers were in-
creased to 3 per cell, making 162 patients. It is
obvious that this is an extremely elaborate in-
vestigation which is unlikely to be carried out in
practice because of the large number of patients
required and the difficulties of filling cells. It
could be simplified by ignoring the factor of age,
thus cutting it down to a $3 \times 2 \times 3$ design with 18
cells requiring a mere 36 (or better still 54)
patients. If it were considered that the age of the

* As far as I know, there is only one that has been
published and is freely available (Clyde, Cramer and
Sherin, 1966). A limited number of multivariate pro-
grammes in the P.L.U.S. system have been developed
in the Department of Psychiatry of the University of
Leeds (Hamilton, 1965).

patients played some part in the response of the
criterion to these influences, it would be possible
to take this into account by carrying out an analy-
sis of covariance. This is practicable even for dis-
proportionate numbers in the cells, though
extremely laborious (also there are some theoreti-
cal difficulties involved).

In general, when on the basis of clinical experi-
ence any factors are considered important, they
should be taken into account when designing the
trial, provided they are qualitative and with only
a few categories. In the above-mentioned example
this applies to the factor of "treatments", i.e. the
different anaesthetic drugs used. If more than
three are to be compared, it would probably be
better in the long run to divide the experiment
into two and compare only some of them in each
one. The same applies to the factor of operative
procedures; for example, in the case of gastro-
intestinal conditions the operations might con-
sist of partial gastrectomy, gastro-enterostomy,
both with or without vagotomy. This makes four
different kinds of operation, but it might be better
to cut it to three, or two if possible. Under certain
circumstances it is possible to pool information
from more than one type of operation, but where
this is done a very clear indication of the extent
of the pooling must be given. When a factor is a
continuous variable, e.g. age, blood pressure,
length of illness, it becomes impracticable to
make use of a sufficient number of groupings. Thus
if the age range is from 20 to 60, dividing this
into 5-year groups would give eight levels of that
factor, multiplying the total number of cases
required by eight. This could be reduced to four
by taking 10-year groups, or even into two by
dividing the patients into those below and above a
certain critical age, but although this makes for
simplification of the investigation, it loses a good
deal of information. The best procedure for con-
tinuous variables, therefore, is to use the method
of analysis of covariance. This, however, intro-
duces a number of complications which cannot
be dealt with here.

Not all factors can be dealt with in this way.
Some cannot be controlled at all, in the sense that
they cannot be introduced as part of the experi-
ment or excluded completely, and others are only
partially controllable. Of the former type, three
may be given as examples: (a) length of operation
and anaesthesia, (b) extent and severity of operation, including such additions as "routine" transfusions, and (c) postoperative treatments. It is obvious that all of these are factors which play an important part in the outcome and which will affect the criteria measured, but in general it is not possible to anticipate their nature and extent and therefore to allocate the patient to an appropriate category. One way of dealing with them is simply to ignore them and to work on the basis that if the investigation includes enough cases then these additional variables will tend to affect equally all the categories and will therefore cancel each other out. Although this is true, it may mean that the numbers of cases required will have to be extremely large, in the first place, to ensure that these additional factors will equalize, and in the second, to make the investigation sufficiently sensitive by using large numbers to counteract the increased variability of the case material arising from the operation of these factors.

A very much better way of dealing with this problem is to devise some method of measuring these factors and to use such measurements as independent variables, for analysis of covariance. This is easy enough for the first, i.e. the length of operation and administration of anaesthetic, but it is certainly a good deal more difficult when one is dealing with the extent and severity of operation and the different kinds of postoperative treatment. Experience in psychological research has shown, however, that there is no great need for finnicky accuracy in this respect. Surgical operations can be graded into three grades of severity and transfusions can be regarded as absent, sufficient and much, and so on. The only requirement is that there should be a reasonable minimum number of cases in the various grades. Obviously, the more cases the better, but in fact two or three are sufficient for analysis. If, however, the numbers are minimal, it behooves the investigator to be cautious in drawing conclusions from his results and to regard them more as tentative indicators for further investigation.

Some factors are partially controllable in the sense that it is possible to make a decision beforehand as to which category the patient belongs, or could belong, and this could be included in the trial. Examples of such are the state of the patient before operation, the type of operation and the nature of the disorder. In the first case, the state of the patient could be included as a factor consisting of such grades as fit, moderately ill, very ill; or some loss of blood and no loss of blood before an operation. Here again the ASA grading system of physical fitness may be of some value. The patient having been allocated to this group at the time of the actual observations, it may appear that this was in error and that he is much more sick than was thought and that he has lost blood whereas it was thought he had not. In that case, no harm is done if he is transferred to the appropriate group when analyzing the data. Much the same sort of thing applies to the type of operation. It may be that once the surgeon has started, he may have to change his mind about what he is going to do. Provided there are categories for different types of operation, then the patient can be switched over to the appropriate group. The same applies for the diagnostic group where it may be that the diagnosis has not yet been adequately established and is settled only during the course of the surgical operation. Although the procedure is the same, it is always useful to have an additional category of "etcetera diagnoses".

MULTIVARIATE CRITERIA

It is unlikely that the investigator will be interested in one variable only. He will want to compare two different drugs on as many variables as possible and will be concerned also with the interaction between these variables. The problems of multivariate statistical analysis will not be dealt with here, nor will the problems concerned with the development of mathematical models of physiological and biochemical processes. To understand the difficulties relating to the designing of a trial, it is useful to visualize the relationship between the data. Thus, each patient will have a number of observations made on him, each one being of a different variable. These can be regarded as being recorded in a row. If these measurements are repeated after an interval of time, they will form another row and a series of such rows will give a rectangular array (or sheet) of measurements. In such an array, each column represents a particular variable and each row represents a given occasion. Since each patient will have such an array, the whole group of
patients will give rise to a rectangular block of measurements, and each row or column for a given patient will then become an array or sheet for all the patients.

It is obvious that measurements on one variable cannot be a substitute for measurements on another, which implies that measurements cannot be taken from one column and put into another. It is also obvious that measurements on one subject cannot substitute for measurements on another subject, i.e. they cannot be transferred from one sheet of measurements to another. Furthermore, many measurements on one subject cannot substitute for one measurement on many subjects and if, for example, two measurements of oxygen saturation in the blood have been taken where one only is called for, then either one of them must be thrown away (which is wasteful of information) or the average of the two must be taken and used as a single measurement. It is not quite so obvious that measurements on one occasion cannot substitute for another occasion. For example, if measurements are taken at 10-minute intervals and one is omitted, it cannot be replaced by another taken, say, 5 minutes later. Furthermore, if a set of observations takes, say, 1 minute and they are repeated every minute, then the last of one set, e.g. pulse rate, could well be regarded as corresponding to the first of the next set, e.g. blood pressure. Sets must, therefore, be adequately separated and every effort must be made to ensure that each set is complete. The investigator must limit the number of observations he makes to well within what is practicable, and preliminary runs should be carried out to establish this. There is one exception to the rule that all measurements repeated on different occasions must be the same for each subject, and that is when the repeated observations are being used to establish some sort of curve representing changes. Provided that the nature of the curve is known, then it is sufficient to make such observations on each patient to establish the parameters of the curve. It is these parameters that will be used for the final analysis and not the actual measurements themselves.

In summary a number of specific problems arise when the general principles of clinical trials are applied to the field of anaesthesia. Some special problems relate to the difficulties of arranging a “double-blind” trial, the selection of appropriate criteria to measure changes and the difficulties involved in designing a trial when the control of many important factors is out of the hands of the investigator. The use of multivariate measurements in appropriate circumstances may offer some solution to these problems.

ACKNOWLEDGEMENT

Thanks are due to Mrs. J. M. Brierly for help in the preparation of the manuscript.

REFERENCES


