Journal of

CHRONIC DISEASES

VOL. 1 APRIL, 1955 NO. 4

Original Communications

THE CONTROLLED CLINICAL TRIAL: THEORY AND PRACTICE*

LOUIS LASAGNA, M.D.

BALTIMORE, MD.

From the Department of Pharmacology and Experimental Therapeutics and the Division of Clinical Pharmacology of the Department of Medicine, The Johns Hopkins University

(Received for publication Nov. 3, 1954.)

The doctor of today is under constant bombardment with claims as to the efficacy of drugs, old and new. It is difficult, if not impossible, to read a journal, attend a medical meeting, or open the morning mail without encountering a new report on the success or failure of some medication. The clinician who would avoid nihilistic rejection or trusting acceptance of all such claims, or capricious decisions as to their merits, is well advised to adopt a yardstick, a set of criteria, that will improve his chances of making sound evaluations. The investigator is equally well advised to do so. The present review will present one such set of guiding principles. In so doing it will pursue two main lines of emphasis. On the one hand, fundamental principles will be accented, not only to provide a framework for subsequent discussion, but to distinguish clearly between the important primary considerations and less essential embellishments and refinements of technique. On the other hand, these same principles will

*Presented on Oct. 30, 1954, at a Fourth Year Therapeutic Conference, The Johns Hopkins University School of Medicine, Baltimore.

be scrutinized in detail and the complexities in their application in actual experiments described. This second line of emphasis is needed to avoid painting a picture of deceptive simplicity and to support the thesis that a doctrinaire insistence on literal and complete adherence to the principles *under all circumstances* can be achieved only by a cavalier dismissal of useful data or by the practice of self-delusion.

It is possible to summarize the fundamentals of the clinical trial[1-4] in the form of five principles. These are: (1) the cooperation of individuals with appropriate skills at all temporal levels of planning and execution; (2) randomization; (3) the "double blind" control; (4) statistical treatment of data; and (5) cautious generalization. Let us consider each principle separately.

A. THE JOINT NATURE OF THE CLINICAL TRIAL

Ideally, the clinical trial represents a fusion of talents, with different individuals contributing to the efficiency and quality of the experiment. This is merely the old principle of employing specialists when needed. Abuse of this principle is inadvisable at most times, but is especially foolhardy when an extensive, expensive, time-consuming project is planned. Although the make-up of such a team should obviously be appropriate to the experiment, a typical group might include a clinician, a pharmacologist, and a statistician (or multiples thereof). There are certain functions performed by the entire group, such as formulating a precise statement of the purpose of the clinical trial. Let us examine some of the particular functions of each member, however. It is currently fashionable in some circles to consider the clinician member of the team as some sort of minor excrescence, "a fifth cousin about to be removed." Such an attitude is arrant nonsense. The clinician is still in most such investigations the prime mover. It is generally he who is aware of the diseases requiring better therapy, of the symptoms inadequately treated. The clinician is also responsible usually for delineating the criteria for success or failure, and for devising methods for the gathering of pertinent data. It is the clinician who provides the all-important background for the clinical trial—knowledge, accumulated over the years, concerning the process under study. An example of the usefulness of this function may help to clarify the point. If one considers the field of antihypertensive drugs, it is apparent how helpful the body of knowledge on hypertension can be in setting up a clinical trial. How much more difficult the task would be, for example, did we not know that elevated blood pressure is a nonspecific manifestation of many disease processes; that even within the field of "essential" hypertension the prognosis varies from decades of unimpaired well-being to the brief downhill course of "malignant" hypertension; that the crucial body areas in this disease are the heart, the brain, and the kidneys. All this information enables the members of the team to focus on the appropriate goals of therapy, and points up the importance of such factors as patient selection and duration of study.

In addition, the clinician is at times in the enviable position of being able to make a real contribution to therapeutics without much outside assistance. I have in mind such disease processes as tuberculous meningitis, or acute staphylococcal septicemia, or acute leukemia, where past experience indicates a uniformly poor prognosis in the untreated case. If, then, a physician uses Drug A on a patient known to have such a disease and the patient is quickly and completely restored to permanent good health, the event is one of major importance. To criticize such a felicitous physician, who calls the case to the attention of his colleagues, because of a lack of placebo controls, or for not having a large series of similar cases, represents a naive display of pharmacologic Philistinism.*

The clinician should also bear the primary responsibility for reminding his associates of the ethical considerations involved in a proposed clinical trial. Let us suppose, for example, that two antibiotics are available, and that the laboratory evidence is conflicting as to whether the two drugs, when used together, are synergistic or antagonistic. In the case of an infectious disease for which there is no effective therapy, it is easy to justify a trial of the combination; the patients have nothing to lose and everything to gain. If no such conditions obtain, there may still be a disease process for which therapy is available but by no means optimal. Pneumococcal meningitis might be considered such a disease; subacute bacterial endocarditis another. Here one may be willing to entertain the risk of some loss of potency in a combination of drugs on the chance that a substantial synergism between the drugs might be demonstrated, with lifesaving effect for some patients. Let us take, however, a disease which is 100 per cent (or essentially 100 per cent) curable with presently available antibiotics. The trial of a new antibiotic on such patients, desirable as it may be from a variety of standpoints (e.g., a search for a cheaper or less toxic drug), must take into account the fact that the patients to be so treated are not likely to appreciate the importance of the data contributed by their own misfortunes to the welfare of unborn generations. Such patients have a good deal to lose by participation in such studies and very little to gain. If it is decided that the points at issue are of such importance that the trial has to be performed, adequate safeguards must be written into the clinical trial so that any patient who seems to be doing poorly on a particular regimen be switched without delay to standard forms of therapy. Such provisions have a good deal of nuisance value for investigators, and have at times ruined whole experiments, but only individuals devoid of empathy can argue that they are never necessary.

The contributions of the other members of the team should not be minimized. The pharmacologist can be most helpful in deciding whether the available data on the pharmacologic action or the toxicity of a compound in animals are such as to warrant its trial in human beings. He can also utilize his specialized talents to analyze what is known about the absorption, distribution, excretion, and fate of the drug, so that rational dosage schedules can be set up. If data on human beings are lacking, he can be invaluable in planning the cautious preliminary work that is so essential.

The statistician's aid is frequently indispensable. It is he who is most attuned to the need for safeguards against bias, for devising an experiment which is

*This is not to deny, however, that happy is the physician who can confirm his own work and twice happy is he who is confirmed by others.

free from systematic errors of allocation or assessment. He can comment critically on how pertinent the intended measurements are to the objective of the trial. An expert statistician can frequently increase the efficiency of a trial, so that the same amount of information can be elicited from a smaller number of patients or procedures. In evaluating data at the end of the experiment he can ensure that the appropriate analytic techniques be applied. Failure to utilize the aid of such specialists can be disastrous, as evidenced by a fiasco of some years ago. An ambitious polio vaccination program was responsible for the vaccination of some 450 children, while 680 in the same community remained unvaccinated as controls. An epidemic visited the area that year, but there was not a single recognized case of polio in either group. This was entirely consistent with the records of previous years, which indicated that only two cases could be expected in such a total number of children. A few simple calculations done (alas!) in retrospect revealed that meaningful data could have been obtained only by having 7,500 to 25,000 children in each of the groups.

Having stressed the *cooperative* nature of the trial, one should also emphasize the phrase "at all temporal levels of planning and execution." Too often, for example, the statistician is called in at the end of a trial, in the hope that the chanting of a few mathematical formulas or Greek symbols over the corpse of an ill-planned experiment will restore the breath of life to the unfortunate victim. One eminent statistician has advised his fellows to eschew any part of a clinical trial in which they have not been active participants from its very inception. Such a remark may be a little extreme, but its point is well taken. In addition to participation at beginning and end, the members of the team should consult with each other during the progress of the trial. It is often difficult or impossible to foresee all possible contingencies, and it may be wiser to scrap an experimental design part way along, if serious errors in planning or execution manifest themselves, than to continue doggedly to the bitter end and find that the therapeutic egg which the investigators were so tenderly hatching had been in reality a billiard ball all the while.

Having said all this, it is only fair to add that the ideal conditions described previously are not always achievable, even with the best of intentions. "Full use of a biostatistician," for example, is easier said than done. Such individuals are relatively few in number and their time is limited even if funds are available to pay their salaries. The outlook is not as bleak as it may seem, however. It is perfectly feasible for a clinician with a working knowledge of statistics, and an appreciation of the principles to be described, to plan and execute a sound clinical trial, and to analyze the data satisfactorily. He may be somewhat inefficient in his methods but is not so likely to commit a basic error fatal to the entire enterprise.

B. THE PRINCIPLE OF RANDOMIZATION

It is difficult to conceive of a rule more important for the clinical investigator than that of avoiding bias in the allocation of cases to different treatment groups. If two drugs (or two procedures, or two surgical techniques) are to be compared, it is obvious that the elimination of all variables other than the treatments in question is a prime goal. The conscious or unconscious distribution of one type of case (the "sickest" ones, for example) to a particular treatment group will introduce another variable into the design which will usually result in a completely useless experiment.* Such a procedure presents potentially different therapeutic challenges to the two treatments. Since it is difficult or impossible to gauge the degree of bias so introduced, or even its direction at times, the result of the study is usually to leave the experimenter no better off than when he began. The best way to avoid such bias is to assign cases by some technique which eliminates the possibility of prejudice. One may employ a randomized sequence (prepared in advance), or flip coins, or use even or odd last numbers on hospital admissions, or alternate cases. The important consideration is not which method is employed so much as the degree to which it succeeds in achieving an unbiased distribution. Thus alternate assignment of admissions to two groups is satisfactory only if the person in charge of admissions is unable to funnel cases selectively into one or another treatment group by the order of admitting them.

It is unfortunate but true that the easiest way to allocate cases in more complicated situations is invariably the least desirable way. For example, if two hospitals are to compare the efficacy of two treatments it is much simpler, administratively, to give all the patients in Hospital Y one drug, and all the patients in Hospital Z the other drug. Such a plan may be disastrous, however, because of the excellent chance that the type of patients, the nursing, the medical staff will differ enough in the two institutions as to affect seriously the results *exclusive* of any possible difference between drugs. During the wartime evaluation of motion sickness preventives it was found that merely the amount of ballast carried by a troopship could affect the incidence of nausea and vomiting in the passengers. In addition, it was established that the section of the ship is an important determinant of the incidence of motion sickness. Thus, not only is it vital to compare medications on soldiers on the *same ship*, but it is imperative that all the treatments under study be allocated at random to the persons *within any given compartment* on the ship.

This principle of randomization is abused with such frequency that examples are easily found. One type of error is illustrated in a certain study of the effectiveness of anticoagulant therapy in the management of patients with cardiac infarction. Patients in this trial were assigned to the drug or nondrug groups on the basis of the day on which they were admitted to the hospital. Thus all patients admitted on odd days were allocated to the anticoagulant group, whereas those admitted on even days were given no anticoagulants. Since this plan is easily recognized by referring physicians, it becomes ridiculously simple to send in patients on a day when the type of therapy is such as the physician considers most appropriate for his patients. Thus, a physician who considers anticoagulants a dangerous and unnecessary measure could avoid sending his patients

*A classic example is the Lanarkshire milk experiment, where one-half the children got milk, but the teachers were permitted to assign the neediest ones to the milk group.[5]

in on odd days, so long as their clinical situation did not demand immediate hospitalization. Another physician who might be "sold" on anticoagulants could adopt just the opposite procedure, or at least attempt to get his "sickest patients" in on "anticoagulant days." The potentialities for bias are obvious. In this study 589 patients were treated and 442 patients were not. Thus, patients were probably not randomly selected since the excess of treated cases is greater than should be expected in a completely random selection of cases. One possibility is that physicians were in general eager to have their patients receive the "advantage" of anticoagulant treatment, and tried to refer patients in on treatment days. Since a critically ill patient would probably be hospitalized as rapidly as possible, it is also conceivable that there may have been a higher percentage of sicker patients in the untreated group.

Another type of error is to use as "controls" all patients who refuse a particular medication or procedure. Thus the postoperative history of a large series of hypertensive patients subjected to sympathectomy has been compared with the course of a large series of patients to whom the operation was tendered but who refused it. Even if one accepts the author's conclusion that the subsequent life history of the operated group was more favorable than that of the unoperated group, it is impossible to say whether the "improvement" is actually due to the surgery, or to the fact that the course of patients who refuse sympathectomy (for whatever reason) is basically less sanguine than that of their more pliable fellow sufferers.

Although the principle of randomization has thus far been discussed only in terms of cases allocated to different treatment groups, it is equally important for the order of administration of treatments to the same patient. For example, for each patient who, serving as his own control, takes Drug W for a period, then Drug Z for a similar period, it is essential for another patient to have this order of drugs reversed. The importance of the order of administration may be illustrated by a study of the effects of hypnotic drugs on psychomotor performance, on the morning after drug administration. One half of the subjects received the placebo on the first night of study, and pentobarbital on the second night, while the other half had this order reversed. The reason for this precaution was not only that the attitude of a subject might change on repetition of the experiment, or that he might become accustomed to the new environment, but that it is well known that most psychomotor tests show a significant "learning effect" with repetition. In order, then, not to underestimate or overestimate any difference between drug and placebo, no one treatment should be allowed to be always first or always second. In this situation, the "learning effect" was appreciated in advance. Regardless of whether the order of administration has been previously demonstrated (or can be reasonably expected) to be important in determining results, however, the wisest procedure is to assume that it will and to proceed accordingly.

A second example can be culled from the literature on coronary vasodilators. One group of investigators observed that outpatients suffering from angina pectoris usually improved during their first few weeks in the clinic regardless of what they were given. Although these investigators did not pursue this observation to its logical conclusion (they always gave the placebo first), it is clear that randomization of order of treatments would be the preferred procedure.

The use of the principle of randomization eliminates the need for separate listing of another principle—that of concurrent comparison. This latter concept is designed to eliminate differences between groups arising from the fact that they were studied at different periods of time. It is very unwise, for example, to compare the results of treatment with veratrum in a group of hypertensives studied in 1954 with an untreated group studied in 1934. Even if no specific measures against hypertension were available now, it is possible that a group of patients treated today would do better in any event because of other advances in therapy since 1934. (Obviously a patient with hypertension is less likely to die of infections today than he was twenty years ago. Similarly, if he goes into shock for any reason today, he is in a much better position, because of the widespread availability of blood and blood substitutes, than he was in 1934.) If the error is compounded by utilizing the 1934 data of another clinic in a different part of the country or the world for comparison, the additional possibility of differences in diagnostic criteria or patient selection (already somewhat present in the first situation) is introduced to wreak further havoc with the validity of the investigator's comparisons.

C. THE "DOUBLE BLIND" TEST

The third major principle is directed to the elimination of "errors of assessment." It is commonly referred to as the "double blind" or "double blindfold" test and consists of attempts to keep both patient and observer unaware of the nature of the medication being given. Thus, if one group of patients is to receive a potent drug and another group is to receive placebos, the patients and observers are kept in ignorance as to which group is which. If the same patient is to receive, on different occasions, a drug and a placebo, neither he nor the person responsible for evaluating the results of therapy is informed of which is the control period and which is the drug period. The bias which these precautions are primarily designed to eliminate is that attributable to "patient enthusiasm" or "investigator enthusiasm." Most patients with disease want to get better, and most investigators are anxious to come up with successful results. Both patient and investigator, therefore, may be tempted (on a conscious or unconscious level) to record improvement in symptoms under treatment. This enthusiasm must therefore be handled by allowing it to diffuse itself out as equally as possible over the active and inactive medications.

There are some situations where prejudice of this sort is very easy to introduce, and others where it is very difficult. For example, pain is known to be present or absent in a patient only if he tells us so. Pain is also composed in part of a psychologic ("emotional") reaction to certain neurophysiologic stimuli. In such a situation, subjective reaction is all-important, and it is easy to see how the results of medication can be affected by suggestion or wishful hinking. On the other hand, if the end point in a clinical trial is survival or death, it is unlikely that there will be much understatement or overstatement on the part of the patient or observer.

The usual way of avoiding errors of assessment is to use placebos. Such pharmacologically inert substances (which some prefer to call "dummy" tablets or injections) are usually described as "indistinguishable" from the medication under study. By this is usually meant that the two medications look, smell, and taste "alike." These active and inactive medications are then designated by code letters or names and their true nature is known only to certain individuals not directly concerned with the accumulation of data. These codes are preferably changed at frequent intervals so that the results of previous trials cannot influence subsequent data.

So far, so good. Let us examine the placebo somewhat more critically, however, since it and "double blind" have reached the status of fetishes in our thinking and literature. The Automatic Aura of Respectability, Infallibility, and Scientific Savoir-faire which they possess for many can be easily shown to be undeserved in certain instances.

First of all, 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 (nonpathologic) origin. But placebos control something else. They also are used to control naturally occurring, that is, spontaneous, changes in the course of disease. Many processes improve sans therapy of any sort. In a recent 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 which 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.

A second misconception about placebos is the belief that they always fool the patient or the observer. With a medication which produces no effects, subjective or objective, other than the one under study, this is possible. Many drugs, however, *do* produce side effects. A good-sized nitroglycerine tablet, for example, will hardly be confused with a lactose tablet by a reasonably alert person once he has had it under his tongue for a few minutes. Most subjects who have never previously received an injection of morphine can quickly tell 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 which 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 morphine from saline solution, but a postaddict can also recognize that it *is* morphine, which complicates matters still further. The perfect placebo, therefore, would be one which would mimic exactly all qualities and effects of the drug under study ***except*** for the effect in question. Obviously, in many cases the achievement of the perfect placebo is an actual impossibility.

The placebo can also fail miserably in eliminating "negative" bias, that is, tendencies on the part of a patient or investigator to deny any effect of a drug, or to deny differences between drug and placebo. Thus, a patient receiving morphine and placebo alternately for pain can, by the simple expedient of calling every dose of medication effective (or ineffective), mask the differences between the most potent and least potent analgesics available. Similarly, let us suppose that a neurologist is convinced that the injection of the stellate ganglion with procaine is a ridiculous treatment for cerebrovascular accidents. Regardless of the fact that someone else sees to it that alternate cases are treated with saline solution and procaine, and that the neurologist is unaware which patient got which, if the neurologist in question is in charge of evaluating the therapeutic results, he can bias the data beautifully by merely writing "no significant improvement" in all cases. The net result will be no difference between drug and placebo. The safeguards against this latter kind of "negative bias" are first of all the predominant tendency, previously mentioned, toward "positive" bias, and second the fact that a failure to demonstrate differences between treatments can only be interpreted as "no difference was demonstrated" ("not proved"), rather than "no difference exists" ("not guilty").

A word should also be said about certain difficulties arising from the use of placebos because of either practical or ethical considerations. If one is trying to evaluate an analgesic drug, for example, the inclusion of placebos in the experimental design is difficult to rationalize from the patient's standpoint, because drugs of established potency are available which will do a reasonably good job of relieving symptoms. Fortunately, this fact can be utilized in the design in a way which will at least partly solve the problem of controls. One can use a standard drug (morphine, for example) and compare the new drug with it at several dose levels. The answer will thus be to the question, "How good is the drug in comparison with morphine?" rather than to the question, "How much better than a placebo is this new drug?" In such a study, every other dose can be of the standard medication so that even if the new drug is ineffective, the timing of doses can be such that a patient does not go very long without getting an effective medication. Such a procedure not only provides comfort for the patient, but also maintains rapport with the medical and nursing staffs responsible for the patients, who soon become aware of difficulties if multiple placebos are given. With a disease like tuberculosis, one of a variety of available drugs can be employed as similar "standards" to serve as yardsticks instead of placebos.

Such techniques are not perfect solutions because it is often extremely interesting and important to know whether a drug is any better than a placebo. In addition, the presence of considerable numbers of "placebo reactors" may result in an underestimation of the optimal dose for the population at large. Further, a number of reports now indicate that the separate handling of data from "placebo reactors" and "nonreactors" may bring out differences between drugs not obvious in the unselected data.[6,7] It has been suggested that a possible solution to the latter problem may lie in the focusing on patients with symptoms of severe degree, since there is evidence that the patients with more severe pain, for example, are less likely to get relief from a placebo than those with less severe pain, and it seems reasonable that severity of other symptoms may likewise be inversely related to placebo success rates.[8] (The extreme of this approach has already been discussed, and is exemplified by such diseases as tuberculous meningitis.) There are dangers in this approach, however. If one studies a new diuretic drug in patients with "intractable" congestive failure which is refractory to the usual therapeutic procedures, a successful outcome is most impressive. A negative result is less easily interpreted. One cannot say that the new drug is completely without effect, because such patients are by definition refractory to digitalis and mercurials, which would thus also qualify as "inactive drugs" in these same patients. Therefore, a negative result in such a difficult therapeutic situation should usually be followed up by studies in less severely ill patients. It is rare to come up with a superdrug, and it would be unfortunate to miss a compound of moderate potency.

Finally, an example should be given of a trial that was conducted without the benefit of a "double blind" control of all aspects, but which nevertheless yielded useful data. The study by the Medical Research Council of Great Britain on the treatment of pulmonary tuberculosis with streptomycin was performed by allocating patients at random to two groups, one group of which received streptomycin, while the other group received only bed rest. This second group was not given placebos, and the physicians in charge of the two groups obviously knew which patients were receiving the drug and which were not. The major safeguard in this study was the reliance on objective criteria, such as fever, sedimentation rate, weight, x-ray changes, and survival rate. As already mentioned, the latter criterion seems hardly liable to biased recording. As an added precaution in studying the x-ray changes, however, the roentgenograms were interpreted by physicians who were unaware of the names and treatments of the patients. Since both the x-ray results and the mortality were strikingly different in the treated and untreated groups, there was little question that a new therapeutic agent was at hand. A purist might say that it is possible that the mere suggestion of potent therapy involved in being stuck four times a day for four months might have produced the beneficial results. To interpret a decrease in mortality from 27 to 7 per cent as a placebo effect in this situation seems a trifle captious, however.*

*Dr. Paul Meier suggests that unfortunately this is just the sort of remark often made to justify a poor experiment. He is, of course, entirely correct, which points up how difficult life can be.

D. STATISTICAL TREATMENT OF DATA

The fourth major principle refers to the analysis of data in a sophisticated fashion so that one can answer such important questions as "What are the chances that the observed differences between treatments may have been due to chance alone?" or "How reliable are my estimates of the potency of these drugs?" Let it be emphasized that a flagrant disregard of one or more of the principles previously described usually renders it unnecessary to apply *any* statistical techniques to the data. For example, if a series of patients is subjected to portacaval shunts, and the first fifty are performed under ether and the next fifty under cyclopropane, it is unsophisticated to make a detailed analysis of the results in the two groups with the notion of comparing the relative safety of the two anesthetics. There are obviously other variables involved with the passage of time (increasing experience and skill of the surgical team, differences in patient selection, and so forth) that would render it impossible to state that any significant differences between groups were actually due to the anesthetic agents. In such a situation it is sounder merely to tabulate the incidence of deaths, complications, therapeutic results, and so forth, and present the results as "our experience during the first two years" versus "our experience during the last two years."

For those of us who are not professional statisticians, it is always most reassuring to have expert help with the final analysis of data. Frequently, one reads papers in which the use of inappropriate statistical techniques has either failed to show significance where such exists or has "demonstrated" significance where none actually exists.

As an example of the latter type of error, one can cite the failure of investigators to make an over-all chi square test or analysis of variance on the total data to find whether individual comparisons of pairs of treatments are warranted. Even if one compares treatments *of equal effectiveness*, there will be a spread of "cure rates" which will vary over a range the size of which is partly determined by the number of treatments. It has been calculated that if one compares the highest and lowest "cure rates" in such an experiment, "significant" differences "at the 5 per cent level" will be "demonstrated" 13 per cent of the time if 3 treatments are used, 40 per cent if 6, and 90 per cent if 20 treatments are run! Obviously, then, to study 5 drugs and 2 placebos and then say that the "best" drug differs from the placebo "at the 5 per cent level" really means little unless one runs an over-all chi square first.*

An example of the inappropriate evaluation of data where just the opposite effect is obtained has been given by Mosteller.[7] If 100 patients each receive 2 drugs at different times, and one analyzes the following incidence of nausea by an ordinary 2 by 2 contingency table:

*As in all areas of human endeavor, there is disagreement among statisticians as to the preferred method of attacking problems. The performance of an "over-all" test followed by t tests is a time-honored and popular procedure, but recent workers (Scheffé, Tukey, and so forth) have proposed substitute tests which seem more reasonable or sophisticated to segments of the profession.

	NAUSEA	NOT NAUSEA	TOTAL
Drug A	18	82	100
Drug B	10	90	100
Total	28	172	200

the results are not statistically significant. Such an analysis, however, is not only predicated on independent samples, and therefore incorrect in these matched data, but is also inefficient since it does not make use of the fact that each person who is nauseated after one drug is more likely to be nauseated after the second drug. Actually, we are not interested in people who are either nauseated with both, or with neither. The only information on differences between the drugs is derived from those individuals who have nausea with one but not with the other. Analyzing the data from this standpoint, we find that 9 of the patients had nausea with both drugs, 81 with neither, 9 after Drug A but not Drug B, and 1 after Drug B but not Drug A. A new table can thus be set up, as follows:

		DRUG A		
		NAUSEA	NOT NAUSEA	TOTAL
DRUG B	NAUSEA	9	1	10
	NOT NAUSEA	9	81	90
	TOTAL	18	82	100

Analysis of these figures by a simple formula gives highly significant differences.

A word should be said about two diametrically opposed tendencies which I think are equally mistaken. Both are the result of an overzealous worship of the magic "$p = 0.05$." This latter figure means that the differences would have been observed only 5 times out of 100 due to chance alone, and that there are 95 chances in 100 that the observed differences are real ones. This common standard has been chosen because it seems like a reasonable compromise between the error of saying something is significantly different when it is not and of saying that something is not significantly different when it is. Let us examine some ways in which this value can be abused. Let us say that as a result of hundreds of observations, one drug is found to be a few per cent more effective than a placebo in relieving cough. The p value for this difference may be < 0.05, but what real meaning does this have? How important is it to know that a drug is "microscopically" better than no drug at all? On the other hand, suppose that one makes a small study on two antitussives and both turn out to be significantly better than a placebo, but in comparing the two drugs against each other, Drug A just misses being significantly better than Drug B (p of 0.06, for example). To dismiss this difference as completely meaningless seems rash. If possible, and if it were important enough, the experiment should be repeated with larger numbers. If the experiment cannot be repeated, and there are no other data available for coming to a conclusion on the relative merits of the two medications, one would be justified in using the *apparently* better of the two.

Statistical techniques are essential in problems of "estimation." If an experiment is run on a diuretic drug, it is one thing to know that it can cause patients to lose fluid, but it is equally or more important to be able to state *how much* fluid can be expected to be lost by, let us say, 95 per cent of patients. For such an estimate we are concerned with putting limits ("confidence limits," so-called) on our estimate of the fluid loss in the population studied. Obviously, a drug which will cause almost all patients to lose 2.5 to 3.5 pounds differs from one which may produce responses varying from 0 to 20 pounds. Predictability of response is usually of considerable importance in the evaluation and use of a drug, and statistical analysis provides a measure of predictability, whether one is concerned with the magnitude of effect of a given drug, or the magnitude of the difference between two drugs.*

E. CAUTIOUS GENERALIZATION

The final principle to be discussed is the principle 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 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 phrases such 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 which has been emphasized recently[9] is the frequent use of volunteers as subjects for studies, the results of which are ultimately applied to more general situations. It is quite apparent that volunteers may differ quite 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. 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.

*For example, an experiment may show two drugs to differ in mean effectiveness by 20 per cent, but it is most helpful to state that from the data one is almost certain that the "true" difference is not less than 10 or more than 30 per cent.

One other source of error is the use of fractions of an experimental population chosen because of certain characteristics which 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 nitroglycerine. 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 outperform nitroglycerine, since the best that any drug could do was to equal its performance.

Obviously, 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 question unanswered.

FINAL COMMENTS

These principles, then, seem the crucial ones for conducting clinical trials or evaluating their results. It is hoped that the picture drawn seems neither too confusing nor overly simple. The principles are simple but their application is frequently not. As in so many fields, a middle course is actually the wisest one. In the field of the clinical trial, the Golden Mean is not "a virtue flanked by two vices" but the only rational approach. I should like to discuss briefly some additional personal sources of humility in this field.

First, there is the fact that frequently a careful clinical trial merely substantiates what has been previously determined by uncontrolled experiments. This is not always so, but many instances can be cited. For example, the carefully done analgesic studies of the group at the Harvard Medical School Anesthesia Laboratory have for the most part defined equianalgesic doses of drugs which are in agreement with those already determined by clinicians at the bedside. Such valid observations obtained by clinicians in an uncontrolled way are particularly likely to be made if a drug is exceptionally potent or toxic, if the disease process being studied is one not likely to improve spontaneously, or if a potent drug is already available for comparison. Although once again one can cite exceptions, it thus behooves investigators who come up with completely negative studies on a time-honored drug to be exceptionally careful in dismissing the medication as inactive. Too frequently the next investigator devises a satisfactory testing situation and corroborates past generations of maligned clinicians.

Finally, for those who like to think that satisfactory clinical trials are new and a part of the pharmacology of the last decade or so, I should like to quote the following, abstracted from an article by Gaddum.[3]

" . . . I took twelve patients . . . (with) scurvy. . . . Their cases were as similar as I could have them. . . . They lay together in one place and had one diet common to all. Two of these were ordered each a quart of cyder a day. Two others took twenty-five drops of elixir of vitriol three times a day upon an empty stomach. Two others took two spoonfuls of vinegar three times a day. . . . Two of the worst patients were put upon a course of sea-water. Of this they drank half a pint every day. Two others had each two oranges and one lemon given them every day. The two remaining patients took an electuary recommended by a hospital surgeon made of garlic, mustard, balsam of Peru and myrrh. The consequence was that the most sudden and visible good effects were perceived from the use of oranges and lemons; one of those who had taken them being at the end of six days fit for duty. The other was the best recovered of any in his condition and was appointed nurse to the rest of the sick."

This experiment was started by James Lind on May 20, 1747.

REFERENCES

1. Hill, A. B.: The Clinical Trial, New England J. Med. **247:**113, 1952.
2. Marshall, E. K., Jr., and Merrell, M.: Clinical Therapeutic Trial of a New Drug, Bull. Johns Hopkins Hosp. **85:**221, 1949.
3. Gaddum, J. H.: Clinical Pharmacology, Proc. Roy. Soc. Med. **47:**195, 1954.
4. Beecher, H. K.: Experimental Pharmacology and Measurement of the Subjective Response, Science **116:**157, 1952.
5. "Student": The Lanarkshire Milk Experiment, Biometrika **23:**398, 1931.
6. Jellinek, E. M.: Clinical Tests on Comparative Effectiveness of Analgesic Drugs, Biometrics Bull. **2:**87, 1946.
7. Mosteller, F.: Clinical Studies of Analgesic Drugs. II. Some Statistical Problems in Measuring the Subjective Response to Drugs, Biometrics **8:**220, 1952.
8. Lasagna, L., Mosteller, F., von Felsinger, J. M., and Beecher, H. K.: A Study of the Placebo Response, Am. J. Med. **16:**770, 1954.
9. Lasagna, L., and von Felsinger, J. M.: The Volunteer Subject in Research, Science **120:**359, 1954.