CRITICAL EVALUATION OF THE EFFICACY OF NEW DRUGS *

By J. H. PANNEKOEK, M.D.

Department of Medicine, St. Geertruiden Gasthuis, Deventer, The Netherlands

Sir William Osler, in his still arresting contribution to the Oxford Loose-Leaf Medicine, 'The Treatment of Disease,' states that: 'In the progress of knowledge each generation has a double labour—to escape from the intellectual thralls of the one from which it has emerged and to forge anew its own fetters.' For this occasion I would paraphrase Osler's sarcastic and rather gloomy statement as follows: Every practitioner has to fight a battle on two fronts; on the one hand against the erroneous views of preceding generations, though retaining what has proved to be right and, on the other, against the ever increasing flow of new therapeutic methods, in an attempt to separate the grain from the abundant chaff. Our main weapon in this fight is a critical attitude based on knowledge of the subject.

Our enemies on the first front are Authority, Tradition and Routine. Today we shall not dwell on this aspect of the battle; suffice it to say that defeat can only be avoided by examining ourselves every day as to the usefulness of what we are doing.

Today we shall have a look at the other front, but with respect to the short time available restrict ourselves to a few aspects only. We cannot discuss the risks inherent to the introduction of any new drug, neither can we discuss the economic and ethical aspects.

It is my task to discuss the criteria to be used in deciding whether a given medicine is effective or not—a discussion which is indeed limited in its scope, although it should be pointed out that the above mentioned criteria do not apply only to the evaluation of drugs, but can also be used, though with a different accent, in evaluating other modes of therapy such as physiotherapy, surgical and dietetic treatment and maybe psychotherapy.

In my opinion, evaluation on the whole is simpler where drug treatment is concerned, although it is still difficult and in some cases even almost impossible.

It may be useful, first, to discuss with you the phases in the establishment of a drug. It is developed in the research laboratories of the large pharmaceutical industries, tested in scores of animal experiments, tried out more or less cautiously as a research preparation on normal subjects and subsequently on patients. Then the birth of the new drug is announced more or less soberly in the scientific press or perhaps even in the Reader's Digest, unless the announcement reaches the medical profession via the radio or television. And then the new drug begins its career.

Jawetz—a well-known American expert in the field of antibiotics—divides this career into four stages, viz.: The first stage is that of enthusiasm. Nothing but praise is heard and the new drug, advertised to the skies, shoots up like a rocket. The indications spread like oil on water and sales curves reach a peak within a very short time. Then, however, critical voices are heard, which initiate the second stage—the stage of initial stabilization, which is at a high level although no longer at the top. The third stage is the stage of disenchantment. Failures are reported and, still worse, there is an increasing number of reports on disagreeable and dangerous side-effects. The life curve of the drug shows a rapid decline to zero and it is highly possible that the drug has to be placed in the ever-growing mausoleum of obsolete drugs. In some cases it escapes this fate, and the disastrous stage is followed by a fourth stage, viz. that of ultimate stabilization at a more or less modest level. Having seen that, just remember that according to Rypkema, 368 new drugs were registered in the Netherlands in 1951 alone!

A sad example is found in the history of the use of antihistamines for the treatment of the common cold. They were introduced in 1949, and the sales in the U.S.A. amounted to 100 million dollars in 1950, but in 1951 there was a complete collapse, following the crushing reports made by the Council on Pharmacy and Chemistry of the American Medical Association and the British Medical

Another trap which makes many victims again and again, is what I should like to call, the false track of the a priori—the attempt to reach conclusions by reasoning alone. If we determine that an improvement in a certain clinical picture results from preceding treatment, then we should establish a causal relationship; this is primarily a problem of natural science. The history of natural science shows that initially attempts were made, analogous to mathematics, to find an explanation apart from any experience, based on thinking only, i.e. an explanation a priori. This has become a complete failure and it is now an established fact that any and every conclusion based on reasoning should be verified, in natural science, by observation, preferably by observations made in experiments. In medicine this fact has not been generally accepted so far. Parsighted minds such as John Hunter were aware of this long ago, as shown by Hunter’s famous exclamation of impatience: ‘Why think? Why not try the experiment?’, with which he addressed his friend and pupil Jenner. Trotter, whom I quoted earlier in this lecture, still considered it necessary to devote, even as late as the thirties, one of his outstanding essays to the question: ‘Has the intellect a function?’ He presents convincing evidence, showing that the part of reasoning in medicine should be far more restricted than is generally accepted. Any and every conclusion, opinion, hypothesis or theory should be verified again and again by experience. And this verification is a tedious and usually less arresting activity than the formation of far-reaching conclusions. Trotter expressed the function of reasoning in the following excellent advice: ‘The wings given to science by reason must be used like those of the ostrich rather than like those of the eagle, and help it along without raising it from the solid earth.’

Perhaps I may elucidate the intention of this fairly abstract discourse by the example of the modern starting-point of the a priori, viz. animal experiments. It can be stated, without exaggeration, that virtually the entire progress made in modern medicine is based on animal experiments. One might even go further and maintain that introduction of a new drug today is not justified unless it has previously been tried out in animal experiments. Since it is far more difficult to collect evidence of the therapeutic effect in man than it is to present pharmacological evidence of its effect on animals, the a priori argument is only too often used instead of an attempt to a therapeutic proof. Excellent work has been done by such investigators as Best and Himsworth in an attempt to demonstrate the favourable effect, under certain experimental conditions, of choline and methionine respectively, on lesions of the liver. In spite of
the fact that the step from animal experiment to the application on man is a large one, the conclusion was almost generally drawn that a high-choline and high-methionine diet, i.e. a high-protein diet, must have a favourable effect on hepatitis. Yet the actual evidence available is extremely meagre.

There has recently been a trend to reduce the distance between the animal experiment and clinical application by means of the development of clinical pharmacology, i.e. the study of the effect of pharmacological agents on man. Factors studied include resorption, distribution in the organism, excretion, the effect on the function of various organs, etc. This approach does in fact provide a much firmer basis for a priori arguments. Yet no mistake should be made here. There is, and there always will be, a difference between the pharmacological and the therapeutic action of a drug.

Now that we are aware of the 'natural history of disease' and have been warned against the lure of the post hoc ergo propter hoc and the a priori evidence, we may safely proceed.

Distinction can be made between three different approaches in the evaluation of drugs, viz. the route of impressionism, the following-up system and the experimental method.

About the first I can be brief. This way is the least difficult, the oldest, the most frequented and the longest. When using this approach one is guided by impressions. One uses a drug in a number of cases and waits for success or failure. An opinion is then formed on the basis of experience or intuition, or both. Wishful thinking, authority, publicity, a priori and post hoc arguments are free to exert their influence. Yet in many cases, most surprisingly, there is a final consensus of opinion which is not always wrong. In some cases, unfortunately, there is simply no other approach possible.

The follow-up method has much in common with what in natural science is called the observational method. A group of patients as a whole is submitted to a certain way of treatment, and the further course of the disease is then systematically followed up. In this manner valuable data can be collected. Unfortunately, the method has the disadvantage of affording no possibility of comparison with the course of the disease under different treatment or without any treatment. Useful results can therefore only be obtained in this way if the clinical course of the affection in question has been established in advance (e.g. in the case of invariably fatal diseases).

This method is completely unsuitable if used in evaluating therapeutic results obtained in diseases with a less sharply circumscribed clinical course. Anyone attempting to form an opinion, with the aid of this method, on the value of drugs in diseases with a variable course (peptic-ulcer, neuritis, migraine, etc.) runs a grave risk of making serious mistakes. Unfortunately ethical considerations sometimes preclude the use of any other approach.

The third method is the surest, the most difficult and the shortest. It is known by many names, but for convenience' sake we can confine ourselves for the moment to Bradford Hill's designation of 'Clinical Trial.' For it was an English investigator who first used this method, and it was by the British Medical Research Council that this method was further developed. It was inspired by the experimental methods used in natural science and its chief characteristic is comparison with findings obtained in controls. This constitutes a criterion which can be used to verify clinical therapeutic investigations.

The British naval surgeon James Lind was the first to make conscious use of this method in 1747. Twelve sailors suffering from scurvy were treated in pairs with several therapeutics recommended for this affection. Two of the twelve were given orange juice and lemon juice. Only these two were cured, and in such a striking way as to show Lind that he had made a valuable discovery. Further experiments confirmed his conclusion, and in 1753 he reported his findings.

It was not until 40 years later that this therapy was introduced by the Admiralty as the standard treatment of scurvy—a quick reaction in the bureaucracy of those days and very favourably contrasting with the attitude of medicine. The profession was not ready to accept Lind's discovery until 150 years later, when his findings were compatible with medical reasoning, as it then was. This shows only too clearly how strong the authority of a priori arguments were with the profession!

A characteristic feature of Lind's method was comparison of the therapeutic effect of fruit juice with the effect seen in controls simultaneously given different treatment. Control group and simultaneity are essential conceptions. However, there are circumstances in which the requirement of the simultaneity may and must be abandoned. One need only mention the use of streptomycin in tuberculous meningitis, in which the principle of 'historical control' was used, because the disease was known invariably to lead to a fatal issue. The control group consisted of all patients who died due to this disease before the introduction of streptomycin. This method of 'historical control' is only effective, however, in the cases with a sharply defined and familiar clinical course, i.e. unfortunately in the majority of cases: a more or less
acute fatal affection. In fact, the type of affection for which the follow-up system can be used.

The arguments in favour of this statement are the following: firstly, our knowledge of the course of the majority of diseases is insufficient as a basis on which an accurate group prognosis can be given. It should be borne in mind that the salient points in these investigations are found in numerous particulars, e.g. mortality, duration of illness, fever curves, complications, etc., and in the majority of cases we do not know all these particulars with sufficient accuracy until the necessary attention is paid to them because we want to try out a new drug. And even when we are well familiar with all these particulars, it is not always possible to use this knowledge. There is no guarantee that the nature of a disease—the notorious genius epidemicus—is not in the process of changing. Only under very special conditions is the use of a 'historical control' justifiable, outside the group of the above-mentioned invariably fatal diseases.

There is a second reason for abandoning the requirement of simultaneity. This is found in the comparative application of drugs in chronic conditions which are virtually stable, e.g. diabetes and hypertension. The same subject can then be used for comparison of the effects of insulin with the effects of oral preparations. This method—developed by Martin in Germany and advocated by Hogben in England—often offers considerable advantages but can only be used in chronic diseases. The various pitfalls concealed here will not be discussed. It will be apparent that the principle of control has not been abandoned here: the patient is his own control.

Returning to the most widespread form of 'Clinical Trial' it can be stated that a group of patients are given the drug to be tested, whereas the control group do not receive this. The selection of the controls is based on a simple principle. Yet experience has shown that it is here that the gravest errors are committed.

It is obvious that, if a reliable decision is to be made regarding differences between patients treated and controls, identity of the two groups involved is indispensable. Statistics have shown that our chance of success is greatest if the division into two groups is purely at random. Unless this is so there is no guarantee that some form of selection does not take place, in which case the views and prejudices of the investigator, whether conscious or unconscious, are of such influence as to cause dissimilarities in the two groups; the results of investigations are in this way rendered questionable. The methods developed in statistics to accomplish division into two groups formed at random will be left undiscussed here. The problems involved, however, are elucidated by means of a few examples.

Laplace, the great French mathematician, suggested the so-called alternating method in 1812. Patients are alternately treated with the drug, tested or not, in chronological order. This means that the investigator knows which patients serve as controls. Some patients, it is sometimes found, are so ill as to make it inadvisable to deprive them of an old, established therapeutic in favour of a new drug of which the effect is still unknown. Such patients are, of course, excluded from the experiment. In the case of the alternating method this decision can be influenced by the fact that the group to which the patient will belong is known; in other words, the investigator's views enter the experiment by the back door. The test is no longer free from prejudiced selection according to the statistician, and in fact any statistical elaboration of material in this manner threatens to become a farce.

The need for great caution is illustrated by the investigation of the American Heart Association into the value of anticoagulants in the treatment of myocardial infarction. These anticoagulants were given to patients admitted on odd days, those admitted on even days serving as controls. At the end of the experiment it was found that the treated group was considerably larger than the control group and included a considerably larger number of 'private patients' with private nurses. This was found to be due to the fact that the arrangements for the experiment became known through rumours, so that several practitioners saw to it that their patients were admitted on an odd day. As a result, the two groups were not comparable and the conclusions reached less reliable.

Nor is it permissible to use patients in one hospital serving as controls for patients in another hospital, without certain provisions. The groups may be dissimilar due to differences in social conditions. Results may furthermore be seriously influenced by differences in nursing. The statistician will even frown upon comparisons between two wards in the same hospital.

In the above examples, there is a trend towards taking the control group composition at random, but there are flaws in the execution of the plan. A great many investigators, however, seem to be completely ignorant of this requirement, as may be shown by the following examples:

An investigator treats a group of ulcer patients by a rest-and-diet cure, and observes that the results are more favourable in fresh cases than in cases of long standing. His conclusions are as follows: (1) this therapy is effective, and (2) the sooner it is started, the greater its efficacy. In fact, the investigator has done this: he compares early
cases (with a better chance of spontaneous cure) with chronic cases (in which part of the possibility of spontaneous cure has been lost). Such a comparison is, of course, unjustifiable and the conclusion—although possibly correct in itself—is no proof.

Another investigator compares the results obtained in a group of ulcer patients, who kept to a prescribed diet, with a group of ulcer patients who gave up the diet, and concludes that keeping to the diet prevents relapses. Again a conclusion which is not justifiable. Can these two groups be compared, and why did not the second group keep to the diet? The results presented by an investigator who ignores the requirement of simultaneous controls at random should be mistrusted, unless he is able to present sound reasons for having deviated from this rule in his special case.

Is it sufficient, therefore, to arrange the experiment so as to ensure that one group is treated with the drug to be tested, while the other group is not? In principle it is, but results, if any, must not be attributed exclusively to the pharmacological effect of the drug tried out. The effect of any drug is a composite one. Besides the pharmacological it has a magic effect, which is based on the patient’s confidence in the efficacy of the drug, and on his confidence—still widespread in spite of everything—in the skill and knowledge of his doctor.

This magic effect is today referred to as the placebo effect, which is often underrated both by the patient and by the physician.

Of late, much has been written about the psychology of the placebo. The most extraordinary things have been discovered in this respect. Many modern drugs are known to produce disagreeable side-effects such as urticaria, fever, etc. A placebo such as lactose has been found capable of producing exactly similar effects. In some situations (post-operative pain) about 40 per cent. of patients are known to respond to a placebo. On the basis of pure intelligence tests it is impossible, and on the basis of the Rorschach test it is only partly possible to distinguish between ‘reactors’ and ‘non-reactors’—two terms for variable conceptions. A ‘reactor’ in certain situations may be a ‘non-reactor’ in different situations. The whole problem becomes even more complicated, due to the fact that some patients have a negative attitude towards any drug; such patients may ruin the results of an otherwise well planned investigation.

Before abandoning this subject, I should like to state as my opinion that, it is not only drugs which have a placebo effect in addition to a specific pharmacological effect. The ‘placebo face’ can be found in all modes of treatment, whether surgical, physical, dietetic, or even psychological.

The consequences of all this for the Clinical Trial are of great importance, for it is impossible to get to know the pharmacological effect of a drug merely by giving the drug to one group of patients and not giving it to another group. The control group must be treated with a placebo, which does not differ from the original drug, either in form and taste, or in mode of administration. Needless to say, the patient should be unaware of the fact that only a placebo is given. It is obvious, considering the origin of the placebo effect, that the therapist should likewise be in the dark as to which patient is given the drug and which the placebo. This arrangement, known as the double blind test, has become an important factor in modern therapeutic investigations. In addition, it is a factor greatly contributing to objectivity in the interpretation of results by the experimenter—and as such, often indispensable. The disadvantages of the double blind test cannot be discussed here.

Certain aspects of modern therapeutic experimentation have been discussed in some detail. There are a great many other aspects besides these. To mention only a few: uniformity of treatment and standardization of the criteria used in determining the success of treatment (mortality, duration, complications, etc.) The time available, however, does not permit of a discussion of all these aspects.

Special mention should be made, however, of one requirement emphasized by Bradford Hill, who set up the rules of Clinical Trial, viz. the homogeneity of the material investigated. For a certain investigation it is often the simplest way to choose a group of patients all suffering from a certain disease to the same degree. The famous streptomycin investigation, made by the Medical Research Council, concerned young subjects suffering from bilateral, exudative cavitary pulmonary tuberculosis—a selection which simplified the investigation, but at the same time limited the conclusions.

The chief characteristics of Clinical Trial—unselected simultaneous controls, and arrangement of the trial in such a manner as to guarantee objective interpretation—have been discussed in detail. In this manner I hope to have provided you with a basis—inasmuch as this was necessary—from which you can make your critical analysis of therapeutic investigations.

Finally, I should like to devote a few words to the significance of statistics in this respect. Professor Bok states that statistics show us the degree of reliability of our conclusions. In my opinion it is obvious that statistical elaboration of the results of therapeutic investigations is indispensable. A note of warning is indispensable as well. A conclusion reached with the aid of statistical methods does not guarantee that the therapeutic experiment was set up in the correct manner. Only too often
is the statistical sauce used to cover a spoiled experimental pudding. ‘No good statistics for bad experiment’ gives expression to the regret of the statistician for abuse of statistical methods. Under the guidance of Fisher, the world of statistics has taken a more intensive interest in the arrangement of the experiment as such. Statisticians would rather be the obstetrician than the coroner of the experiment, and rightly so. More and more statisticians cooperate throughout the experiment, which to me is as it should be.

We have come to the end of the difficult road of therapeutic experimentation and to finish up, the question remains of what we have accomplished. We can bring to the fore the positive side of our results, or we can throw light on that which has not yet been accomplished. It is the latter approach that I should prefer in this paper.

Firstly, as a rule the therapeutic experiment is arranged on the basis of a certain dosage. It is by no means sure that, with a different dosage, the experiment would not have had different results, possibly more favourable for the new drug. Secondly, if a homogeneous group of patients are used for the experiment, as indicated by Hill (e.g. the exudative form of pulmonary tuberculosis in the streptomycin experiment), then the results in patients of a different type (e.g. with caseous foci in the above-mentioned case) may be completely different.

These two disadvantages can be removed by continuing the experiment with different doses and a different type of patient.

The main disadvantage of therapeutic experiments is, however, that the results are only too often unsuitable for generalization. Only very few groups of patients can be regarded as really representative of all patients suffering from the disease in question, both now and in the future. It should be borne in mind, for instance, that most investigations so far made concerned hospital patients—who are different from the patients seen by the general practitioner. An attempt may be made to reduce this disadvantage by repetition of the experiment elsewhere, or by co-ordinated experiments simultaneously made in different places. In spite of these limitations, however, and in spite of all the difficulties, modern Clinical Trial should be regarded as a model and a criterion of all therapeutic investigations.

I should like to summarize what I have said in a few questions which we should ask ourselves in studying a therapeutic investigation. These questions, freely adapted from Mainland, are: Who is the author? Does he give evidence of sufficient knowledge of the natural course of the disease? To what extent is his patient group representative? Who are his controls? Are these in fact unselected? Has the placebo effect been accounted for, and does the paper give a guarantee that interpretation of results has been objective? Are the results obtained significant statistically?

A study of therapeutic reports in this manner will, in my opinion, lead to the conclusion that much is stated which cannot stand critical analysis. Was Osler right after all, when he predicted that thralls would be exchanged for fetters?

I would not attempt to answer this question. It is an established fact, however, that in the therapeutic field unstable factors still abound. It is better to admit this fact than to act as the ostrich, hiding its head in the sand. Now, as in the future, errors and mistakes are bound to be made.

As Goethe would have it: ‘Es irrt der Mensch, so lang er strebt.’ Let us try, then, to confine the number of errors to a minimum.

Deventer, October 10, 1955.

BIBLIOGRAPHY

MAINLAND, D., and STEELE, J. M. (1954), 'Statistics in Medical Research,' in Methods in Medical Research, vol. 6, Chicago.
MARK, R. E. (1952), 'Wege vergleichender Therapie in der inneren Medizin,' Berlin.