1

Reading the Medical Literature: Assessing Reports on Clinical Intervention Studies

Renée J Robillard

Steven C Ullery

The ability to read the medical literature critically is a skill that has become increasingly important for health care professionals for three main reasons. The first is the enormous proliferation in medical information. The number of medical, allied health, nursing, and pharmacy journals published on paper and on-line has increased dramatically; health-related websites, many of which do not contain peer-reviewed data, have multiplied; and interest in medical news has blossomed in the popular media in response to baby boomers' concerns about health and aging.

The second reason is the growing sophistication of patients about health issues, resulting largely from their acquisition of medical information from sources other than their physicians, especially the news media and the Internet.

The third is the expanding influence of cost-containment initiatives in health care. Increasingly, third-party payers are requiring that the treatments they pay for be as effective and as inexpensive as possible. Often, evidence for efficacy and cost saving is found in the

medical literature, reflecting an approach to medical practice called evidence-based medicine, which emphasizes the need to move beyond clinical experience and physiologic principles to rigorous evaluation of clinical actions.[1]

Unfortunately, even the medical information published in leading medical journals can be irrelevant, inconclusive, unoriginal, confusing, misleading, poorly presented, or a combination of these distressing characteristics. Reading the medical literature therefore becomes a reader-beware situation: to promote the health of your patients and to protect yourself against problems caused by giving flawed clinical advice, you must not simply believe everything you read. But many health care professionals have little training in critically appraising medical articles. Even those who faithfully attend journal clubs may not feel completely confident about their ability to assess what they read.

In the past few years, several worthwhile guides to reading the medical literature have been published (see notes). This chapter consolidates some of the information in those publications, concentrating specifically on clinical intervention studies, that is, investigations in which researchers treat patients for a condition and then assess the results of the treatment. The intervention may be a drug, an operation, psychological therapy, or an educational program. Excellent sources of information on evaluating other kinds of studies (investigations of screening or diagnostic tests, meta-analyses and systematic reviews, economic analyses, observational cohort studies of prognosis or exposure to an agent, epidemiologic cross-sectional surveys, and case-control studies) include the books by Greenhalgh, Crombie, and Elwood and the "Users' guides to the medical literature," a series of *JAMA* articles by the Evidence-Based Medicine Working Group.[2] In addition, the website of the Evidence-Based Medicine Resource Center (http://www.ebmny.org) has an extensive bibliography of print and electronic publications on finding and interpreting the most useful clinical literature.

SELECTION OF ARTICLES

Health care professionals read scientific articles to update their general knowledge of a discipline, to acquire information for research, or (most frequently) to answer a specific clinical question. Those still in training may read because they have been assigned articles in a course. Because your time to read is limited, you want to avoid wasting it on publications that do not meet your needs. Yet computerized database searches of the literature may yield hundreds of articles, most of which will not provide substantive information on the topic in which you are interested. It is therefore important to limit the number of articles that you must find and review to get the high-quality information you want. Fortunately, there are several new ways to do this relatively quickly and efficiently, many of which were reviewed by Hunt et al.[3] These include consulting "prefiltered" evidence-based resources such as the journal *ACP Journal Club*, the collection of the Cochrane Library, and the series *Clinical Evidence;* and using the free Web-based PubMed system to make searching more productive.[4] The medical librarians or information specialists at your institution are the best sources of information on all the searching resources available where you work and how to use them; brief descriptions of some of these resources are given below.

The purpose of *ACP Journal Club* (and its computer-searchable CD-ROM form, *Best Evidence*), which is published bimonthly by the American College of Physicians–American Society of Internal Medicine (ACP-ASIM), is to "select from the biomedical literature articles that report original studies and systematic reviews that warrant immediate attention by physicians attempting to keep pace with important advances in internal medicine" and to provide commentary on each article.[5] More than 150 top journals are reviewed to identify articles that qualify as the "best literature" according to an extensive set of criteria (for example, 80 percent follow-up for a randomized trial). These articles are then abstracted, and commentary on each is

provided by an expert who discusses the context of the article, any methodologic problems with the study the article describes, and recommendations for clinical practice. Authors of the articles are given the opportunity to review the abstract and commentary before publication.

The Cochrane Library, which is available on both CD-ROM and the Web, includes the Cochrane Database of Systematic Reviews, the Database of Reviews of Effectiveness, the Cochrane Clinical Trials Registry, and a handbook for reviewing medical literature. The material in the Cochrane Library is provided by the Cochrane Collaboration, an international group that prepares, maintains, and disseminates systematic reviews of health care interventions in all fields. For the database of systematic reviews, Cochrane Collaboration reviewers search MEDLINE, EMBASE, and other sources for reports on the most important randomized studies of a topic. The reports selected are then reviewed according to strict and stated criteria. Before you do a MEDLINE search on your topic, check to see whether a Cochrane review of it has recently been written. If so, it could save you a great deal of time and effort in your search for answers to clinical questions.

Clinical Evidence is a printed book-length "compendium of evidence on the effects of common interventions"[6] that is produced by the ACP-ASIM and the BMJ Publishing Group. It is updated and expanded every six months. *Clinical Evidence* is not a textbook or a book of guidelines; it simply provides summaries of the "best" available evidence on clinical interventions or else states that there is no good evidence. The summary page for each clinical topic provides a list of questions addressed in the summary, the interventions covered, and whether or not they have been found to be effective. The publishers of *Clinical Evidence* believe that the series is complementary to the Cochrane Collaboration's systematic reviews, in that it takes the information provided by those reviews and other "high-quality" sources and puts it in one place in a concise format that is easy for busy clinicians to consult.

For example, in the therapy category, clicking on "sensitivity" will retrieve articles on your topic associated with the terms "randomized control trial" or "drug therapy" or "therapeutic use" or "all random." Clicking on "specificity" will yield articles described by the terms "all double and all blind" or "all placebo." Detailed information on PubMed's clinical queries feature is provided on the Web at http://ncbi.nlm.nih.gov:80/entrez/query/static/clinical.html.

CLINICAL INTERVENTION STUDIES

Randomized control trials (RCTs), controlled (but not randomized) clinical trials, case series, and case reports are all clinical intervention studies. The RCT is considered to provide the strongest evidence, the case report the weakest. Remember, however, that not all RCTs are unbiased (see Questions about Methods), primarily because many trials that are labeled RCTs have defects in design that prevent them from fulfilling the strict criteria for RCTs. Those criteria are the focus of this chapter.

Studies of drugs (and, in some cases, of medical devices), especially investigations that are performed to fulfill regulatory requirements, may be divided into three or four phases, which are done in order.[7] The purpose of a phase I trial of a drug is to determine a dose range that is well tolerated and produces no major side effects. Only small numbers of patients or healthy volunteers are enrolled in this kind of trial and there is usually no control group. A phase II trial is designed to provide preliminary information on whether a drug has any therapeutic value at all, that is, on whether it is effective in alleviating disease when used at a tolerable dosage. A phase II trial may or may not be controlled, but even if it is, the number of patients enrolled is usually too small to detect any but the largest treatment effects. A phase III trial is usually an RCT. When well designed, it provides definitive evidence of the efficacy of a drug and detects the most common side effects. It is this kind of study that persuades a

published investigations may be made. At best, case series and case reports suggest a testable hypothesis; the information they provide should not be used in selecting therapy for other patients.

QUESTIONS TO ASK ABOUT AN INTERVENTION STUDY

Let us say that your search of the literature on the clinical issue in which you are interested, assisted by consulting prefiltered resources, is now finished, and you have downloaded or photocopied the published reports on the best RCTs that seem relevant to the issue. You are now ready to examine closely the reports on your desk. This kind of scrutiny is most easily done by asking a series of questions about each paper. The questions can be divided into categories according to the section of the paper in which the answers will most likely be found. You can generally answer the preliminary general questions by looking at the paper's title, list of authors, Abstract, and Introduction. Perusal of the Methods section should answer your questions about the design of the study and the statistical analysis that was done. The Results section is usually the place to go for answers about outcomes, follow-up, and statistical findings. The clinical implications of the study, the consistency of its results with those of earlier studies, and the study's limitations are best discerned by examining the paper's Discussion section.

Preliminary General Questions

Is the topic of the paper somewhat original?

Few biomedical papers describe truly groundbreaking investigations but, as Greenhalgh notes, research may enhance our knowledge of a clinical issue by providing a larger or longer study, more rigorous methods, or findings in a different patient population. Does the paper you are examining look as if it offers a new approach to the issue it addresses?

Do the authors have a solid track record?

If you attend conferences in your field often and do a fair amount of professional reading, you may have an impression regarding whose work is careful and whose is substandard. On the other hand, some excellent papers may have been written by authors unknown to you. You can always do an author PubMed search to find out what the authors have written about similar topics and what journals (prestigious or not) have published their research.

Is one of the authors a statistician, or is a statistician's contribution acknowledged?

The fact that a statistician was involved does not guarantee soundness of the study design, but it does provide evidence of dependability. The universal availability of personal computers and statistical software for them has produced a great many do-it-yourself statistical analyses, especially by researchers who do not have easy or inexpensive access to a professional statistician. Although statistical software can efficiently crunch numbers and output means, standard deviations, and *P* values, it cannot replace the wisdom provided by a statistician in designing a study and choosing the most appropriate statistical analyses.

Who sponsored the study?

Government agencies and private impartial foundations cannot provide all the research funding sought today. Therefore, many studies are funded by a pharmaceutical or medical-device company, or other organization with a special interest in the findings. The leading scientific journals have policies requiring disclosure of sources of funding for research studies reported in them. Manuscripts describing industry-sponsored studies are peer-reviewed by consultants to the journals in the same manner as other manuscripts. If they pass muster, and if the topics they address are considered by the journals' editors to be of interest to readers, the manuscripts are published. It

is then up to you to decide whether the corporation acknowledged as having provided funding for a study may have influenced the findings in any way. This decision should be based on a close reading of the article.

Occasionally, authors do not disclose funding sources to a journal's editors. This is extremely risky because the authors' professional reputation will be severely compromised if the editors and the authors' colleagues discover this omission later. If you know that the research described was supported by a for-profit organization or that the authors have another (especially financial) relationship with such an organization that is not disclosed in the paper, consider why the relationship was hidden and whether that should affect your view of the paper.

Is the site of the study described sufficiently similar to where you work?

Although you sometimes read a paper for general information about an issue, you most often read for special information that *you* can use immediately to help *your* patients. If you practice family medicine in a village clinic in Central America, it may not be worth your time to read about a new screening test that uses magnetic resonance imaging, because your patients probably do not have ready access to this technology. Similarly, a paper describing the efficacy of a new antihypertensive agent in a study that enrolled only white patients may not be of much interest if most of your patients are African American. An article on a drug that may prevent second heart attacks may not be useful if you practice in the student health service of a university. As Dans and colleagues note, when clinicians are considering whether a study is applicable to their practice, they "first must decide whether the biology of the treatment effect will be similar in patients they are facing; second, their patients' risk of a target event which the treatment is designed to prevent; third, the adverse effects that may accompany treatment; and fourth, their own ability to deliver the intervention in a safe and effective manner."[8]

What was the aim of the study? What hypotheses did the researchers test? Are the conclusions reached (assuming they are valid) important to you?

In the Abstract or the Introduction section of a paper (or both), the aim of the study, according to Crombie, should be "phrased as a hypothesis to be tested or as a question to be answered. The absence of such a statement can imply that the authors themselves had no clear idea of what they were trying to find out. If this were the case it is likely that they did not find out much of interest."[9] In addition, the question asked in the study must be appropriate for the issue being addressed and relevant to your practice. You may discover that the study did not ask the "right" question. For example, a Dutch study published in the *New England Journal of Medicine* that compared conventional open anterior surgery for inguinal hernia repair with laparoscopic surgery was criticized in an accompanying editorial for addressing "the wrong clinical question," in that it compared *tension-creating* conventional open repair with *tension-free* laparoscopic repair.[10] Because most North American surgeons use *tension-free* conventional open repair, however, the editorialists pointed out that the question in which US and Canadian surgeons are most interested—whether tension-free laparoscopic repair provides better results than tension-free open repair—was unfortunately not asked by the study. Of course, if you are a surgeon who has never used laparoscopic techniques to repair inguinal hernias and you have no intention of ever doing so, you may choose not to read about the Dutch study simply because you consider it irrelevant to your practice.

Questions about Methods

Evaluating the Methods section of a paper describing an intervention study is the most important part of your critical analysis; according to Greenhalgh, "strictly speaking, if you are going to trash a paper, you should do so before you even look at the results."[11] It may also be the most intimidating part of your appraisal, chiefly because the Methods section usually contains the information on study design

results by being responsible for an observed association or by masking an association. For example, in a study indicating that a new antihypertensive agent reduces blood pressure significantly, the results are affected by the confounding factor that the patients in whom the agent was apparently effective also lost weight during the study. Sound research studies use data-analysis techniques that control for possible confounding variables.

How were the patients chosen? Were they sufficiently similar to your own patients for the study to be of interest?

The inclusion and exclusion criteria for enrolling patients in the study should be given in the Methods section of the paper you are evaluating. Without this information you cannot determine whether the study's results might be applicable to *your* patients. The patients in the study may have been generally sicker or generally healthier than your patients; they may have belonged to ethnic, socioeconomic, or age groups different from those of most of your patients; or they may have received different attention during the study than what you routinely give your patients. The study's subjects, unlike your patients, may have had nothing wrong with them other than the condition being addressed or, also unlike your patients, they may not have been smokers or consumers of alcohol.[13] Keep in mind, however, that well-designed clinical research often requires that the subjects of a study be somewhat unlike "usual" patients. For example, a study of the effects of heredity on the development of lung cancer might exclude smokers, to eliminate the possibility that lung cancer in the subjects with a putative lung cancer gene was caused by smoking rather than by the gene.

The subjects in a study should be chosen with the most important clinical issues in mind. A trial of a new treatment for osteoporosis that excluded women or an investigation of a novel technique for repairing groin hernias that excluded men might seem to have limited usefulness.

sign patients to study groups by using a computer-generated list of random numbers. Unacceptable methods of randomization include use of the last digit of a patient's date of birth or medical record number, toss of a coin, sequential assignment, and date of a patient's visit to a clinic. These methods do not produce true randomization, chiefly because clinicians in the study may know in which study group patients would be placed before they make the final decision to enter them into the randomization process. Subconsciously or not, the clinicians may then keep patients they believe will benefit from the intervention from being place in the nonintervention group.

Was the study blind?

Like randomization, blinding is important in minimizing bias in an RCT. Patients who know they are receiving a new treatment, or clinicians who know they are administering one, may believe they have observed an improvement in the patients' condition even if one has not actually occurred. A study may have a single-blind, double-blind, or triple-blind design, depending on how many of the three partners in the investigation (that is, patients, clinicians, and data analysts) do not know which subjects are receiving which treatment (active or placebo) during the study's observation period. Generally, the greater the number of partners without this knowledge, the less biased the study. Of course, it is not always possible to blind one or more of the parties. For example, a study comparing two kinds of knee operations could not include blinding of the surgeons to the type of procedure performed. A study comparing radiation with chemotherapy in treating breast cancer could not blind either the patients or those administering the treatment. Both studies, however, could include blinding of those who analyze the clinical outcomes—perhaps primary care physicians or nurses.

The Methods section should indicate how the blinding was done (for example, by giving some patients placebo pills that looked and tasted like the pills containing the active agent that was studied), who had control of the blinding code during the trial (that is, who knew

which patients were receiving a placebo and which the active agent), and when the blinding code was broken.

Were enough patients studied to reach valid conclusions?

Because clinical trials, especially RCTs, are expensive, time-consuming, and difficult to administer, researchers are motivated to enroll as few subjects as possible. However, studies with a large number of subjects tend to produce the most conclusive results become the effects of most interventions are small (rarely does an effect prove to be as dramatic as, say, treating appendicitis with surgery) and a large number is needed to discern a small treatment effect. Determining how many subjects are enough to answer the research question adequately requires making a sample-size (power) calculation before the investigation begins.

The specific type of power calculation done by a statistician varies according to the situation. Four factors are usually taken into consideration: the magnitude of the difference to be detected between subject groups, that is, the difference that is clinically important; the risk of an alpha (type I or false-positive) error, that is, of concluding that a treatment is effective when it really is not; the risk of a beta (type II or false-negative) error, that is, of concluding that a treatment is ineffective when it really is effective; and the characteristics of the data, especially the outcome measures, such as the rate of posttreatment events being tracked and variations among patients.[14]

A report on an RCT should include a justification of the study's sample size in the Methods section, the Discussion section, or both. Reports that do not mention sample size probably describe studies that enrolled too few patients. Such reports should be approached with caution; "underpowered studies are ubiquitous in the medical literature, usually because the authors found it harder than anticipated to recruit their subjects."[15] If you need to know for certain whether a study you are reviewing had an adequate sample size, a statistician may be able to do a power calculation for you by using data from the published report.

If a power calculation is given in the study report, you can check whether the stated probabilities that type I and type II errors occurred are close to generally accepted statistical standards. For a type I error, the alpha level is conventionally set at 0.05 to minimize the chance that a treatment a researcher finds to be effective really is not. For a type II error, the beta level is customarily 0.2; this is much larger than the alpha level because it is typically considered not as problematic to conclude that a truly effective treatment is ineffective than to claim that a worthless treatment is useful. However, the relative settings for these error rates depend on the objective of the study. For instance, if a disease is often fatal, if no satisfactory alternative treatments are available, and if a new intervention apparently does not have serious side effects, researchers might legitimately accept a relatively high risk of finding that a treatment is effective when it is not (large alpha error) in order to decrease the possibility of missing a useful treatment (small beta error).

Were the outcome measures (end points) appropriate?

Outcome measures indicate how a treatment helps patients. An ideal outcome measure (that is, one that is least likely to introduce bias) is well defined, specific, objective, widely accepted as being clinically important, directly observed by an independent observer, and recorded in a comprehensive database.[16] Outcome measures include death, prolongation of survival, disability, morbidity (disease recurrence or prevention), physiologic variables, physical and psychological comfort of the patients, patient satisfaction, and financial cost.

The outcome measures used in the study you are examining may vary enormously from the ideal. On the one hand, death is a well-defined, specific, objective, clinically important, and easily observed measure (though cause of death may not be). In contrast, recovery is ill defined, nonspecific, subjective, and difficult to record. Suppose a study found that a certain drug increased the concentration of a "beneficial" blood enzyme (that is, enzyme concentration was the outcome measure used) in seriously ill patients but did not prolong

If a power calculation is given in the study report, you can check whether the stated probabilities that type I and type II errors occurred are close to generally accepted statistical standards. For a type I error, the alpha level is conventionally set at 0.05 to minimize the chance that a treatment a researcher finds to be effective really is not. For a type II error, the beta level is customarily 0.2; this is much larger than the alpha level because it is typically considered not as problematic to conclude that a truly effective treatment is ineffective than to claim that a worthless treatment is useful. However, the relative settings for these error rates depend on the objective of the study. For instance, if a disease is often fatal, if no satisfactory alternative treatments are available, and if a new intervention apparently does not have serious side effects, researchers might legitimately accept a relatively high risk of finding that a treatment is effective when it is not (large alpha error) in order to decrease the possibility of missing a useful treatment (small beta error).

Were the outcome measures (end points) appropriate?

Outcome measures indicate how a treatment helps patients. An ideal outcome measure (that is, one that is least likely to introduce bias) is well defined, specific, objective, widely accepted as being clinically important, directly observed by an independent observer, and recorded in a comprehensive database.[16] Outcome measures include death, prolongation of survival, disability, morbidity (disease recurrence or prevention), physiologic variables, physical and psychological comfort of the patients, patient satisfaction, and financial cost.

The outcome measures used in the study you are examining may vary enormously from the ideal. On the one hand, death is a well-defined, specific, objective, clinically important, and easily observed measure (though cause of death may not be). In contrast, recovery is ill defined, nonspecific, subjective, and difficult to record. Suppose a study found that a certain drug increased the concentration of a "beneficial" blood enzyme (that is, enzyme concentration was the outcome measure used) in seriously ill patients but did not prolong

their survival. Despite the study's positive findings, you may not want to use the drug because you feel the enzyme outcome measure is not clinically important.

In evaluating another study, you may conclude that the ability to return to work is an inappropriate outcome measure because most of the subjects were retired. A trial of an analgesic agent in which male patients were asked only whether the drug decreased their pain may not reveal that the agent also caused erectile dysfunction. Thus, any recommendations the researchers make will be based on an incomplete documentation of outcomes.

Your clinical experience is the key to determining the relevance of the outcome measures used in a trial. If you are unfamiliar with the issue studied, you can review several similar investigations to see which outcome variables are usually assessed. Be skeptical about studies with end points that are radically different from those typically used.

It is also a good idea to check for claim inflation. For example, the Methods section of an article describing the outcome of a surgical procedure states that the researchers recorded only the time until patients were able to get out of bed on their own after the operation. This measure is relatively specific and objective because it is recorded on nursing charts. However, when you come to the Discussion section, you find that this measure has been inexplicably and hyperbolically transformed to mean "recovery from surgery." In another report, the Methods section states that the getting-out-of-bed-on-one's-own measure was used throughout the study to mean "recovery from surgery." This may be more honest—but is it clinically accurate?

Was follow-up long enough?

The duration of follow-up should be appropriate for the clinical question being addressed and sufficiently long to detect major adverse or beneficial events, or to give a strong indication that such events are extremely unlikely to occur in the future. A study of a medication used to relieve postoperative pain may correctly have a follow-up time of only a week or two, because that is roughly how long patients need

group. However, statistical methods can be used to adjust for such baseline differences between groups. Check to see that this adjustment was made, but remember that the stronger the relation between the basic characteristics of the subjects and the outcome of the study, and the smaller the sample size, the more the differences between the study groups will weaken any assumption made about treatment efficacy.[17]

Careful study reports usually contain a table or figure (perhaps in Methods, but often in the Results section) listing the basic characteristics of the study groups and "usually giving the mean or median for the principal measurements together with an indication of how the subjects vary (for example, the standard deviation or interquartile range)."[18] Reports with no reference to the comparability of study groups may describe trials in which the groups were not really comparable because of faulty randomization that resulted in bias in group assignment.

It is important to remember that a big difference in the number of subjects in the study groups does not necessarily indicate a flaw in the study's design. Several statistical methods will take care of this type of situation; an example is an analysis based on proportions. The researchers should explain the difference, however, because it may be related to the issue of study dropouts (see Questions about Results).

2. Was an obscure or exotic statistical test used? Were several statistical tests used? Was there a lot of subgroup analysis in a study with a relatively small sample size? Are the statistical tests mentioned in the Methods section the same as those discussed in the Results section?

Asking these questions will help you to judge whether a study's investigators engaged in the practice of *P*-value fishing, that is, analyzing and reanalyzing data until a significant *P* value (usually <0.05) emerges. Even if you know little about statistics, you can (simply because you have read many reports on clinical studies) recognize the names of the common statistical tests. If the study report "describes a standard set of data that has been collected in a standard way but the test used is unpronounceable and not listed in a basic statistics textbook, you should smell a rat."[19] The odor may be less offen-

are included in the group to which they were originally randomly assigned, even if they did *not* receive their assigned treatment (or nontreatment) because they stopped coming for office visits, did not take their medication consistently, did not have the operation for which they were scheduled, or were inadvertently given a placebo instead of an active agent (or vice versa). Often the subjects who do not comply with a study protocol are those who are doing especially poorly; thus, excluding them from the data analysis would introduce bias because only the results in the subjects who did well would be considered. An ITT analysis "preserves the value of randomization: prognostic factors that we know about, and those we don't know about, will be, on average, equally distributed in the two [study] groups, and the effect we see will be just that due to the treatment assigned."[23]

The ITT design was originally used in pharmaceutical trials in which a drug was compared with a placebo or no treatment; therefore, a large therapeutic effect in the patients given the drug was anticipated. In such a situation, an ITT protocol is a very conservative approach to data analysis. Because the subjects are analyzed in the group to which they were originally randomly assigned, any noncompliance with the treatment will serve only to reduce the possible observed differences between groups. Thus, when a large difference is observed despite these conditions, it is very convincing.

In other circumstances, however, an ITT protocol is not appropriate. In an investigation to determine whether two treatments have the same efficacy, an ITT protocol can make demonstrating "equivalence" easier. Consider a trial designed to show that a standard drug and a new drug for arthritis have equivalent anti-inflammatory activity. Such a study might be conducted if the new drug is much less expensive than the standard drug. Suppose an ITT protocol is used and that the new drug is really not as effective as the standard drug. Then any problems that occur in the randomization process (for instance, patients randomly assigned to receive the standard drug actually get the new drug) will enhance the ability to show that the

drugs are equivalent by reducing the difference in outcomes between the two groups of patients. Thus, the trial will not correctly demonstrate that the new drug is inferior.

An ITT protocol is also not the best choice for a phase II trial, in which the principal objective is to ascertain whether the agent being studied has even a hint of therapeutic value. Also, if the goal of a study is to evaluate the safety of a treatment, an "as-treated" analysis is more appropriate than an ITT protocol.

Questions about Results

The Results section of a paper describing an intervention study usually contains tables and figures as well as text, and all should be examined carefully to help answer the following questions (as well as to see whether the numbers in them add up correctly!). As with the questions about methods, some of those about results relate to statistics. Again, even if you have limited statistical expertise, you can spot red flags indicating problems with the study's data analysis.

Do the Results section and the Methods section match?

The descriptions of a study's groups of patients, outcome variables, and clinical and statistical evaluations that appear in the Results section of a paper should match those given in the Methods section. This seems obvious, but sometimes researchers refrain from mentioning test results that did not appear to add much information to their findings. Yet the very fact that little information was obtained (the test results were perhaps inconclusive) might be quite important.

Was follow-up as complete as possible? Were all patients accounted for? Who dropped out of the study and why?

These questions should be asked because patients who drop out of a research study are usually different from those who remain in the study. If the dropouts are not accounted for, the study will be biased in favor of the results in the patients who completed it. Patients withdraw from studies for several reasons: incorrect entry into the

trial (it is discovered only after enrollment that the patient did not meet the inclusion criteria), a suspected adverse reaction to the intervention being studied, loss of interest in participating, clinical reasons (for example, pregnancy), loss to follow-up (the patient moves), or death.[24] Patients who drop out cannot simply be forgotten; instead, researchers are obligated to try to find out exactly why they disappeared and to report their findings in the paper describing the study.

Some trials appear to have a substantial number of patients lost to follow-up. How can you tell if this number is excessive? Guyatt, Sackett, and Cook suggest that one should assume that all patients lost from the treatment group did badly and that all lost from the control group did well, and then recalculate the outcomes under these assumptions.[25] If the conclusions of the trial change, the study's claims may be based on weak evidence.

How are outliers handled?

An outlier is an unusually high or low value (representing, for example, a result on a laboratory test) that is most likely to appear in a table in the Results section of the paper you are reviewing. If you see such a number, check to see whether the authors explain or comment on it. Reasons for outlying results include individual variations among patients and errors in measurement, interpretation, calculation, or proofreading. In statistical analyses, a few outlying values "can pull against the bulk of the data, creating misleading effects."[26] Thus, if outliers are present, researchers should explain how they were dealt with during the data analysis.

Were changes made in the study protocol after the trial began, to save time or money or because of untoward events?

In general, intervention studies should not change horses in midstream or return to the stable before the end of the race. There are exceptions to this rule. One is a study that is stopped because the patients in the treatment group began to experience serious adverse effects clearly related to the therapy. A second is an investigation in

which the patients in the standard-treatment or placebo group are obviously doing so much worse than those receiving a new treatment that it would be unethical to withhold the new therapy from them any longer. Sometimes a study is stopped because the patients in the standard-treatment group are clearly doing better than those in the new-drug group. For example, early in 2000, one part of a large National Heart, Lung, and Blood Institute study of treatments for hypertension was stopped because one of the drugs being tested (doxazosin, an alpha-adrenergic blocker) was observed to be less effective than a traditional diuretic (chlorthalidone) in reducing cardiovascular events.[27]

If a legitimate protocol change was made in a study, the reasons behind it should be thoroughly explained in the paper. Some studies are shortened or otherwise altered because a sponsor stopped sending money, a researcher needed to rush publication because he was up for tenure, or the only data collector in the study with whom the subjects felt comfortable quit her job and patient follow-up visits dropped precipitously. None of these are scientifically acceptable reasons for modifying a study protocol, and of course the paper is unlikely to say that the protocol *was* changed—for these or any similar reasons. However, if many data seem to be missing from the paper's Results section, or a high subject dropout rate (more than 15 percent) is noted, you may feel some skepticism about the study.

Are both P *values and confidence intervals reported?*

Most clinicians understand that a *P* value is a measure of the probability that an observed difference in outcomes between groups was a result of chance. A low ("statistically significant") *P* value suggests that there was a true difference, in that it indicates that it is unlikely that the researchers made a mistake by observing a difference where none existed. The traditional cutoff point for significance is 0.05; however, the value used is a matter of choice and is often lower.

Reporting only a *P* value in a study may be insufficient, however. In fact, according to Yancey, "when statisticians talk to each other . . .

dence in those results if the "significant" *P* value is set lower than 0.05. On the other hand, the presence of CIs might indicate a degree of sophistication about statistical analysis particularly and study design in general that helps to assure you that the researchers really knew what they were doing.

Are the effects of the intervention expressed in terms of benefits and drawbacks for your patient?

When you recommend a treatment to a patient, she or he usually does not want to know whether it has produced a statistically significant difference in an outcome measure in a clinical trial. Instead, your patient and all other patients want to know how much better off they personally will be if they take the treatment, particularly if it has unpleasant or potentially dangerous side effects. Writers of papers on intervention trials should do their best to help you help your patients to understand the risks and benefits to them of a certain therapy. For this reason, researchers should report *all* the adverse effects of a treatment they observed, perhaps in a table in the Results section. Most treatments have some adverse effects, although they may be minor. If researchers report that no adverse effects occurred, they should indicate that they conducted a careful search for such problems that came up empty, and they should describe that search.

Researchers can also present values for relative risk reduction, absolute risk reduction, and number needed to treat. Calculation of these values answers the following questions.

1. How large was the treatment effect?

Treatment effect can be represented by the absolute risk reduction, which is the difference between the proportion of subjects with an adverse outcome (death, morbidity) in the control group and the proportion with an adverse outcome in the treatment group; by the relative risk, which is the risk of an adverse outcome in patients in the treatment group relative to that in control patients; or by the relative risk reduction, which is the complement of the relative risk

ing, association is only a clue, meaning more study or confirmation is needed."[30] If a paper on a clinical trial does conclude with claims of causation, you may evaluate those assertions by asking the following questions.

Are the results plausible?

Do the results make biologic sense? Check to see whether the researchers have provided a biologic justification for their findings with respect to what is known about the disease process and the intervention agent. Fletcher, Fletcher, and Wagner relate the following story:

> Some years ago, medical students were presented a study of the relationship between the cigarette smoking habits of obstetricians and the vigor of babies they delivered. Infant vigor is measured by an Apgar score; a high score (9–10) indicates that the baby is healthy, whereas a low score indicates that the baby might be in trouble and require close monitoring. The study suggested that smoking by obstetricians (not in the delivery suite!) had an adverse effect on Apgar scores in newborns. The medical students were then asked to comment on what was wrong with this study. After many suggestions, some finally said that the conclusion simply did not make sense. . . . It was then acknowledged that, although the study was real, the "exposure" and "disease" had been altered for the presentation.[31]

The lesson of this tale is that readers of a research paper must remember to think outside or beyond that paper. That is, they must avoid getting so involved in the authors' methods, assumptions, and conclusions that they ignore their common sense and clinical experience regarding causation.

The timing of the treatment effect reported in a paper should also be examined for plausibility. For example, you may have some doubts about a study claiming that a new antibiotic eliminated all pathogens from the blood within six hours. In addition, in most situations, the

treatment studied should be found to be effective in a wide range of subjects, that is, in patients of both sexes and various ages. If it is not, the evidence for causation is not consistent.

Are the results consistent with those of other studies?

Well-written study reports include a concise, relevant literature review in the Discussion section. Here the researchers discuss studies similar to theirs and indicate whether the findings of their study confirm or diverge from those of previous investigations. A study with results that are completely different from those of similar studies requires a particularly careful evaluation, especially if the other studies have a more powerful study design. If you suspect that the literature review is unbalanced (only studies with similar results are mentioned), you may find it helpful to read reviews on the topic addressed by the study to get a better overview of the findings of earlier investigations.

Have the authors discussed possible limitations of the study?

Few studies are definitive, because most have limitations that may have produced bias. Not all sources of bias in a given study are known or suspected by the researchers, but some certainly are. For example, the sample size may have turned out to be too small to answer the research question, according to the researchers' own power calculation; or the data-collection method may have introduced a possibly confounding variable. For instance, in a study of an appetite suppressant, it may be that most of the subjects who took the agent were weighed midmorning and most of those who took a placebo were weighed right after lunch. Thus, the time of the weighing may have confounded the study's finding that those who took the drug lost weight more quickly than those who did not.

Researchers should include an honest description of the limitations of their study in the Discussion section of their study report, mentioning such factors as selection bias, follow-up method and time, and sample size. If the researchers omit this explana-

treatment studied should be found to be effective in a wide range of subjects, that is, in patients of both sexes and various ages. If it is not, the evidence for causation is not consistent.

Are the results consistent with those of other studies?

Well-written study reports include a concise, relevant literature review in the Discussion section. Here the researchers discuss studies similar to theirs and indicate whether the findings of their study confirm or diverge from those of previous investigations. A study with results that are completely different from those of similar studies requires a particularly careful evaluation, especially if the other studies have a more powerful study design. If you suspect that the literature review is unbalanced (only studies with similar results are mentioned), you may find it helpful to read reviews on the topic addressed by the study to get a better overview of the findings of earlier investigations.

Have the authors discussed possible limitations of the study?

Few studies are definitive, because most have limitations that may have produced bias. Not all sources of bias in a given study are known or suspected by the researchers, but some certainly are. For example, the sample size may have turned out to be too small to answer the research question, according to the researchers' own power calculation; or the data-collection method may have introduced a possibly confounding variable. For instance, in a study of an appetite suppressant, it may be that most of the subjects who took the agent were weighed midmorning and most of those who took a placebo were weighed right after lunch. Thus, the time of the weighing may have confounded the study's finding that those who took the drug lost weight more quickly than those who did not.

Researchers should include an honest description of the limitations of their study in the Discussion section of their study report, mentioning such factors as selection bias, follow-up method and time, and sample size. If the researchers omit this explana-

tion altogether or do not cover limitations, you may well question their findings.

Do the study's findings have clinical importance, regardless of whether they have statistical significance?

This question, one of the most important you can ask about a study, is a summary query related to several specific questions discussed earlier. A clinically important finding is one that has implications for patient care, particularly *your* care of *your* patients, whereas a statistically significant finding represents a conclusion that there is a low probability that an observed event occurred by chance. As Lang and Secic observe:

> Statistical significance essentially reflects the influence of chance on the outcome; clinical importance reflects the biological value of the outcome. In general, small differences between large groups can be statistically significant but clinically meaningless. A difference of 0.02 kg in the weights of two groups of adults is not likely to have any clinical importance even if such a difference would have occurred by chance less than 1 time in 100 ($P<0.01$) or even less than 1 time in 100,000 ($P<0.00001$).
>
> It is also true that large differences between small groups can be clinically important but not statistically significant. In a study of 20 patients in which even 1 patient dies, the death is clinically important, whether or not it is statistically significant.[32]

Other differences between clinical importance and statistical significance pointed out by Lang and Secic are that statistics are derived from groups, whereas medicine is practiced on individuals; statistical answers are probabilistic, whereas medicine requires committed decisions; and statistical analysis always requires measurement, whereas medicine sometimes requires intuition. A well-designed study incorporates a balance, so that clinically relevant differences are found to be statistically significant, whereas differences that are not relevant are found not to be significant.

YOU, YOUR PATIENTS, AND FLAWED STUDY REPORTS

There remains a "wide chasm between what a trial should report and what is actually published in the literature."[33] Several reviews have described serious defects in many of the study reports published in top biomedical journals.[34] Medical journal editors and formal and informal groups of clinicians, researchers, statisticians, information specialists, university administrators, and government officials are currently addressing these problems in several ways: (1) by developing statistical and nonstatistical guidelines for reporting clinical trials and laboratory studies; (2) by taking steps to mitigate publication bias (the tendency for papers that report positive results to be accepted for publication more often than those reporting negative results); (3) by increasing the use of statistical consultants; (4) by articulating policies on ethics in scientific publications; and (5) by educating readers through the use of print and electronic media. In the meantime, it is up to you to carry the torch of biomedical literature evaluation to illuminate the risks and benefits of therapeutic options for the patients who depend on you.

NOTES

1. Guyatt G. H., and Rennie, D. Users' guides to the medical literature. *JAMA* (1993) 270:2096–97; and Oxman, A. D., Sackett, D. L., and Guyatt, G. H., for the Evidence-Based Medicine Working Group. Users' guides to the medical literature. I. How to get started. *JAMA* (1993) 270:2093–95.

2. Greenhalgh, T. *How to Read a Paper: The Basics of Evidence Based Medicine*. London: BMJ Publishing Group, 1997; Crombie, I. W. *The Pocket Guide to Critical Appraisal: A Handbook for Health Care Professionals*. London: BMJ Publishing Group, 1996; Elwood, J. M. *Critical Appraisal of Epidemiological Studies and Clinical Trials*. 2nd ed. Oxford: Oxford University Press, 1988; and Guyatt and Rennie, Users' guides.

3. Hunt, D. L., Jaeschke, R., McKibbon, K. A., for the Evidence-Based Medicine Working Group. Users' guides to the medical literature. XXI.

19. Greenhalgh, *How to Read a Paper.*
20. Crombie, *Pocket Guide.*
21. Fletcher, Fletcher, and Wagner, *Clinical Epidemiology.*
22. Crombie, *Pocket Guide.*
23. Guyatt, Sackett, and Cook. Users' guides II A.
24. Greenhalgh, *How to Read a Paper.*
25. Guyatt, Sackett, and Cook. Users' guides II A.
26. Crombie, *Pocket Guide.*
27. United States National Library of Medicine. Clinical alert: NHLBI stops part of study—high blood pressure drug performs no better than standard treatment. Available at http://www.nlm.nih.gov/databases/alerts/bloodoo.html.
28. Yancey, J. M. Ten rules for reading clinical research reports. *Am J Surg.* (1990) 159:533–539.
29. Guyatt, G. H., Sackett, D. L., and Cook, D. J., for the Evidence-Based Medicine Working Group. Users' guides to the medical literature. II. How to use an article about therapy or prevention. B. What were the results and will they help me in caring for my patients? *JAMA* (1994) 271:59–63.
30. Cohn, *News and Numbers.*
31. Fletcher, Fletcher, and Wagner, *Clinical Epidemiology.*
32. Lang and Secic, *How to Report Statistics.*
33. Begg, C., Cho, M., Eastwood, S., Horton, R., Moher, D., Olkin, I., Pitkin, R., Rennie, D., Schulz, K. F., Dimel, D., and Stroup, D. F. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. *JAMA* (1996) 276:637–639.
34. Yancey, Ten rules; Altman, D. G. The scandal of poor medical research: We need less research, better research, and research done for the right reason. *BMJ* (1994) 308:283–284; Hall, J. C., Mills, B., Nguyen, H., and Hall, J. L. Methodologic standards in surgical trials. *Surgery* (1996) 119:466–472; Moher, D., Dulberg, C. S., and Wells, G. A. Statistical power, sample size, and their reporting in randomized controlled trials. *JAMA* (1994) 272:122–124; and Pocock, S. J., Hughes, M. D., and Lee, R. J. Statistical problems in the reporting of clinical trials: A survey of three medical journals. *N. Engl. J. Med.* (1987) 317:426–432.