Research Methodology: What the Clinician Should Know

Lawrence Garfinkel, M.A.

Many clinicians find it difficult to fully evaluate the findings of research papers published in medical journals. Theoretically, the journal's review board has selected studies that are well designed and well carried out and rejected the poorer ones. This may not, however, be necessarily true. Some reviewers who are familiar with the literature and techniques employed in their specialties, but not with research methodology, may accept a manuscript with methodologic or statistical flaws.

This paper is not concerned with case reports, descriptions of techniques, review articles or those which reflect opinions of the author, but rather with the basic elements of good research design, particularly as they pertain to experimental and epidemiologic studies.

The following should be kept in mind when evaluating a research study:

1. statement of the problem and formulation of the hypothesis;
2. review of the literature;
3. pretesting;
4. design of the study;
5. analysis of the data;
6. interpretation of results.

Statement of the Problem

The investigator should clearly state the problem and the hypothesis he is testing. Usually, this is put in the context of a larger problem of clinical management or research interest. The clinician must determine for himself the importance of the hypothesis being tested, and how it now or eventually may relate to his needs. Many researchers are concerned with extremely esoteric problems and it is sometimes difficult to understand how the investigation could conceivably be linked to clinical problems in human beings. This is not to imply that basic research is unnecessary. But it is equally true that not all medical research investigations can, even eventually, supply information that is relevant to man.

A distinction must be made between "hypothesis testing" and "hypothesis seeking." In the former, the investigator has developed a hypothesis from the work he or others have done in the past and designed a study specifically to test this hypothesis. In the latter, the investigator tests a group of variables (drugs, habits, etc.) against one or many diseases to learn if any of the variables are related. It is important to realize that in a large study, in which many variables are being tested, at least some, just by chance, will seem to be related, although in reality they are not.

It is perfectly legitimate and useful to conduct a research investigation by testing a number of variables, but the investigator cannot treat his findings as though they were the results of hypothesis testing. The proper procedure is to design another study that tests the new hypotheses, or at least to report the findings as suggestions for new areas of research, not as definite conclusions.
Review of the Literature

It has been frequently stated that no research investigation exists in a vacuum; there is always a background of data to draw on which deals with the hypothesis being tested. The investigator should refer to this information, and its relevance to the problem. Unfortunately many investigators refer obliquely to the work of a colleague, or reproduce in their own bibliography the references from another paper on the same subject, without reading the articles. Instances occur where the author misquotes another paper or cites findings that do not exist. How can the average reader of a medical journal know if a paper is being misquoted? It is admittedly difficult, and the burden of checking the references really belongs to the editorial staff of the journal.

The researcher's function is to synthesize the results of other investigations relevant to his work and report them briefly. One of the hallmarks of a good paper is how carefully this has been done.

Pretesting

Unless the degree of error in measurement is already known, as in some standard tests, the instrument being used must be pretested to demonstrate its level of error. For instance, a questionnaire should be pretested to determine how subjects will respond. A mechanical device should be tested on the same subject several times in the same way to learn the magnitude of error of the device itself. If measurement is by the subjective opinions of a panel of evaluators, there must be some pretesting to determine whether each evaluator can duplicate his own readings, as well as the level of agreement or disagreement among panelists.

Researchers refer to the reliability and the validity of a study. Reliability is the degree to which results can be duplicated. How small, for example, is the experimental error in measurement? Or to what degree can evaluators duplicate each other's findings? In studies where large differences between groups are being tested, the reliability of the measuring instrument may not be as crucial as when the differences between groups are small.

Validity refers to the degree to which the results of a study can be extrapolated to a more general group. Are the subjects of a study so unrepresentative that the results cannot hold for the entire group? Can the results of a study on a mouse population be generalized to a human situation? Can a mathematical model be extended to the situation in man?

Design of the Study

One of the most important elements in designing any study is to try to remove any bias that might affect results—or if bias cannot be completely eliminated, to be aware of its magnitude. Bias may be unconscious; it may pertain to the selection of subjects, the method of testing, the attitude of the tester, and many other factors. Because the investigator is
aware of possible bias, many use methods of random assignment in selecting, designating and testing subjects. "Double-blind" methods are often used where the tester and evaluator are unaware of which group the subjects are in.

Some investigators test too many different types of subjects at the same time. It is always best to design a study making groups as homogenous as possible. For example, if both age and sex are relevant to the investigation, it is preferable to limit the study to a narrow age range in one sex in order to maximize the number of usable subjects. Otherwise, the investigator may find that he has too few subjects in each age/sex group for a statistically appropriate analysis. In some instances, a technique called "matched pairs analysis" is employed in which subjects in different groups are matched on all variables except the factor being studied.

The investigator must be aware of the problems of sampling. Selection of patients in a hospital must be done carefully, so that their characteristics do not give a biased result. In most epidemiologic retrospective or case-control studies, the investigator selects subjects with a specific disease, and determines the proportion with a certain characteristic (habit, use of drug, etc.). He then compares this percentage with that of a control group not having the disease, which has usually been selected from another group of hospital patients. The selection of the control group is often difficult as well, since a hospital population is not necessarily representative of the general population. Thus, the distribution of the characteristic being tested may differ from that of the population at large. For example, in some early retrospective studies of cigarette smoking and lung cancer, patients with heart disease were included in the control group because the investigators did not yet know that cigarette smoking is related to coronary disease. Some investigators, aware of the difficulty in choosing proper controls, select two to three separate groups or repeat the study using different hospital settings, evaluators and controls.

The term "relative risk," used in evaluating results of a retrospective study, is derived by dividing the percentage in the case group with the characteristic being studied by a similar percentage in the control group. For example, in a recent study of antihypertensive drugs and breast cancer, there were 11 reserpine users in 150 breast cancer patients, or 7.3 percent, and 13 reserpine users in 600 surgical controls, or 2.2 percent. Thus the relative risk was 7.3 ÷ 2.2 or 3.3 to one.

In a prospective or follow-up study, the problem of choosing a proper control group does not arise since the investigator obviously has no idea which subjects will die; analysis involves comparing morbidity or mortality rates of groups with and without the factor being studied. However, a prospective study is generally more expensive than a retrospective study and takes longer to complete.

Another major problem confronting the investigator is the number of subjects to include in a research investigation. In general, the greater the number of subjects, the less likely it is that chance variations will affect the results. If the researcher expects very large differences in the groups being compared, it may not be necessary to include as many subjects as in groups where small differences are expected. Possibly the most frustrating experience in carrying out a research study is to find, at the end, that not enough subjects were included for a statistically satisfactory conclusion.

Analysis of the Data

In evaluating data, the clinician should bear in mind that in all statistical testing, the investigator is asking: "Do the differences observed between groups
represent a chance variation, or a real difference?" The statistical test (t test, chi-square test, F test) applied to the data gives the probability that the differences were due to chance. That is, if differences in means (averages) are shown with the notation "p less than .05" then the probability of the observed difference resulting from chance is five in 100 or one in 20. Generally, in most biomedical studies with a p less than five percent, the investigator rejects the hypothesis that the differences are accounted for by chance, and selects the alternative hypothesis that a real difference exists.

If critical decisions are involved in a research study, the investigator may set limits on the probability of a difference being due to chance of p less than .01 or even .001 (one in a thousand).

To avoid drawing unwarranted conclusions it must be remembered that statistical testing is only a tool and that the major principles of good research design (selection of proper subjects, elimination of biases, standardization of testing methods, etc.) are more important than a statistically significant test. With poor research design, a statistically significant difference in groups may be meaningless. On the other hand, when large numbers of subjects are studied, very small differences, which may have no clinical or practical significance, can be statistically significant.

Interpretation

In commenting on the results, the careful investigator considers all factors that may bias the results and reports the precautions he took to eliminate them. In a narrow sense, the results of the investigation are limited to the subjects being tested. In extrapolating to larger groups, the investigator must review all other pertinent work and assimilate his conclusions into the entire body of information. He should repeat his initial hypothesis and the results of the tests. Unexpected results should be stated cautiously and further testing suggested.

As an example, some years ago a researcher was investigating the month and year of birth of lung cancer patients. He discovered that a relatively large number of those in his series were born in March. When a chi-square test was applied to the data, he found that more lung cancer deaths occurred in March than could be attributed to chance and that the differences in month of birth were statistically significant. He then cited other data which showed that children whose last months of gestation occurred in the winter were deprived of vitamin A, causing an irreversible metaplasia in the critical stage when the lungs begin to function, predisposing the individual later in life to pulmonary diseases, including cancer.

This is an example of "hypothesis seeking" rather than "hypothesis testing." The hypothesis (stated in the paper as a conclusion) arose out of the data—it was not being tested by the data. In this case, the author based his conclusions on a total of 330 male lung cancer deaths. Soon after this paper appeared, several others were published which showed contradictory evidence from larger lung cancer series; there was no difference in frequency of lung cancer cases by month of birth. Thus although the difference in month of birth observed in the first study was "statistically significant," the conclusions were in error due either to chance or faulty original data.

This study emphasizes an important point in comprehending and evaluating epidemiologic studies. We are constantly reading of new epidemiologic findings in the lay press, some of which forecast imminent environmental doom for large segments of the population. These studies should be interpreted in their proper context. Obviously the original paper must be read and analyzed carefully before one can intelligently make an evaluation.