STATISTICS IN CLINICAL RESEARCH: SOME GENERAL PRINCIPLES

By Donald Mainland
Department of Anatomy, Dalhousie University, Halifax, Nova Scotia, Canada

A discussion of statistical methods in clinical research by one who is neither a professional statistician nor a clinician must appear rather incongruous. There may be some advantages in this incongruity, however, for there may be something of interest in the point of view of one who is using statistical methods daily in the laboratory and who is also in daily contact with clinicians. Such a person sees, by contrast, the difficulties of clinical research, and he sees how clinicians who possess keenly critical minds and are interested in scientific medicine are handicapped by lack of acquaintance with statistical ideas.

The statements in this paper will be affected also by experience in an undergraduate course of Quantitative Medicine given jointly by Dr. C. B. Stewart, Professor of Epidemiology, and the Anatomy Department of Dalhousie University. Students' difficulties seem to be the same as those of clinical research workers, and they concern principles rather than arithmetic. The following remarks, therefore, may interest both clinical workers and the statisticians whom they may consult.

The Scope of Statistical Methods

First, a few remarks about the term "statistical methods." It was suggested that this paper should discuss the pitfalls encountered by clinical workers because they fail to apply statistical methods. One could paraphrase that and refer to the pitfalls encountered if one fails to conduct an experiment correctly.

This may seem like an exaggerated claim for statistical methods, but we should remember that all investigation is concerned with differences. For example, we wish to find differences in speed of recovery of patients after two different treatments and to distinguish these differences from those due to other factors that also affect speed of recovery. We are investigating differences, *i.e.*, variation, and the methods of investigating variation are statistical methods.

Investigating variation means far more than applying statistical tests to data already obtained. Such tests are much more common in medical literature than they were twenty years ago, but there has not been a corresponding increase in the use of statistical reasoning. Any medical journal provides evidence of this. For example, a casual inspection of the recent issues of one of the best-known weekly journals of general medicine has provided numerous instances of this lack, and yet that journal pays more attention to statistical methods than do many others. Three of these examples may be mentioned.*

* No bibliographic references are given for these examples because there is no intention to criticize individuals. Moreover, if any group is to be criticized, it is not the clinical investigators but the laboratory scientists who taught them.

An obstetrician, treating placenta praevia, recorded a conservative method that gave much better results than had been reported by other workers. Two weeks later, a critic pointed out that the method of examination employed (vaginal examination under anesthesia) had very likely enabled the obstetrician to diagnose correctly as placenta praevia certain conditions that others, not using that method of diagnosis, would have called "antepartum hemorrhage of unknown origin," and patients so classified usually do well with little or no treatment. The different observers, therefore, were probably sampling different populations.

In the second example, two workers described a method of opening superficial veins for the withdrawal of blood without causing them to become blocked during healing and therefore useless for future transfusions. They stated that the veins, examined subsequently, remained patent in 100 per cent of the cases, but they did not state how many they had observed. Therefore, no one can tell what different percentages of success might be obtained by further use of the same method.

The third example is from an investigation of ketosteroid excretion in the urine before and after a brain-sectioning operation for the treatment of mental disease. Correct statistical tests were applied to the results and, 3–4 months after the operation, a very significant reduction in steroid excretion was found. The observer concluded that a certain part of the brain probably influences steroid metabolism, but did not indicate that he had considered other possible causes, such as other elements in the operation apart from the brain sectioning, or some other factor in the care of the patients. The investigation was not planned to allow for the various differences between the pre- and postoperative states of the patients.

Statistical ideas, to be effective, must enter at the very beginning, *i.e.*, in the planning of an investigation. These ideas, however, lacking in so many published reports, are even less frequent in the unpublished efforts of clinicians to assess the value of their treatments. There must be something wrong with so-called "scientific" medical education when a young physician says that he has obtained promising results by treating migraine with histamine and yet cannot understand why a professor of pharmacology should ask about controls.

This is our reason for giving a course in quantitative medicine. It is admittedly hard to assess our results, but we felt that it was somewhat hopeful when a student asked recently: "Why do not the editors of medical journals employ someone to scrutinize papers before they are published?" It is of interest that the class at that time had not been shown any statistical tests but had merely been guided in the examination of published statements.

Another question asked at the same time was: "How can I believe anything I read or hear when there are so many sources of error?" That question leads to the next part of the course, in which we consider the requirements for an adequate sample. A few elementary comments on that topic may be appropriate here.

The Requirements for an Adequate Sample

Let us consider a very simple type of investigation. We wish to take two samples of patients with the same disease and apply one treatment to one sample and another treatment to the other sample. We avoid some misleading differences by standardizing our methods of observation and our other techniques, but there are many factors still left, apart from the two treatments, that will produce differences in the outcome.

First, there are what may be called *major factors.* People differ from each other according to sex, age, and racial stock, and these three factors affect the course of disease and the response to treatment. Another major factor is the severity of the disease, including the presence or absence of complications. Sometimes environmental factors, *e.g.*, economic status and occupation, are obviously major, and so, sometimes, are psychological factors.

All such major factors may be called "recognizable" factors, for which allowance can be made by separating the patients into different classes: males and females, children and adults, patients with and without complications, and so on. In all cases, however, there is the possibility of other major factors, for which allowance cannot be made in this way because they are quite unknown. For example, by passing catheters into the heart to study its output, it has recently been found that congestive heart failure is really a complex group of conditions with different responses to digitalis.[1] Doubtless, many other diseases will be found to have a similar complexity, but meanwhile we have to discover a treatment for them as they are now labeled. Also, we must be sure that, in our present ignorance, we do not vitiate our results by a preponderance of one type of the disease in one of our samples. Likewise, there may be hidden environmental factors, and we must avoid the risk of all such hidden bias.

Finally, there are, affecting every patient, innumerable *minor factors,* known and unknown—anatomical, physiological, biochemical, psychological, and environmental, *e.g.*, small differences in the virulence of bacteria, differences in details of medical and nursing technique, in investigational measurements, and in the criteria by which we diagnose and by which we assess the state of the patients after treatment. Some of these factors will affect the outcome and our judgment in one direction and some in the opposite direction. When we are selecting our samples, some of the factors will still be in the future, but we must allow for them at the very beginning.

Purposive Sampling. Now, therefore, we ask: "How can we allow for these three sets of factors—the major recognizable, the major unknown, and the minor factors?" The first step, as already mentioned, is to separate the individuals into the recognizable major classes appropriate to the problem. This purposive sampling reduces the variation; *i.e.*, if we now take two samples from any one class, there will be less difference between them than if we sampled the original heterogeneous collection. Thus, any differences in the effects of the two treatments will stand out more clearly.

This is all very elementary, but even at this level there are misconceptions. We obtain from simple physics and chemistry the notion that we

should make everything as alike as possible except the factor that is being tested, the difference between our two treatments. Perhaps we can demonstrate known facts in this way, but it is not the way in which actual investigations are carried out, even in physics and chemistry. We could, indeed, continue indefinitely to equalize the various factors that would or might influence the outcome, but there would still remain factors which, being unknown, could not be equalized.

Moreover, thinking of how we are going to use the results, we see that extreme equalization is usually undesirable. We wish to know, for instance, whether treatment A or B is better when applied within certain rather broad groups, *e.g.*, to old men with a certain severity of the disease and having also some other condition, such as arteriosclerosis. We should aim, therefore, to reduce the variation by purposive sampling, not as far as possible, but as far as convenient and useful; *i.e.*, we should divide the patients into the appropriate major groups.

Randomization. In each major group, each patient will still be the resultant of a different set of major and minor factors. Therefore, the patients will differ in their speed or completeness of recovery, even if the two treatments are exactly alike in their effects. These various differences cannot be equalized in the two treatment samples. Hence, we must allocate the patients in such a way that we can tell what allowance to make for the inequalities. The only way to do this is to make chance decide for us; that is, we must allocate the treatments strictly at random. This does not mean haphazard choice or the acceptance of samples as random because we cannot think of any reason why they might not be random. Even in such simple procedures as choosing animals from cages, "experience has shown," in the words of Yule and Kendall,[2] "that the human being is an extremely poor instrument for the conduct of a random selection." We must employ an automatic method, such as coin tossing, card shuffling, or, what is quicker and more convenient, the use of tables of random numbers. The initial randomization is often enough, but a small supplementary one may be needed if, later, an unexpected choice has to be made, for example, between hospital beds.

Assessment of Results. Having made chance operate in the selection of samples, we can, after the experiment, use our knowledge of chance to assess the results, because we know how often various differences in the results would occur by chance, *i.e.*, if there were no difference between the treatments.

We may be able to say, for example: "This difference between the results in the two samples would occur so often by chance that we do not feel confident that it indicates a difference between the treatments." Our verdict is "Not proven—not significant." On the other hand, we may be able to say: "This difference is due either to chance or to a difference between the treatments; but such differences are so rarely due to chance that we believe it probably indicates a difference between the treatments." Our verdict is "Significant."

In contrast, let us consider our verdict if we had not randomized. We

could then say: "The difference is due either to chance or to something else. Such differences rarely being due to chance, we believe that it is probably due to something else." But this "something else" may be either the difference between the treatments or some bias due to unknown factors, or perhaps treatment plus bias. Because we did not randomize, we have no way of telling.

Statistical Significance. In these verdicts, one word requires comment: "significant." Some workers seem to think that "statistical significance" is something imposed on us by the mathematicians. On the contrary, it has been introduced and used by those who have to act on the results of statistical tests. To call events "significant" simply means that they would so rarely occur by chance that we feel justified, for the purpose in hand, in believing that they were probably due to something more than chance. For most purposes, it has been found satisfactory to apply the term in such a way that, of all events that are actually due to chance, only the rarest 5 per cent will be so labeled. Let us see what this implies.

In the course of a lifetime, a worker will be engaged in many investigations in which events, *e.g.*, differences between samples after treatment, are due entirely to chance. If he adopts the 5 per cent rule he will, in 5 per cent of such investigations, mistakenly say that something more than chance is probably operating. It might therefore be suggested that he should always demand a higher standard, *e.g.*, a 1 per cent error. We must remember, however, that many of his investigations will, unknown to him, involve what may be called "real" differences, *i.e.*, not due solely to chance, and, if he insisted on the higher standard, he would miss more of these real differences.

No one, however, is bound to adhere to the 5 per cent rule. The standard should depend on the particular problem; but, whatever standard we adopt, we should know exactly what it means, and we should set our standard before, and not after, the results have been obtained.

If we understand the meaning of statistical significance, we shall see that, although the verdict is automatic when once we have set our standard, this does not absolve us from further thought. We must remember that chance may have allotted one of the major factors, previously discussed, predominantly to one treatment group. This may be one of the unknown factors, or it may be a factor that we can detect by scrutiny and further analysis of the records of the experiment—a factor such as the occurrence of a complication during the course of treatment.

Again, the effects that we attribute to a certain treatment may really be due to something that is commonly or constantly associated with the treatment, and further investigation may be desirable to disentangle the causal relationships.

The Status of Unplanned Observations

Having seen in outline the requirements for adequate samples, we may now ask how far these requirements are met by data from nature's experiments in disease, from hospital records, and from clinicians' incidental observations. Such data are often obviously not worth analyzing in detail,

and in even the best of them, except possibly some of the very simplest, there must remain doubt regarding interpretation, because the sampling is not known to have been random. This applies even to observations in which a careful worker compares his present results from one treatment with his previous results from another treatment. These unplanned observations may be the only information available as a basis for action, and they may form a useful basis for planned experiments; but we should never forget their inferior status.

Medical Progress without Planned Experiments. Such a condemnation naturally raises the question: Has not clinical medicine made great progress without these statistical methods? As a partial answer to this question three comments could be made:

(1) Most of the main advances in medicine have been due to some method, such as chemotherapy, that has produced an effect strikingly different from previous experience and so rapid as to leave no doubt regarding causal relationships.

(2) As soon as we start to explore the limits of such a new method, or to compare different modifications of it, planned experiments are necessary.

(3) In the less spectacular parts of medicine, one may perhaps believe that, despite lack of proper experiments, there has been progress, whereby poorer methods have been gradually replaced by better ones, the observers having by luck avoided serious bias in their samples. This may indeed be so, but anyone tends to be skeptical who has heard debates at medical meetings and has witnessed, for a quarter of a century, many apparently capricious changes in medical beliefs and fashions (*e.g.*, in the treatment of burns). Having seen the success of properly planned experiments elsewhere in applied biology, he tends to advocate them in medicine also.

The Moral Problem. The word "experiment" brings us to a problem peculiar to human medicine—the moral problem. It is the physician's duty to do his best for his patients, and, if he believes that there is some evidence in favor of a certain treatment, he will feel bound to use it. If, however, he is acquainted with the requirements for valid proof, he will often see that what looked like evidence is not evidence at all, and he will feel free to experiment. During the experiment, of course, he will sometimes feel impelled, for therapeutic reasons, to alter the treatment in one or more patients. Even then, however, it may be possible to use the data obtained up to the time when the treatment has to be changed.

The very fact that these difficulties exist shows the importance of careful planning and analysis of results by modern methods, which enable us to extract all the available information even from small samples.

Other Statistical Procedures

A very simple experimental design has been discussed in order to bring out the essential principles. It is not the purpose of this paper to discuss designs in detail. It is desirable to mention, however, that more complex designs, first developed in agriculture and other sciences, are very applicable to clinical research. Such are the factorial design and the incomplete block

design. These designs not only yield more information from a given number of patients than do the simpler designs, they also provide information about the effects of one factor in the presence of others, which simpler designs cannot do. (For a brief exposition of the principles of factorial design as applied to medical research the reader is referred to a recent article by Greenwood.[3])

The only types of calculation mentioned so far have been tests of significance. Equally important are statistical estimates, of which three examples may be given:

(1) *Regression Coefficients.* These have very wide application. For instance, if we are comparing days required for wound-healing under two or more different treatments on different groups of patients, regression coefficients will enable us to make allowance for differences between the groups in respect of age and any other measurable feature that may influence the speed of healing.

(2) *The Numbers of Individuals Required to Demonstrate a Certain Result If It Could Be Demonstrated at All.* These estimates are very desirable and should be made either from records available before an investigation is started or at any early stage of the investigation. If such estimates had been made, some investigations would never have been started, because it would have been seen that, in view of the limitations of time, facilities, money, or numbers of patients, the investigation would be useless.

(3) *Confidence Limits.* No competent laboratory worker imagines that his results, in a chemical analysis for example, have any meaning unless he has estimated their possible error. Yet it is not generally realized that clinical data require similar estimates before they can form the basis for a sound conclusion, or even for a rational opinion. For instance, a surgeon may feel fairly well pleased if, performing a certain operation on 30 patients, he has had unsatisfactory results in only two of them. If, however, he continues to use this operation on patients of the kind represented by his sample of 30, he may find in the long run, as the result of chance alone, a higher or lower proportion of unsatisfactory results. From his present evidence, if he adopts the usual standards of judgment, he should not feel confident that his ultimate proportion of unsatisfactory results will be less than 22 per cent, and he would be safer to set this percentage at 27. On the other hand, he need not be surprised if, without having improved his technique, he finds that his ultimate proportion of unsatisfactory results is only 0.8 per cent. (Such confidence limits are easily obtained by reference to tables and graphs.[4])

Sources of Guidance in Statistical Methods

Although a full reading list will not be attempted here, a few suggestions on articles and books may be helpful to investigators. Greenwood's[3] article, already mentioned, presents a clear picture of the functions of the statistician in medical research. Meleney's[5] report on the prevention of infection gives a valuable account of the difficulties met in this kind of research and of methods to overcome them. The British Medical Research

Council's[6] investigation of the streptomycin treatment of tuberculosis can be taken as a model of experimental design in therapeutic trials.*

Among the numerous statistical textbooks, the medical investigator should, in general, disregard those that do not concentrate on the methods developed by Professor R. A. Fisher. The books by Bradford Hill[7] and Albritton[8] can be specially recommended for exposition of principles and elementary methods in the medical field, but for further information the medical investigator has to use books prepared for other biological workers (Snedecor[9] and Mather[10].)

Whatever articles or books are read, however, it is doubtful if anyone can safely start using statistical methods without the personal help of someone who has used them, and medical workers sometimes have difficulty in knowing whom to ask for help. They may assume that a mathematician would be most appropriate, but, even if a mathematician specializes in the statistical branch of mathematics, he is not thereby fitted to give guidance in the application of the methods.

An economist may do much statistical work, but he is unlikely to appreciate the problems of experimental design and of the treatment of small samples that a medical investigator must face. If a statistician in public health or epidemiology has given attention to these problems, he can be very helpful, but statistical techniques that are useful in those branches of medicine are mostly large-sample methods. For example, the standard error of the binomial, $\sqrt{Npq}$, is widely used in public health statistics, but, in order to make it dependable for use with small samples, somewhat complicated additional calculation is needed, unless tables and graphs[4] are used.

A medical research worker, therefore, may have to seek rather far for help, and he is often more likely to find it among workers in applied science, especially agriculture, than in medical faculties or laboratories of pure science.

Finally, it must be stressed again that, whatever sources of help are found and whatever techniques are employed, the investigator himself has to grasp the principles of statistical reasoning. The remarks in this paper, although they have dealt largely with simple topics, may have served to illustrate the fact that modern statistical principles are not something that we can take or leave as we wish, for they comprise the logic of the investigator in all fields, including the field of clinical research.

References

1. McMichael, J. 1948. Pharmacology of the failing human heart. Brit. Med. J. **2:** 927.
2. Yule, G. U. & M. G. Kendall. 1940. An Introduction to the Theory of Statistics. Griffin. London, England.
3. Greenwood, M. 1948. The statistician and medical research. Brit. Med. J. **2:** 467.
4. Mainland, D. 1948. Statistical methods in medical research. I. Qualitative statistics (enumeration data). Can. J. Res., E. **26:** 1.

* For a discussion of this investigation see the article in this monograph by Dr. D.D. Reid, Statistics in Clinical Research, Ann. N. Y. Acad. Sci. 52 (6): 931.

5. MELENEY, F. L. & A. O. WHIPPLE. 1945. A statistical analysis of a study of the prevention of infection in soft part wounds, compound fractures, and burns, with special reference to the sulfonamides. Surg. Gynec. Obst. **80:** 263.
6. Medical Research Council. 1948. Streptomycin treatment of pulmonary tuberculosis. Brit. Med. J. **2:** 769.
7. HILL, A. B. 1945. Principles of Medical Statistics. The Lancet. London, England.
8. ALBRITTON, E. C. 1948. Experiment Design and Judgment of Evidence. Edwards. Ann Arbor, Mich.
9. SNEDECOR, G. W. 1946. Statistical Methods Applied to Experiments in Agriculture and Biology. Iowa State College Press. Ames, Iowa.
10. MATHER, K. 1946. Statistical Analysis in Biology. Methuen. London, England.