Chapter 4

EXPERIMENTAL DESIGN AND STATISTICAL PROBLEMS

LOUIS LASAGNA AND PAUL MEIER

NO SINGLE experiment can supply all the information that one would like to have. To carry out a worthwhile and successful experiment one must first determine a particular question—or set of questions—which might reasonably be answered by the proposed clinical trial and whose answers will at the same time be of enough value to make the experiment worthwhile.

In work with drugs, the most frequently asked question is: "How good is Drug A?" This question is a good deal more complex than it sounds. First of all, it implies that the investigator has a clearly defined class of patients for whom the drug is to be evaluated, and that there are enough suitable patients available for the study. Second, the question implies that the experimenter knows what therapeutic effects he is looking for, and how to measure them. Third, it suggests that he knows what dosage of Drug A should be used, and the length of time for which to give it. Fourth, since the investigator usually means not "how good" but "how much better than something else," the question implies that the investigator has chosen one or more reference standards, e.g., no treatment, placebo therapy, or a standard drug.

ASKING THE QUESTION

A common error in planning an experiment is to ask too much. A battery of questions covering all aspects of the response to

treatment may be planned, and attempts may be made to analyze differences in response relative to age, sex, and a host of other characteristics. The result of asking too many questions is often a drop in the morale of the investigative staff, lack of co-operation from the patients, and a general loss in the accuracy of the records, in addition to the problems raised in attempting to analyze too many factors. The investigator will usually find that an experiment of reasonable size does not give very precise results on a host of interrelated characteristics, and that he has to be content with a few basic findings. If these findings are of sufficient value, much has been accomplished.

It should not be inferred that all interesting questions are simple ones, or vice versa, but rather that the problem of asking a question that is at the same time of interest and capable of being answered by a clinical study is one that requires considerable thought and often considerable ingenuity. The remaining problems are largely a matter of technique—it is in asking the correct questions that the investigator must demonstrate his ability as a scientist.

PLANNING THE EXPERIMENT

The formulation of the questions to be answered must be made explicit by the protocol of the study. Because asking the correct question is crucial, it is advisable to spend a great deal of time in planning the experiment. It is often wise to pool the talents of several experts, hoping to anticipate and avoid as many pitfalls as possible. While it is often desirable for such experts to be full-fledged partners in the venture, it is not always necessary. A typical group might include a clinician, a pharmacologist, and a statistician.

The clinician is often responsible for initiating the study. He should provide a realistic appraisal of the clinical problems likely to be encountered, the number of patients available over a given period of time, variability in patient material and prognosis, and ethical and legal considerations. For example, a urologist may feel, on the basis of past experience, that the male patients in his clinic (because of obstructive uropathy) present a considerably more severe challenge for chemotherapy than do the female

patients, whose "bladder colds" often improve spontaneously or on nonspecific "therapy." Or, an internist may remind his nonmedical colleagues of the tremendous variability in prognosis for patients suffering from "essential hypertension," and of how certain patients may be expected to go for years or decades without serious impairment, whereas others have an excellent chance of developing serious complications or of dying within a year or two. Such considerations obviously should be weighed with care before entering into a trial. A one-year study to determine whether an antihypertensive agent prevents mortality in obese, fifty-year-old women with mild asymptomatic hypertension is probably doomed to failure (most of the patients would probably live *without* treatment); a study of similar length on antileukemic drugs in a small number of adult patients with acute leukemia might give very useful information.

The clinician member of the team must also be constantly aware of ethical and medicolegal restraints. It is he, for example, who should know that pneumococcal pneumonia is at present admirably treated with any one of a handful of antibiotics, and that a completely new antibiotic suggested for clinical trial in this disease ought to carry with it the real hope of distinct advantages in ease of administration, cost, or toxicity before a trial is warranted. Where current therapy is already excellent, the investigator ought to be extraordinarily cautious with new agents; on the other hand, where present treatment is far from adequate (as, for example, in Friedländer's pneumonia) the clinical trial of a possibly useful new drug may be warranted on less compelling evidence.

The pharmacologist can be helpful in a varity of ways. He can help decide whether animal toxicity data are adequate to undertake a clinical experiment. When information is available regarding absorption, distribution, excretion, and fate of a drug, he can help set up rational dosage schedules. If data on human beings are lacking, he can be invaluable in planning the careful preliminary work that is essential. If a pharmacologist is not readily accessible locally, the pharmaceutical company that supplies the new drug can often substitute. The drug company responsible for a new agent usually can provide at least a good guess as to recommended dosage. As many data as possible should be ob-

tained from the manufacturer, with special emphasis on acute and chronic toxicity studies and the nature of any expected side effects. These data will be furnished either on the basis of animal experiments or preliminary human work on the drug or its congeners.

The statistician member of the team has the responsibility for anticipating the form that the experimental results will take and for insuring that the design of the study will permit a meaningful analysis. Although an experienced statistician may be able to make valuable contributions to the study in other respects, his primary function is a technical one. The statistician should make clear the relationship between the number of patients studied and the precision of the comparisons that will be made. By such devices as matching or use of covariance techniques he may be able to increase the efficiency of a trial so as to improve precision or reduce the number of patients required. He should know the limitations of new techniques as well as their advantages and be able to protect his colleagues from overenthusiasm for interesting techniques that may be inappropriate.

The statistician may prove useful in a way not strictly within the bounds of his technical competence. He usually has a heightened awareness of the many kinds of bias that may afflict a clinical study. The biases of observers, of patients who accept or do not accept a treatment, of patients who become lost from observation, are all familiar troubles to the statistician experienced in clinical trials.

In any event, since the type of statistical analysis appropriate at the end of an experiment is contingent upon the way in which patients were selected and the way in which the data were collected, it is desirable to get statistical advice early rather than late.

The clinical trial should ideally be a co-operative venture not only in its planning stages or at its completion, but at various intervals during its life history. During the progress of the trial, serious errors in planning or execution may manifest themselves. In such event, it may be wiser to scrap an experimental design rather than plod doggedly onward, collecting useless data.

Even with the most outstanding group of consultants there must be one investigator who takes primary day-to-day responsi-

bility for the conduct of the study, and he must be one who can understand and synthesize the advice of the consultants. Despite the considerable advantages described above as accruing from a panel of experts from different fields, it should be emphasized that if an investigator appreciates the basic principles of the clinical trial and has a working knowledge of simple statistical concepts, he can often plan and execute a sound experiment, when specialized advice is not available.

DOSAGE

Ideally, one would like to use optimal doses of a drug, i.e., amounts that would work well in most patients while producing a minimum of side effects. Since such a well-educated guess is rarely possible, several alternatives present themselves. The simplest design asks such a question as: "How does Sleepo, a new hypnotic, in a dose of 0.5 g., compare with a standard medication?" The standard medication might be either placebo or (in this case) such a drug as pentobarbital. It usually turns out best to use *both* of these "standards," because one would like to say at the end of the experiment, not only whether Sleepo is better than a sugar pill, but also how it compares with a medication commonly used for producing sleep. If one uses only a pentobarbital control and Sleepo turns out about as effective as pentobarbital, there always remains the nagging possibility that the "success rates" observed were really only placebo rates, and that the experimental population or the design was incapable of discriminating between drugs.

More detailed information about the dose-response curves for Sleepo and pentobarbital could be obtained by utilizing two or more dose levels for each drug. For example, one could use the "usual recommended dose" of each, and also double this dose. With two points on the dose-response curve of each drug we would have a more comprehensive view of the relative efficacy (and toxicity) of the two drugs.

Ideally, the determination of the optimum therapeutic dose of a drug to be evaluated would require estimation of the dose-response curve *for each patient*. For example, the dose of digitalis in congestive heart failure can be variable. Often, it can be de-

termined only by pushing the dose to the point of toxicity before eliminating the possibility of effectiveness, so that the dose decided upon for a given study in a group may be somewhat unreliable for a particular individual. It is, nevertheless, valid in such a situation to estimate the *population dose-response curve* and thus arrive at an *average* effective dose. If the individual variability is great, it does not interfere with statistical validity but indicates merely that in the definitive experiment a larger sample must be studied. Another instance where the optimum dose for study is difficult to determine a priori is when analgesics are used for a short-lived complaint such as postpartum pain. Here no possibility for individual titration of doses is present, since very few doses will be required for any patient.

In many cases, however, there is no compelling reason why establishment of dose-response curves cannot more closely approximate clinical practice. There has developed a harmful tradition that controlled trials must of necessity utilize rigid dosage schemes. There is, in fact, no reason why patients in a double-blind experiment cannot have their dosage raised if beneficial effects have not been achieved, or lowered if toxic phenomena appear. In most cases it is also helpful to have acquired some feeling for dosage, duration, and timing before embarking on a formal experiment with a drug.

SAMPLE SIZE AND SEQUENTIAL PLANS

How many patients should one study? At times this question is extremely simple, since the study may arbitrarily run for one year (for administrative purposes) or may be limited to 30 patients because only so many are being seen in the clinic where the experiment is to be run. At other times the problem becomes highly complex. In all cases, it is advisable to appraise the situation as realistically as possible, making the best possible guesses as to numbers of patients available, therapeutic performance likely to be achieved by the standard drug, performance expected of the new drug, etc. One can then, by a few simple calculations,[1] calculate how many patients will be required, e.g., to show that a new drug relieves 20 per cent more patients than the standard drug's usual 60 per cent. Occasionally, such calculations indicate

that a study had better not be started, since the small number of patients available almost precludes the establishment of significant differences between treatments.

When, in 1951, the field trial of gamma globulin as a prophylactic against poliomyelitis was planned, advance calculations indicated that something on the order of 50,000 vaccinated and an equal number of control children would be required.[2] The results at the end of the trial indicated that the use of appreciably fewer children would almost surely have led to inconclusive results.

Lately there has been considerable interest in *sequential designs.* This type of design makes allowance for "controlled peeking" at one's data so that experiments can be terminated as soon as a difference between treatments is securely established. Such "peeking" vitiates the use of classical statistical techniques in the analysis of one's data, and new techniques have been developed for this purpose.

The primary advantage conferred by a sequential design is that, when the outcome of an experiment is clear early in the game, it is possible to quit early and to draw conclusions with a known and predetermined risk of error. Where serious ethical considerations are involved (such as the withholding of a potentially beneficial drug from patients who may otherwise die) the advantage in a scheme that permits early and secure conclusions to be drawn is obvious. Likewise, the ability to cease testing an obviously inferior product with known risk of error, but without completing a preset number of trials, may eliminate much wasted effort. More generally, *on the average* the amount of clinical material consumed by a sequential scheme per clinical trial will be less than that required by fixed sample-size schemes.

However, sequential schemes have a number of disadvantages to counterbalance these benefits: (1) A sequential design is feasible only when the clinical response to treatment can be determined with little delay. If enough subjects can be gathered within a week but the result of treatment is observable only months later, there is no way to make profitable use of a sequential scheme. (2) A sequential design will generally require considerable flexibility in administrative arrangements. If a large-scale clinical trial is to be set up, it may be necessary to hire a special staff and

to provide for other services. With fixed sample-size plans such needs can usually be estimated in advance with fair accuracy. By its very nature the sequential design makes such planning difficult. (3) The theoretical development of sequential designs is still fairly rudimentary. At present only a few special designs are available for use and the construction of others would involve considerable laborious calculation. The available formulas for determining confidence limits are only approximate and more exact results may require somewhat tricky and unfamiliar calculations. Eventually, further research may eliminate such objections, but at present these restrictions on the use of sequential designs are substantial. (4) Finally, the information provided by current sequential schemes is generally more limited and more difficult to interpret than the information provided by fixed sample-size schemes.

Suppose that we are able, through a sequential design, to stop experimenting after studying only a few patients. We may then be 95 per cent certain that Drug A is in fact better than Drug B, but we will have only a crude idea of how much better. If we are really interested in the magnitude of the effect, a fixed sample-size scheme is more likely to serve our purpose. When a new drug is found to be better than a standard, a fairly good measure of the degree of improvement is almost essential, and the investigator is likely to continue the experiment until an adequate assessment can be made. In this event the "sequential" feature of the original plan is really irrelevant. However, when the new treatment is found to be inferior to the standard, one may not care to see just how bad it really is, and the sequential design may then result in a considerable saving. In any case, the data provided by a sequential plan are usually less suitable for informal manipulation by the experimenter than are the data collected in a fixed sample-size experiment. If one does not expect to be content with the simple "Yes" or "No" answer for which the sequential design was constructed, a fixed sample-size design may be much more satisfactory.

ALLOCATION OF SUBJECTS

No high-powered analysis of data will compensate for data of poor quality or dubious validity. A prime assumption in the statis-

tical analysis of the comparative performance of two drugs is that there has been no bias in the allocation of patients to one particular treatment group. It is impossible to interpret a "statistically significant difference" between drugs if the therapeutic challenge has been made more difficult for one treatment by purposeful shunting of even occasional patients into that group. In such a case, Drug A may look better than Drug B either because it *is* better than B, or because the patients getting Drug A were less ill than those getting B. Since there is no way to distinguish between these alternatives, such an experiment is usually worse than useless.

The "random" assignment of patients required for a valid comparison must not be interpreted in the lay sense of "haphazard." Schemes such as alternation of subjects or odd and even serial numbers are notoriously subject to bias in assignment of patients. That they are often a minor gain in convenience is poor payment for the loss of an objective measure of error. Although an investigator may convince himself that there is no conceivable way in which the serial number could affect the characteristic under study, this state of affairs may merely be a tribute to his lack of imagination. It cannot be emphasized too strongly that the rationale for confidence statements (*vide infra*) is completely dependent on the employment of true randomization. With the aid of random number tables[3,4] the process requires little effort. (It is surprisingly difficult to make coin flips or card draws satisfactorily random. The use of random number tables is, in practice, much more satisfactory.) However, the procedures are not always obvious and they depend on the experimental design used. Therefore, the investigator must instruct himself on the proper procedure or bring in a consultant to aid in the randomization.

PRECISION OF MEASUREMENT

Once the patients are properly allocated to treatment groups, attention must be turned to errors of measurement. An experiment excellent in all other respects may fail because measurements are made with insufficient care or because of inadequate record-keeping.

If measurements are not carefully made, real differences between drugs may be missed. Randomization and double-blind

techniques will prevent the experimenter from frequently claiming differences that are not present. However, they will not help him to control poor precision of measurement. Unbiased but sloppy measurements may lead to the correct assertion that "no significant difference was found." However, confidence limits may show that the actual difference could be substantial, even though not detectable in the experiment as conducted. A more subtle error resulting from poor measurement technique will not necessarily be detected by calculating confidence limits. This error occurs when the measurements are *all* biased toward a common value. If the observer fails to recognize the signs of improvement among some patients and marks nearly all patients "not improved," a substantial real difference between drugs may go undetected, and there may be no indications of excessive measurement error.

Proper record forms and recording techniques can be of great value in reducing measurement error. One of the most important principles of good record-keeping in clinical trials is that the observer should not be required to give all his information in narrative form. One must always allow for comments, lest an important item of response be ignored, but the more objective and systematic the form—using checklists, scales, and so forth—the more likely one is to get reproducible results. On the other hand, long checklists and innumerable questions defeat their own purpose. Although the form should be as objective as possible, it must also be as brief as possible. As mentioned in Asking the Question, a long series of questions will result in poor morale, inattention, and perhaps even unco-operativeness, so that the result may be a loss rather than a gain in the quality of measurement. Pretesting the form is an essential but frequently overlooked part of the preparation for an experiment.

When the techniques of measurement are crude and unreliable it may at times be helpful to use several observers. The several observers may be used to check the reproducibility of the measurement and they may also provide somewhat different kinds of data. For example, in studies on hypnotic drugs, one observer may be the patient, who is answering the question, "How long does a patient estimate it takes him to fall asleep after a dose of Drug A?" Another observer may be a technician, or nurse, or

doctor, who is answering the question, "How long does it take for Drug A to produce a behavioral state interpreted by an objective observer as sleep?" The two sets of observations need not coincide, since the questions are different, but they can both provide useful information about drug action and its evaluation.

Also, on occasion, it will be found that one set of data collected by one observer or group of observers is worthless, whereas the data provided by other kinds of observers are interpretable. In such a situation, multiple sources of data can avoid the disaster that would ensue if complete reliance were laid on a single observer. For example, let us imagine that some psychiatric patients are being followed both by interviewing psychologists and psychiatrists and by ward attendants filling out rating scales. If the attendants spend only short periods of time on any one ward or shift, and are frequently changing wards or the time during which they make observations (e.g., day to night), it may be almost impossible to make sense out of rating scales, since in one week the evaluation of a patient would be made by one observer, and the next week by a second observer. It is unlikely that different attendants will have exactly the same scale of values, and at the least one can expect a loss of precision. Each psychologist or psychiatrist, on the other hand, can interview the patient or test him at his convenience and presumably will maintain a more or less constant set of values throughout the course of the experiment.

BIAS AND CONTROLLED EXPERIMENTS

In the above paragraphs it was remarked that imprecise measurements might make it impossible to detect a substantial actual difference between drugs. An even more serious error, however, is to find and declare statistically significant differences that arise from sources other than true drug effects. Most patients with disease want to get better, and most investigators have some sort of prejudice about any given drug—usually in wanting to come up with successful results, but at times in the opposite direction. This enthusiasm (or lack of enthusiasm) must be allowed to diffuse itself out as equally as possible over the medications under study. The major reliance here must be placed on "blind" tech-

nique. Usually, this is "double blind," i.e., patient and observer are both unaware of the nature of a particular medication. At times, a "single blind" technique suffices, if the end point to be determined (such as death) is not particularly amenable to overstatement or understatement, or if the patient records the data himself under circumstances in which the experimenter cannot influence him.

Whether the treatments being compared are active drugs or active drugs and placebos, it is necessary (for successful deception) to have tablets or capsules or injections that are as indistinguishable in physical appearance as possible. The medications are then designated by code letters or numbers (preferably a different one for each patient) and the code is known only to certain individuals not directly concerned with the performance of the trial.

One misconception about placebos is the belief that they always fool the patient or the observer. With a medication that produces no effects, subjective or objective, other than the one under study, this is possible. Many drugs, however, do produce side effects. Most subjects who have never previously received an injection of morphine can quickly distinguish 15 mg. of this drug from an injection of saline solution. They may not experience euphoria, but they will very likely be dizzy, or nauseated, or sleepy. (Indeed, an experimental population that could not distinguish between the two would be useless for many purposes.) Therefore, if such a subject can voluntarily affect some measurement being studied, it is a simple matter to bias the results regardless of "placebo controls." It should be stated, however, that there are all degrees of recognition of medications. Thus a novice to narcotics can distinguish that morphine is "some sort of drug," but a "postaddict" may also recognize that it is morphine, which complicates matters still further. The perfect placebo, therefore, would be one that would mimic exactly all qualities and effects of the active drug except for the effect under study. Obviously, in many cases the achievement of the perfect placebo is an actual impossibility.

Extremists have suggested that even partial breaking of the double blind deception in this way destroys the usefulness of this stratagem. Fortunately, there are some happy facts about clinical

trials that negate much of this criticism: (1) it is rare for a drug to produce side effects in all subjects, (2) it is rare for a placebo *not* to produce side effects in some subjects, (3) the production of side effects (usually unpleasant) is not tidily correlated with desired therapeutic effect, (4) analysis can always be performed to investigate the possible correlation of side effects and therapeutic success, (5) coding can (and should) be done individually so that one code "break" does not bias all other subsequent observations, and (6) even an obviously detectable placebo can still serve as a control for spontaneous improvement.

There is a mistaken notion that placebos merely control "suggestion" or "suggestibility." There is no question that this is partly their function, although it is not true that symptoms can be improved by a placebo only if they are of psychologic origin. But placebos control something else. They also are used to control both external factors common to all groups (such as changes in milieu) and naturally occurring, that is, spontaneous changes in the course of the disease. Many processes improve without therapy of any sort. In a study of hypnotics in preoperative patients, it was found that 70 per cent of placebo-treated individuals fell asleep satisfactorily. This superficially implies a rather suggestible group. Quite the contrary seemed to be true, however, for the percentage of patients falling asleep successfully in a similar group *receiving no medication at all* (drug or placebo) was almost identical with that in the placebo group. The placebo rate in this instance was thus primarily a reflection of the ability of a group of such patients to fall asleep under certain specific conditions regardless of medication. Such a situation can be appreciated only if one contrives an experiment so that there are control periods (or groups) when nothing of any sort is given and that may be compared with a placebo-treated period (or group). Such a design involves more subjects, time, and effort, and may not be justified by the interests of the experimenter, but it behooves the latter to be cautious in his interpretations of "placebo effects" if he does not include an "untreated" group.

It has been suggested that it may be desirable to use "active placebos," i.e., substances that are devoid of therapeutic effect but that produce side effects mimicking those of the drug under study. When available, an active placebo may be a useful adjunct

to an experiment, but one is still well advised to include an "inactive" (or old-fashioned) placebo as one treatment, since it is not always possible to guarantee a priori that the drugs in the "active" placebo will be free of beneficial (or deleterious!) effect on the parameter under study. In many situations one may properly be interested only in comparing active drugs (e.g., a new drug with a standard drug) rather than drug and placebo; if the side effects of the drugs compared are similar, much of the above concern will disappear.

It has been asserted by some that the double blind technique in some way interferes with doctor-patient relationships, and by others that the technique in some way defeats the investigator and renders him unable to differentiate active from inactive drugs. This argument is apparently based on the observation that certain investigators are capable of discerning therapeutic benefit from an "active" drug only when they know what they are giving, and fail completely to discern such benefit when kept unaware of the nature of the medication. Reasons for this interesting state of affairs are not usually proffered (although certain explanations come readily to mind) nor is there any suggestion offered as to how one ought otherwise to control investigator bias, a proved hazard in clinical work.

In point of fact, of course, the double-blind technique has been used successfully on many occasions to distinguish between useful and useless drugs of many types. It seems to us that investigators should appreciate the following facts. First, the double-blind technique (even when properly used) is only one aspect of a good experiment and is not sufficient to guarantee the quality of the total design or its application. Second, the failure of a significance test to "establish a difference" between two drugs does not prove that they are the same; whether or not the difference is significant, an investigator of moderate inquisitiveness will want to calculate confidence limits to see how large or small the actual difference might reasonably be. In this way he may find that his experiment was too insensitive to detect the anticipated difference. (It would be to the investigator's advantage, of course, to plan his experiment in such a way as to avoid this outcome.) Finally, if an investigator already "knows" that a drug is active, and is unwilling to entertain evidence to the contrary,

there is hardly any point in doing an experiment to "determine" whether the compound is better than a placebo.

TEMPORAL EFFECTS; DRUG-ORDER INTERACTION

In many experiments it may seem desirable to test two or more medications on the same subject. There are two different kinds of reasons for performing such an experiment.

First, it may be suspected that Drug A is best for one kind of patient and Drug B for another, but the distinguishing characteristics of the two types are not known in advance. If this proved to be the case, one might undertake to investigate the differences between the two types of patients in order to learn how to classify individuals as A or B types before starting treatment.

The other reason for testing two or more medications on the same subject is to gain precision. The use of each patient "as his own control" is a special instance of the pairing or matching device often used to reduce experimental error. In this case matched pairs are created by pairing each individual tested at a given time with the same individual tested at a later time. If individuals tend to be more or less constant in a measurable characteristic over time but to differ from one another, the degree of difference in the performance of two drugs may be much more easily established with paired comparisons than otherwise.

However, if patients are to be studied with more than one medication, it is important to be aware of and make allowance for the possible effects of order of administration of drugs. For example, in postoperative pain the severity of discomfort changes with time, and the magnitude of effect of a drug given on the second day may be quite different from the effect it would have if given on the first day. Familiarity with the experimental conditions and the learning process may produce average results on the second trial quite different from those on the first trial. At times, the first medications given to patients in a trial (as in certain outpatient studies on coronary vasodilators) are most efficacious; at other times (as in postoperative pain), the opposite is true.

Because of these systematic differences between first and

second trials, the experimenter must first decide what quantity to use as his basic measure of response to treatment. A reasonable choice for the measure of the effect of Drug A, say (and the one most often taken), is the average of the response to Drug A when A is given first and the response to Drug A when it is given second. Of course, the experiment must be designed to present the drugs in both orders and the analysis must take account of this feature of the design.

When the investigator is seeking to distinguish individuals who respond favorably to one of the Drugs A and B but not to the other, his purpose is exploratory rather than confirmatory. He is looking for leads rather than for precise quantitative results, and any findings that suggest the existence of "A-type patients" and "B-type patients" would require checking in another trial of Drugs A and B separately in each type. Provided that the orders of presentation AB and BA are both used, it should be possible to distinguish the effects of two types of patients from simple learning effects.

Where responses to treatment are "irreversible" or where a patient can be studied once only, the self-controlled trial is obviously impossible. In addition, the therapeutic challenge is often variable from time to time in a given individual so that "each person his own control" at times becomes a quasi-meaningless charade. There are cases in which a comparison between different individuals may be far more useful than comparisons within individuals.

Also, an experiment requiring two trials on each subject is often much more difficult administratively than one requiring only a single trial for each. For example, a subject who finds the first treatment objectionable may refuse the second, and the experiment may thereby suffer a substantial number of dropouts. Such losses may, in addition to lowering precision, introduce biases in the comparison that are difficult to evaluate. (This problem is discussed in the section Statistical Analysis.) Although there are certainly instances in which the testing of each subject with both drugs may yield a substantial gain in precision, the problems introduced by such a design may be severe and, if the object is solely to increase precision, the actual gains to be anticipated should be calculated and weighed against the possible disad-

vantages. In most kinds of studies the saving in number of patients will probably be less than 50 per cent.

PAIRING; MATCHING

When subjects are variable in some characteristic that may affect their response to treatment it may be worthwhile to subdivide the group to be tested into subgroups that are more or less homogeneous and to make comparisons within subgroups. For example, in comparing the ability of two forms of tetracycline to elevate blood levels it may be advisable to group the subjects by weight. If we take this idea to the ultimate limit we will pick out subgroups consisting of two subjects each, that is, we would pair off the two heaviest individuals, assigning one chosen at random to Drug A and the other to Drug B, and likewise for the next heaviest pair and so on. If the response to treatment is correlated with weight, the pairing will result in greater precision for the comparison of the drugs. If we had three drugs to compare, we would match up triples instead of pairs, and similarly for larger numbers of treatments. This matching procedure has been found very useful in agricultural experiments; designs using this procedure are known as "randomized block" designs—a "block" in this case being the collection of individuals grouped together, and the "randomized" referring to the random allocation of treatments to the individuals within a "block."

Two factors are primary in determining the utility of matching in the clinical evaluation of drugs: the correlation of the matching variable with the measure of response, and administrative considerations.

It is sometimes thought that the existence of a well-established correlation between some individual characteristic and the response variable makes matching on that characteristic imperative. In fact, matching will be worthwhile only if the correlation is high.[5] More precisely, if we have exact matching and if the correlation between the matching variable and the response variable is R, the fractional reduction in the number of subjects required for equivalent precision with a purely randomized design is R^2. Thus, if the correlation coefficient is 0.5, a matched design could yield a saving of 25 per cent of the subjects required by a design

using purely random allocation. If R is appreciably less than 0.5, there is very little to be gained by matching. Unfortunately, the correlation coefficient is almost never known in advance, even approximately.

In contrast to agricultural experiments, the formation of "blocks," (i.e., matched pairs, triples, etc.) in a clinical trial may be administratively very difficult. Often the characteristics of the subjects are not known in advance of the start of the trial. For example, in studies of postpartum pain, subjects are accepted and assigned to treatment as they arrive and only one, or at most a few, will be under consideration at any one time. In such a case matching by the physical characteristics of the patient is practically impossible. In other cases such matching is possible, but time-consuming, and damaging to the adequate supervision of other details of the trial. Occasionally there are natural groupings that may conveniently be used as blocks. For example, in semi-private wards with two beds per room, one may choose the two patients in a room as a block on the ground that environmental factors may play some role, or simply on the ground that it is administratively convenient to do so. On the other hand, there are many situations in which the matching variable is subject to temporal change so that individuals who seem well matched when the trial begins are not at all well matched later on. In cases where substantial losses from observation can be anticipated, many matched pairs will be broken by loss of one member and the matching process becomes pointless. Sometimes there are individuals who are hard to match and are therefore dropped from observation. If there are many of these, the loss in precision may outweigh the gains from matching.

There are, of course, many variables that one might match upon, such as severity and duration of disease, age, etc., but to match for more than one or two variables poses practical problems that are often insuperable.

Unfortunately, there is little published evidence on the value of matching in clinical trials. As indicated earlier, the correlation between the response variable and the matching variable must be rather high (of the order of 0.5) before matching results in any great gain in precision. Thus it may be anticipated that in cases where good evidence of high correlation does not exist, there will

probably be only minor gains in precision as a result of matching. In view of our general uncertainty about the possible gains in precision, the decision whether to match at all, or on what variable to match, will generally be controlled by the problems of administration. If, in a given situation, it is as easy to match as not, then one may very likely wish to do it. If, on the other hand, matching is administratively difficult—and this is often the case—it should be done only if there is clear evidence that it will result in worthwhile gains.*

STATISTICAL ANALYSIS

If the experiment has been well designed and conducted and has not run into special difficulties, the statistical analysis should create no problem. However, there are a few basic principles that deserve consideration.

First, the nature of the analysis is entirely dependent on the design of the experiment. (This assertion may seem to be a truism, but it is surprising to see how often it is overlooked. For example, one can find many instances in which a sample of paired observations is analyzed as if the observations were independent.) Any experimenter would do well to be sure he knows how to design, make random allocations for, and analyze the simplest kinds of experimental designs—in particular the completely randomized and the randomized-block designs. If, for some reason, a more complex design is needed, he should seek the advice of a competent statistical consultant. It is true that the complex designs and methods for their analysis are carefully described in a number of texts, but the necessary assumptions are not always clearly set forth and, even when they are, the likelihood that the clinical study in question will not materially violate the assumptions is better decided jointly by the clinician and the statistician than by either alone.

Unhappily, it is likely even under the most favorable circumstances that a clinical trial will in one way or another give rise to

* In very small samples, there may indeed be some loss owing to matching on an irrelevant variable because of a reduction in degrees of freedom in estimating error. In samples where 20 or more pairs are involved, however, this loss will be trivial.

some special problems. Not all these can be anticipated, and again the advice of a statistical consultant familiar with the details of the study may be extremely valuable.

Perhaps the most common difficulty is the loss of some patients from observation. Such losses may occur for many reasons—uncooperativeness, toxicity of the drug, death, etc. This situation is generally covered in articles by a remark such as the following: "Seventeen patients failed to complete the course of treatment—fifteen on Regimen A and two on Regimen B. The analysis is restricted to the 113 patients who completed the treatment schedule." The analysis then presented usually takes no account of the lost patients. Now it may be that this type of analysis is the most reasonable under the circumstances, but such a study is by no means equivalent to a study that began with 113 patients and had no losses. The fact of losses introduces a new source of bias, possibly great enough to vitiate the results completely. Worse still, the experimenter may sometimes be unable to tell if his results remain valid or not. For example, if significantly more subjects are lost from the group treated with Drug A than from the group treated with Drug B, one may suspect that Drug A is in some way more objectionable than Drug B and, since more of the likely-to-be-affected group has been selected out (lost) from the group on Drug A, the remaining parts of the two groups are not strictly comparable, and no amount of statistical manipulation will make them so. In studies that use only one treatment for each individual one can decide in advance how to score lost cases; in general they will be given scores indicating unsatisfactory responses.

In doing this we slightly alter the question asked, since we were originally concerned with the effect of the drug when given as planned, and with the results observed according to plan. However, it is usually better to modify the question somewhat and get straightforward evidence on the modified question than to get biased evidence about the original question. There is no wholly satisfactory method for dealing with losses other than to avoid them. Thus, although complete freedom from losses is often an impossible goal, it is worth great effort and expense to keep the number of losses at the absolute minimum.

Now, supposing that the issues of loss and other kinds of bias are settled, what form should the analysis take? It is a pleasure to

observe that the habit of presenting only an analysis-of-variance table showing the existence of significant differences—an unfortunate practice common in some areas of research—is not frequent in the literature on drug trials. Although it is important to show that Drug A has been shown beyond a reasonable doubt (say at a 5 per cent or 1 per cent level) to be better than Drug B, this is far less than half the story. The most valuable information is that which specifies how much better A is than B, and this is best given in the form of an estimate with confidence limits. Even if the drugs are *not* shown to be different, confidence limits enable one to set bounds on how big or small the difference might be. More attention to this situation would avoid the common pitfall of concluding from a negative result that Drug A is in fact not appreciably better than Drug B, or that a previous experiment, yielding significant results, is necessarily in conflict with the present negative finding. A small experiment is likely to give negative results, that is, nonsignificant differences, but confidence limits so broad as to be consistent with very substantial true differences.

All too frequently investigators are led to hope that by making a sufficient number of measurements they will turn up something "significant." However, there are many instances in which at least several of the measurements are of serious interest. If each measure is analyzed by itself, two kinds of questions are likely to arise.

One may find that although no single measure shows a significant difference between treatments, all or most of the measures tend in the same direction. Alternatively, one may find that one or two measures show "significant" differences, but most do not, and one may question whether the multiplicity of measurements is responsible for the appearance of this "significant" result. How should one test for over-all significance of the difference between groups in such cases? The standard statistical technique for this purpose is that of Discriminant Function Analysis.[6] Here we find that linear function of the measures which most sharply distinguishes between the groups and, allowing for the freedom of choice for this function, we test its significance. (Help from a statistician is advisable for anyone attempting this analysis for the first time.) Although this procedure works well if we are concerned solely with establishing significance, the discriminant

function itself is rarely of clinical interest, and confidence limits derived from it will have little meaning. A more satisfying analysis results if one or two measures—possibly combinations of the primary measurements—are chosen in advance and receive top billing in the analysis. Such measures should reflect the areas of greatest interest. If they turn out to show only small differences one may be content to play down a significant difference (possibly caused by the multiplicity of measurements) if it relates to a comparatively obscure characteristic of the response.

This is not the place to discuss the details of calculation. Depending on the level of his background in statistics, the researcher may turn to Hill's monograph,[7] to any variety of texts such as that by Dixon and Massey,[8] or to the original, and in many ways still the best, general source, R. A. Fisher's *Statistical Methods for Research Workers.*[6] Here we will only remark that the restriction to normal distributions often cited can usually be ignored. Although the usual confidence limits are exact only if the distribution of observations is normal, the approximation is adequate for most purposes, even when the distributions are markedly nonnormal. Even binomial observations (success or failure) may be treated by standard methods, using the numerical value 1 for success and 0 for failure. In the ordinary two-sample comparisons this leads directly to the standard formulas. In the case of matched observations the techniques are not as familiar, but tests of significance are given by Cochran[9] and, for the special case of paired comparisons, by Mosteller.[10]

Finally, when more than two drugs are compared in the same experiment, it is wise to make some allowance for the fact that there is more play for chance events, and that, if we use the usual confidence procedures, the probability of making at least one erroneous statement in our comparisons of the several drugs is greater than 5 per cent. To compensate for this effect, the confidence limits may be slightly widened. Appropriate discussion may be found in a paper by Dunnett.[11]

CAUTIOUS GENERALIZATION

The final principle to be discussed is that of cautious generalization. The problem of transposition of data and conclusions from one experimental setup to others is a fundamental one, of

course, for many areas of scientific endeavor. The action of morphine in patients with increased intracranial pressure or severe pulmonary disease may be at least quantitatively different from its effects in normal individuals. The action of digitalis in patients with congestive heart failure is hardly the same as its effects in a normal individual. The pain of childbirth is in a number of ways different from the pain of chronic headache. Many such examples could be given to illustrate this point. Reports of the efficacy or inefficacy of a medication should always be qualified with such phrases as "under the conditions of the experiment," "with these dosages," "in these patients," and so forth. There are many instances of disagreement in the medical literature on results obtained with various drugs. Some of these differences are easily explained by experimental technique, dose, duration of study, and so forth. Many are not, however, and it would seem appropriate to be generous in such instances, assume the honesty of the investigators, and look elsewhere for explanations.

One possible source of error is the frequent use of volunteers as subjects for studies, the results of which are ultimately applied to more general situations. It is apparent that volunteers may differ markedly from the nonvolunteers in their own group, let alone from individuals in other groups. This is not to say that such studies are useless; indeed, they are the only studies possible in many circumstances. If the differences between volunteers and nonvolunteers are along a nonrelevant parameter, such differences are probably unimportant. One must, however, be careful to use such data for what they are worth and rely on subsequent confirmation in other situations before generalizing broadly.

One other source of error is the use of fractions of an experimental population chosen because of certain characteristics that make an experimental study easier for the investigator than if a random sample were chosen. An example is the choice of 52 patients of a group of 3,000 patients with angina pectoris for a study on vasodilator drugs. This small sample was chosen because its members responded consistently (from an electrocardiographic standpoint) to exercise and because they showed favorable response to nitroglycerin. Making a study on such patients is perfectly permissible, but it may not tell much about the average patient with angina pectoris, who obviously did not qualify for the study. In addition, the method of selection made

it inevitable that no drug studied would surpass nitroglycerin in efficacy, since the best that any drug could do was to equal its performance.

The most satisfactory situation with regard to a drug is to have its performance tested carefully by a variety of independent observers in different clinics with different types of patients under different conditions. If such studies are in general agreement as to the drug's efficacy, there is little room for doubt. On the other hand, a single negative or positive study, no matter how carefully performed, must always leave some questions unanswered.

REFERENCES

1. Mainland, D., and Sutcliffe, M. I.: Statistical Methods in Medical Research. II. Sample Sizes Required in Experiments Involving All-or-None Responses, *Canad. J. M. Sc., 31:*406, 1953.
2. Hammon, W. M., Coriell, L. L., and Stokes, J., Jr.: Evaluation of Red Cross Gamma Globulin as a Prophylactic Agent for Poliomyelitis. 1. Plan of Controlled Field Tests and Results of 1951 Pilot Study in Utah, *J.A.M.A., 150:*739, 1952.
3. Fisher, R. A., and Yates, F.: *Statistical Tables for Biological, Agricultural and Medical Research.* Oliver & Boyd, Ltd., Edinburgh, 1953.
4. The Rand Corporation: *A Million Random Digits with 100,000 Normal Deviates.* The Free Press, Glencoe, Ill., 1955.
5. Cochran, W. G.: Matching in Analytical Studies, *Am. J. Pub. Health, 43:*684, 1953.
6. Fisher, R. A.: *Statistical Methods for Research Workers.* 11th ed., Oliver & Boyd, Ltd., Edinburgh, 1950.
7. Hill, A. B.: *Principles of Medical Statistics.* 6th ed., Oxford University Press, New York, 1955.
8. Dixon, W. J., and Massey, F. J., Jr.: *Introduction to Statistical Analysis.* McGraw-Hill Book Co., New York, 1951.
9. Cochran, W. G.: The Comparison of Percentages in Matched Samples, *Biometrika, 37:*256, 1950.
10. Mosteller, F.: Clinical Studies of Analgesic Drugs. II. Some Statistical Problems in Measuring the Subjective Response to Drugs, *Biometrics, 8:*218, 1952.
11. Dunnett, C. W.: A Multiple Comparison Procedure for Comparing Several Treatments with a Control, *J. Am. Stat. Ass., 50:*1096, 1955.