Proof and Policy from Medical Research Evidence

Cynthia D. Mulrow
University of Texas Health Science Center–San Antonio

Kathleen N. Lohr
University of North Carolina School of Public Health–Chapel Hill

Abstract When judging the benefits and harms of health care and predicting patient prognosis, clinicians, researchers, and others must consider many types of evidence. Medical research evidence is part of the required knowledge base, and practitioners of evidence-based medicine must attempt to integrate the best available clinical evidence from systematic research with health professionals' expertise and patients' rights to be informed about diagnostic and therapeutic options available to them. Judging what constitutes sound evidence can be difficult because of, among other things, the sheer quantity, diversity, and complexity of medical evidence available today; the various scientific methods that have been advanced for assembling, evaluating, and interpreting such information; and the guides for applying medical research evidence to individual patients' situations. Recommendations based on sound research can then be brought forward as either guidelines or standards, and criteria exist by which valid guidelines and standards can be developed and promulgated. Nonetheless, gaps and deficiencies exist in current guidelines and in the methods for finding and synthesizing evidence. Interpreting and judging medical research involves subjective, not solely explicit, processes. Thus, developments in evidence-based medicine are an aid, but not a panacea, for definitively establishing benefits and harms of medical care, and the contributions that medical research evidence can make in any clinical or legal situation must be understood in a context in which judgment and values, understanding of probability, and tolerance for uncertainty all play a role.

Scope

Many types of evidence must be considered when judging the benefits and harms of medical care and forecasting the prognoses of patients

Journal of Health Politics, Policy and Law, Vol. 26, No. 2, April 2001.

Table 1 Types of Evidence Involved in Medical Judgments

• Medical research	• Society's values
• Particulars of patient situations such as course and severity of illness, concurrent mental and physical disease, education, beliefs, social resources, and finances	• Patients' readiness to accept and adherence to recommended diagnostic, therapeutic, and/or monitoring strategies
• Medical providers' experiences, beliefs, and skills	• Health care systems' rules, resources, and financing

(Table 1). This article addresses one form of evidence and answers the question "What constitutes sound medical research evidence?"

Specifically, as the first in a series of papers prepared for the workshop on "'Evidence:' Its Meanings and Uses in Law, Medicine, and Health Care," we address the evolution and current concepts of medical research evidence and methods that are used to synthesize and judge such evidence. Further, we offer an overview of the status of medical evidence, evidence-based medicine, and clinical practice guidelines in medicine. We review the history, development, and current meaning of evidence in medicine, as well as how medical evidence is currently manifested in guidelines.

The primary definition of "evidence" given in *Webster's New World Dictionary* (1988) applies: the data on which a conclusion or judgment may be based. It is accepted that medical data often are limited. Medical research inadequately addresses many health-related situations that confront patients, practitioners, health care systems, and policy makers. The gaps between what research evidence shows will likely benefit or harm, and what patients and the public receive or are exposed to, can be large (Haynes 1993). We do not focus on such gaps and the reasons behind them (e.g., inadequate decision support systems at the point of care, rapidly evolving complex knowledge, competing priorities and limited resources, conflicting values, errors, or insufficient skills and communication). Rather, we address methods for judging and summarizing health care evidence from the ideological perspective of the medical profession.

Evolution of Ideas about Medical Evidence

Both the diversity and quantity of medical evidence increased during the twentieth century. In the first half of the century, advances in medical research were based primarily on basic, physiologic, and reductionist

approaches (Annas 1999; Porter 1997). Units of study focused on cells, organs, and animals. By the second half of the century, two major developments changed the face of medical research. First, revolutionary advances in our understanding of molecular and cellular biology prompted scientists to initiate remarkable new avenues of study, such as the Human Genome Project (HGP). Second, the branch of medicine known as epidemiology spawned new research designs for use with human participants, most notably the advent of the clinical trial (Bull 1959; Lilienfeld 1982; Porter 1997; Williams 1999). These new tools to answer important scientific questions raised the bar for medical research that was directly applicable to medical care of patients (Williams 1999).

Concomitant with the development of new research designs, increasing medical research of all types was seen. In the 1990s, more than two million articles were published annually in more than 20,000 biomedical journals, more than 250,000 controlled trials of health care therapies had been conducted, and more than $50 billion was being spent annually on medical research (Ad Hoc Working Group for Critical Appraisal of the Medical Literature 1987; Michaud and Murray 1996; Cochrane Collaboration 1999).

Not surprisingly, the medical profession's beliefs concerning evidence have evolved in the United States, influenced in large part by swings in their philosophies of how to approach health care (Figure 1). In the late 1700s, Dr. Benjamin Rush, a signer of the Declaration of Independence and the "founding father" of American medicine, urged practitioners and patients alike to be "heroic, bold, courageous, manly, and patriotic" (Payer 1988; Silverman 1993: 5). Rush's followers sanguinely believed in direct, drastic intervention: "When confronted by a sick patient, providers gather their purges and emetics, bare their lancets, and charge the enemy, prepared to bleed, purge and induce vomiting until the disease is conquered" (Silverman 1993: 6). A hundred years later, this "do everything you can, anything is possible" approach was replaced with a more nihilistic philosophy espoused by the famous North American physician and writer Oliver Wendell Holmes (not to be confused with the little-known son and Justice of the same name!). As a reaction to medicine's unbridled use of treatments such as purging, blistering, mercury, and arsenic, Holmes (1988: 6) espoused "doing nothing because doctors did more harm than good." A renowned early-twentieth-century American physician, William Osler, mirrored Holmes's message: "Most remedies in common use are likely to do more harm than good" (Thomas 1983: 15).

Thus, in the early 1900s, treatment of disease was a minor part of

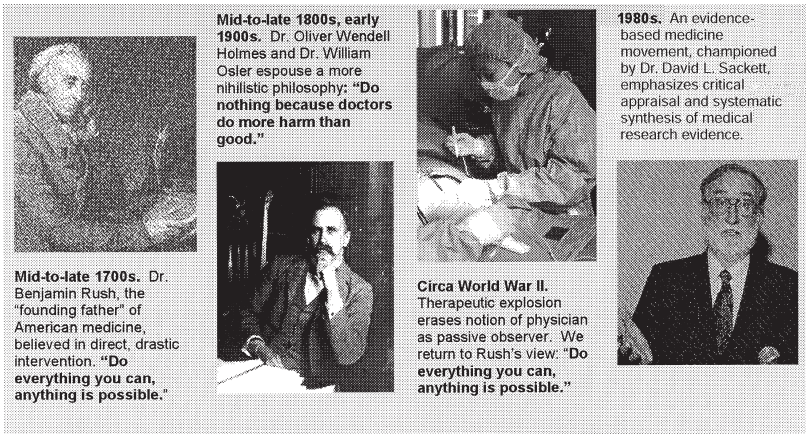

Figure 1 Historical Trends in North American Philosophy toward Medical Care

American medical curricula. Rather, the focus was on accurate diagnosis, prediction of course of disease, and doctors standing by as compassionate family friends and advisors (Porter 1997; Williams 1999). A therapeutic explosion around the time of World War II erased any notion that doctors would remain passive observers, sitting with a magazine of largely blank cartridges; a feverish and soaring optimism hit American medicine (Gordon 1994; Porter 1997; Williams 1999). We returned to Rush's "do everything you can, anything is possible" dogma.

Diagnostic and treatment strategies were adopted with little thought given to the need for careful observations in adequate numbers of patients and for comparisons of outcomes between persons given an intervention or diagnostic test and those not given the intervention or test. Potential harms of diagnostic and therapeutic approaches often were not studied, and innovations were adopted enthusiastically and uncritically. Fueled by recognition of some treatment disasters, an underlying value system firmly embedded in scientific inquiry and experiment, marked variation in the practice patterns of medical professionals, and new types of medical research and dissemination strategies, leading North American physicians propagated "evidence-based medicine" during the last decades of the twentieth century.

Evidence-based medicine is defined as the conscientious, explicit, and judicious use of current best evidence in making decisions about health care (Sackett et al. 1997). Evidence-based practice, building on the orig-

inal definition, is said to be "an approach to decision making in which the clinician uses the best evidence available, in consultation with the patient, to decide upon the option which suits that patient best" (Muir Gray 1997: 9). The latter concept does emphasize the role of patients in shared decision making about their health care. Thus, practicing evidence-based medicine involves integrating the medical professional's expertise and the patient's right to choose among diagnostic and treatment alternatives with the best available external clinical evidence from systematic research.

Best available external clinical evidence is taken to mean clinically relevant evidence, often from the basic sciences of medicine, but especially from patient-centered clinical research into the accuracy and precision of diagnostic tests, the power of prognostic markers, and the safety, efficacy, and effectiveness of therapeutic, rehabilitative, and preventive regimens (Sackett et al. 1997). Although evidence-based medicine has provoked antagonism and skepticism among some academics and practicing physicians, many of its underlying principles reflect the medical profession's current understanding of sound medical evidence (Naylor 1995; Feinstein and Horwitz 1997; Lohr, Eleazer, and Masukopf 1998). Moreover, evidence-based medicine stresses a structured critical examination of medical research literature; relatively speaking, it deemphasizes average practice as an adequate standard and personal heuristics.

Assembling, Evaluating, and Interpreting Medical Research Evidence

Medical research evidence can be simple and straightforward or complex and conditional. The latter, common instance poses a tremendous challenge to consumers, health care providers, and policy makers who try to understand what scientific evidence is valid. Moreover, understanding the causes of diseases, benefits and harms of diagnostic or therapeutic strategies, and prognoses of patients often requires accumulating and critiquing data from multiple studies and disciplines (Hulka, Kerkvliet, and Tugwell 2000).

When evidence is not simple, and when there is a lot of it, we can use frameworks and trained experts to assemble, sort through, and integrate evidence. Scientific methods for assembling, evaluating, and interpreting medical research evidence have been developing rapidly (Light and Pillemer 1984; Eddy 1992; Cook et al. 1995; Cook, Sackett, and Spitzer 1995; Mulrow and Cook 1998; Cochrane Collaboration n.d.). The principles

Table 2 Principles for Assembling, Evaluating, and Interpreting Medical Research

- A priori explicit statements of questions being addressed
- Systematic, explicit rather than selective, "file drawer" searching for pertinent research
- Systematic sorting of relevant from irrelevant research using preset explicit selection criteria
- Systematic critique of the validity of individual pieces of medical research based on the quality of the research methodology
- Critique of the generalizability of pieces of research based on characteristics of participants involved in research studies and characteristics of the agents or strategies tested in the research
- Integration of bodies of evidence based on sources of evidence, research design, directions and magnitudes of clinical outcomes, coherence, and precision
- Extrapolation of research findings to particular situations based on preset criteria
- Continual updating and integrating of evidence (perpetual revision)
- Open attribution and statement of conflict of interest by those who do research synthesis

behind these methods are to avoid bias in finding, sorting, and interpreting data, and to be comprehensive and current (Table 2).

The methods that one uses to assemble and critique relevant evidence vary depending upon the question that is asked. Table 3 displays broad concepts of types of studies to look for and ways to critique and interpret them, depending upon whether the question relates to harm, diagnosis, prognosis, or treatment.

Interpreting and Judging Medical Research

Practitioners of evidence-based medicine and developers of clinical guidelines and standards may need to address the quality and strength of medical research at three levels. First (and arguably simplest) is evaluating the *quality and applicability* of individual studies. In this effort, one attempts to understand how well research studies have been designed and conducted as well as whether results apply to specific or general populations of patients. Second is evaluating the *strength and applicability* of a body of evidence about a specified clinical question. In the second effort, one judges how much credence and reliance to place on a collection of individual studies. The third consideration involves the *intensity* of recommendations, and so pertains more to experts developing authoritative

Table 3 Examples of Types of Relevant Research and Methods of Critique and Interpretation

Harm	Diagnosis	Prognosis	Treatment
Assemble Relevant Research			
• Case reports with challenge designs • Cohort studies • Case-control studies • Controlled trials	• Diagnostic test studies	• Cohort studies • Controlled trials	• Controlled trials
Critically Evaluate Evidence			
• Appropriate temporal relationship? • Appropriate follow-up duration? • Dose-response gradient? • Positive rechallenge test? • Comparison groups similar? • Exposure measured appropriately? • Outcome measured appropriately? • Strong and precise association? • Biologically plausible association? • Research sponsorship clear?	• Test performed appropriately? • Independent, blind comparison to appropriate standard? • Appropriate spectrum of patients? • Standard applied regardless of test result? • Diagnostic power and precision? • Research sponsorship clear?	• Representative patient sample? • Follow-up long and complete? • Objective outcome criteria applied blindly? • Adjustment for known prognostic factors? • Validation set if testing predictive power? • Likelihood of outcomes over time? • Prognostic estimates precise? • Research sponsorship clear?	• Randomized with concealed allocation? • Outcome assessments unbiased? • Groups treated equally except for intervention strategy? • Few withdrawals and dropouts? • Intention-to-treat analysis? • Tested intervention similar to practice? • Trial participants markedly atypical? • Research sponsorship clear?
Know How to Interpret			
• Relative risk • Relative odds • Odds ratios • Probability tests • Confidence intervals • Meta-analysis	• Sensitivity • Specificity • Likelihood ratio • Probability tests • Confidence intervals • Meta-analysis	• Absolute terms (five-year survival rate) • Relative terms (size of risk from a prognostic factor) • Survival curves • Probability tests	• Relative risk reduction • Absolute risk reduction • Number needed to treat • Probability tests • Confidence intervals • Meta-analysis

Table 4 Sources and Designs of Research

Primary Studies in Humans
- Randomized, controlled trials
- Nonrandomized, controlled trials
- Cohort or longitudinal studies
- Case-control studies
- Cross-sectional descriptions and surveys
- Case series and case reports

Nonhuman Studies
- In vitro (laboratory) studies
- Animal studies

Syntheses
- Systematic reviews including meta-analyses
- Decision and economic analyses
- Guidelines

guidelines containing recommendations than to experts assembling systematic reviews of the evidence. The force or intensity with which a recommendation is made often reflects the strength of evidence and the level of net benefit expected for the health service in question.

Interpreters and judges of medical evidence are faced with multiple sources and various research designs (Table 4) (Mulrow and Cook 1998). These can include laboratory experiments, observations in a single patient or groups of patients, studies in humans with cases (persons with condition of interest) compared to controls (persons without the condition of interest), and controlled trials of one diagnostic or therapeutic strategy compared to another.

Although in some situations the evidence will be clear, in many other situations judges of medical research are faced with murky, dubious, narrow, conflicting, or irrelevant evidence. They use judgment to weigh types of evidence based on study methodology and precision and magnitude of results. As exemplified in Table 3, all pieces of evidence are not equal; their value depends on the specific question and context.

Numerous rating schemas exist—in the form of checklists and scales—that can help delineate the types of research that are most appropriate to answer particular questions. There are also multiple rating schemes for appraising particular study designs such as randomized trials. These are approaches chiefly for grading the quality of individual studies, but their reliability, validity, feasibility, and utility are today largely either unmeasured or quite variable (Sacks, Chalmers, and Smith 1983; Schulz et al. 1994; Guyatt et al. 1995, 1998; Moher et al. 1995;

Hadorn et al. 1996; U.S. Preventive Services Task Force 1996; Lohr and Carey 1999; SIGN 1999a, 1999b).

The value of any single piece of medical research evidence is derived from how it fits with and expands previous work and from the study's intrinsic properties (Cooper 1984: 79–113). Integrating an entire body of relevant medical research, and then assessing the strength of that collection of research, is usually more important than critiquing a single piece of research evidence. This often requires piecing together heterogeneous items of direct and indirect evidence. (Medical evidence is considered indirect if two or more bodies of evidence are required to relate the exposure, diagnostic strategy, or intervention to the principal outcome.)

Integrating evidence is invariably a subjective process, dependent on the skills and values of the individuals who are trying to synthesize multiple pieces of diverse medical evidence. Individuals summarizing medical research make judgments about the relevance, legitimacy, and relative uncertainty of particular pieces of evidence, the importance of missing evidence, the soundness of any models for linking evidence, and the appropriateness of conducting a quantitative summary (Mulrow, Langhorne, and Grimshaw 1997). Conclusions of any synthesis of indirect research evidence are inferential and based on a combination of facts, arguments, and analogies. An important pitfall to avoid is confusing lack of high-level evidence with evidence against effectiveness: absence of proof is not the same as proof of absence.

Several frameworks can help guide, standardize, and make explicit the process of synthesizing bodies of medical research evidence (Hill 1965; Naranjo et al. 1981; Cadman et al. 1984; Pere et al. 1986; Sox et al. 1989; Woolf et al. 1990; Woolf 1991; Eddy, Hasselblad, and Shachter 1992; Huff 1992; NHMRC 1995; Fleming and DeMets 1996; Cook et al. 1997; Mulrow, Langhorne, and Grimshaw 1997). An example of a classic framework for assessing a body of evidence relating to harm is given in Table 5 (Hill 1965). Some of these criteria are similar to those noted in Table 4 regarding critical evaluation of individual pieces of evidence relating to harm. However, the framework for synthesizing a body of evidence and for designating the strength of that evidence has significant differences; a hierarchy of relevant valid evidence (e.g., experimental evidence in humans) and an emphasis on consistent and coherent results across multiple types and sources of evidence are apparent.

In the end, those compiling medical research evidence may be able to define and assign only relatively subjective classifications of the strength

Table 5 Framework for Synthesizing Body of Evidence Relating to Harm

- Experimental evidence in humans with exposed and unexposed participants?
- Strength or magnitude of association?
- Consistency of association across studies?
- Specificity of association?
- Appropriate temporal sequence (exposure occurred before harm)?
- Plausible based on existing biological and physiological understanding?
- Dose-response relationship?
- Coherence of evidence across multiple types and sources of evidence?

of evidence on a given question—such as "excellent" to "poor" or "strong," "moderate," or "weak." For example, "good" evidence may exist when data in individual studies are sufficient for assessing the quality of those findings, when data across studies are consistent, and when they indicate that the intervention in question is superior to alternative treatments. By contrast, evidence may be only "fair" when information from individual studies can be graded but is subject to challenge on quality grounds and/or when reasonably acceptable data across studies are inconsistent in their findings. Finally, a body of evidence may be characterized as "poor" when the number of relevant studies is minimal, when the quality of individual studies is highly suspect because of flaws in design or conduct, or when the evidence is so conflicting that no logical or defensible conclusions can be drawn.

Applicability of Medical Research Evidence to Populations or Individuals

Much research evidence applies to probabilities of occurrences in groups or populations and not in individual patients. In either instance, accurate prediction or proof of causality (or both) applicable to real-life settings is difficult and relies on judgment regarding the magnitude of probability and uncertainty (reasonable doubt) that one considers as acceptable proof. For example, even therapies that are "proven effective" will not work in every patient, and therapies or exposures that are "proven harmful" will not harm every patient to whom they are given.

Guides for applying medical research evidence to the individual patient situation call for the following actions (Glasziou et al. 1998; Ross 1998): (a) stratify research findings according to an individual's charac-

teristics (often not possible); (b) ask whether the underlying pathophysiology and presence of comorbid conditions in the individual patient situation are so different that the research is not applicable; (c) assess whether the intervention or exposure in the real-life setting approximates that tested in research; (d) estimate benefits and harms from research obtained from groups, but apply those estimates based on established knowledge of the individual's characteristics or risks; and (e) take into account individual preferences, competing priorities, and resources.

Recommendations Based on Evidence: Guidelines versus Standards

Medical recommendations based on research evidence can be formed as guidelines or standards. Clinical practice guidelines are "systematically developed statements to assist practitioner and patient decisions about appropriate health care for specific clinical circumstances" (Institute of Medicine 1990: 38). Methods of formulating guidelines may differ in several respects, including methods for identifying, appraising, and ranking relevant research evidence; models for integrating indirect evidence; methods for incorporating experience and opinion; whether harms, costs, and values are explicitly considered; and sponsorship (ibid.).

The four critical concepts to understand about the creation of defensible guidelines are (1) that the development process is open, documented, and reproducible; (2) that the resulting product or products can be of use to both clinicians and patients; (3) that the concept of "appropriateness" of services is well reflected in the guideline (where appropriateness means essentially that the potential health benefits of the service exceed the potential harms or risks by a margin sufficiently large that the service is worth providing); and (4) that the guideline relates specifically to clearly defined clinical issues.

Explicit criteria have been available for a decade to use in assessing the soundness of practice guidelines and in directing the development of new guidelines from systematic reviews of evidence (Woolf 1992; Carter et al. 1995; Cluzeau and Littlejohns 1999; Shaneyfelt, Mayo-Smith, and Rothwangl 1999). Such criteria emphasize two broad attributes of guidelines: that they be credible with practitioners, patients, payers, and policy makers, and that the developers be accountable for the conclusions they draw from the evidence and for the recommendations they base on those conclusions.

Important criteria concerning the process of guideline development

call for developers to ensure the clarity of what they have written, that they have used a multidisciplinary approach, that they have dated their work and identified a point in the future when the guidelines ought to be revisited in the light of possible new evidence, and that the entire process be documented. Equally important criteria about the substance of the guideline reinforce the views that the clinical scope of the guideline be explicit, that the guideline provide for appropriate flexibility for clinical decision makers when medical evidence is not clear-cut, and that the guideline have acceptable reliability and validity.

Arguably the most important attribute of guidelines is validity. That is, guidelines should, when followed, lead to the health and cost outcomes expected for them. Elements of their validity consider the substance and quality of the research evidence cited, the ways that such evidence is evaluated, the strength of the collective body of evidence in question, the intensity or force of recommendations in light of the strength of evidence, and judgments about likely net benefits to patient populations. In some instances, empirical evaluations of the validity and utility of specific guideline recommendations may be available.

Whether created or adapted locally or nationally, most guidelines are an amalgam of clinical experience, expert opinion, and research evidence (Institute of Medicine 1992; Woolf 1999). In the United States, there are literally thousands of practice guidelines. Not surprisingly, some of these vary in content and conclusions, conflict with one another, or both.

Guidelines most often apply to the general and not the particular. They require extrapolation to individual circumstance. Whether individual circumstances warrant a different standard can be judged only case by case. Following evidence-based guidelines may generally but not always assure good medical care; diverging from guidelines does not always signal poor care (Mulrow 1996; Weingarten 1997; Woolf et al. 1999).

Unlike a guideline, which is a recommendation for best practices, *standards* are practices that are medically necessary and services that any practitioner under any circumstance would be required to render (Brook 1991; Leape 1995; Eddy 1996). Guidelines are meant to be flexible and amenable to tailoring to meet individual circumstances; standards are meant to be inflexible and should always be followed, not tailored (Eddy 1996). Formulating standards rather than guidelines requires a higher bar. One needs to consider the relative effectiveness and harms of a wide variety of diagnostic and treatment options for multiple possible medical conditions that a patient or population may face. One also needs to assess feasibility and costs of those options.

Evidence-based guidelines that focus on single conditions likely will inform, but not determine, standards of medical care that our society deems necessary. Likewise, research evidence can and should inform standards of care, but research evidence in and of itself will invariably be inadequate to establish standards because standards will require priority setting based on cost and value judgments.

At the present time, consumers, health care providers, judges, and policy makers lack ready, scientific means for comparing the relative effectiveness and harms of various types of medical care (Woolf 1999). Such information is critical for setting priorities and standards. An irony of our medical information age and of evidence-based medicine is that we have thousands of studies and systematic summaries of those studies that focus on effects of specific exposures or treatments on particular outcomes. Although valuable, this narrowly focused repository of data provides a piecemeal rather than an integrative approach when choosing among competing priorities and setting the standards that are most likely to improve health.

Moreover, we have little scientific work from the perspective of defining global or national health goals and examining the relative effectiveness of various strategies for achieving those goals. A recent suggestion regarding the creation of a bibliographic research evidence collection center, paired with a simulation modeling program, could aid better estimation of the potential benefits and harms of competing health care strategies (ibid.). Such projections could help policy makers, clinicians, and patients give due priority to the strategies most likely to improve health. Regardless, we need greater emphasis on formulating broader evidence-based guidelines and standards that at least (a) address clusters of conditions (e.g., cardiovascular disease or cancer) rather than single specific conditions and (b) define and translate harms as well as they define and translate benefits. For evidence-based medicine, a final irony may be that these more integrative approaches are sorely needed, yet they rely on more assumptions than do simple but less integrative techniques.

All these factors point to an important conclusion about the role of evidence-based practice and guidelines in the courts today. The gaps and deficiencies in current guidelines make them difficult to apply as the definitive information for legal or judicial decision making, just as they may often be difficult to implement in medical decision making. The field of evidence-based medicine is progressing rapidly in clinical substance and methodology, but the day has not yet come when it undergrids all that is or could be done in medicine or the medicolegal context.

Summary

Medical research is continually evolving and accumulating; yesterday's precedent may be today's anachronism. Interpreting and judging medical research evidence involves explicit as well as subjective processes. Although neither research evidence nor its synthesis is always neutral and objective, we do have evidence-based techniques that aid comprehensive collation, unbiased and explicit evaluation, and systematic summarization of available research. For example, hierarchies of types of research evidence that are relevant for different types of questions have been developed. In addition, techniques exist by which to appraise the relevance and validity of individual pieces as well as bodies of research evidence and to link them to guidelines and standards.

Such developments in evidence-based medicine are an aid, not a panacea, for definitively establishing benefits and harms of medical care and prognoses of patients. First, interpreting and judging continually evolving medical research involves subjective processes that are inherently dependent on the "eye of the observer." Second, although methods of rating and integrating research evidence are evolving and being tested, any single or uniform "best method" for such a complex task is unlikely to be available in the near future (if ever). Third, guidelines, even when based firmly on high-quality research, are not always relevant or valid for individual situations; nor, usually, are they adequate for establishing medical necessity across different conditions. Fourth, much research applies to groups of patients or populations and not to individuals. Fifth, for both medicine and law, accurate prediction and/or absolute proof of causality applicable to individuals or to real-life settings are difficult, if not impossible, in many instances. Finally, the contributions of medical research evidence to proof or policy for any given clinical (or legal) situation will come in a context in which judgment and values, understanding of probability, and tolerance for uncertainty all have their place.

References

Ad Hoc Working Group for Critical Appraisal of the Medical Literature. 1987. Academia and Clinic: A Proposal for More Informative Abstracts of Clinical Articles. *Annals of Internal Medicine* 106:598–604.

Annas, G. J. 1999. Burden of Proof: Judging Science and Protecting Public Health in (and out of) the Courtroom. *American Journal of Public Health* 89:490–493.

Brook, R. 1991. Health, Health Insurance, and the Uninsured. *Journal of the American Medical Association* 265:2998–3002.

Bull, J. P. 1959. The Historical Development of Clinical Therapeutic Trials. *Journal of Chronic Diseases* 10:218–248.

Cadman, D., I. Chambers, W. Feldman, and D. Sackett. 1984. Assessing the Effectiveness of Community Screening Programs. *Journal of the American Medical Association* 251:1580–1585.

Carter, A. O., R. N. Battista, M. J. Hodge, S. Lewis, A. Basinski, and D. Davis. 1995. Report on Activities and Attitudes of Organizations Active in the Clinical Practice Guidelines Field. *Canadian Medical Association Journal* 153:901–907.

Cluzeau, F. A., and P. Littlejohns. 1999. Appraising Clinical Practice Guidelines in England and Wales: The Development of a Methodologic Framework and Its Application to Policy. *Journal of Quality Improvement* 25:514–521.

The Cochrane Collaboration. 1999. The Cochrane Library, Update Software (Oxford, U.K.: Update Software): Issue 4.

———. N.d. Cochrane Reviewer's Handbook 4.0. Available on-line at www.cochrane.dk/cochrane/handbook/handbook.htm.

Cook, D. J., N. L. Greengold, A. G. Ellrodt, and S. R. Weingarten. 1997. The Relation between Systematic Reviews and Practice Guidelines. *Annals of Internal Medicine* 127:210–216.

Cook, D. J., G. H. Guyatt, A. Laupacis, D. L. Sackett, and R. J. Goldberg. 1995. Clinical Recommendations Using Levels of Evidence for Antithrombotic Agents. *Chest* 108:227S–230S.

Cook, D. J., D. L. Sackett, and W. O. Spitzer. 1995. Methodologic Guidelines for Systematic Reviews of Randomized Controlled Trials in Health Care from the Potsdam Conference on Meta-Analysis. *Journal of Clinical Epidemiology* 48:17–71.

Cooper, H. M. 1984. The Analysis and Interpretation Stage. In *The Integrative Research Review: A Systematic Approach*, ed. H. M. Cooper. Beverly Hills, CA: Sage.

Eddy, D. M. 1992. *A Manual for Assessing Health Practices and Designing Practice Policies: The Explicit Approach*. Philadelphia: American College of Physicians.

———. 1996. *Clinical Decision Making: From Theory to Practice. A Collection of Essays from the Journal of the American Medical Association*. Sudbury, MA: Jones and Bartlett.

Eddy, D. M., V. Hasselblad, and R. D. Shachter. 1992. Meta-Analysis by the Confidence-Profile Method: The Statistical Synthesis of Evidence. Boston: Academic.

Feinstein, A. R., and R. I. Horwitz. 1997. Problems in the "Evidence" of "Evidence-Based Medicine." *American Journal of Medicine* 103:529–535.

Fleming, T. R., and D. L. DeMets. 1996. Surrogate Endpoints in Clinical Trials: Are We Being Misled? *Annals of Internal Medicine* 125:605–613.

Glasziou, P., G. H. Guyatt, A. L. Dans, L. F. Dans, S. Straus, and D. L. Sackett. 1998. Applying the Results of Trials and Systematic Reviews to Individual Patients. *ACP Journal Club* A–15–16.

Gordon, R. 1994. *The Alarming History of Medicine: Amusing Anecdotes from Hippocrates to Heart Transplants*. New York: St. Martin's.

Guyatt, G. H., D. J. Cook, D. L. Sackett, M. Eckman, and S. Pauker. 1998. Grades of Recommendation for Antithrombotic Agents. *Chest* 114:441S–444S.

Guyatt, G. H., D. L. Sackett, J. C. Sinclair, R. Hayward, D. J. Cook, R. J. Cook, et al. 1995. Users' Guides to the Medical Literature IX: A Method for Grading Health Care Recommendations. Evidence-Based Medicine Working Group. *Journal of the American Medical Association* 274:1800–1804.

Hadorn, D. C., D. Baker, J. S. Hodges, and N. Hicks. 1996. Rating the Quality of Evidence for Clinical Practice Guidelines. *Journal of Clinical Epidemiology* 49:749–754.

Haynes, R. B. 1993. Some Problems in Applying Evidence in Clinical Practice. *Annals of the New York Academy of Sciences* 703:210–224.

Hill, A. B. 1965. The Environment and Disease: Association or Causation? *Proceedings of the Royal Society of Medicine* 58:295–300.

Holmes, O. W. 1988. *Medical Essays, 1842–1882*. Boston: Houghton Mifflin.

Huff, J. 1992. A Historical Perspective on the Classification Developed and Used for Chemical Carcinogens by the National Toxicology Program during 1983–1992. *Scandinavian Journal of Work and Environmental Health* 18(supp1.):74–82.

Hulka, B. S., N. L. Kerkvliet, and P. Tugwell. 2000. Experience of a Scientific Panel Formed to Advise the Federal Judiciary on Silicone Breast Implants. *New England Journal of Medicine* 342:812–815.

Institute of Medicine. 1990. *Clinical Practice Guidelines: Directions for a New Program*, ed. M. J. Field and K. N. Lohr. Washington, DC: National Academy Press.

———. 1992. *Guidelines for Clinical Practice: From Development to Use*, ed. M. J. Field and K. N. Lohr. Washington, DC: National Academy Press.

Leape, L. L. 1995. Translating Medical Science into Medical Practice: Do We Need a National Medical Standards Board? *Journal of the American Medical Association* 273:1534–1537.

Light, R. J., and D. B. Pillemer. 1984. *Summing Up: The Science of Reviewing Research*. Cambridge: Harvard University Press.

Lilienfeld, A. M. 1982. *Ceteris Paribus:* The Evolution of the Clinical Trial. *Bulletin of the History of Medicine* 56:1–18.

Lohr, K. N., and T. S. Carey. 1999. Assessing "Best Evidence": Issues in Grading the Quality of Articles for Systematic Reviews. *Joint Commission Journal on Quality Improvement* 25:470–479.

Lohr, K. N., K. Eleazer, and J. Masukopf. 1998. Review. Health Policy Issues and Applications for Evidence-Based Medicine and Clinical Practice Guidelines. *Health Policy* 46:1–19.

Michaud, C., and C. J. L. Murray. 1996. Resources for Health Research and Development in 1992: A Global Overview. In *Investing in Health Research and Development*. Report of the Ad Hoc Committee on Health Research Relating to Future Intervention Options. Geneva: World Health Organization.

Moher, D., A. R. Jadad, G. Nichol, M. Penman, P. Tugwell, and S. Walsh. 1995. Assessing the Quality of Randomized Controlled Trials: An Annotated Bibliography of Scales and Checklists. *Controlled Clinical Trials* 16:62–73.

Muir Gray, J. A. 1997. *Evidence-Based Healthcare. How to Make Health Policy and Management Decisions*. London: Churchill Livingstone.

Mulrow, C. D. 1996. Critical Look at Clinical Guidelines. In *The Scientific Basis of Health Care Services*, ed. M. J. Peckham and R. Smith. London: BMJ.

Mulrow, C. D., and D. Cook, eds. 1998. *Systematic Reviews: Synthesis of Best Evidence for Health Care Decisions*. Philadelphia: American College of Physicians.

Mulrow, C., P. Langhorne, and J. Grimshaw. 1997. Integrating Heterogeneous Pieces of Evidence in Systematic Reviews. *Annals of Internal Medicine* 127:989–995.

Naranjo, C. A., U. Busto, E. M. Sellers, P. Sandor, I. Ruiz, E. A. Roberts, et al. 1981. A Method for Estimating the Probability of Adverse Drug Reactions. *Clinical Pharmacology and Therapeutics* 30:239–245.

National Health and Medical Research Council (NHMRC), Quality of Care and Health Outcomes Committee. 1995. *Guidelines for the Development and Implementation of Clinical Practice Guidelines*. Canberra: Australian Government Publishing.

Naylor, C. D. 1995. Grey Zones of Clinical Practice: Some Limits to Evidence-Based Medicine. *Lancet* 345:840–842.

Payer, L. 1988. *Medicine and Culture: Varieties of Treatment in the United States, England, West Germany, and France*. New York: H. Holt.

Pere, J. C., B. Begaud, F. Haramburu, and H. Albin. 1986. Computerized Comparison of Six Adverse Drug Reaction Assessment Procedures. *Clinical Pharmacology and Therapeutics* 40:451–461.

Porter, R. 1997. *The Greatest Benefit to Mankind: A Medical History of Humanity from Antiquity to the Present*. London: HarperCollins.

Ross, J. M. 1998. Commentary on Applying the Results of Trials and Systematic Reviews to Individual Patients. *ACP Journal Club* A–17.

Sackett, D. L., W. S. Richardson, W. Rosenberg, and R. B. Haynes, eds. 1997. *Evidence-Based Medicine: How to Practice and Teach EBM*. London: Churchill Livingstone.

Sacks, H. S., T. C. Chalmers, and H. Smith Jr. 1983. Randomized versus Historical Assignment in Controlled Clinical Trials. *New England Journal of Medicine* 309:1358–1361.

Schulz, A. F., I. Chalmers, D. A. Grimes, and D. G. Altman. 1994. Assessing the Quality of Randomization from Reports of Controlled Trials Published in Obstetrics and Gynecology Journals. *Journal of the American Medical Association* 272:125–128.

Scottish Intercollegiate Guidelines Network (SIGN). 1999a. *Grading Systems for Recommendations in Evidence-Based Clinical Guidelines*. Edinburgh: SIGN.

———. 1999b. *SIGN Guidelines: An Introduction to SIGN Methodology for the Development of Evidence-Based Clinical Guidelines*. Report no. 39. Edinburgh: SIGN.

Shaneyfelt, T. M., M. F. Mayo-Smith, and J. Rothwangl. 1999. Are Guidelines Following Guidelines? The Methodological Quality of Clinical Practice Guidelines in the Peer-Reviewed Literature. *Journal of the American Medical Association* 281:1900–1905.

Silverman, W. A. 1993. Doing More Good Than Harm. In *Doing More Good Than Harm: The Evaluation of Health Care Interventions*, ed. K. S. Warren and M. Nosteller. New York: New York Academy of Sciences, pp. 5–11.

Sox, H. C., Jr., S. Stern, D. Owens, and H. L. Abrams. 1989. *Assessment of Diagnostic Technology in Health Care: Rationale, Methods, Problems, and Directions*. Washington, DC: National Academy Press.

Thomas, L. 1983. *The Youngest Science: Notes of a Medicine-Watcher*. New York: Viking.

U.S. Preventive Services Task Force. 1996. *Guide to Clinical Preventive Services*. 2d ed. Baltimore, MD: Williams & Wilkins.

Weingarten. S. 1997. Practice Guidelines and Prediction Rules Should Be Subject to Careful Clinical Testing. *Journal of the American Medical Association* 277:1977–1978.

Williams, G. H. 1999. The Conundrum of Clinical Research: Bridges, Linchpins, and Keystones. *American Journal of Medicine* 107:522–524.

Woolf, S. H. 1991. *Manual for Clinical Practice Guideline Development*. Publication no. 91-0007. Rockville, MD: U.S. Department of Health and Human Services, Public Health Service, Agency for Health Care Policy and Research.

———. 1992. Practice Guidelines, a New Reality in Medicine II: Methods of Developing Guidelines. *Archives of Internal Medicine* 152:946–952.

———. 1999. The Need for Perspective in Evidence-Based Medicine. *Journal of the American Medical Association* 282:2358–2365.

Woolf, S. H., R. Grol, A. Hutchinson, M. Eccles, and J. Grimshaw. 1999. Clinical Guidelines: Potential Benefits, Limitations, and Harms of Clinical Guidelines. *British Medical Journal* 318:527–530.

Woolf, S. H., R. N. Battista, G. M. Anderson, A. G. Logan, and E. Wang. 1990. Assessing the Clinical Effectiveness of Preventive Maneuvers: Analytic Principles and Systematic Methods in Reviewing Evidence and Developing Clinical Practice Recommendations. A Report by the Canadian Task Force on the Periodic Health Examination. *Journal of Clinical Epidemiology* 43:891–905.