
CME review article

This feature is supported by an unrestricted educational grant from AstraZeneca LP

The reading, writing, and arithmetic of the medical literature, part 3: critical appraisal of primary research

Lyndon Mansfield, MD

Objective: To offer suggestions that will help clinicians improve their scrutiny of the medical literature and apply these suggestions to their own medical writings.

Data Sources: Literature searches began at the National Library of Medicine's online database and were traced to primary sources.

Study Selection: Studies were selected for their ability to provide historical background, discuss the aspects of study design and statistical analysis, and explore important facets of reading and writing medical manuscripts.

Results: Physician readers will become more proficient in their skills as both users and creators of the medical literature.

Conclusions: Reading and interpretation of the medical literature requires a set of skills that can be learned. Similarly, good medical writing skills can be developed. Achieving these skills will enhance the clinician's practice of medicine.

Ann Allergy Asthma Immunol. 2006;96:7–16.

Off-label disclosure: Dr Mansfield has indicated that this article does not include the discussion of unapproved/investigative use of a commercial product/device.

Financial disclosure: Dr Mansfield has indicated that in the last 12 months he has served as a speaker's bureau member and has received research support from AstraZeneca LP.

Instructions for CME credit

1. Read the CME review article in this issue carefully and complete the activity by answering the self-assessment examination questions on the form on page 2.
2. To receive CME credit, complete the entire form and submit it to the ACAAI office within 1 year after receipt of this issue of the *Annals*.

INTRODUCTION

Critical appraisal is the careful evaluation of the quality and credibility of published research—encompassing the rationale, design, execution, interpretation, presentation, and application of research findings.¹ Ideally, it is what peer reviewers do. However, it takes time, training, and perspective to do well, and it requires not merely reading or understanding an article but studying and “dissecting” it. In view of the large number of studies that have found high error rates in the published literature, it appears that critical appraisal is not done often enough or well enough. Thus, readers must evaluate the quality of an article on their own.

The problem with the scientific literature that critical appraisal addresses is nicely illustrated by the following quotes from 2 authorities in the field of biomedical publications. Frederick Mosteller, PhD, retired professor of biostatistics

and health policy management at Harvard University, noted, “When a well-done trial or experiment or observational study is fairly, honestly, and thoroughly reported, it will have so many warts and footnotes and exceptions that it may be hard for the uninitiated to believe that the work was of high quality.”² Drummond Rennie, a senior editor of *JAMA*, wrote, “There seems to be no study too fragmented, no hypothesis too trivial, no literature citation too biased or too egotistical, no design too warped, no methodology too bungled, no presentation of results too inaccurate, no argument too circular, no conclusions too trifling or too unjustified, and no grammar and syntax too offensive for a paper to end up in print.”³

Critical appraisal is the skill that allows readers to distinguish between the “warts and footnotes and exceptions” that can accompany even high-quality research reports and the “fragmented . . . trivial . . . biased . . . warped . . . bungled . . . inaccurate” elements that can plague poor research reports. In a larger sense, critical appraisal is the first step in practicing evidence-based medicine (EBM). (In fact, EBM is in large part literature-based medicine.)

Western Sky Medical Research, and Department of Pediatrics, Texas Tech Regional Health Science Center, El Paso, Texas.

Received for publication November 11, 2004.

Accepted for publication in revised form February 28, 2005.

Evaluation of research reports should begin with a healthy skepticism regarding research quality. Assume that research is flawed, inadequately documented, and poorly presented until you have evidence to the contrary, and do not be swayed by the reputations of the researchers, their institutions, and the journals in which their reports are published. In this article, the last in a series of 3, I review some of the factors that should be considered when evaluating research reports. I offer suggestions that will help clinicians improve their scrutiny of the medical literature and apply these suggestions to their own medical writings. All referenced material is based on literature searches that began at the National Library of Medicine's online database and were traced to primary sources. Studies were selected for their ability to provide historical background, discuss the aspects of study design and statistical analysis, and explore important facets of reading and writing medical manuscripts.

APPROACHES TO EVALUATING RESEARCH QUALITY

Hierarchy of Evidence

Critical appraisal is not the first or only method used to determine research quality. In 1984, Green and Byar⁴ proposed a "hierarchy of evidence" for grading research (Table 1). In their schema, research designs are ranked by level of confidence in the validity of results. For example, results from randomized controlled trials generally carry more weight than results from cohort trials, which in turn carry more weight than results from case-control studies.

Randomized trials, however, have drawbacks. They are time-consuming and expensive, and they cannot always be used to study relationships of interest, such as differences between medical and surgical treatments, rare conditions, and chronic diseases. In addition, results from a well-conducted retrospective trial may be more credible than results from a poorly conducted randomized trial. Research design, by itself, is not sufficient for evaluating study quality.

Quality Scores and Checklists

Another approach to determining research quality involves quality scores, which are used to weigh the presence or absence of certain features of research design, research activities, and research reporting. Dozens of such scoring mechanisms have been developed,⁵ but results from 2 recent studies of the value of quality scores have found them wanting.

Table 1. Hierarchy of Evidence From Least to Most Valid

Anecdotal case reports of single patients
Case series without controls
Case series with historical controls
Analyses of clinical databases or registries
Case-control (retrospective) studies
Cohort (prospective) studies
Single randomized controlled trials
Confirmed randomized controlled trials
Meta-analysis of randomized controlled trials

In the first study, Balk et al⁶ evaluated the presence or absence of several "quality features" in studies identified in a large systematic review of the literature. They found no association between any feature and systematic differences in the direction or magnitude of an effect.

In the second study, a meta-analysis of 17 trials, Juni et al⁷ used 25 different scaled checklists to identify high-quality trials. Agreement among the checklists in identifying high- and low-quality studies was low, and both the direction and size of the pooled effect of the high-quality studies depended on which checklist was used.

Checklists are an outgrowth of quality scores. Authors and journal editors use checklists not to measure study "quality" but to ensure that an article addresses at least the most important elements of the study being reported. The CONSORT (Consolidated Standards of Reporting Trials) statement for reporting results from randomized trials is probably the most important of these checklists.⁸⁻¹⁰ Checklists proposed for other trial designs include the QUORUM (Quality of Reporting of Meta-Analyses) statement for meta-analysis of randomized controlled trials,¹¹ the MOOSE (Meta-analysis of Observational Studies in Epidemiology) checklist for meta-analysis of observational studies,¹² and the STARD (Standards for Reporting of Diagnostic Accuracy) statement for reporting characteristics of diagnostic tests and details of test development.¹³ Quality checklists are essential in the critical appraisal of research: the process for determining the credibility of research and research findings.

Research Sponsorship

Corporate sponsorship of research studies has the potential to affect the quality and credibility of an analysis. A large pharmaceutical company that sponsors a study can control the study design, investigators, and sites, as well as interpret findings and present conclusions without independent review. Many journals recognize this potential conflict, and address it in their guidelines for authors. It is common for journals to require a statement to the effect that "the authors had complete control over what they published without interference from the sponsoring agency." Such statements are usually provided on the title page of a submitted manuscript and reiterated in the cover letter to the editor, but journal editors are not required to publish the statement with the article. Industry-sponsored research is not de facto biased. In many cases, it is of higher quality than that conducted in academia. Readers should look to the editorial guidelines of each individual journal to find out how this issue is addressed.

KEY POINTS OF CRITICAL APPRAISAL

In critical appraisal of articles, the more that is known about research designs and activities, statistical analysis, measurement techniques, and the medicine relevant to the research, the better. I describe several key points herein. Details are available elsewhere.¹⁴⁻¹⁸

The issue that drives critical appraisal is the ability to evaluate how much a study is affected by error and bias. Error

is inherent in all studies. Types of error include *random error*, the biological variation in the variable itself; *sampling error*, which results from studying only a sample of the population of interest; and *measurement error*, which is introduced through imprecision in measurement instruments. Another type of error, *bias* or *systematic error*, occurs when some aspect of the study uniformly pushes results away from the “truth.”

Random error, sampling error, and measurement error can be quantified or estimated to determine their effect on results. Bias or systematic error, however, is much more difficult to detect, quantify, and prevent. Common biases are described herein.

Pragmatic vs Explanatory Trials

If research is to be evaluated appropriately, its underlying intent must be clear.^{19,20} At one end of the spectrum are *explanatory* or *efficacy studies*, which are performed to understand a disease or biological process. Such studies are usually conducted under “optimal” or “laboratory” conditions that allow tight control over patient selection, treatment, and follow-up. Results from such studies may provide insight into underlying biological mechanisms, but often they do not generalize well to clinical practice, an environment that cannot be tightly controlled. For example, allergy-challenge test results are likely to differ from community-acquired exposure trial results simply because the circumstances of exposure are different.

At the other end of the spectrum are *pragmatic* or *effectiveness studies*, which are performed to guide decision making. These studies are usually conducted under conditions similar to those of community medical care. Results from such studies may be distorted by factors for which controls are not implemented. Although the explanatory power of these results may thereby be limited, their applicability to clinical practice may be enhanced. For example, patients in a pragmatic trial are more likely to be heterogeneous in demographic and clinical characteristics than are patients in an explanatory trial (usually the latter must meet strict inclusion criteria). As a result, treatment effects or group differences are often smaller in pragmatic studies than in explanatory studies.

Some researchers try to satisfy both explanatory and pragmatic criteria and, as a result, do neither well. For purposes of critical appraisal, criteria used to evaluate an explanatory trial differ substantially from those used to evaluate a pragmatic trial.

Sampling Bias

If sampling for a clinical trial were only as easy as sampling a dinner wine! Unlike wine sampling, in which we are fairly sure that a taste is representative of the remainder of the bottle, a sample in a clinical trial may or may not be representative of the population of interest.

There are many types of sampling bias. One of the most important is *referral-filter bias*, which occurs as certain pa-

tients drop out of the sample pool while moving through the health care system. For example, patients with severe asthma successfully treated by local specialists are unlikely to be included in an asthma study conducted at a tertiary care facility, which accepts only those patients whose asthma is refractory to most common therapies. Similarly, because of referral-filter bias, a sample of patients at the tertiary care facility is unlikely to be representative of patients with asthma in general. For this reason, the CONSORT statement recommends that the source of patients should be identified.^{8–10}

Volunteer bias refers to the fact that patients who volunteer for clinical trials often differ from those who do not. That is, volunteers are more likely to take risks and to have more confidence in the health care system. This bias may be affected by study characteristics. For example, some patients may not volunteer for a study in which they have an equal chance of being assigned to a minimally invasive outpatient procedure or to an open surgical procedure that requires hospitalization.

Assignment Bias

In case-control and cohort studies, groups are defined by their diagnosis or by their exposure to a given condition or intervention. Not surprisingly, the most common source of bias in these studies arises from differences in case definitions or from ascertaining the nature or degree of exposure. For example, how is exposure to secondhand smoke measured, and how do we know if the measurement is accurate? In retrospective studies, *recall bias* can be a problem because memory is selective. Patients’ reports as to quantity and duration of exposure are subject to recall bias. Even when patients are followed forward in time, ascertaining the degree of exposure can be difficult.

In experimental trials, the gold standard for group assignment is random assignment. In *random assignment*, each patient has a known and usually equal probability of being assigned to either the treatment group or the control group. Simple random assignment does *not* ensure that the groups will be similar at baseline or that they will be of equal size. (However, strategies such as blocking and stratification can ensure these conditions.) Random assignment *does* ensure that any differences between groups—known *and* unknown—are the result of chance and not bias.

Random assignment alone is not sufficient to prevent bias. Each study’s allocation schedule, a list of patient numbers and group assignments, must be concealed from individuals enrolling and treating patients. To ensure *allocation concealment*, investigators restrict access to the schedule: for example, it may be kept at a central data coordination center and seen only by the supervising statistician. When an investigator calls to enroll a patient in the study, the statistician, not the investigator, assigns the group. Allocation concealment prevents investigators from manipulating group assignments by, for example, waiting to enroll a particular patient until a desired group assignment becomes available. Indeed, effect

sizes reported in some studies with inadequate or un concealed allocation have been up to 30% higher than effect sizes reported in similar studies with adequate allocation concealment.²¹

Allocation concealment keeps each patient's group assignment secret before assignment. In contrast, *blinding* keeps this information secret after assignment. Blinding prevents bias during data collection and guards against situations in which patients in the treatment group are watched more closely for adverse effects or improvements than are patients in the control group. In higher-quality studies, the success of blinding is evaluated.⁸⁻¹⁰

Measurement Bias

Measurement bias occurs when measuring instruments systematically overvalue or undervalue a variable. In assays or laboratory tests, this bias is prevented by calibrating the instrument against an accepted reference standard and reporting the details in the "Methods" section of the publication. When measurement involves researcher judgment, as in typing cells or interpreting medical images, measurement bias can be quantified by testing the intrarater and interrater reliability of the judges.

A more subtle form of measurement bias is *lead-time bias*, which is introduced into a time-to-event or survival analysis when the beginning of the period under study depends on how that beginning is determined. For example, sensitive equipment that can detect disease at its earlier stages increases time to event not by prolonging the event but by moving the beginning of the event backward.

Analytical Bias

Bias can also occur in analyzing and interpreting study results. For example, most studies have a few dropouts. When the dropout rate is more than approximately 15%, however, results should be interpreted with caution.¹⁸ However, patients may drop out of a study *because* of the intervention: it is too painful, has adverse side effects, requires too much time, and so on. Thus, analyzing only those patients who completed the study can bias the results by ignoring the negative effects of the intervention. Thus, in *intent-to-treat analysis*, all patients (study completers and dropouts) are analyzed within the group to which they were assigned; in contrast, *on-protocol analysis* includes only study completers. Better studies may report both types of analysis: intent-to-treat analysis to determine whether the treatment itself might be harmful and on-protocol analysis to determine whether the planned protocol was effective.

Another form of analytical bias is associated with the *baseline risk of an event*. The event rate in the control group is often used as a surrogate for baseline risk. In trials in which the event rate in the control group is low, the treatment is unlikely to decrease the incidence of the event in the treatment group, regardless of how good the treatment might be. That is, if the goal of treatment is to prevent death, someone

in the control group has to be dying simply to establish that death was a possibility that could be prevented.

Confounding, the third type of analytical bias, occurs when the effect of an exposure or treatment on an outcome is distorted by the effect of another, usually unmeasured variable. Researchers studying the effects of coffee, for example, may be confounded by the fact that many coffee drinkers are also smokers. If the variable of smoking is recognized in advance, however, data on smoking behavior can be collected, and the effect of this behavior on the outcome being studied can be statistically "controlled." The problem occurs when data are not collected on potential confounding variables, usually because the effects of these variables are not anticipated.

Reporting Bias

As mentioned in the second article in this series, reporting bias may occur with *subgroup analysis* and *post hoc analysis*. That is, investigators may selectively report study results other than those for the primary comparison. Also mentioned in that article is the difference between absolute risk reduction (ARR) and relative risk reduction (RRR). ARR is the difference between the treatment group event rate and the control group event rate, and RRR is ARR as a percentage of the control group event rate. Thus, if 20% of control group members and 5% of treatment group members experience adverse events, then ARR is 15% (ie, 20%–5%) and RRR is 75% (ie, $[20\% - 5\%] \div 20\%$).

There are 2 problems with reporting results in terms of RRR. First, because RRR is usually more impressive than ARR (eg, 75% vs 15%), investigators may use the larger percentage to "strengthen" select results. In fact, some drug advertisements present RRRs for treatment effects and ARR for adverse effects, maximizing the perception of effectiveness and minimizing the perception of risk.

The other problem is that, as with reporting any percentage, the base rate is not always obvious. To take our example again, if only 0.2% (vs 20%) of control group members and only 0.05% (vs 5%) of treatment group members experience adverse events, we could still report 75% RRR. Although this percentage is still accurate, it refers to such a small absolute difference that it may be clinically meaningless.

EVIDENCE SYNTHESIS

To synthesize evidence is to convert what is known about a topic into useful knowledge. In the past, clinicians received little guidance in this area, which is why traditional medical care is sometimes called *experience-based medicine*. Now the emphasis is shifting to EBM. Evidence synthesis in EBM often involves 2 relatively new methods: *systematic review of the literature* and *meta-analysis*.

Narrative vs Systematic Literature Reviews

In a traditional *narrative review of the literature*, an "expert" selects articles deemed most relevant to a problem, draws some conclusions about the problem (based on these articles and on personal experience), and summarizes these conclu-

sions in a review article (Table 2). In a *systematic review*, strict criteria are used to conduct a thorough search of the literature, relevant articles are identified, and data are compiled into evidence tables and are interpreted in the context of all relevant studies. In contrast to a narrative review, a systematic review is a distinct, reproducible method.

Both narrative and systematic reviews may have the same potential difficulties—quality evaluation problems, publication bias, and language bias. The problems associated with objectively assessing the quality of a study have already been described. Objective evaluation may be easier in randomized trials than in other types of studies, because features such as random assignment, allocation concealment, blinding, and intention-to-treat analysis are standard and can be easily evaluated. Case-control and cohort studies have far more variation in design features and are thus much more difficult to evaluate objectively.

One reason for conducting a systematic review is to identify higher-quality studies on a topic on the assumption that valid results are more likely to come from such studies than from lower-quality studies. As mentioned, however, objective quality evaluation remains elusive. Still, studies vary in rigor of design and execution, and identification of higher-quality research is usually possible.

Publication bias refers to the fact that publication is more likely with “positive” studies (ie, studies with statistically significant results favoring authors’ hypotheses) than with “negative” or inconclusive studies. Publication bias is also affected by sample size: larger studies are more likely to be published, regardless of their results. Publication is less likely with smaller, negative studies, and the number of such studies must be sufficiently large if their results are to be used to challenge previously accepted conclusions.

Until recently, it had been difficult to find reference to any data that contradicted the conclusions of a published study or were unfavorable to a particular pharmaceutical product. Clinical trial registries are now in place to make such data more accessible. It is too soon to tell how effective these registries will be in maintaining up-to-date records of all clinical trials, regardless of their findings.

Language bias refers to the possibility that relevant studies published in nonnative languages may not be identified or

included in a review. English is commonly accepted as the language of science. Given the difficulties inherent in identifying, retrieving, and translating large numbers of foreign-language publications, many reviews are limited to studies published in English and indexed in the major scientific databases.

Meta-Analysis

Meta-analysis extends a systematic review by statistically combining the numerical results from multiple trials into a single outcome measure and confidence interval (Fig 1).²³ By combining the results from individual studies, meta-analysis greatly increases the overall sample size, which increases the statistical power of the analysis, as well as the precision of estimate of the treatment effects. As in systematic reviews, meta-analysis is conceptually attractive because results are more credible when several high-quality studies, performed in different settings, report similar findings.

Meta-analyses may have several purposes, including the following¹⁸:

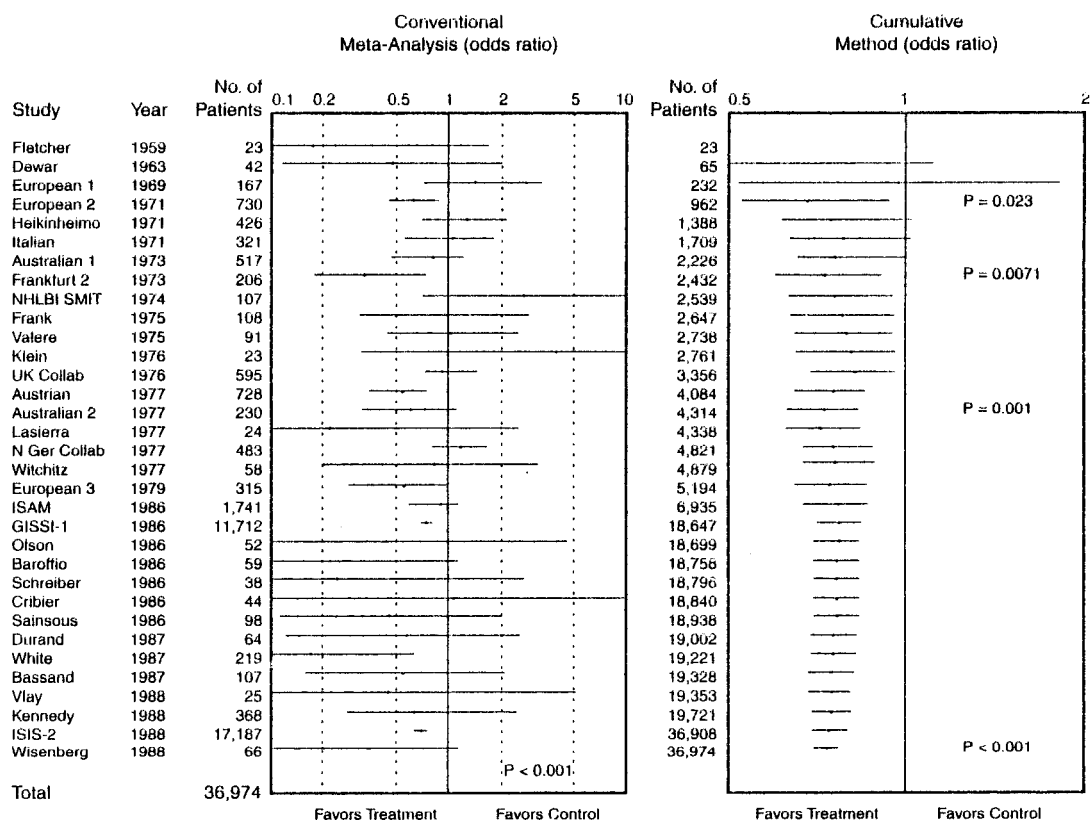
- Summarizing a large and complex body of literature on a topic
- Resolving conflicting reports in the literature
- Clarifying or quantifying the strengths and weaknesses of studies on a topic
- Documenting the need for a major clinical trial
- Avoiding the time and expense of conducting a clinical trial
- Increasing statistical power by combining many smaller studies
- Improving the precision of an estimated treatment effect
- Detecting smaller treatment effects than have been reported
- Investigating variations in treatment effects through subgroup (stratified) analysis
- Improving the generalizability of known treatment effects

Because the data for meta-analyses usually come from articles published in scientific journals, the quality of the meta-analysis depends heavily on the quality of these studies, how well their findings are reported, and whether they come to the attention of the meta-analyst. Although most authori-

Table 2. Differences Between Narrative and Systematic Literature Reviews*

Step	Narrative Review	Systematic Review
Literature search	Often limited, opportunistic, and not reproducible	Extensive, systematic, and reproducible; governed by a written protocol
Article selection	Informal and subject to individual bias	Criteria based and reproducible; governed by a written protocol
Data abstraction	Informal and subject to individual bias	Systematic and reproducible; governed by a written protocol
Data aggregation	Informal and subject to individual bias	Reproducible; structured evidence tables
Data interpretation	Informed expert opinion	Evidence-based opinion informed by access to all the evidence; sometimes meta-analysis

* Data are from Lang.²²



Source: Lau and others (1992).

Figure 1. Standard "forest plot" of results of meta-analysis. The point estimate for each study, usually expressed as an odds ratio or a risk ratio, is indicated by a dot. The horizontal line extending from the point estimate is the 95% confidence interval. The pooled estimate and its confidence interval are traditionally shown at the bottom of the plot.

ties now prefer systematic reviews to narrative reviews, meta-analysis has generated more debate.

The focus of the debate is whether and how numerical results should be combined statistically. Critics argue, correctly, that a good meta-analysis of poorly conducted trials still yields a poor result. They also argue, with some justification, that the variability in the clinical, methodologic, and statistical aspects among studies should preclude combining them. For example, can results from studies performed on children be legitimately combined with those of studies performed on adults? What about studies with different end points, designs, follow-up periods, or settings or studies conducted in different countries, in patients with different diagnoses, or at different stages of disease progression? Can these results be legitimately combined?

Proponents argue that because meta-analysis intentionally acknowledges these differences, it makes explicit the process of synthesizing research that is missing in traditional narrative reviews and experience-based medicine. In addition, if the results of a variety of studies all point in the same direction, the treatment effect is more likely to be real and

robust and less likely to be caused by the unique conditions of an individual study.

EVIDENCE-BASED MEDICINE

As already mentioned, EBM is largely literature-based medicine. The 3 building blocks of literature-based medicine are (1) primary research studies, (2) systematic reviews and meta-analytic syntheses of these studies, and (3) clinical practice guidelines, treatment recommendations, and decision analyses, which are based on syntheses of best evidence and on explicit considerations of patients' options and preferences.¹⁵ I have described the need for critical appraisal in evaluating the quality of primary research studies and have commented on the value of evidence synthesis in the form of systematic reviews and meta-analysis. Rather than review the role of clinical practice guidelines, however, I close with some thoughts on EBM.

As a field of study, EBM is the "conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients. The practice of evidence based medicine means integrating individual clini-

Table 3. Steps in Practicing Evidence-Based Medicine*

1. Ask the clinical question. The clinical question should incorporate 3 elements:
 - a. The patient. Describe the patient as a member of a population in terms of age, sex, and ethnic group. Describe the clinical problem in terms of the patient's disease or general health condition.
 - b. The intervention. The intervention may be clinical examination, prevention, prognosis, cause, differential diagnosis, diagnostic tests, or self-improvement.
 - c. The expected outcome. Ask, "What can I hope to accomplish?" "Have all clinically relevant options been considered?" "What could the intervention really affect?"
2. Find the best evidence. The best evidence is found by following these steps:
 - a. Translate the clinical question into a usable search strategy.
 - b. Select an appropriate database resource.
 - c. Enter the search strategy according to the rules of the database selected.
 - d. Browse the located records to identify which are best.
3. Critique the evidence. Located articles must be evaluated according to a variety of criteria, including:
 - a. Is this evidence valid? Is it true, accurate, correct?
 - b. Is this evidence important? Is it useful in clinical practice? Strategies for determining validity and importance vary according to the intervention being considered—diagnosis, causation, treatment, prevention, prognosis, or continuing education.
4. Apply the evidence. Integrate the evidence into clinical practice.
5. Evaluate the performance. Because evidence-based medicine is process oriented, you need to consider and evaluate your performance as you progress from step 1 through step 5.

cal expertise with the best available external clinical evidence from systematic research."²⁴ Few would argue that valid evidence should not help guide medical decision making. However, in practice, finding current, valid evidence that is applicable to a given patient is not so easy.

First, although some studies have reassuringly found that most patients do, in fact, receive evidence-based care, at least in some settings,^{25,26} many common diagnostic and therapeutic procedures may not have been adequately studied for a given patient population. Second, much of the published evidence is not of high quality.^{5,9,18} Third, evidence still has to be interpreted, which can be a subjective process.²⁷⁻²⁹ Finally, applying evidence obtained from groups to decisions for individuals is not always straightforward. Thus, EBM supplements experience-based medicine—and does not, should not, and cannot replace it.¹⁵

On the other hand, EBM has much to recommend it. Medical science is distinguished from other forms of healing, because it is supposed to rely on scientific evidence and reasoning. The practice of EBM can improve the quality and consistency of care.¹⁵ It seeks to make the best use of existing knowledge and, in so doing, to reduce uncertainty about treatment decisions (uncertainty is the most common cause of medical errors).¹⁷ The explicit purpose of EBM is to improve patient care; it is not intended to be "cookbook medicine" or "cost-containment medicine."

EBM can also be thought of as a process of lifelong, self-directed, problem-based learning in which caring for one's own patients creates the need for clinically important information about diagnosis, prognosis, therapy, and other clinical and health care issues, in which its practitioners do the following:

- Convert these information needs into answerable questions
- Track down, with maximum efficiency, the best evidence with which to answer them

- Critically appraise that evidence for its validity, importance, and usefulness
- Integrate the appraisal with clinical expertise and apply the results in clinical practice
- Evaluate their own performance¹⁵

The steps for practicing EBM are well defined (Table 3). The value of habitually following these steps is clear for several reasons. Physicians are ethically bound to provide the best care possible, and, in medicine, best care should be consistent with scientific evidence. Thus, new evidence can and should change the way that medicine is practiced.¹⁵ However, new evidence traditionally diffuses slowly through the medical community. Physicians who practice EBM should be able to access new evidence more quickly.¹⁵ In addition, clinical knowledge and skills deteriorate over time, and EBM can slow, if not stop, such deterioration.¹⁵

CONCLUSIONS

There is far more to critical appraisal of primary studies, evidence synthesis, and EBM than can be included in a single article (see Appendix for additional resources for conducting systematic reviews). However, the concepts described herein and the general skepticism that governs critical appraisal—indeed, all scientific inquiry—should provide a sound introduction to the practice.

ACKNOWLEDGMENT

I thank Tom Lang for his efforts and assistance in preparation of this article.

REFERENCES

1. Lang T. Interpreting and reporting public health and medical research: techniques and 13 key questions. Continuing education course handout presented at: Annual Meeting of the American Public Health Association; 2003; San Francisco, Calif.
2. Mosteller F, Gilbert JP, McPeck B. Reporting standards and

- research strategies for controlled trials. *Control Clin Trials*. 1980;1:37–58.
3. Rennie D. Guarding the guardians: a conference on editorial peer review. *JAMA*. 1986;256:2391–2392.
 4. Green SB, Byar DP. Using observational data from registries to compare treatments: the fallacy of omnimetrics. *Stat Med*. 1984; 3:361–370.
 5. Moher D, Cook DJ, Jadad AR, et al. Assessing the quality of reports of randomised trials: implications for the conduct of meta-analyses. *Health Technol Assess*. 1999;3(12):i-iv, 1–98.
 6. Balk EM, Bonis PAL, Moskowitz H, et al. Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials. *JAMA*. 2002;287: 2973–2982.
 7. Juni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA*. 1999;282: 1054–1060.
 8. Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. *Lancet*. 2001;357:1191–1194.
 9. Moher D, Schulz KF, Altman DG, for the CONSORT group (Consolidated Standards of Reporting Trials). The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. *Ann Intern Med*. 2001;134:657–662.
 10. Moher D, Schulz KF, Altman D, for the CONSORT group (Consolidated Standards of Reporting Trials). The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. *JAMA*. 2001;285: 1987–1991.
 11. Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF, for the Quorum group. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUORUM statement. *Lancet*. 1999;354:1896–1900.
 12. Stroup D, Berlin J, Morton S, et al. Meta-analysis of observational studies in epidemiology: a proposal for reporting. *JAMA*. 2000;283:2008–2012.
 13. Bossuyt PM, Reitsma JB, Bruns DE, et al, for Standards for Reporting of Diagnostic Accuracy. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Ann Intern Med*. 2003;138:W1–W12.
 14. Sackett DL. Bias in analytic research. *J Chronic Dis*. 1979;32: 51–63.
 15. Guyatt G, Rennie D, eds. *User's Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice*. Chicago, Ill: American Medical Association; 2002.
 16. Elwood JM. *Critical Appraisal of Epidemiological Studies and Clinical Trials*. New York, NY: Oxford University Press; 1998.
 17. Eddy DM. *A Manual for Assessing Health Practices and Designing Practice Policies: The Explicit Approach*. Philadelphia, PA: American College of Physicians; 1992.
 18. Lang T, Secic M. *How to Report Statistics in Medicine: Annotated Guidelines for Authors, Editors, and Reviewers*. Philadelphia, PA: American College of Physicians; 1997.
 19. Schwartz D, Lellouch J. Explanatory and pragmatic attitudes in therapeutical trials. *J Chronic Dis*. 1967;20:637–648.
 20. Simon G, Wagner E, Vonkorff M. Cost-effectiveness comparisons using real-world randomized trials: the case of new anti-depressant drugs. *J Clin Epidemiol*. 1995;48:363–373.
 21. Schulz K, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA*. 1995;273:408–412.
 22. Lang T. Systematic reviews as research assignments for training physicians. *Acad Med*. 2004;79:1067–1072.
 23. Hunt M. *How Science Takes Stock: The Story of Meta-Analysis*. New York, NY: Russell Sage Foundation; 1997.
 24. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. *BMJ*. 1996;312:71–72.
 25. Ellis J, Mulligan I, Rowe J, Sackett DL. Inpatient general medicine is evidence based. *Lancet*. 1995;346:407–410.
 26. Geddes JR, Game D, Jenkins NE, Peterson LA, Pottinger GR, Sackett DL. What proportion of primary psychiatric interventions are based on evidence from randomised controlled trials? *Qual Health Care*. 1996;5:215–217.
 27. Kahneman D, Slovic P, Tversky A, eds. *Judgment Under Uncertainty: Heuristics and Biases*. New York, NY: Cambridge University Press; 1982.
 28. Plous S. *The Psychology of Judgment and Decision Making*. New York, NY: McGraw-Hill; 1993.
 29. Fischhoff B, Lichtenstein S, Slovic P, Keeney D. *Acceptable Risk*. New York, NY: Cambridge University Press, 1981.
- Requests for reprints should be addressed to:*
 Lyndon Mansfield, MD
 1901 Arizona Ave
 El Paso, TX 79902
 E-mail: immunman@pol.net

APPENDIX

RESOURCES FOR CONDUCTING SYSTEMATIC REVIEWS

Books

Cooper HM, Hedges LV, eds. *The Handbook of Research Synthesis*. New York, NY: Sage; 1994.

Egger M, Davey Smith G, Altman D. *Systematic Reviews in Health Care: Meta-Analysis in Context*. London: BMJ Books; 1995.

Articles

Badgett RG, O'Keefe M, Henderson MC. Using systematic reviews in clinical education. *Ann Intern Med*. 1997;126: 886–891.

Bero LA, Jadad AR. How consumers and policymakers can use systematic reviews for decision making. *Ann Intern Med*. 1997;127:37–42.

Cook DJ, Greengold NL, Ellrodt AG, Weingarten SR. The relation between systematic reviews and practice guidelines. *Ann Intern Med*. 1997;127:210–216.

Cook DJ, Mulrow CD, Haynes RB. Systematic reviews: synthesis of best evidence for clinical decisions. *Ann Intern Med*. 1997;126:376–380.

Counsell C. Formulating questions and locating primary studies for inclusion in systematic reviews. *Ann Intern Med*. 1997;127:380–387.

Egger M, Smith GD, Sterne JA. Uses and abuses of meta-analysis. *Clin Med*. 2001;1:478–484.

Hasselblad V, Mosteller F, Littenberg B et al A survey of current problems in meta-analysis. discussion from the Agency for Health Care Policy and Research interPORT Work Group on Literature Review/Meta-Analysis. *Med Care*. 1995;33:202–220.

Hunt DL, McKibbin KA. Locating and appraising systematic reviews. *Ann Intern Med*. 1997;126:532–538.

Ioannidis JP, Lau J. Pooling research results: benefits and limitations of meta-analysis. *Jt Comm J Qual Improv*. 1999; 25:462–469.

Juni P, Altman DG, Egger M. Systematic reviews in health care: assessing the quality of controlled clinical trials. *BMJ*. 2001;323(7303):42–46.

Lau J, Ioannidis JPA, Schmid CH. Quantitative synthesis in systematic reviews. *Ann Intern Med*. 1997;127:820–826.

McQuay HJ, Moore RA. Using numerical results from systematic reviews in clinical practice. *Ann Intern Med*. 1997; 126:712–720.

Meade MO, Richardson WS. Selecting and appraising studies for a systematic review. *Ann Intern Med*. 1997;127:531–537.

Mulrow CD, Cook DJ, Davidoff F. Systematic reviews: critical links in the great chain of evidence. *Ann Intern Med*. 1997;126:389–391.

Sacks HS, Berrier J, Reitman D, Ancona-Berk VA, Chalmers TC. Meta-analyses of randomized controlled trials. *N Engl J Med*. 1987;316:450–455.

Savoie I, Helmer D, Green CJ, Kazanjian A. Beyond Medline: reducing bias through extended systematic review search. *Int J Tech Evaluate Health Care*. 2003;19:168–178.

Sterne JA, Egger M, Smith GD. Systematic reviews in health care: investigating and dealing with publication and other biases in meta-analysis. *BMJ*. 2001;323(7304):101–105.

Internet Sites

<http://www.mnsinc.com/solomon/MetaA/MApage2.html>

<http://www.sghms.ac.uk/phs/staff/jmb/meta.htm>

British Medical Journal Meta-Analysis Series (Nov 22, Dec 6, Dec 13, 1997; Jan 3, Jan 10, Jan 17, 1998).

1. Egger M, Davey Smith G. Meta-analysis: potential and promise. <http://www.bmj.com/archive/7119/7119ed.htm>

2. Egger M, Davey Smith G, Phillips N. Meta-analyses: principles and procedures. <http://www.bmj.com/archive/7121/7121ed.htm>

3. Davey Smith G, Egger M, Phillips AN. Meta-analysis: beyond the grand mean. <http://www.bmj.com/archive/7122/7122ed2.htm>

4. Egger M, Davey Smith G. Meta-analysis: bias in location. <http://www.bmj.com/archive/7124/7124ed2.htm>

5. Egger M, Schneider M, Davey Smith G. Meta-analysis: spurious precision. <http://www.bmj.com/archive/7125/7125ed2.htm>

6. Davey Smith G, Egger M. Meta-analysis: unresolved issues and future developments. <http://www.bmj.com/archive/7126/7126ed8.htm>

7. Egger M, Sterne JAC, Davey Smith G. Meta-analysis software. <http://www.bmj.com/archive/7126/7126ed8.htm>

Objectives: After reading this article, participants should be able to demonstrate an increased understanding of their knowledge of allergy/asthma/immunology clinical treatment and how this new information can be applied to their own practices.

Participants: This program is designed for physicians who are involved in providing patient care and who wish to advance their current knowledge in the field of allergy/asthma/immunology.

Credits: ACAAI designates each Annals CME Review Article for a maximum of 2 category 1 credits toward the AMA Physician's Recognition Award. Each physician should claim only those credits that he/she actually spent in the activity. The American College of Allergy, Asthma and Immunology is accredited by the Accreditation Council for Continuing Medical Education to sponsor continuing medical education for physicians.

CME Examination

1–5, Mansfield L. 2006;96:7–16.

CME Test Questions

1. A well-done clinical trial must have:

- no exceptions
- exceptions
- inaccurate results
- a bad study design

2. Critical appraisal is a skill:

- to distinguish high-quality research
- To distinguish poor-quality research
- a and b

3. The hierarchy of evidence is:

- an organization of intellectuals
 - a means to give a value to the type of evidence presented
 - a term that means higher-level evidence is always more true than lower levels
4. Industry-sponsored research:
- is less believable than academic studies
 - is not peer reviewed prior to publication
 - has to follow the same publication guidelines as other research
5. Sampling biases include all but:
- referral filter bias

-
-
- b. volunteer bias
 - c. assignment bias
 - d. randomization bias
6. Analytical bias is not:

- a. on protocol analysis
 - b. baseline risk of an event
 - c. confounding
 - d. reporting bias
-