REVIEW ARTICLE

PROBLEMS IN THE EVALUATION OF DRUGS IN MAN*

By Walter Modell, M.D.

Director, Clinical Pharmacology, Associate Professor of Pharmacology, Cornell University Medical College, New York

It is a standard cliché in the teaching of therapeutics, but it is nevertheless a fact, that every treatment should be a fresh experiment in therapy. The details of the clinical experiment and how it must be conducted to elucidate what it proposes to find out is, nevertheless, not ordinarily a part of our medical curriculum. However, therapeutics has now progressed to a point where the issues of what comprises the clinical experiment is a practical, indeed, a vital matter.

The reality and immediacy of this problem was brought home to us in New York when it was reported that 5 per cent of the patients in one of our outstanding teaching hospitals was there in consequence of a reaction to medication[1-7].

I find it difficult to plunge further into this argument without appearing to say that, despite the high quality of the technical knowledge and training which makes up modern medical education, I seem to think that physicians sometimes cannot tell whether they are helping patients by means of the drugs they give them or instead by some highly personal communication they make to the patient or that an apparent improvement is merely a chance development[8,9].

I am assuming that here, as in the United States, a physician's training provides him with a substantial background in the basic and clinical sciences and a well-developed skill in patient examination, and that the combination enables him to discover what is wrong with his patient, to decide whether his patient is getting better or worse, and to make a shrewd guess as to the outcome of the case. What is quite another matter, is that he probably has neither been taught nor encouraged to discover precisely why, as a consequence of all the factors which enter into the complex that make up the physician's ministrations, the patient does get better or does get worse. By this I mean to imply more specifically that the physician, who learned as a medical student to determine whether and how a drug raised or lowered blood pressure in a cat, is usually not prepared to say after the administration to his patient of the drug whose action in the cat is so well documented, whether his patient's blood pressure rose or fell or refused to do either as a consequence of the medication, the medicating, the medicator or as a result of some other circumstance. This, however, is the basic problem in the clinical evaluation of drugs, one which is seriously neglected in the training of the medical student.

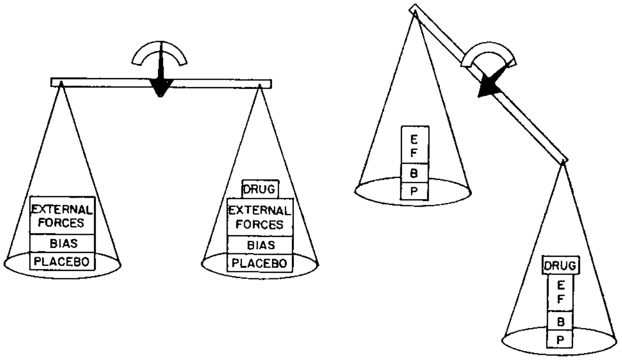
* *Based on a Special University of London lecture given at King's College, Strand, April* 30, 1959.

Nor is he taught how to examine a publication on therapy critically. Yet publication of a fallacious, poorly documented or inadequately analysed statement is almost always unfortunate for, regardless of its truth or substance, by virtue of publication alone, a medical statement, even a Letter to the Editor, acquires authority and, what is even more likely to lead to trouble is, that while it is relatively easy to have it published, it is far more difficult to erase a published blunder or delusion from the minds and memory of the medical profession, especially if it is a hopeful one. There are serious implications here, for regardless of what their real merit eventually turns out to be, drug manufacturers will continue to advertise the laudatory published statements long after substantial evidence to the contrary is adduced, to the confusion of the physician who tries to understand what is going on and to the misfortune of the patients of the physicians who do not. Unless the physician himself has some knowledge of the standards by which drugs may be properly evaluated and of the fundamentals of good design in drug evaluation, only the always well-intentioned but, alas, sometimes fallible, journal editor stands between him and reliance on shaky or spurious claims.

Criteria for separating substantial claims from the insubstantial must be set and methods of drug evaluation must be developed which not only differentiate between the good and the bad, but which also distinguish between the good and the better. This is not to say that there are yet no high standards and no good methods, for we have both, but the necessity for using them is not always recognised, and sometimes they are misused—viz, the large number of papers on new drugs containing claims which soon prove to be meretricious.

I would like to pursue, therefore, an examination of the many factors which influence patient response after the administration of a drug and which make it difficult to determine why he reacts as he does, and how one goes about the business of distinguishing between an alteration in the patient's physiologic state as a consequence of the direct effect of a medicament and as a consequence of one of the many influences in everyday life which impinge on physiologic function and which are generally briefly disposed of by our calling them chance occurrences. These must be distinguished if one is to know whether a particular medication is any better, or worse, than a simple lactose tablet.

The issues for resolution in the clinical evaluation of drugs are basically the same as for well-designed experiments in all other experimental disciplines and can be stated simply enough: identification and control of all the factors which may interfere with or assist in making observations and in collecting and evaluating data. This is what I should like to consider here in outlining the problems in clinical evaluation. For this purpose I propose to use as a model, scales which weight the evidence for and against drug action. In a proper clinical evaluation the effects of drugs *per se* are matched against all other influences which tend either to prevent the action of the drug from swinging the balance in the proper direction or to tip it in the other direction and simulate an expression of drug action and, in either case, to provide answers which are not spurious and due to some other active force, and which otherwise may be misinterpreted as being evidence of action or lack of action of the drug[10–18].



Factors which Influence Data in Clinical Evaluations

The factors which influence data in clinical evaluations, especially when the effects of drugs on subjective responses are involved, may be enumerated as follows: (1) the pharmacodynamic action, (2) the dosage, (3) the subject, (4) the controls, (5) placebo actions, (6) bias, (7) the forces extraneous to the experiment, (8) the collection of data and (9) the sensitivity of the method.

Pharmacodynamic Action

When objective measurement of effects is possible, pharmacodynamic actions present the least difficulty in their evaluation. When pharmacodynamic actions are potent, reproducible, and are not significantly influenced by psychic forces, evaluation is also relatively simple. The action of a mercurial diuretic, for example, lends itself to dependable and relatively precise measurement and a true bioassay may be relatively easily performed in man because a measureable loss of weight, due to the effect on the drug on oedema, can be precisely translated into diuresis, an action which usually is not appreciably compromised by external forces[19,20].

Drug actions which must be evaluated in terms of subjective responses, and especially those which are not in themselves impressive, are far more difficult to evaluate. Thus it is exceedingly difficult to demonstrate differences in the analgesic effect of drugs of the order of effectiveness of aspirin, while it is simple to prove that morphine is effective, and still easier to prove that general anaesthetics are even more potent pain-relieving agents[15,21,22].

The measurement of the useful actions of hypotensive drugs is difficult, even though blood pressure can be measured precisely, because of the tendency to wide spontaneous variation in blood pressure in patients with hypertension as well as because blood pressure is also readily altered by immediate circumstances, tension, position, strain and rest[23,24].

Each drug, therefore, requires careful consideration with respect to the method most appropriate to the observation and measurement of its action[25,30].

Dosage

The proper evaluation of drug action requires the use of the proper dosage. It is obvious enough that when dosage is too low, regardless of the pharmacodynamic actions or potency of the drug, clinical evaluation will not reveal any difference between the drug and a placebo, and when the dosage is too high any therapeutic effect will be obscured by the toxic effect. This is why we have so little precise and dependable information about the actual usefulness of the tranquillisers in the treatment of the everyday simple anxieties that we all suffer, while there is substantial information about their value in the treatment of schizophrenia. In the latter condition large enough doses are used to observe and estimate or actually measure effects. In the former we use token doses and hope for subliminal effects. These of course defy ordinary methods of estimation or measurement and, as a result, we not only do not know which tranquilliser is better for the run-of-the-mill anxiety but we also do not really know whether many are any good at all[28].

Neither toxic nor token dosage may be used in clinical evaluation; dosage must be carefully chosen. If a single dose is used it is usually preferable that it be one which, by preliminary examination, is found to be on the sensitive portion of the dosage-response curve of the drug. There are some designs which use threshold or ceiling doses, and although these often appear to have practical as well as logical advantages, a serious disadvantage lies in the fact that effects on the extremes of dosage-response curves are difficult, sometimes impossible, to evaluate precisely. The best way to compare drugs is to use a series of graded doses of each. This provides a more substantial basis for comparison than that of single doses, no matter how well-chosen the latter may be, for it also serves as a sort of built-in measure of the sensitivity and the discriminating powers of the method—an internal operational control.

The Subject

In much the same way that some species of laboratory animals are superior to others for particular experiments in the laboratory the choice of a suitable subject is often a critical matter for an investigation in man. Thus, while the best subject will tend to make the method more sensitive, unsuitable subjects may make the method so insensitive that it is unable to detect the particular drug action under investigation and, therefore, regardless of the effectiveness of the drug, provides only negative answers.

The argument that the patient with the disease for which the drug is ultimately intended is the best subject for the evaluation of the drug is not always correct. When he is the only possible subject, he is, perforce, the best subject, but there are also situations in which there is a choice, and sometimes he may be unsuitable. The choice rests first of all on the purpose of the investigation, whether it is to define the action of a drug in man, that is to say, a pharmacologic examination or to predict the value of a drug in the treatment of a particular disease, that is, a therapeutic evaluation. These have fundamentally different goals and it may not be assumed or even implied that the results of one type of investigation invariably applies to the other.

Some amplification is necessary. The investigation of the pharmacodynamic actions of a drug in man is a necessary preliminary to its therapeutic evaluation. Although the first may have implications of utility in particular disease states, indeed provide clues to therapeutic usefulness which are ultimately borne out by subsequent experience, sometimes the clinical trial of the drug fails to confirm the suggestive pharmacologic findings in man as well as in animals. Basic pharmacologic information on the effects of the drug in man is essential for its therapeutic exploration but only the therapeutic evaluation of the drug in the patient with the disease will provide the conclusive evidence of its value in that disease. Each approach serves a particular and indispensable purpose in the final evaluation of the therapeutic potential of the drug, but only rarely can both be carried out simultaneously. The therapeutic evaluation of a drug cannot, therefore, be carried out reliably in any subjects but those for whom its use is proposed. And if the evaluation is to have predictive value, the group of subjects must represent a fair, or random, sample of the patients who suffer from the disease.

But, the problem in the choice of subjects for the exploration or evaluation of the pharmacodynamic actions of a drug in man is a fundamentally different one[31–35]. Since the disease state *per se* is not being examined here, in so far as it is possible, the complex of factors which make up the disease, as well as all the other known and unknown factors which influence and which may obscure the examination, measurement and evaluation of changes in man's functional state as the result of drugs, must be identified and eliminated or controlled in some way.

Whereas the utility of a diuretic in the treatment of congestive failure can be predicted only after an experience with a representative sample of patients with congestive failure, the effects of diuretics on particular electrolytes, or the influence of electrolyte load on diuretic action can be much better explored in healthy subjects. This is not to say that the normal subject is always the best for pharmacological investigation. There are situations, for example, the effect of a drug on cardiac arrhythmias which demand the afflicted patient for the pharmacologic as well as the therapeutic exploration. There are also situations in which there appears to be a choice. For example, drugs for motion sickness may be investigated in voyagers aboard ship or in healthy subjects using the Barany test[36–38]. Additional examples will not further clarify the idea that, in the exploration of the pharmacologic action of drugs in man, there are drugs which are so highly specific that only the patient with the disease for which the drug is intended can be the subject, and there are drugs for which both the patient and the normal man are suitable subjects for evaluation, in which case the choice may be based on convenience alone, and finally, there are drugs for which the normal man appears to be clearly the more desirable subject for pharmacologic evaluation than the patient.

Subjects must be selected in a manner which insures the ability of the group as a whole to discriminate between "active" and "inert" agents; that is to say that the subjects must be sufficiently sensitive to the drug action under investigation to be able to appreciate differences of practical significance. While it is usually desirable that the group of subjects be a representative one, above all it must be sensitive in the sense that it must be able to detect an effect should it develop. Where large numbers do not provide this by chance alone, efforts may have to be made to impart this quality to the group by deliberate elimination and selection.

The question may arise whether to use in-patients or out-patients as subjects for a particular study. The former provide the advantages which ensue from a relatively protected and more or less constant environment while the latter provide the advantage of large numbers at low cost as well as little work in patient care. However, there are other features which bear on the choice of one over the other which are not so apparent. The patient in the hospital has little incentive to physical or mental activity and, as a matter of fact, he is likely to spend a great deal of this time in bed and to be especially pleased if he can escape some of the ennui of his hospital stay by napping a large part of the day as well as sleeping through the entire night. The out-patient, on the other hand, is far more active and, through exposure to the usual activities of living, is regularly challenged throughout the day by one stress or another. I have observed that the latter patients appear to appreciate the actions of sedative drugs more than the former and, I am inclined to believe that this may be due to the difference in the daily living experience and frequency of stress challenge in each case. This difference may also be important in other areas of drug evaluation.

Subjects likely to give misleading results must not overwhelm the group. Where sex makes a difference, the group may be selected accordingly. That dosage-response may be unusual in the very young must sometimes be taken into account. The potentially high side-effect liability of patients with renal disease, and the elderly in general, must be considered. In studies involving subjective criteria, exceedingly phlegmatic subjects desensitise the method by not reacting, while exceedingly neurotic and over-reactive or suggestable patients tend to compromise the sensitivity of the method through wide swings of mood and attitude as the result of placebo as well as of active medication. In no event should the number of unusual, abnormal, or resistant subjects be excessive and, there are suggestions in the literature that, given the proper basis for their elimination and, with the standards established at the outset, any or all such may properly be removed before the study is begun.

Knowledge of participation in the exploration or any kind of special examination of a new drug, to say nothing of the highly charged concept of being the subject of an experiment, seems to exert special psychic pressure on the subject and make him act in something other than his usual manner, to be overly introspective, to try to help the investigator or, in some instances, to react with fear or resentment. Some patients will tolerate discomforts with less difficulty when they know they are participating in an investigation; on the other hand, some may be less tolerant. But however they may react, information that they are participating in an experiment alters the subject and, thereby, adds another element to patient reaction and, in that way, desensitises the method. In some trials it is not possible to keep the fact from the subject but, generally, the optimal subject is one who does not know that he is participating in an experiment[13].

This leads to the important practical problem of the acquisition of subjects for studies, always an important and often a difficult and limiting one, especially when large groups are necessary. It would seem at first that the most reasonable and just method is to call on volunteers. But the volunteer is not a normal subject; he is a *volunteer*, and he may present the problems just cited based on this fact. In addition, the highly co-operative and willing volunteer may provide a set of highly personal psychological problems. In an interesting analysis of this problem by Lasagna and Von Felsinger, it was shown that the volunteer is an unusual subject and often clearly undesirable on that account[39].

Collecting an indiscriminate group of subjects for drug evaluation and dealing with the problems of chance differences in subjects merely by the process of randomisation simply equalises the influence of a large amount of dead weight for, in order to overcome the spurious swinging of our balance, an equal number of unsuitable subjects are put on both pans. However the choice of the subject is finally made, it should be made with the idea in mind that his proper selection has a great deal to do with the ultimate sensitivity of the method.

The Controls

In no discipline can an experiment be pursued without a control. Even those experiments which the experimenters, themselves, presume to be without controls, are nevertheless, controlled. The control is the only basis for a comparison and, thus, there is a control implicit in every judgment on a drug. Somehow or other the statement is made or implied that one drug is better or worse or equal to some other drug or to no treatment at all. The only important question about the control is, therefore, not whether one has been used, but rather whether it is sound as a basis for the comparison on which the judgment is based.

It is tempting to use the easy way out; to use what has been called the historical control, that is to say, a recounting of previous personal or recorded experience as a basis of comparison. Often, this is not recognised as a control—and understandably so—for it is a treacherous one. No method of drug examination is more likely to lead to erroneous conclusions. It has none of the safeguards provided by other controls, the elimination of placebo effects and of bias and the natural course of events and chance as the effective forces in apparent drug action. It also fails utterly to provide comparable bases for examination of control and experimental groups. Only in the case of the disease in which an irrevocable or unquestionably characteristic course has been established, and particularly when the condition is rare, is the historical control justified.

The classic experiment employs separate groups for control and treatment, but this provides significant data useful for statistical validation only when the groups are formed by random selection and when they are extremely large or extraordinarily homogeneous[40]. In man, this poses a serious practical problem. Since the human equivalent of litter-mates in the laboratory, that is, sets of identical twins, are too rare to be hoped for, proper matching of control and experimental groups is essential. When the number of subjects is sufficiently large and patients are put in either group by a process of random selection, chance alone will insure a proper balance. Matching of control and experimental human subjects by actual selection, however, is an insuperable task.

An alternate method which is acceptable, is to give each patient the medicaments and placebo serially so that each subject serves as his own control. This is the so-called cross-over design. When the condition of the patient during the control and the experimental periods is similar, if not identical, there is a valid basis for comparison. In clinical evaluations this plan is often the more satisfactory because smaller numbers are needed, hence the study can be conducted in less time and with less cost. On the other hand, when the subjects have progessive disease, it may be inappropriate to compare the effects at two stages of the disease and, in such a case, only separate control and treated groups are acceptable. Another difficulty with the second plan is that each subject is required to participate in the entire course of the experiment and, in the usual study, a discouragingly large percentage often fail to do so, thereby increasing the number of original subjects necessary for the study and necessitating a design which will not collapse if a subject defaults.

Placebo Actions

The term "placebo" has taken on many implications not within the philologic meaning of the word, as for example, "negative placebo actions". As the word is currently used in clinical evaluations, it includes a large series of visceral, somatic and psychic responses to the physician, to his presence, to his words, to his ministrations, and to his medications. Such an action is inherent in all medications regardless of whether they are useful, hazardous, impotent, inert, unpleasant, inadequate, or inappropriate or for that matter, new or old, as long as the medication is prescribed by the physician himself. To be certain that these are not the only effects of drugs under examination, it is essential to have a basis of comparison of drug effect with "pure" placebo effect. To provide this, one must also give an inert medication which is otherwise identical with the drug under examination. It has been suggested by Gaddum that such an inert material is more properly called "dummy" than "placebo"[41]. But, "dummy" or "placebo", an inert control medicament must be given in all clinical studies to distinguish between the effects of the *act* of drug administration and the pharmacodynamic effects of the medicament itself[42–52].

Such a measure provides the only defence against the suggestion that results reported after the administration of a drug are due to placebo actions rather than to the pharmacodynamic action of the drug itself. In using placebo for control it is well to recognise that in the analogy provided by our chemical balance the placebo is not restricted to one pan and the drug action to the other. Since placebo action is inherent in every act of medicating by the physician there is, in fact, placebo in both pans, and the scales merely measure the difference between them. That is to say, placebo effect is being exerted on both pans at all times and the only measurement is of that which the drug may provide in excess of its inherent placebo action and, in the event that the two do not summate, it measures merely drug action which is not masked by placebo action.

Bias

In addition to the considerable psychic force exerted by the administrator of a drug if he be an accredited member of the medical profession, the so-called placebo action of drugs, the hopes of the patient and the therapist alike, as well as any bias either may have with respect to treatment or experiment, also exert considerable force on patient response after the administration of drugs and, therefore, on the art of the collection of the data. Therefore, these must also be reckoned with in all clinical evaluations.

The patient may want to get better to the extent that he is inclined to see good effects after administration of any new medication, and colour his subjective responses accordingly. On the other hand, he may find compensations in his illness and wish to preserve his complaints, hence be inclined to depreciate pharmacodynamic effects, sometimes miscalled "negative" placebo action. The physician's knowledge of the nature of the medicament is exceedingly important, for regardless of how much he tries, if he knows the identity of the medicament, he may nonetheless relay this information to the patient. In addition, his understandable bias may lead him to interpret, hence modify, data along preconceived lines as he collects it and, as a result, there may be substantial apparent effects from accumulated bias. The importance of the unconscious communication of the physician was proved in a study in which patients could not detect the difference between placebo and aspirin unless the physician prescribing them himself knew which was which[53]. The standard procedure is not only to use placebo and drug which are identical in appearance, but also to keep both the physician and the subject ignorant of which is in use at the time of the prescribing, questioning, and examining.

Of all the devices to insure valid data, none seems to have attracted so much attention and to have evoked so much controversy as this so-called double-blind technique. It is a philosophically sound, as well as practical, control device to use in clinical evaluations to deal with the tendency of conscious and unconscious bias to obscure and distort the effects of drugs. It deals with the influence of the physician's bias (his professional purpose to help his patient as well as his preconceived ideas and prejudices about the medication and his unconscious communication to his patient) on his observations by blinding him, that is, keeping him ignorant of whether he is giving or has given his patient placebo or active drug. So also are the effects of the patient's bias (his hopes and his anxieties) on his estimates of his subjective responses dealt with by blinding him, that is, keeping him ignorant of whether he is receiving or has received active drug or inert tablet of identical appearance; hence the double-blindness. What is important to remember in this connection is that the myopia and astigmatism of the physician and the subject due to bias are corrected only in the sense that blinding will compensate for them, and that nothing has been added to increase the visual acuity of either observer.

The question arises whether the double-blind control must invariaby be used in clinical evaluations. This has been reviewed by Gold[54–57]. Much as we favour its use, occasionally it does not seem feasible. There are instances in which the drug promptly reveals itself by its unmistakable side effects and automatically removes one or both blindfolds by an action other than the one for which the drug is being examined. How could one use the double-blind control in a study comparing a general anaesthetic and a placebo? Page and Corcoran point out that, although the physician may remain blind, with many hypotensive drugs the patient's blindness soon vanishes because the drugs have obvious effects in addition to the hypotensive action and, in such a case, only the physician remains blind[16,58]. Despite such difficulties, there are sometimes devices for circumventing them.

Perhaps because of its dramatic qualities, the double-blind technique has attracted widespread attention. It has also apparently been widely assumed that it is a complete method of evaluation it itself, instead of being only a control device. Indeed, it is often called the double-blind *test*. Many seem to believe that all that is necessary for a good clinical evaluation is to use the double-blind technique and, regardless of all other details, inevitably and automatically, the results obtained will be valid. Since it is relatively easy to apply the double-blind technique, some are using this as the only control measure in the design of their clinical evaluations. In many publications it is stated in the title itself that the double-blind control was used, not only as if the use of a control in an experiment was so exceptional as to be worthy of special mention, but also as if to indicate in advance of reading that a special type of insurance has been taken out to guarantee that the results about to be recounted were bound to be above reproach[53,59–63]. The enthusiasm for this catchphrase has grown so that "triple-blind studies" have already been described and I read recently that "a five-way blind cross-over was carried out"[64]. It would seem that the fascinating notion is developing that if there is sufficient blindness it will ultimately lead to some sort of occult vision.

Unfortunately for those of us involved in clinical evaluation of drugs, the double-blind control does not provide either a simple or a complete solution to our problems any more than a control is all that is needed by experimenters in any other discipline for the complete design of an experiment. Nor does it eliminate bias as an element in the method; it merely deals with it by equalising its effects so that, as weighed in our scales, unequally distributed bias alone will not account for the apparently decisive evidence.

Forces External to the Experiment

There are a large number of extraneous influences which affect the state of the subject's physical, functional and psychic state—a change in the course of his illness, a happy experience, a lost job, a family quarrel, a seasonal allergic state, a change in the weather, a turn in world affairs —which may also influence his response to drugs. That is to say, there are changes in the subject's state which develop after the administration of a drug and therefore may appear to be responses to drug action. These may be both objectively and subjectively recorded responses[65,66].

To a limited extent, these factors may be reduced by removing the patient from his home to the protective atmosphere and routine of the hospital where those influences that disturb are likely to remain relatively constant compared with the much more labile scene in the usual home under the best of circumstances. This is not always so, however; some patients may find the hospital environment disturbing rather than protective and restful and, for them, this change makes them poorer subjects for drug evaluation.

The tendency of external forces to influence response can be dealt with by prescribing medication and placebo by a scheme of random distribution so that the disturbing forces affect the apparent response to placebo and drug alike, and being spread equally they appear to favour neither. It is to be pointed out that, by this scheme, the influence of external events on the apparent response of patients to drugs is not eliminated, but is spread equally, that is, divided equally between both sides of the scales, so that the scales do not swing by virtue of extraneous forces alone. It is also to be noted that where it is possible to reduce these forces, for they can never be entirely eliminated, less is then placed in each pan, and, to that extent, the scales are less burdened with dead weight.

Collection of Data

When objective measurement is possible and differences can be expected to be large, there is relatively little difficulty in the collection of data. However, when the patient must communicate his subjective experience, it is quite another matter. There are few experimental procedures in which so vital a part of an experiment as the collection, storage and interpretation of observations is left in the hands of an interested, biased, and untrained assistant, yet this is precisely what is being done when the patient-subject is asked to report and summarise his experiences after a period of medication. How can he help having his total recollection affected more by recent events than those farther back in his memory.

What happened two days or two weeks ago, or even two hours before questioning can be coloured by the patient's annoyance over some action of the receptionist in the out-patient waiting room.

The daily report-card system was designed to deal with this defect in the interval report system by decreasing the interval between the recording of respones to one day[67]. When these records are kept by the patient himself, however, the improvement in methodology is more apparent than real. At the very best, it substitutes a 24-hour recall for, say, a two-week recall, but it still has the same fundamental deficiency. In practice, it is often no better than the longer interval system. I have observed patients filling out their "daily" report cards while sitting on the benches waiting to be called to turn in cards that were supposed to be filled out faithfully each night during the two-week interval between examinations. While this method of data harvesting obviously supplies more data than a longer interval report system, it has not been subjected to an analysis which proves that the data itself or the answers derived through its use are more substantial. Other improvements have been suggested; having the subject mail a postcard each night or telephoning a report each night, but even these compromises leaves the data subject to the caprice of the patient for too long[68].

Not so long ago, two of us separately examined the effects of aspirin in the relief of pain, each using a different method of collecting data, but in all other details, a similar design[15]. The drug and the doses were the same, there was the same use of placebo and double-blindness, randomisation, and so on. One of us used a two-week report card system while the other questioned the patient during the course of the action of the drug.

The first method provided a large number of cards, hence a large amount of data. Statistical analysis of the data provided by this method showed no significant difference between the analgesic effects of placebo and aspirin, hence the answer that aspirin was without effect on arthralgic pain. In the second method, the statistical analysis of the data not only indicated a significant difference between the analgesic action of aspirin and placebo, but it also described parameters for aspirin action, a fine dosage-response curve and a curve of action. The only difference between the two methods was that the second collected patient responses directly from the patient as the course of drug action developed while the first permitted the patient to keep the data in his possession, subject to all the events of life which affected him, until he communicated it to us at some later date.

In general, any device which leaves the discrimination of the reaction to drugs at the mercy of patient recall provides the setting for outside influences on, and tampering with, data. Every effort should be made to minimise the period between the experience with the drug and the recording and collecting of the data; the data should be taken out of the patient's hands as soon as possible and, thereby, kept as nearly as possible in its original form.

We may now consider the effect that the collection of data has on our balance. As we allow time to alter data through our control process of randomisation, we once more burden both sides of the balance with yet more dead weight, weight immaterial to the problem at hand, which is the weighing only of pharmacologic actions.

Sensitivity of the Method

As in methods for chemical analysis, every design for drug evaluation requires a demonstration that its sensitivity is appropriate for the distinction it chooses to make. A scale of sensitivity should indicate first, the ability of the method to detect the drug action *per se*, and second, the increments in effects which it can distinguish. Without the first, a negative answer cannot be defended, and without the second, a positive answer has no quantitative meaning.

A negative answer is valid only if it is demonstrated at the same time that the method can also appreciate the effects of a standard and similar drug. The ability of a method to discriminate increments in effects can be indicated by its capacity for dosage-response when a series of graded doses of the standard or experimental drug is used. Such a scale of sensitivity gives positive results quantitative meaning.

The internal control just described is not only essential to establish the propriety as well as the sensitivity of the method as such, but it may not be eliminated in further clinical evaluations with the same method (as can be done with impunity in some other disciplines, for example, methods of chemical analysis can usually be repeated many times without rechecking accuracy and sensitivity). The need for a continuing test of sensitivity comes not from instability of the method, but from variations in the population of subjects which, at one time or another, may make them more or less able to discriminate between "active" and "inert" agents.

INTERPRETATION OF THE DATA

The current literature has placed overwhelming emphasis on one item in the design of a proper method of clinical evaluation, the double-blind technique, and relatively little on all the others. Unless all means of control are considered and given their proper importance in the design of clinical evaluations, improper and erroneous conclusions will be drawn from data that are supplied by studies which use the double-blind control just as well as from those which do not.

Which way the scales, which we have used as the model for methods of drug evaluation, swing, that is, whether drug action *vis-à-vis* chance is favoured, depends, of course, on the relative weight in one or the other pan. Whether the swing is meaningful or misleading depends on whether the weight which swings it is due to a specific action of the drug or to any of a myriad of forces which influence man's behaviour and his mental, physical and visceral activity. When such a model is used, one way to prevent swings of the balance by factors other than the intrinsic pharmacodynamic action of the drug itself, is to accept and spread their influence equally on both sides of the balance, thereby causing no disturbance in balance by their weight. Nothing more than this is accomplished by the control devices of placebo, double-blindness and randomisation; they merely prevent chance or biased swings of the balance in either direction.

What is rarely taken into account in clinical evaluations is how much weight is necessary to make the balance swing at all, that is, the basic sensitivity of the method. Whatever the original sensitivity of the balance, consider what is done with it in the usual design for clinical evaluation. Consider that the scales are not empty at the outset of the evaluation, merely in balance. We place equally on both pans, placebo action of drugs, bias, the influence of diverse extraneous factors such as weather political events, family stresses, and a number of other vagaries of human experience that tend to mould or alter man's functional state and his response to drugs. It is to be repeated, these are not removed as interferences; they are preserved and spread equally over both pans of the balance by the process of randomisation and by the control of double-blindness. The balance is thereby dead-weighted with a large amount of material which is foreign to the specific problem at hand. No matter how sensitive originally, such a procedure makes the balance less sensitive just as an analytical balance sensitive to a fraction of a milligram under usual conditions is no longer swung out of balance by milligrams when dead-weighted with several kilograms on each pan.

Ultimately, therefore, the sensitivity of a method of clinical evaluation is a function of the relative weight of the pharmacodynamic force under investigation and the weight of the nonessential interfering forces which are treated by equalising them; the greater the former with respect to the latter the more sensitive the method and, vice versa, when the latter becomes relatively heavier, the method becomes proportionately less sensitive. To the extent that dead-weighting grossly desensitises the scales, this process can lead to erroneous interpretations in the sense that it indicates no differences whenever it is used to weigh forces which it can no longer sense.

Of the disturbing factors already discussed, some are subject to choice and, in that sense, the disturbance may be eliminated. Thus it may be possible to choose the proper dosage range, the most sensitive subjects, and the appropriate control. Some factors which cannot be eliminated may be modified; the removal of the patient from the home to the constant environment in the hospital may reduce the external variables. The collection of data on the spot reduces the treachery of patient recall. Finally, there remain some disturbances which cannot be reduced, removed or modified; bias and placebo actions. For those which cannot be eliminated there is only the double-blind control and randomisation to spread the prejudicial factors equally.

In the studies with aspirin cited briefly above, the reason for the discrepancy in the results by the two methods used is to be found in their relative sensitivity; one method provided a false negative answer because, in order to prevent a false positive answer, interfering forces were dealt with only by the desensitising process of acceptance and balancing out, thereby becoming too insensitive for its task, whereas the second was sufficiently sensitive and gave a precise positive answer because it had eliminated the interferences to a practical degree.

A great danger in interpreting clinical evaluations lies in failure to recognise the meaninglessness of the negative answer when the method is not sufficiently sensitive for the purpose. The failure to demonstrate statistically significant differences between drugs or treatments is frequently misinterpreted to mean that no real differences exist. However reasonable the latter may seem from the data, an assertion that the drug or treatment effects are identical is not easily proved. Statistical tests of significance merely tell us the likelihood that whatever differences are noted in the data are due to chance. Thus, when the differences are statistically significant we are assured that this is unlikely to be a chance occurrence, and we may then, with a measurable degree of confidence, rightly or wrongly (for the statistics themselves do not validate the basic data), ascribe the results to essential differences in the effects of the drugs. Differences which are statistically insignificant could result simply from an inadequate trial or from an insensitive method of evaluation which statistical analysis may not indicate.

It is well to remember that statistical analysis proves nothing about the original validity of the data—it is merely a device for establishing the betting odds on the reproducibility of the results obtained by the same method, the predictability of similar conclusions with future experience under the same conditions. Statistical prognostication is always based on the assumption that the data used were worthy of collection; statistical analysis of poor data is tantamount to attempting to make a silk purse out of a sow's ear. Only when the design provides built-in controls, showing an ability to discriminate meaningful effects or to show graded effects with graded doses of the drugs, can any valid inferences be drawn from negative results (that is to say, statistically insignificant diffrences or, if you will, significant indifferences).

It should be made clear also that, although statistical procedure presently seems to have assumed an especially prominent position in reports on drugs, fundamentally this is not at all new. As with the use of controls, no matter how an experiment is planned, how the terminology seems to intrude, or how the results are expressed, statistical analysis is inseparable from clinical evaluation of drugs. It is a biologic fact that all physiologic reactions and failures to react exhibit some degree of individual variability and, as a consequence, any statement about the pharmacologic or therapeutic action of a drug has implicit in it the statement that this is not a chance occurrence. It is, therefore, a statement based on either a calculation or a guess of statistical significance; the only question which remains is its quality and its applicability.

However the experiment is designed, if the signicance of differences that are indicated by the data is to be established with a degree of assurance, the data must almost certainly be subjected to statistical analysis.

It is good practice, therefore, to plan the collection of data in such a way as to simplify subsequent analysis and interpretation[69–79]. This is not to say that the mere statement of statistical significance insures correctness of their interpretation. If the data are inappropriate or improperly collected, as illustrated by the results of the study with aspirin, despite their statistical significance their interpretation may be erroneous.

CONCLUSIONS

Clinical evaluations are so beset by external disturbing forces that every possible control measure must be applied if valid and durable results are to be obtained. It has been pointed out that the selection of the proper dosage range is vital and that the selection of the proper subject is equally critical in the design for clinical evaluation. As far as possible all external disturbances must be eliminated. Data must be collected promptly and before any tampering has occurred. Treatments must be randomised. In addition to the use of the placebo control, the double-blind control should also be used whenever and wherever it is feasible. There is no conceivable disadvantage in the application of the double-blind control, only protection against spurious data, but it must not be used as a means of avoiding the elimination of bias and interfering psychic factors. I would like to emphasise as strongly as possible that its use will not validate otherwise poorly designed experiments. While it will prevent false positive interpretations, used in a poorly designed experiment, it will not prevent a false negative interpretation.

Each clinical evaluation must be sensitive enough to detect what it proposes to discover, and each experimental design must have built into it an indicator that it is capable of such detection. A negative conclusion is without merit unless there is incorporated in the clinical evaluation a demonstration that the method is competent to indicate a positive effect when it is present, i.e., an internal control. It is suggested that in clinical evaluations another, demonstrably effective, drug always be used in addition to the placebo control, to indicate this essential competence of the method.

Beyond this, there is the problem of the sensitivity of the method, the increments in effect which it can distinguish. Few clinical evaluations indicate what differences in effect they can discriminate. Yet in evaluations in all other disciplines, it is standard procedure to provide such a scale. Clinical evaluations cannot escape this requirement; the complete clinical evaluation must include a built-in sensitivity scale, and, through the use of graded doses, a demonstration of the increments in pharmacodynamic effect which the method can distinguish. When differences between standard and unknown or placebo are indicated, the sensitivity of the method to distinguish differences is thereby at hand to indicate the quantitative significance of the differences.

The definition of the effects of many drugs and the proof of the superiority of one drug over another require investigational designs which are based not only on the principles laid down here but which are also designed with due regard to the particular drug, the particular subject, and the particular circumstance under which they must be conducted. There is yet no standard method—there are basic requisities, essential controls, and some well-established procedures—but each different pharmacodynamic action of a drug may need a different subject, a different control, a different circumstance, or a different design for its proper evaluation.

REFERENCES

1. Barr, *J. Amer. med. Ass.*, 1955, **159**, 1452.
2. Friend and McLemore, *New Engl. J. Med.*, 1956, **254**, 1223.
3. Kirsner, *Ann. int. Med.*, 1957, **47**, 666.
4. Moser, *New Engl. J. Med.*, 1956, **255**, 606.
5. Rising, *Postgrad. Med.*, 1958, **24**, 200.
6. Schiffrin, *ibid.*, 1958, **24**, 305.
7. Wooley, *Perspectives in Biol. and Med.*, 1958, **1**, 174.
8. Coleman, Menzal and Katz, *J. chron. Dis.*, 1959, **9**, 1.
9. Dowling, *Arch. intern. Med.*, 1957, **100**, 529.
10. Goodwin and Rose, *J. Pharm. Pharmacol.*, 1958, **10**, Suppl., *24T.*
11. Lasagna and Meier, *Ann. Rev. Med.*, 1958, **9**, 347.
12. Menzel, Coleman and Katz, *J. chron. Dis.*, 1959, **9**, 20.
13. Modell, from *The Relief of Symptoms*, Saunders, Philadelphia, 1955.
14. Modell, from *Drugs of Choice*, C. V. Mosby Company, St. Louis, 1958.
15. Modell and Houde, *J. Amer. med. Ass.*, 1958, **167**, 2190.
16. Pannekoek, *Postgrad. med. J.*, 1957, **33**, 396.
17. Sabshin and Ramot, *Arch. Neurol. and Psych.*, 1956, **75**, 362.
18. Travell, *Amer. J. phys. Med.*, 1955, **34**, 129.
19. Modell, *Ann. int. Med.*, 1944, **20**, 265.
20. Modell, Gold and Clarke, *J. Pharmacol.*, 1945, **84**, 286.
21. Houde and Wallenstein, *J. Amer. geriatrics Soc.*, 1956, **4**, 167.
22. Wallenstein and Houde, *Fed. Proc.*, 1953, **12**, 377.
23. Murphy and Schulz, *Postgrad. Med.*, 1956, **19**, 403.
24. Shapiro, *J. Amer. med. Ass.*, 1956, **160**, 30.
25. Ambrus, Ambrus, Bauer and Noell, *J. Pharmacol.*, 1957, **119**, 129.
26. Gottschalk, Kapp, Ross, Kaplan, Silver, Macleod, Kahn, Van Maanen and Acheson, *J. Amer. med Ass.*, 1956, **161**, 1054.
27. Hill, *Brit. med. Bull.*, 1951, **7**, 278.
28. Hoch, from *Drugs of Choice*, C. V. Mosby Company, St. Louis, 1958.
29. Mushin and Mapleson, *Brit. J. Anaesthesia*, 1957, **29**, 249.
30. Raymond, Lucas, Beesley and O'Connell, *Brit. med. J.*, 1957, **2**, 63.
31. Lasagna and Imboden, *Fed. Proc.*, 1956, **15**, 451.
32. Lasagna, Von Felsinger and Beecher, *J. Amer. med. Ass.*, 1955, **157**, 1006.
33. Laties and Weiss, *J. chron. Dis.*, 1958, **7**, 500.
34. Loomis and West, *J. Pharmacol.*, 1958, **122**, 525.
35. Von Felsinger, Lasagna and Beecher, *J. Amer. med. Ass.*, 1955, **157**, 1113.
36. Gay and Carliner, *Bull. John Hopkins Hosp.*, 1949, **84**, 470.
37. Glaser and Hervey, *Lancet*, 1951, **2**, 749.
38. Report of Study by Army, Navy, Air Force Motion Sickness Team. *J. Amer. med. Ass.*, 1956, **160**, 755.
39. Lasagna and Von Felsinger, *Science*, 1954, **120**, 359.
40. Clark, *Edinb. med. J.*, New Series, 1935, **42**, 1.
41. Gaddum, *Proc. R. Soc. Med.*, 1954, **47**, 195.
42. Beecher, *J. Amer. med. Ass.*, 1955, **158**, 399.
43. Beecher, *Amer. J. Physiol.*, 1956, **187**, 163.
44. Beecher, *J. Amer. med. Ass.*, 1955, **159**, 1602.
45. Beecher, Keats, Mosteller and Lasagna, *J. Pharmacol.*, 1953, **109**, 393.
46. Fischer and Dlin, *Amer. J. med. Sci.*, 1956, **232**, 504.
47. Friend, O'Hare and Levine, *Amer. Heart. J.*, 1954, **48**, 775.
48. Houston, *Ann. int. Med.*, 1938, **11**, 1416.
49. Lasagna, Laties and Dohan, *J. clin. Invest.*, 1958, **37**, 533.
50. Moyer, *Arch. intern. Med.*, 1956, **96**, 608.
51. Wolf, *J. clin. Invest.*, 1950, **29**, 100.
52. Wolf and Pinsky, *J. Amer. med. Ass.*, 1954, **155**, 339.
53. Batterman and Grossman, *ibid.*, 1955, **159**, 1619.
54. Cornell Conference on Therapy, *Amer. J. Med.*, 1954, **17**, 722.
55. Cornell Conference on Therapy, *ibid.*, 1947, **2**, 296.
56. Cornell Conference on Therapy, *N.Y. St. J. Med.*, 1946, **46**, 1.
57. Gold, *Amer. J. Med.*, 1952, **12**, 619.

58. Corcoran, Dustan and Page, *Ann. int. Med.*, 1955, **43**, 1161.
59. Aravanis and Luisada, *Ann. int. Med.*, 1956, **44**, 1111.
60. Bello and Turner, *Amer. J. med. Sci.*, 1956, **232**, 194.
61. Bepler and Rogers, *ibid.*, 1957, **234**, 459.
62. Hailman, *J. Amer. med. Ass.*, 1953, **151**, 1430.
63. Koteen, *Ann. int. Med.*, 1957, **47**, 978.
64. Abstract, *Antibiotic med. and Clin. Therapy*, 1958, **5**, 615.
65. Wolff, *Ann. int. Med.*, 1947, **27**, 944.
66. Wolff, *Proc. Assoc. Res. nerv. and ment. Dis.*, 1950, **29**, 1059.
67. Greiner, Gold, Cattell, Travell, Bakst, Rinzler, Benjamin, Warshaw, Bobb, Kwit, Modell, Rothendler, Messeloff and Kramer, *Amer. J. Med.*, 1950, **9**, 143.
68. Chernish, Gruber and Kohlstaedt, *Proc. Soc. exp. Biol., N.Y.*, 1956, **93**, 162.
69. Armitage, *Amer. J. Public Health*, 1958, **48**, 1395.
70. Bross, *J. chron. Dis.*, 1958, **8**, 349.
71. Bross, *Ann. int. Med.*, 1955, **43**, 442.
72. Fisher, from *Design of Experiments*, Oliver and Boyd, Ltd., London, 1949.
73. Hume, *Lancet*, 1957, **2**, 1049.
74. Jellinek, *Biometrics*, 1946, **2**, 87.
75. Luykx, *J. Amer. med. Ass.*, 1949, **141**, 195.
76. Mainland, *Ann. rheum. Dis.*, 1955, **14**, 337.
77. Mainland, *Amer. Heart. J.*, 1958, **55**, 644.
78. Marshall and Merrell, *Bull. John Hopkins Hosp.*, 1949, **85**, 221.
79. Symposium, *Biometrics*, 1952, **8**, 206.

Lasagna, *J. chron. Dis.*, 1955, **1**, 353.
Lasagna, Mosteller, Von Felsinger and Beecher, *Amer. J. Med.*, 1954, **16**, 770.
Page and Corcoran, *Circulation*, 1956, **14**, 868.
Pepper, *Amer. J. med. Sci.*, 1943, **206**, 703.
Woolmer, *Proc. R. Soc. Med.*, 1959, **52**, 98.