Med. Decision Making
Vol. 1, No. 3, 1981

0272-989X/81/030265-11$3.70

Deciding Whether the Conclusions of Studies are Justified: A Review

*Daniel A. Albert, M.D.**

Critical review of original studies is a major source of medical information for the physician; it necessitates a clear understanding of the research objective of the study. The design of the study, whether experimental or observational, should be appropriate to answer the research question and, further, the study should employ accurate and precise measurements on a suitable group of subjects. The critical reader, aware of the principles of study design and analysis, can assess the validity of the author's conclusions through careful review of the article.

A guide to the analysis of an original article is presented here. The reader is asked to formulate questions about a study from the abstract. The answers to these questions, which form the basis for acceptance or rejection of the author's conclusions, are in general available in the body of an article. This technique is demonstrated with two fictional studies about patients with angina.

One of the most important skills a physician should have is the ability to critically analyze original contributions to the medical literature. Specifically, a physician should be able to assess both the validity of the authors' conclusions and the applicability of these findings to his patients. The reader should depend neither on the authors nor on the journal editors to perform these tasks for him.

In spite of the importance of critical analysis, little formal instruction is offered in either undergraduate or postgraduate medical training to develop this skill. This article, therefore, offers an epidemiologically-oriented approach for the clinician to use in evaluating original medical studies.

Practitioners often perceive two major limitations in their own ability to

*Department of Pharmacology, Harvard Medical School and The Sidney Farber Cancer Institute, Boston. Address requests for reprints to Dr. Albert, Room 830D, Sidney Farber Cancer Institute, 44 Binney Street, Boston, Massachusetts 02115, USA.

critique. The first is a lack of statistical training; this engenders uncritical acceptance of numerical data and undue awe of "p values." However, the statistical operations in general medical journals are usually elementary, and with very little effort can be easily understood by clinicians [1,2]. The second perceived deficit is unfamiliarity with new or specialized technology. This deficit consigns the formal review of original articles to academicians working within the field of study. However, intimate technical knowledge is rarely necessary to determine the clinical import of original articles.

Most physicians approach an article through a casual reading of the abstract. If the subject is of particular interest they may then read further into the body of the paper. In fact, one should be prepared to approach the body of an article with questions derived from the abstract. The answers to these questions will, in general, be present in the body of the article, and will form a basis from which the reader can assess the validity of the author's conclusions. I will present two fictional abstracts, one describing a study which evaluates a new therapeutic modality, the other describing the analysis of a risk factor in disease production. I will analyze these abstracts to demonstrate the major points that a physician must obtain from the body of the paper when critically analyzing an original study. The format of this guide will follow the format used in most original articles.

Abstract 1. The Treatment of Angina with Anginol

We investigated the efficacy of anginol in a university clinic population with established coronary insufficiency. Forty-four patients were selected from a clinic population and randomized into two groups, one of which received anginol for four weeks and the other of which received placebo. The anginol-treated group had significantly ($p < 0.01$) greater exercise capacity, symptom relief, and feelings of well-being. We conclude that anginol is efficacious in the treatment of symptomatic coronary artery disease.

Abstract 2. The Role of Marital Dysfunction in the Pathogenesis of Angina

The role of marital problems in the etiology of symptomatic coronary artery disease was investigated. A recent heart association questionnaire identified 1,000 respondents with symptomatic chest pain (angina) in Stress City, New York, population one million. These respondents, as well as an age- and sex-matched group of people without angina, were later questioned for evidence of marital discord. Over 23% of angina patients reported some form of marital problem compared to 7% of persons without chest pain ($p < 0.001$). Almost half of those with angina and marital discord reported exacerbation of symptoms with increasing marital difficulties. We conclude that marital problems are a significant cause of angina.

Introduction

The introduction to a paper states the general purpose and background of the study being reported. From it the reader should be able to formulate a

one-sentence statement of the objective of the study, against which the conclusions of the study can be compared. Occasionally a study is purely descriptive, but much more often it will suggest a relationship between features of a disease process. If descriptive, the critique will center on the accuracy and completeness of the description. If an association is hypothesized it must be testable and quantifiable; further, it should be clear whether or not the association is thought to be causal. Both of our abstracts claim that certain associations exist—the first claims that anginol therapy results in symptomatic benefits, and the second claims that a social stress, marital discord, contributes to the production of angina. These hypotheses should be clearly stated in the research questions which the authors sought to investigate.

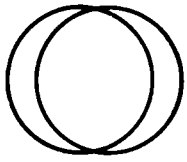
Materials and Methods

The materials and methods section addresses two separate issues: Who or what is being studied, and what measurements are being made?

Laboratory experiments customarily deal with well-characterized *in vitro* systems or populations of identical inbred strains of animals. However, this luxury is not available to clinical researchers who must select their study group from a larger population. The selection process may result in a study group which does not reflect the characteristics of the general population to which the investigator wishes to refer his findings [3]. For example, university medical center patients may be sicker or medically more complicated than community hospital patients; generalizing from studies of the former to the care of the latter may be misleading. The selection process has introduced a *transition bias* [4]. Bias refers to any process which produces study results that differ systematically from true values [5]. Studies on human populations should clearly delineate eligibility requirements as well as categories of patients that were excluded, in order to determine the applicability of results to the population of interest.

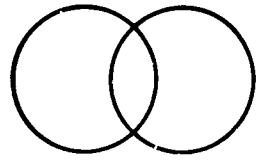
Although both of our fictional studies are investigating the nature of angina, it is clear that the two study groups are quite different. The therapeutic trial, conducted in a university setting, would likely involve patients whose disease is well documented and severe. They may also have concurrent disorders that are prognostically important. The characteristics of the base clinic population, the process by which patients were selected to be in the study, the eligibility requirements, and all prognostically important factors should be explicitly mentioned. Whereas the first study is likely to involve quite ill patients, the second is likely to include people convinced they have disease but for whom no documentation exists. Even an excellent questionnaire is unikely to achieve the high degree of specificity of diagnosis (few false positives) present in the first study. These studies may represent extremes in specificity of diagnosis and it is up to the reader to decide whether any important differences exist between the study group and his or her own patients.

The materials and methods section must also be scrutinized for the accuracy and precision of the observations being reported. Accuracy (occasionally called *validity*) is how well the measurement made reflects the quantity sought. For example, cuff blood pressures follow intra-arterial blood pressures closely, and are therefore reasonably accurate measures of intra-arterial pressure. However, resting electrocardiograms are abnormal in the absence of coronary artery disease and vice versa, and therefore do not accurately assess coronary artery disease. This may be easily visualized using Venn diagrams, as in Figure 1.



ABNORMAL CUFF BLOOD PRESSURES

ABNORMAL INTRA-ARTERIAL BLOOD PRESSURES



ABNORMAL RESTING ELECTRO-CARDIOGRAMS

CORONARY ARTERY DISEASE

Figure 1.

Precision or *reliability* refers to the reproducibility of each measurement. Many laboratory tests are quite precise and split-sample determinations yield nearly identical results. Unfortunately, the same can not be said for most measurements made by clinicians. Observer variability has been well documented in many clinical situations, such as roentgenographic interpretation [6]. In situations where precision cannot be assumed, the reader should look for measurements performed blind to the patient's status in the study and preferably more than once by different observers. In lieu of these precautions, some measure of observer variability should be sought.

In many clinical situations, it is hard to find measurements that are both precise and accurate. Investigators often measure what can be measured easily, thereby sacrificing accuracy for precision. These "proxy" measures, such as the sedimentation rate for disease activity in rheumatoid arthritis, appear attractively "hard" compared to "soft" clinical measurements, but often are, in fact, misleading. In many cases the creation of clinical indices, such as the New York Heart Association Functional Classification or the Apgar score, in conjunction with observations by several trained observers blinded to the patient's status within the study, achieves precision while retaining accuracy.

These caveats about observations are clearly more applicable to the therapeutic trial than to the risk factors analysis. Although the three quantities

measured in the first study, exercise capacity, symptom relief, and feelings of well being, are intuitively reasonable, they are quite subjective. Even exercise tolerance, the most objective, may be influenced dramatically by encouragement. Under these circumstances the reader is justified in demanding that measurements be made by an observer who is unaware of whether the patient being measured is an experimental subject or control, and further that the patient be unaware of his status as an experimental subject or a control (double blind). If a functional classification is used it should have been validated on a population whose characteristics were established by another, more reliable means.

Although observations in the strict sense are not being performed in the marital discord study, the two questionnaires used both have that function. The first questionnaire is an attempt to include subjects with angina and to exclude others; the second questionnaire must separate subjects with from those without marital discord. No matter how reasonable the questionnaires appear, some measure of their accuracy and precision should be available to the reader.

Results

This section of a paper presents both the data obtained and the mathematical manipulations of the data. These must be reviewed in the context of the architecture or design of the study [4].

Study designs fall into two major categories: (1) experiments, which involve deliberate manipulations of subjects; and (2) surveys or observational studies, which analyze the results of natural events. Experiments require a well defined, simple, planned maneuver (either therapeutic or diagnostic), and a means of randomly allocating the maneuver among the subjects to generate an experimental and a control group. The most common defect in experimental design is the lack of an appropriate control group. Surveys by definition cannot satisfy the maneuver criteria, as nature determines both the complexity and the allocation of the maneuver. But since the investigator does determine the choice of a control group, the reader should have evidence that the subjects and controls are comparable in as many ways as possible.

Most surveys test hypotheses about the causes of disease, and these surveys fall into three categories: (1) cohort (prospective) studies, where a group of people without a particular disease is well characterized and then followed for the evolution of disease over time; (2) cross-sectional studies, which compare the characteristics of a study group and control population during one time period to detect disease associations; and (3) case-control (retrospective) studies, which typically identify a group of diseased patients and compare them to an appropriate non-diseased group, to evaluate differences in pathogenic attributes. Cohort studies are the standard of validity

for surveys, but they are costly, time consuming, and require a high prevalence of disease before they are feasible. The validity of causal associations demonstrated by all types of surveys is a matter of considerable debate [7,8,9].

The trial of anginol in Abstract 1 is an experiment comparing anginol to a placebo. The reader should examine the text for evidence that the randomization was properly performed and resulted in two groups approximately equal in size, in severity of disease, and in all other factors (such as sex and age) that could affect the results (often called "confounders"). Further, it should be clear whether the control group received only a placebo or placebo in addition to baseline therapy and if so what baseline therapy.

The risk factor study in Abstract 2 is a survey. Because patients with the disease were identified first and a comparison group was then chosen and the risk factor assayed for, the study is retrospective. It is the directionality of the study—from the diseased patients to the factor—that makes this study retrospective, not the calendar dates. Cohorts of patients that existed in the past can still be studied prospectively from records, i.e., a historical cohort. By necessity, all case-control studies are retrospective and all cohort studies are prospective. Cross-sectional studies may be either, depending on whether diseased patients or a cohort of subjects was identified first. The determination of a study's directionality may be quite difficult; if a study group with a particular disease and a control group without are followed forward in time to assess their prognosis, the study is a cohort prospective study. In this case the outcome is not the presence of the disease but the prognosis of diseased versus non-diseased people. The study described in Abstract 2 is a case-control retrospective assessment of marital discord in patients with angina.

The directionality of the study design is crucial because only prospective studies can justify the assertion that the factor truly preceded the outcome. This *antecedent-consequent relationship* is pivotal to eventual claims of causality. Because of the design of the study in Abstract 1 as a trial, causality is suggested if the results are positive and the other aspects of the study satisfy the critical analysis. On the other hand, even if the study in Abstract 2 is performed impeccably, its retrospective design makes assertions of causality less convincing. Causality has been successfully inferred from case-control studies, however, such as the identification of maternal rubella as a cause of congenital malformations [10]; and conversely, even carefully constructed trials have failed to generate unequivocal causal relationships, such as in an experiment to determine whether anticoagulants lower mortality in acute myocardial infarction [11]. All in all, the design of our first study, a trial, is better suited to answer the question of whether the drug truly caused the improvement than the second study, a case-control investigation, is suited to answer the question of whether discord leads to angina. To answer

the question of whether marital discord causes angina one would need a cohort survey which defines a population free of disease, identifies those with and without discord, and then follows the entire population for the occurrence of angina. The answer would derive from assessing the difference in the incidence of angina between discord and non-discord groups.

Assertions of causality in all types of studies must rest on not only a statistical association between the risk factor and the disease, but also on the plausibility of a causal link, its coherence with other data, appropriate temporal sequence, geographic distribution, and a dose-response relationship (the greater the exposure, the greater the risk). Further proof of causality is afforded by experimentally excluding other competing hypotheses, or by reducing disease by reducing the exposure. These principles form a sort of Koch's postulates to establish causality [12].

One must be wary of hypothesized causes that may in fact be passenger phenomena. For example, ice in drinks at a cocktail party is not the cause of inebriation, it simply accompanies the alcohol. Likewise, although blacks may have a greater incidence of tuberculosis than whites, it is more likely that factors related to socioeconomic status rather than genetic predisposition are responsible for this association. These are examples of a secondary or confounding effect being confused with a cause. Factors that commonly confound results include age, severity of disease, sex, race, socioeconomic status, utilization of medical care, and associated disease processes [13]. To evaluate confounding effects in the anginol study, the reader should look for a table (or the equivalent) comparing all possible variables that might affect the outcome of treatment with anginol in the subject and control groups. Likewise, the reader should scrutinize the marital discord study for measures of known risk factors of coronary artery disease, to assess their influence on the results. These risk factors should be equivalent in the subject and control groups or they may distort the results. Also, if the outcome varies with severity of disease (or age or sex) one must be certain that the subjects and controls had equal severity of disease, as well as that the drug was effective in every severity (or age or sex) category. This is known as *analysis by stratification*, and insures homogeneity of the comparison groups [4]. A topical example is the stratification of patients with coronary artery disease by one-, two-, or three-vessel disease to judge the effectiveness of bypass surgery [14].

Once the study design has been evaluated, the numerical manipulations may be interpreted. Most of these manipulations, such as averages and ranges, are easily comprehended. Occasional fallacies arise, especially with rates and distributions, but they rarely pose real difficulties for the clinician [15,16]. However, difficulties often do arise in assessing outcome differences between subjects and controls. Both surveys and experiments produce data that quantify the difference in outcome generated by a maneuver.

For a risk factor in a survey of disease causation, this may be visualized as

		outcomes D	$\overline{D}$
Population	(subjects)RF	a	b
	(controls)$\overline{RF}$	c	d

where: RF = risk factor
$\overline{RF}$ = no risk factor
D = diseased
$\overline{D}$ = not diseased

Similarly, the results of a test in a diagnostic experiment may be represented as

		outcomes D	$\overline{D}$
Population	(subjects)T	a	b
	(controls)$\overline{T}$	c	d

where: T = positive test
$\overline{T}$ = negative test

Lastly, the results of therapy in a disease treatment experiment would be

		outcomes D	$\overline{D}$
Population	(subjects)Rx	a	b
	(controls)$\overline{Rx}$	c	d

where: Rx = treated
$\overline{Rx}$ = untreated

The results of most studies can be reduced to tables of this sort, and our two examples are not exceptions:

Study 1

	Improved	Not Improved
Anginol	a	b
No Anginol	c	d

Study 2

	Angina	No Angina
Discord	a	b
No Discord	c	d

One would like to prove that the risk factor, or test, or therapy was truly the source of the difference in the outcomes between the controls and the subjects. That is, that the maneuver in question was the only important reason a and b were different from c and d. This difference in outcome could arise from four sources: (1) inherent differences between the groups; (2) differences in the way the two groups were studied; (3) random variation; and (4) effects of the maneuver [17]. Since the author would like us to conclude that (4) is the true source, we must be satisfied that neither (1), (2), nor (3) is the source of this difference. Condition (1) can be dismissed as a problem if an accurate randomization process was used in an experiment, or if a suitable comparison or "control" group was chosen in a survey. Likewise, (2) can be eliminated if the study design applied similar procedures to the subjects and their controls, and if there was no misclassification of patients due to

inaccurate measurements. There also should be no selective loss of patients from any of the four boxes. Finally (3), the question of whether the numerical differences obtained could have occurred by chance alone, is addressed by tests of statistical significance. The probability that the difference would have occurred by chance alone, usually reported as a value of p, is most commonly calculated using the chi-square test for categorical data, or the t test for means. If the probability that the difference would have occurred by chance alone is less than one out of twenty ($p < 0.05$), the difference is called "statistically significant" at that level. It should be obvious that statistical significance is generally a necessary but not sufficient condition to demonstrate that the maneuver generated the difference in outcomes, i.e., it satisfies condition (3), but not (1) or (2). A study which has demonstrated a statistically significant difference between subjects and controls that was due to the factor being investigated must also achieve clinical significance.

Clinical significance is a difficult judgmental assessment which combines the importance of the outcome and the effectiveness of that treatment for that outcome. One mathematical handle for assessing effectiveness is the *therapeutic* or *risk ratio* $(a \times d)/(b \times c)$ calculated from the two-by-two table. If the ratio is high and the outcome is important, the study is likely to have clinical significance. This may be true even if the absolute risk to an individual is low. Witness that the addition of oral contraceptives to a woman who is prone to thromboembolic disease (by genetics or the environment, e.g., cigarettes) is of clinical significance (i.e., risk ratio is high and the outcome is important) even though the absolute risk to any individual woman is very low. Conversely, a study that shows a low ratio may still have clinical significance if the absolute risk is high and the outcome is important. Smoking has a clinically significant effect on heart disease, even though the relative risk is not high (about 2:1). In claims of therapeutic efficacy the reader must be especially aware of the importance of the outcome. Even if the therapeutic ratio is very high the study may not show efficacy if the outcome is not important—a drug that may be spectacularly effective in getting massively obese people to lose two pounds is unlikely to be a great therapeutic advance. Therapeutic efficacy should also be viewed in the context of the monetary costs and the side effects of the therapy. Both diagnostic and therapeutic efficacy can be rigorously evaluated using standard techniques of decision analysis [18]—a topic beyond the scope of this paper.

Discussion

In the last section of the paper the authors describe the associations which have been demonstrated and the implications of their conclusions for diagnostic or therapeutic practice and for further research. The critical reader must address these questions: Have the conclusions of the study answered the questions initially posed in the Introduction? If so, are the conclusions valid? Validity means that (a) the study design was not seriously flawed (i.e., no bias was introduced by the choice of the subjects, in the allocation

or handling of the groups, or in making measurements), (b) the numerical results show both statistical and clinical significance, (c) the conclusions explain the data better than other competing hypotheses, and (d) the conclusions do not assert more than the study design permits. The Discussion section should also be examined for the author's own reservations about his study and its conclusions, as well as for plausible explanations of any pertinent contradictory evidence from the literature.

Finally, if the reader is convinced that the conclusions of the study are valid, he must determine whether the results are applicable to the patients in his practice. This consideration rests on how comparable the study population is to the reader's patients.

Summary

The critical reader should entertain the following questions while critiquing a study:

Introduction—What is the objective of the study? Is it purely descriptive or has a testable hypothesis been formulated?

Materials and Methods—Who or what is being studied and what population did they come from? Are the measurements accurate and precise?

Results—If the design was experimental, was the maneuver randomly allocated and was a suitable control group generated? If the design was observational (cohort, cross-sectional, or case-control), were the comparisons appropriate? If a causal association is implied, is it plausible and have all likely confounders been dealt with? Are the numerical manipulations reasonable? Are numerical differences in outcome due to the maneuver being investigated rather than inherent differences in the groups or the way they were studied? Are both statistical and clinical significance demonstrated?

Discussion—Are the conclusions appropriate to the objectives and to the design of the study? Are they valid? To what other populations can the conclusions be generalized?

A final note—a thorough critique will leave no article untouched. Most studies lend more or less credence to their conclusions, and rarely does a study prove something with certainty. Likewise, only rarely should a study's results be completely dismissed because of deficiency in its design. Each study must be viewed in the context of the literature to which it contributes and final judgment often rests with the weight of further evidence.

Acknowledgments

The author wishes to thank Dr. David A. McKay for his guidance in developing this paper and Drs. Dan Davis, Michel Ibrahim, and Edward Wagner for their critical comments.

References

1. Altman DG: Statistics in medical papers. J R Coll Physicians Lond 2:274–280, 1977
2. Feinstein AR: Clinical biostatistics XXV. A survey of the statistical procedures in general medicine journals. Clin Pharmacol Ther 15:97–107, 1974
3. Feinstein AR: Clinical biostatistics IV. The architecture of clinical research (cont) Clin Pharmacol Ther 11:595–609, 1970
4. Feinstein AR: Clinical biostatistics. St. Louis, C.V. Mosby Co., 1977, pp 17–27, 71–88, 385–430
5. Murphy EA: Resolution, precision, and accuracy. Theory and practical aspects. In, The logic of medicine. Baltimore, Johns Hopkins Univ Press, 1976, pp 58–80
6. Lusted LB: Introduction to medical decision making. Springfield, Charles C. Thomas, 1968, pp 98–128
7. Sartwell PE: Retrospective studies. A review for the clinician. Ann Intern Med 81:381–386, 1974
8. Susser M: Causal thinking in the health sciences. New York, Oxford University Press, 1972
9. Rothman K: "Causes". Am J Epidemiol 104:587–592, 1976
10. Gregg NM: Congenital cataract following German measles in the mother. Trans Ophthalmol Soc Australia 3:35–46, 1941
11. Gifford RD, Feinstein AR: A critique of methodology in studies of anticoagulant therapy for acute myocardial infarction. N Engl J Med 280:351–357, 1969.
12. Hill B: The environment and disease. Association or causation? Proc R Soc Med 58:295–300, 1965
13. Murphy EA: Confounding. In, The logic of medicine. Baltimore, Johns Hopkins University Press, 1976, pp 263–281
14. Brymer JF, Buter TH, Walton JA, Willis PW: A natural history study of the prognostic role of coronary angiography. Am Heart J 88:139–143, 1974
15. Campbell SK: Flaws and fallacies in statistical thinking. Englewood Cliffs, Prentice-Hall, 1974
16. Huff D: How to lie with statistics. New York, Norton, 1954
17. Colton T: Statistics in medicine. Boston, Little, Brown, 1974, pp 253–322
18. Weinstein MC, Fineberg HV: Clinical decision analysis. Philadelphia, W.B. Saunders, 1980