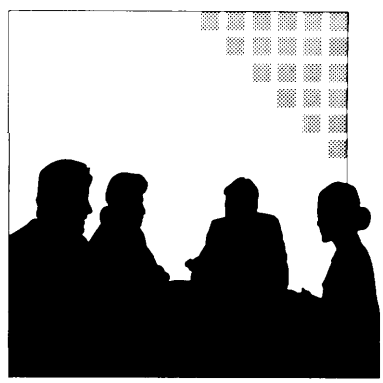
The Development of Clinical Practice Guidelines

7

Clinical practice guidelines are increasingly being viewed as promising tools for promoting cost-effective and appropriate care (201-206,944). Of particular interest are guidelines that are based on the recommendations of panels of experts or representatives sponsored by various organizations. At least 1,500 such guidelines exist, issued by groups as diverse as physicians' professional associations, health care insurers, and the federal government (628).

The focus of this chapter is on federal activities, with a special interest in the efforts of Agency for Health Care Policy and Research (AHCPR) because it represents the latest guideline effort and one of particular interest to Congress. Selected, well-established private guideline efforts are also discussed here to put federal activities in a broader context. Chapter 8 looks beyond the guideline development methods and assesses the potential for guidelines to change clinical practice.

Six federal and four private-sector guideline development efforts form the basis for discussion here. Detailed descriptions of each guideline effort are presented in appendix C.



FEDERAL GUIDELINE EFFORTS

1. **Agency for Health Care Policy and Research.** Since it was established in 1989, AHCPR has issued 11 guidelines, with another 10 under development. Guidelines are produced by panels of 15 to 18 members that are notable for their emphasis on including both consumer representatives and nonphysicians, as well as physicians from a variety of disciplines. AHCPR's guidelines generally address the clinical management of broad health conditions, such as cancer pain and heart failure, and take up to three and a half years to complete.

2. **NIH Office of Medical Applications of Research.** National Institutes of Health's OMAR has issued over 100 Consensus Development Statements since the program's inception in 1977. The primary mission of the Consensus Development Program is to identify and then disseminate to clinicians clinically relevant findings emerging from NIH research, and most topics for conferences are suggested by Institutes of NIH. OMAR's process is unusual for its brevity and its format. Although panel members receive some background information, the recommendations are developed over the course of a single, three-day Consensus Development Conference that includes substantial public input.
3. NIH National Heart Lung and Blood Institute. NHLBI has sponsored detailed guidelines on three medical conditions: high blood pressure, high cholesterol, and asthma. Unique to NHLBI's effort is its guideline panel structure; the guidelines are issued by very large panels (20 to 50 members) that are overseen by coordinating committees made up of representatives of professional societies, voluntary health agencies, and consumer organizations. The coordinating committees have an educational focus; they help promote the guidelines as well as perform other educational functions.
4. **NIH National Cancer Institute.** NCI has previously produced a number of guidelines on cancer prevention and management, but recently it has decided not to make explicit recommendations at all (305). Instead, NCI now issues evidence-based informational statements through its computerized PDQ (Physician Data Query) database. Standing "editorial" panels, which include both NCI staff and outside experts, review and interpret the literature and periodically update the statements on the database.
5. **CDC Advisory Committee on Immunization Practices.** The ACIP, probably the best known of the many groups within the Centers for Disease Control and Prevention (CDC) that issue clinical practice guidelines, comprises a 12-member standing committee that makes recommendations regarding immunization doses, schedules, and other issues with input from liaison representatives from professional societies and other federal agencies. Unlike most other federally sponsored guidelines, those of the ACIP generally are formally endorsed as government policy.
6. **U.S. Preventive Services Task Force.** The USPSTF, convened by the Office of Disease Prevention and Health Promotion (ODPHP), [1] was the first federally sponsored guidelines panel to rate the quality of the scientific evidence behind its recommendations and to link its recommendations directly to that evidence. It is unusual in that it limits group judgment to interpreting the evidence; the personal opinions of panel members are not considered relevant to the guidelines. The first Guide to Clinical Preventive Services, published in 1989, reviewed evidence of the effectiveness of 169 preventive services. It is now being updated and augmented.

PRIVATE EFFORTS

1. **American College of Physicians.** Since 1981, ACP has developed more than 160 guidelines through its Clinical Efficacy Assessment Project (CEAP). Its guideline recommendations, like those of USPSTF, are rated according to the level of evidence supporting them, although the panels do not exclude a role for expert opinion. CEAP panels comprise only internists (the membership of ACP). Their process is unusual for its heavy reliance on consultant-produced reviews of the evidence as the basis for guidelines.
2. **AMA Diagnostic and Therapeutic Technology Assessment Program.** The American

[1] ODPHP is located within the Department of Health and Human Services, under the Assistant Secretary for Health.

Medical Association's DATTA program, in existence since 1982, uses an expert panel to provide forma] guidance regarding the safety and effectiveness of individual technologies (e. g., lung transplantation, Teflon™ preparations for urinary incontinence). Unlike other efforts that produce clinical guidelines. the DATTA process relies primarily on a mailed survey of the opinions of an expert panel; there is no interaction among the panel members.

3. Harvard Community Health Plan (HCHP). Practice guidelines in the form of clinical algorithms—structured flowcharts of decision steps and preferred clinical management pathways (see box 7-1)—are developed as part of this health maintenance organization's quality improvement program. As of early 1994, over 30 clinical topics had been completed or were under development (e.g., asthma, colon cancer screening, depression), most created by an internal panel of HCHP clinicians. Unlike most other guidelines efforts reviewed here, HCHP panels specifically consider cost-effectiveness during the algorithm development process.
4. **RAND Corp.** RAND has developed a method for using forma] group processes to rate the appropriateness of indications for medical and surgical procedures (e. g., hysterectomy. coronary angiography). The ratings have been used both retrospectivcly, to assess the appropriateness of care, and prospectively in precertification programs. The process includes nine-member multispecialty clinical panels that review background material on the literature and rate each possible indication for a procedure on a 9-point appropriateness scale, using a highly structured process of group interaction. Median ratings are used to describe the group judgments, and levels of agreement and disagreement arc formally defined.

GUIDELINE ISSUES RELATED TO DEVELOPMENT

Overview

The diverse federal and private efforts to develop clinical practice guidelines, discussed in this chapter and described in greater detail in appendix C, share a number of features. Most groups devel - oping guidelines have in common the objective of improving clinical decisionmaking in some way by providing clinicians, and sometimes the public. with information. All assign the basic task of creating or endorsing the guideline recommendations to a panel of appointed experts or representatives. In the case of guidelines issued by federal agencies, the guideline panels are virtually always groups of external advisors; most agencies issue, but do not formally endorse, the guidelines created by these groups.

All of the guideline efforts also have some process for identifying potential guidelinc topics, for extracting relevant background information from the scientific literature to which panel members can refer, and for eliciting judgments and (usually) additional opinions or experiences about the literature and the topics from panel members. Most also convene panel members in person to discuss recommendations. Guidelines are usually issued as a book, an article, or statement that includes recommendations to clinicians regarding clinical practice.

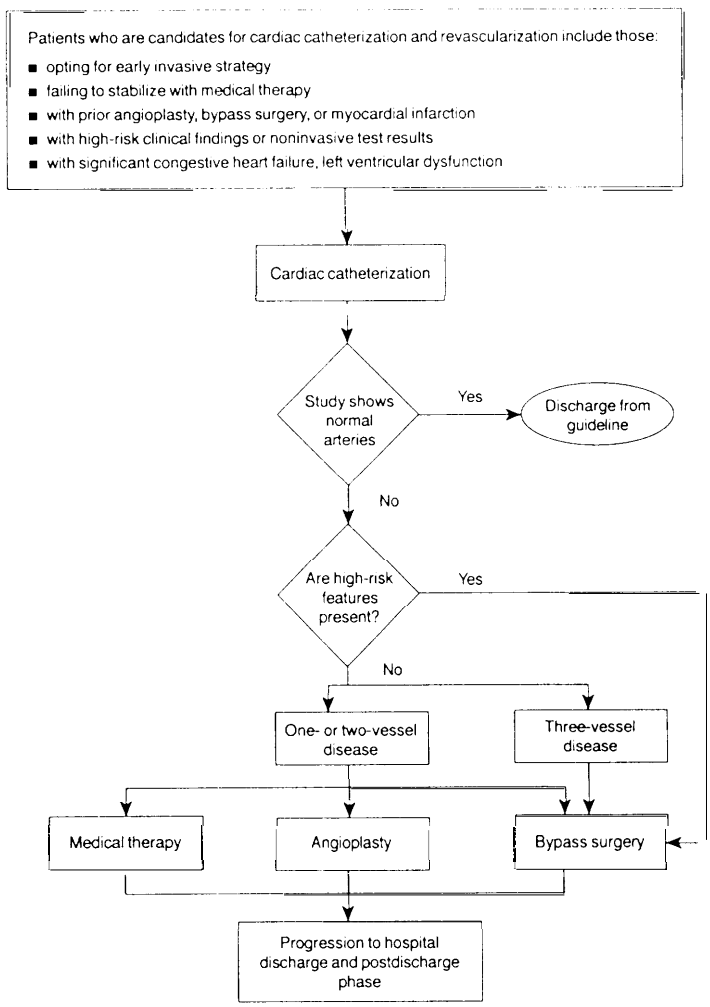
Despite the similarity in the basic structure of guideline development activities. the actual methods of the different groups vary considerably. Major features of guideline development that tend to distinguish one approach from another am:

- **The way in which guideline topics are selected.** Some agencies (e.g., AHCPR) have statutory direct ion regarding guidelines topics. Others (e.g., ACP and AMA's DATTA program) generate topics internally by various means, while still others (e. g., OMAR and RAND) primarily generate guidelines on topics proposed or endorsed by external sponsors.
- **The characteristics of guideline panels and the processes and criteria for selecting panel members.** Guideline panels usually include between 10 and 20 individuals. Some are homogeneous, including only members of a particular group (e.g., a professional society), while others include a range of individuals such as health care providers, methodologists, and

BOX 7-1: Presenting Guideline Recommendations Through Clinical Algorithms

Algorithms are powerful tools for making explicit the relationship between clinical states and diagnostic and therapeutic decisions where there is diagnostic certainty (e.g., if positive strep test, then antibiotic therapy) or diagnostic uncertainty with a probably benign outcome (e.g., if probably viral throat infection, then culture and wait to treat). Algorithms enable the clinician to practice a defined standard of care and may be translated into protocols or chart audits (490)

The Harvard Community Health Plan develops clinical algorithms as part of its quality assurance program. Clinicians seem to prefer algorithms over prose descriptions of the decisionmaking process (490). The National Heart Lung and Blood Institute Illustrates its recommendations with algorithms, and some of the Agency for Health Care Policy and Research panels have also used algorithms to illustrate their recommendations and to identify patient counseling and decision points (figure 7-2) (321). An example of AHCPR's algorithm for management of patients undergoing cardiac catheterization is described below.



SOURCE Off Ice of Technology Assessment, 1994, based on sources as shown Full citations are at the end of this report

consumers. Few organizations sponsoring guidelines. however, have detailed, documented rules regarding panel composition.

- **The scope and perspective of the guidelines.** Some guidelines consider the relative benefits and harms of a wide variety of the alternative clinical approaches for a particular condition or complaint (e. g., AHCPR and NHLBI), while others target particular procedures or technologies and describe their appropriate uses (e.g., DATTA and RAND). Almost all guideline panels consider the safety and effectiveness of interventions, but increasingly, guideline panels are addressing broader issues such as cost-effectiveness, patient preferences. and aspects of health system organization that affect the use of the interventions under consideration. Guidelines also differ in whether they are targeted to specialists, primary care providcrs, or other potential users.
- **The processes used to extract evidence and other information from the scientific literature, experts, the public, and other sources. Some** guideline processes emphasize exhaustive literature searches and syntheses, while others work without a formal analysis of the quality of evidence available to them, or descriptive information on the current state of medical practice.
- **The group processes used to consider evidence and produce agreement on recommendations.** Many panels have fairly loose, free-flowing discussions through which they debate evidence, opinions. and recommendations. Others, however. emphasize formal ways to structure the interaction and judgments of panel members.
- **The degree to which the methods used by panels are explicit, documented, and available.** The processes of some guideline groups are described in great detail within the guidelines themselves and in professional journals (e.g.. HCHP, RAND), Other groups have not published any descriptions of how their guidelines were developed (e.g., CDC's ACIP. NHLBI).
- **The extent to which guideline recommendations are linked directly to scientific evidence.** Some, guideline panels rely primarily on scientific evidence as the basis for recommendations (e.g.. USPSTF and ACP), while others rely on the opinions and judgment of experts to make recommendations when evidence is lacking (OMAR, RAND, AHCPR). In at least one case, prescriptive recommendations are no longer made at all; NC I recently decided to provide informational statements to physicians, which simply interpret existing evidence. rather than specific recommendations for practice.
- Administrative features of the process, such as whether guideline panels are "standing" or ad hoc and the extent of administrative oversight of guideline activities.

These features are discussed in more detail below.

Choosing Guideline Topics

Many organizations that are developing guidelines publish the criteria and process they use to select topics. Criteria frequently cited as being used to select topics for guideline development include:

- **Public health impact—the** prevalence, incidence, and severity of the condition in question and the potential for interventions to prevent the condition or ameliorate symptoms.
- **Cost of procedure—a** procedure might be costly as a single unit (e. g., organ transplantation) or because it is commonly performed, for example, as part of population screening (e.g.. colonoscopy).
- **Availability of evidence—for some** technologies **there is** good evidence on which to base judgments (e.g.. several randomized clinical trials), while for others only descriptive clinical experience and opinion are available.
- **Variation in clinical practice—may** reflect clinician uncertainty or genuine differences in schools of thought in the rnanagement of certain conditions.

- **Controversy —may be over alternative interventions for a condition, who should deliver care. or where a service should be delivered.**
- **New versus established technologies+ stab**lishing guidelines on a new or emerging technology could forestall inappropriate use.

In general, however, these criteria serve more as loose guides than as part of a systematic prioritization process. For example, many groups select topics based on the level of controversy and availability of evidence, but most do not try to assess the state of clinical practice or the quality of evidence for a particular guideline topic *before* a guideline topic is selected. Instead, a guideline topic is selected usually through some sort of nomination or survey process, and then a panel is assembled to focus the assessment and begin to identify relevant evidence.

Federal agencies often have congressional mandates that give some direction to their selection of topics. AHCPR's guidelines effort, for example, is specified by its authorizing statute, which directed the agency to examine issues of relevance to the Medicare and Medicaid populations (Public Law 101 -239). When reauthorized in 1992, AHCPR was further directed to consider clinical treatments or conditions that were costly, for which there was significant variation in the frequency or the kind of treatment provided, and for which inappropriate use of health care resources was likely (Public Law 102-410).

The authorizing legislation also specified that AHCPR, created in December 1989, had to issue at least three guidelines by January 1991 (Public Law 101-239). To reach this deadline, AHCPR initially selected topics for which guideline development was already underway (798). Since 1993, AHCPR has published a list of possible topics for guideline development in the *Federal Register* and elsewhere and solicits comments and recommendations for new topics (812). AHCPR has also recently brought representatives of groups together to discuss potential topics for guidelines (53). The Institute of Medicine (IOM) is currently conducting a study for AHCPR on setting priorities for guideline development (81 3).

Guideline efforts within NIH emphasize the role of disseminating research findings of the Institutes to clinicians. Topics for OMAR's Consensus Development Conferences are suggested by the Institutes themselves. In addition, two of the institutes, NHLBI and NCI, issue their own guidelines or statements on topics within their domains. NHLBI focuses on only a few clinical conditions that fall within its purview (i.e., high blood pressure, cholesterol, and asthma), while NCI issues statements on topics across the spectrum of cancer management (i.e., screening, treatment, and supportive care).

AHCPR, OMAR, NHLBI, and NCI all frequently cover preventive services, such as screening and immunizations, in their guidelines. In addition, two other federal guideline efforts reviewed here—CDC's ACIP and ODPHP's USPSTF—focus exclusively on preventive practices. USPSTF covers the full range of preventive services provided in clinical settings, while ACIP makes recommendations relating to immunization practices.[2]

Some private guideline sponsors have developed more systematic ways to solicit opinions on potential topics for guidelines from practicing clinicians. ACP, for example, surveys its members to help identify topics of interest to practicing internists as part of its CEAP program. Topics are selected for AMA's DATTA program in part through a survey of DATTA subscribers.[3] HCHP develops guidelines based on nominations made by HCHP clinicians and also considers health plan data to identify practices for which there is variation.

[2]CDC also issues other guidelines. Topics generally focus on achiev ing national health objectives as stated in *Healthy People 2000* and in CDC's mission statements (709).

[3]Questions for DATTA evaluations are considered from a variety of sources (e.g., physicians, patients, third-party payers, peer review ers). The survey is one of several mechanisms used to identify topics (see appendix A).

Organizations vary greatly in the type of topic selected for guideline development. Some focus on selected conditions and complaints while others address specific procedures or technologies (945). AHCPR has assumed an interdisciplinary perspective to examine alternative approaches to diagnose and manage selected chronic conditions and complaints (e.g., pressure ulcers, pain. depression). For example, they examined medical, surgical, and behavioral interventions in their urinary incontinence guideline. Such an approach is attractive to primary care providers and patients in that it provides an assessment of the relative value of competing approaches that may never before have been compared with one another in a single document. These comprehensive and interdisciplinary guidelines are generally more demanding to produce from a methodologic point of view (e.g., nomenclature, measurements, and outcomes of interest may differ across disciplines) and may therefore take more time and be more costly than a more narrowly focused guideline. Some of the AHCPR guidelines, for example, have taken over three years to develop.

Other efforts are more narrowly focused on the circumstances for which particular technologies are most appropriately used. RAND researchers have developed appropriateness ratings for expensive, commonly used surgical procedures (e.g., coronary angiography, hysterectomy).4 OMAR has tended to focus on technologies emerging from the NIH research arena and in (he process of diffusing into clinical practice (e.g., using antibiotics to treat peptic ulcers). The AMA DATTA program evaluates primarily new and emerging technologies of interest to specialists (e.g., lung transplantation). These more narrowly focused assessments have their attractions: they can target a technology about which there is uncertainty within the practice community. they can be targeted to certain groups of clinicians, they are attractive to insurers and health planners, and they can often be completed relatively quickly. Most OMAR and DATTA assessments, for example, are completed within one year.

Selecting Guideline Panels

The diversity of guideline efforts is reflected in the composition of guideline panels, which vary in size and include a range of individuals from technical experts to consumers (609). Most guideline panels range in size from 10 to 20 members. NHLBI panels are unusually large, including as many as 50 members, with smallcr subcommittees formed to address specific subtopics.

The background and perspectives of the individuals involved in guideline setting likewise vary considerably, across both different guideline efforts and different panels within a particular effort. Federally sponsored guideline groups are often relatively diverse, including nonphysicians and consumer representatives. The AHCPR and NHLBI panels, for example, usually include a range of health care providers and at least one consumer representative. In contrast, private physician groups have generally confined panel membership to physicians. The ACP CEAP and AMA DATTA panels, for example, include only physicians, some of whom have methodologic expertise.

A potential threat to the validity of a guideline is selecting panel members who share a particular bias. A biased group could be assembled quite inadvertently by selecting certain types of members. In research on panels using the RAND method. all-surgical panels rated more procedures appropriate and had more agreement about appropriateness than a mixed panel composed of surgeons and non surgeons (448,683). Within mixed panels. surgeons rate the appropriateness of surgical procedures substantially higher than do nonsurgeons (585). This finding is consistent with others showing that physicians who perform a given interven -

[4] More recently RAND has looked at less invasive procedures (e.g., spinal manipulation for low-back pain) (689).

(ion frequently are more likely to judge it as beneficial (578).

This user bias is not surprising, as one would expect those who perform a procedure to be more committed to its value. What is striking is the magnitude of the effect. In one example, a panel of surgeons assessing carotid endarterectomy rated 70 percent of cases as appropriate, whereas a multidisciplinary group found only 38 percent of procedures to be appropriate (448).[5]

User bias is not the only source of panel difference; different backgrounds and cultural assumptions also matter. Panels in the United States and the United Kingdom (with the same physician specialty composition) came to different conclusions when assessing the appropriateness of treatments for coronary disease. The U.S. panel judged more indications appropriate, and had better agreement among members, than did the U.K. panel. When the ratings of the two panels were applied to two groups of patients who had had the procedure in question, the U.S. panel judged 17 percent and 27 percent of the procedures as inappropriate, whereas 42 percent and 60 percent were judged as inappropriate by the U.K. panel counterparts (85).

Panel sponsorship, composition, and the inherent interests of different groups of clinicians can be major factors in the acceptance of guidelines. (138,582). Some eye surgeons, for example, disagreed with some aspects of AHCPR's cataract guideline. The AHCPR panel reviewed evidence on the criteria used to determine when cataract surgery might be indicated and found no evidence to support the use of some of the preoperative tests now in use. The AHCPR panel recommended that a patient's level of visual dysfunction rather than certain other tests be used as a criterion for surgery. The surgeons contended that the federal guidelines were intended to reduce the number of Medicare patients who would be eligible for cataract surgery. In another example, the American Psychological Association, the principle professional society representing psychologists, failed to endorse AHCPR's depression guideline, in part because of its perceived emphasis on medical therapy at the expense of psychotherapy (618).

To enhance guideline credibility, most clinical practice guideline efforts have panels drawn heavily from clinician groups whose practices will be affected by the guideline. AHCPR and NHLBI, for example, solicit nominations for panel membership from health care professional organizations. The guidelines developed at HCHP are written by the very clinicians who will ultimately use them. (The research suggesting that clinicians are more likely to believe guidelines in whose development they participated is discussed further in chapter 8,)

The desire to appoint guideline panels that are credible to the clinicians whose practices will be the most affected presents a dilemma for policymakers, because it also creates the potential for biased guidelines when developed by enthusiasts. Guideline panels that are intended to represent affected clinicians are likely to comprise a disproportionate number of users.

Another limitation of homogeneous clinician panels, particularly panels comprising primarily physicians, is the inability of such a panel to represent nonphysician concerns. It may be easier to consider interventions outside of the usual purview of medical specialists, and issues such as patient preferences and concerns, with a more heterogeneous panel. The inclusion of nurses and a psychologist on AHCPR's urinary incontinence panel, for example, probably facilitated the panel's consideration of interventions such as biofeedback techniques. An important recommendation of the urinary incontinence panel was

[5] The RAND process rates indications as appropriate, equivocal, or inappropriate. The proportion of *inappropriate* cases fell from 31 percent to 19 percent when a multidisciplinary instead of an all-surgical group rated the procedures. Of note is that both panels rated the same 12 to 13 percent of cases as inappropriate. This indicates that there appears to be a consensus regarding inappropriateness for a small subset of cases (448).

that such techniques be considered seriously as treatment options that could be alternatives to surgery. which has inherent risks and complications (802). Heterogenous panels too have their limitations, however. While consumer representatives and panel members representing fields such as ethics may play important roles in setting the guideline agenda and expressing the possible personal and social consequences of guideline recommendations, they often lack a technical background and so may not be able to fully participate in panel deliberations regarding the interpretation of medical and epidemiological evidence under consideration.

Despite the importance of panel membership on the scope and recommendations of a guideline, few organizations have strict criteria or rules regarding panel membership. Government-sponsored panels on occasion have been accused of bias for including enthusiasts for particular interventions (312). Of the processes reviewed for this report, only RAND and AMA's DATTA set a limit for the number of panelists that perform the procedure under consideration. AHCPR's reauthorizing legislation required that panel members who derive their primary income from procedures under consideration be limited on the panel but did not specify what the limit should be. Some groups attempt to screen panelists for potential conflicts of interest (e.g., AHCPR, NIH, ACP). OMAR seeks a chairperson and panelists who are neutral (388).To try to assure neutrality, the publications of candidate panelists are scrutinized to ensure that they have not published extensively on the conference topic (378).

Defining the Scope of Guidelines

The intended audience for a guideline is an important determinant of the guideline's scope. Guidelines are typically directed at physicians, but issues of importance to other health care providers (e.g.. nurses, chiropractors), patients, and payers have broadened the focus of some guideline efforts. Federal guideline efforts have generally focused on primary care clinicians and increasingly have been directed to patients. AHCPR, NHLBI, and NCI, for example, direct their guidelines and statements to both clinicians and patients.

Safety and effectiveness are issues addressed in almost all guideline efforts. Given that guidelines are policy statements about the appropriate distribution of clinical resources, however (see chapter 6). some observers argue that unrealistic or even undesirable recommendations can be made when factors such as cost, health care system constraints. and patient preferences are not considered in the process of examining alternative clinical approaches. For example. a recent recommendation of the NIH Consensus Development Program-that all infants be screened for hearing impairment within the first three months of life (preferably before discharge from the hospital)—has been criticized, in part because many practical implementation and cost issues were not fully addressed (59,1 46). Similarly, some argue that implementing the NIH Consensus Panel recommendation that primary care physicians refer elderly patients suffering from sleep disorders to centers for sophisticated testing would be prohibitively expensive, because the condition is very prevalent (701). While some policy makers argue that guideline developers need to consider the health policy implications of their guidelines while they are being developed, scientists involved in guideline development have often expressed their discomfort in assuming this role more explicitly (633).

Cost and Cost-Effectiveness

The role of cost projections and cost-effectiveness analysis in developing practice guidelines is controversial (945). Many federal guidelines have included assessments of the guidelines' likely impact on health care costs, and have included some informal discussion of existing evidence of cost-effectiveness, but none reviewed by OTA for this report has routinely included formal cost-effectiveness analyses in the recommendation making process. Increasingly, guideline developers have included resource assessments in their guidelines,

but groups differ in how much and how explicitly they allow costs to influence their recommendations.

Most guidelines issued by AHCPR to date have included statements about some of the anticipated changes in health spending that would associated with guideline implementation, but they have not explicitly considered the relative cost-effectiveness of alternative interventions when making guideline recommendations. Nor, often, have the anticipated savings from implementing an AHCPR guideline been compared directly and quantitatively with new costs that the guideline would impose (e. g., by encouraging the use of certain services). The recently released AHCPR guideline on heart failure, for example, was promoted with a discussion of the fact that its implementation could result in savings of $2 billion per year (448a). The guideline as promoted, however, did not present quantitative estimates of offsetting new costs. A few AHCPR panels have not considered cost explicitly at all (e.g., the guideline on management of HIV infection).

HCHP guidelines panels do frequently and explicitly consider costs and cost-effectiveness in their deliberations. NHLBI has added some discussion of costs in its most recent guideline on managing high cholesterol (857). The USPSTF, NCI, and RAND panels sometimes review evidence of cost-effectiveness, but they have explicitly excluded cost as a criterion for their recommendations or judgments regarding appropriateness.

Even when panels do attempt to incorporate cost considerations in guideline development, cost data are often not available for all interventions under consideration. AHCPR has commissioned a study on sources of cost data for guideline development, and the agency reports that it is assessing the adequacy of the cost analyses included in 10 of the guidelines it has sponsored (821).

The IOM has concluded that every clinical guideline should include information on the health and cost implications of alternative management strategies, but that every guideline need not be based on formal judgments of cost-effectiveness. They reasoned that this charge may be too great for individual guidelines panels and that perhaps guideline developers were not always the right source of such judgments (376). The Institute did not explore in detail exactly how cost-effectiveness considerations should be integrated into guideline development, however.

Patient Health Status and Functioning

There is great interest in formulating clinical recommendations based on health status assessments that are of interest and relevance to patients (376). Measures such as maintenance of physical, cognitive, and social functions and alleviation of pain and discomfort are especially relevant when developing guidelines for non-life-threatening chronic illnesses. NCI, for instance, maintains information on supportive cancer care (e.g., managing cancer-related nausea and pain) as part of its computerized PDQ database for patients and clinicians (359). While functional outcomes are widely regarded as important, information on them is often unavailable because many clinical studies have not included them as outcome indicators.

Sometimes functional patient outcomes are considered, but there are limits to how accurately they can be assessed. All guideline panels examined in this report comprise primarily clinicians, and many studies of patient outcomes are based on clinicians' assessments. As discussed in chapter 3, however, clinicians' and patients' assessments of outcomes and their importance can be very different. Nor do clinicians' value judgments, used throughout the guidelines process, necessarily reflect patient and societal values (884).

It is also unclear how much weight the outcomes are given. When clinicians rate appropriateness as part of the RAND process, for example, a variety of patient outcomes may be considered. RAND general] y defines appropriateness to mean that "the expected health benefit (increased life expectancy, relief of symptoms, reduction in anxiety. improved functional capacity, etc.) exceeds the expected health risks (mortality. morbidity, pain produced by the procedure)" (688). What

mix of outcomes physicians use in their assessments and what relative weights these outcomes are assigned is unknown. The meaning of such assessments is even open to question, because some evidence suggests that physicians are often poor judges of levels of patients' discomfort and functional status.

Patient Preferences

Patient preferences are measures of sat is faction or desirability that people associate with the presence of symptoms and functional limitations that can affect quality of life (268). Incorporating patient preferences into guidelines is of great interest, but how to measure and usc patient preferences are subjects of ongoing research and debate. The quantification of patient preferences (also called patient utilities) is an active area of research, but there have been few attempts to formally integrate such patient preferences in the guideline process. A notable exception was Oregon explicit incorporation of patient preferences when initially prioritizing health services to establish a benefit package under its controversial Medicaid reform plan (788).

AHCPR and NHLBI have included patient representatives on guideline panels. and AHCPR and the NIH Consensus Development Program routinely hold public forums as part of the guideline process where public concerns and questions can be aired. AHCPR has pioneered a very pragmatic way to include patient preferences into their guidelines. Clinical algorithms are used to portray recommended management strategies, and the algorithms include points in the decisionmaking process where physicians and other caregivers need to discuss with patients or families their preferences for particular options (32 1,376). Assessments of patient preferences were made in the AHCPR guideline on benign prostatic hyperplasia. Wide variations in preferences were found, leading the panel to conclude that patients preferences need to be elicited as part of the treatment decisionmaking process (819).

Recognizing patient preferences in electing treatment options is clearly desirable for patients, but it does not necessarily lead to decreases in observed variation or more standardized practice. In one study, patients who had experienced acute upper gastrointestinal bleeding almost always preferred to have diagnostic endoscopy rather than less invasive tests or no testing. because they found the information it conveyed comforting even though it would not affect their management or prognosis. The researchers concluded that the current rate of diagnostic endoscopy is higher than would be expected based on physicians preferences but is quite consistent with patient preferences (1 77).

Identifying and Synthesizing Evidence

All of the guideline efforts reviewed here include some mechanism for identifying and synthesizing the existing literature relevant to the guideline topic, so that it can be considered and discussed by panelists. The way in which this is carried out, however. varies considerably.

In many cases, extensive literature reviews are conducted as part of the guideline development process, sometimes at great expense. At AHCPR, for example, literature reviews have taken up to nine months and have cost up to $235,000 (376).

Once the evidence is amassed, different strategies can be used to synthesize it. Often panel members assess the literature themselves. The NHLBI guideline process, for instance, leaves literature reviews to panel subcommittees with no set methods or criteria established to ensure uniformity within a guideline.

Panels sometimes are assisted by a methodologist trained in epidemiology or statistics. AHCPR panels, for example. have generally benefited from the assistance of methodologists assigned to the panel to construct evidence tables. A number of analytic techniques have been developed to synthesize clinical evidence (e.g., meta-analysis). but the techniques are not usually used because they arc time consuming and differences in study characteristics often preclude their use.

A few guideline panels actual] y rate the quality of available evidence and give the most weight to high-quality studies (e.g., AHCPR, NCI.

BOX 7-2: U.S. Preventive Services Task Force Strength of Recommendations, Design Categories, and General Criteria of Effectiveness

Strength of recommendations

- A There is good evidence to support performing the preventive service,
- B: There is fair evidence to support performing the preventive service
- C There is poor evidence to support performing the preventive service, but recommendations may be made on other grounds.
- D There is fair evidence to discontinue performing the preventive service
- E, There is good evidence to discontinue performing the preventive service

Study design categories

- I Randomized controlled trials
- II: Controlled trials without randomization
- **II-2** Cohort or case-control analytic studies
- **II-3** Multiple time series, dramatic uncontrolled experiments
- **III:** Opinions of respected authorities, descriptive epidemiology

General criteria **of effectiveness**

Screening tests

Accuracy and reliability of screening tests

Effectiveness of early detection

A, Treatment efficacy

B Asymptomatic period

C Benefits of early detection

D. Acceptability

Counseling interventions

Efficacy of behavior change in risk reduction

Efficacy and Effectiveness of counseling patients about health behaviors

Immunizations/ chemoprophylaxis

Efficacy and effectiveness of agent

Efficacy and effectiveness of counseling

SOURCE S H Woolf and H C Sex, "The Expert Panel on Preventive Services Continuing the Work of the USPSTF," *American Journal of Preventive Medicine* 7(5)326-330, 1991

USPSTF, ACP). Generally, priority is given to study methods that are less prone to bias, with evidence from randomized controlled trials rated highest and evidence from sources such as case reports or expert opinion rated lowest (box 7-2) (1 44,207,871).

Different groups use different rating schemes and even among the panels sponsored by a single agency, rating systems may vary. The different guidelines issued by AHCPR, for example, have employed different systems to rate evidence (501). Efforts to develop a uniform rating system are complicated by the incorporation of aspects of both the design and quality of the study. Some observers have questioned whether it is always appropriate to give more credence to clinical trials than to other study designs, because a well-done case-control or other quasi-experimental study may sometimes be superior to a poorly conducted randomized clinical trial (376). While explicit rating systems are useful as guides, expert judgment is often still needed to assess the value of many studies.

The USPSTF has sometimes used "causal pathways" to frame the evaluation of evidence (44,717). For example, if evidence is lacking on the association between a preventive service and the outcome of interest (e.g., the impact of screen-

ing adolescents on future scoliosis-related morbidity), the panel examines evidence along the causal pathway (e.g., the relationship between screening and diagnosing scoliosis early, and then the relationship between early intervention and subsequent health outcomes such as back complaints, disability. and psychosocial effects) (figure *7-1).*

There are recognized deficiencies in the body of literature available for review. An examination of the literature available on six medical and surgical procedures, for example, revealed:

- few randomized controlled trials on which to rate the procedures' appropriateness,
- incomplete and contradictory information on the indications for and efficacy of the procedures,
- almost no data on costs and utilization, and
- data on complications that failed to specify patients' symptoms or the relationship between complications and reasons for doing the procedure (245).

Consequently, results from studies other than randomized trials often must be used in the guidelines process. Evaluating such studies is another example of the importance of judgments in interpreting evidence. The AHCPR cataracts panel, for example, considered claims data findings that suggested an increased risk of a serious complication (retinal detachment) in some cataract patients (391). This research lent support to two of the guideline recommendations: that the indications for the procedure be clearly documented in the chart; and that the laser procedure should not be scheduled at the same time as the original cataract procedure (806).[6] Researchers, concerned that limitations of the data used in this study might have led to a misleading finding of complications, are now attempting to collect detailed primary data (724). The AHCPR panel on benign prostatic hyperplasia rejected the use of findings about the risks of transurethral resection of the prostate (TURP) from administrative data, judging that the data source was likely biased (140.8 19).

Recognizing that interpretation of evidence may be a matter of judgment, some panels have used formal processes to assess the reliability of these judgments. The AHCPR cataracts panel, for example. had multiple reviewers rate the content and methodology y of research articles and assessed interrater reliability.

Techniques To Aid Group Interaction and Decisionmaking

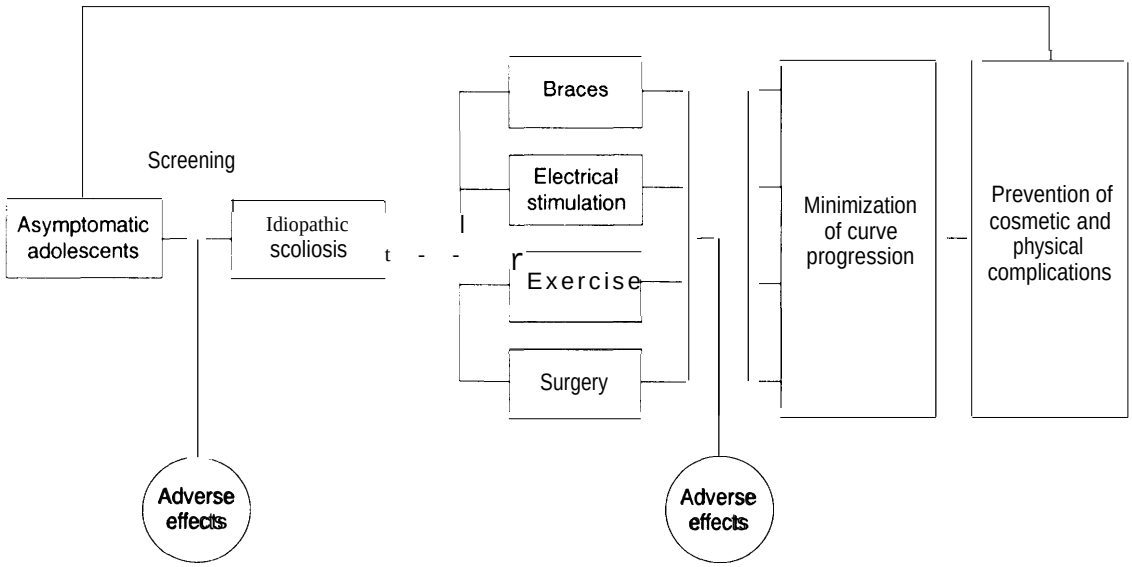
Guideline panels must incorporate information, exchange ideas and opinions, and finally reach some level of agreement on practice recommendations. Group composition and aspects of group process become increasingly important as the availability and strength of evidence declines (469).

How these essential aspects of the guideline development process are accomplished varies considerably. Many guideline processes are informal and have been organized around a series of loosely defined steps:

- A group of appointed experts or representatives is assembled.
- Available literature is collected and summarized (by staff or others) and then reviewed individually by panel members.
- Ideas, opinions. and interpretations of the literature are exchanged in meetings that follow a "roundtable" format. A chairperson facilitates the meetings, often with an appeal to the evidence as it is described in the review.
- Recommendations are made and agreed on, sometimes by a vote (often with a requirement of majority or unanimous agreement).
- Recommendations are reviewed by outside experts and practitioners and then reconsidered by the group.

[6] This last recommendation was partly based on evidence that the laser procedure is not routinely needed, but it was reinforced by the new claims-based information concerning retinal detachment.

FIGURE 7-1: Examining the Evidence on Screening for Idiopathic Scoliosis Along a Causal Pathway



Linkage: step in causal pathway	Evidence codes	Quality of evidence
1 Accuracy of screening tests. evidence that physical examination of back can detect curves.	II-2	Fair significant Interrater variation, poor reference standard, lack of evidence form physician screening
2. Adverse effects of screening evidence that screening is associated with an Increased risk of complications,	III	Poor most postulated adverse effects have not been evaluated in studies.
3. Effectiveness of early detection evidence that persons detected through screening have better outcomes than those who are not screened	II-3	Poor uncontrolled studies based on time trends after initiation of screening, failure to control for confounding temporal factors
4 Braces,	II-2, II-3	Poor selection bias, lack of internal control groups (most studies), inadequate follow-up, small sample sizes, lack of health outcome measures.
5. Lateral electrical surface stimulation.	II-2, II-3	
6. Exercise.	I, II-3	
7, Surgery	II-2, II-3	
8 Curve progression evidence that curves detected on screening are destined to progress to curves of clinical significance,	II-3	Fair significant number of patients unavailable for followup, variable measures of progression,
9 Complications of curve progression evidence that persons with scoliosis are more likely to experience back complaints, psychosocial effects, disability.	II-3	Poor studies generally lack control groups, have high attrition rates, include mixture of patients with different problems, and use variable measures to judge outcome
10 Adverse effects of treatment evidence that treatment is associated with an increased risk of complications,	III	Poor most postulated adverse effects have not been evaluated in studies,

SOURCE U S Department of Health and Human Resources, Public Health Service, Off Ice of the Assistant Secretary for Health, Off Ice of Disease Prevent[on and Health Promotion, U S Preventwe Services Task Force, Washington, DC, 1993

This approach typifies, for example, the NHLBI and ACIP guideline processes.

Formal Group Processes

A very few organizations issuing guidelines use formal. structured interactive group techniques to orchestrate the guideline process and to make explicit recommendations. Of the groups reviewed here, two-HCHP and RAND, both private guideline developers—use formal group processes. A range of group process techniques have been developed to facilitate group decisionmaking.[7] Some methods are best suited for identifying problems and establishing objectives. Others are designed to help conceptualize alternatives, while still others are tailored to groups that need to make choices among a range of alternatives (538). For the production of any guideline, then, different group processes might be used at different stages of the development process.

Two group processes extensively studied and used to develop clinical practice guidelines are the Nominal Group Technique (NGT) and the Delphi technique (box 7-3) (366). These methods help ensure participation of all members, and they provide explicit decisionmaking rules. Group judgments achieved through either the NGT or Delphi technique generally improve judgments relative to those derived by taking the average of individual judgments, but neither technique clearly seems to outperform the other (660). Both techniques are superior to informally interacting groups in generating new ideas (156).

Formal structured methods can potentially improve group performance by organizing complex information for group consideration, facilitating agreement and decisionmaking, and increasing personal satisfaction of group participants (156). In the absence of a formal process, groups may not perform optimally because one or a few individuals can easily dominate discussions. thereby suppressing the consideration of a balanced set of options. Informal group discussions can also sometimes lack focus and be time-consuming and unproductive. Because much of the expense of guideline development lies in the assembling of experts, methods to make their time together more efficient are desirable.

A potential barrier to using formal methods to structure group process is their unfamiliarity to clinicians. Also, the relative value of these techniques has not been assessed in the context of practice guidelines.

Decision Support Systems

Another way to structure the guidelines process is to use a structured, quantitative framework for integrating and weighting medical and other scientific data. Such support systems can make unwieldy problems more manageable by structuring thought processes, clarifying interrelationships among important factors, and integrating complex data (681). Decision support systems require explicit definitions of the problem, assumptions, events, and outcomes. Such a process helps to assure that relevant factors are considered. and it enables others to review and check the reasoning behind decisions.

Decision support systems help overcome the inherent human limitations of processing information and making judgments (68 1). For example. most people:

- cannot consider more than three to seven alternatives concurrently;
- have a limited cognitive capacity to revise judgments; and
- have biases that affect judgments (e.g., people consistently overestimate the probabilities of events familiar to them and underestimate the probabilities of unfamiliar events).

These limitations in judgment affect the assessment of probabilities, integration of new or contradictory information, estimation of the validity of evidence, and assessment of preferences and values (68 1,942).

[7] Much of the developmental work on group processes has occurred in nonmedical settings (855).

BOX 7-3: Formal Group Process Techniques Used in Developing Guidelines

Nominal Group Technique

The Nominal Group Technique (NGT), developed by Andre Delbecq and Andrew Van de Ven in 1968, has been used widely in human services organizations, business, and as part of evaluative research, The NGT splits problem solving into two phases, an idea-generating phase and a decisionmaking phase A different group process is used for each phase, In the first phase, each member of the group Individually makes a list of ideas for group consideration. This aspect of the process gives the technique its name—individuals participating in the "nominal" group process are a group "in name only" (i e., nominal) and do not initially interact verbally All individually generated ideas are then recorded on a flip chart for the group and are openly discussed In the second decisionmaking phase of the process, individuals vote on priority ideas and a group decision is mathematically derived through rank ordering or rating

The Delphi Technique

The Delphi Technique was created at the RAND Corp. in 1950. It was originally used to forecast technological developments, thus, like the Delphic oracle, it was used to look into the future The technique was designed to help groups of experts identify a range of possible program alternatives, explore underlying assumptions or information leading to different judgments, and to reach consensus on complex Issues,

Unlike the NGT, the Delphi technique does not require that participants meet face-to-face, Generally the technique IS typified by the followlng process:

- A questionnaire is dlstributed by mail to a respondent group,
- Respondents independently answer the questionnaire and return it,
- Responses are summarized and a feedback report is developed for each respondent
- Respondents receive the feedback report with a new questionnaire and independently evaluate their earlier responses,

Other factors that affect judgment relate to how questions are framed, For example, clinicians appear to make different judgments in evaluating an individual patient as compared with considering a group of similar patients. Physicians seem to give more weight to the personal concerns of patients when considering them as individuals and more weight to general criteria of effectiveness when considering them as a group (629).[8]

In another example of the importance of framing questions, RAND researchers first asked panelists to rate the appropriateness of certain scenarios for endoscopy and cholecystectomy, specifying that the patients in the scenarios had no comorbidities. When the panelists were asked to rerate the scenarios, with the patients described as having high comorbidity, only of few of the scenarios originally designated as appropriate re-

[8] In one experiment, physicians were asked to make clinical decisions after reading two clinical scenarios. The scenarios were identical except that in one they were asked to evaluate an individual patient and in another they evaluated a comparable group of patients. Physicians were more likely to recommend additional testing and recommend therapy when they evaluated the individual versus the group scenarios (629).

BOX 7-3 continued: Formal Group Process Techniques Used in Developing Guidelines

- Respondents are asked to independently vote on priority ideas Included in the second questionnaire and return their responses
- A final summary and feedback report is sent to the respondents and to decisionmakers

The Delphi process often varies according to whether the respondent group is anonymous, whether open-ended or structured questions are used to obtain Information for the respondent group, how many Iterations of questionnaires and feedback reports are needed, and what decision rules are used to aggregate the Judgments of the respondent group The theory underlying the Delphi technique is that Improvements in judgment with each Delphi Iteration occur because the most knowledgeable panelists confidently retain their Judgments and anchor the median close to the true value, while less knowledgeable panelists change their Judgments to be closer to the median If this in fact occurs, the median response should move toward the truth over rounds of the Delphi process

Those who maintain their same judgments over Iterations may not in fact be more knowledgeable, but instead be dogmatic and intransigent If so, the convergence of opinion observed over Iterations may just reflect the influence of a dominant individual To alleviate this effect, substantive feedback to panelists must include not just median ratings but also justifications for ratings based on the evidence at hand There is a tendency for group judgment to converge over time and without appropriate feedback, such convergence could represent an artifact of the method rather than true convergence of opinion

SOURCES Office of Technology Assessment 1994 based on information from A L Delbecq A H Van de Ven and D G Gustafson *Group* Techniques *for Program Planning A Guide to Nominal Group and Delphi Processes* (Glenview IL Scott Foresman and Company 1975) G Rowe G Wright and F Bolger Delphi A Reevaluation of Research and Theory *Technological Forecasting and Social Change* 39235-251 1991

mained so designated (13 percent for endoscopy, 33 percent for cholecystectomy) (405).

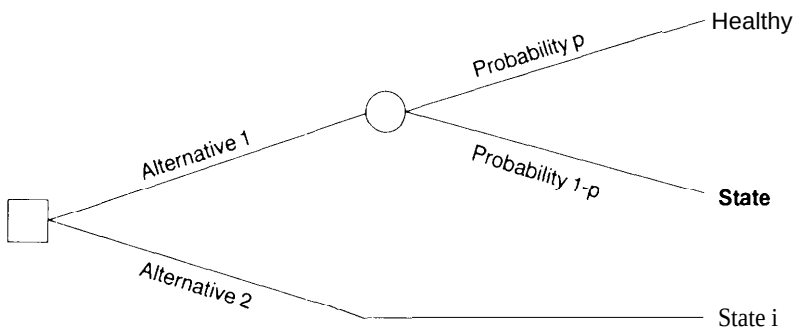
One of the most common decision support systems used in developing guidelines is *decision analysis,* a useful structure for determining the preferred course of action under conditions of uncertainty *(547),* Decision analysis provides a framework for specifying the probability that a particular clinical state exists and quantifying the value of the various outcomes of a decision (see chapter 3 for a more detailed description of this technique). A decision analysis attempts to answer the questions. “Is it more desirable that I do this or that’?” and “If this is so, what is the probability that that is so’?” It is used during the guideline panel’s attempt to consider all the relevant information. Steps involved in decision analysis include defining all possible outcomes of interest. quantifying their probability of occurring, and sometimes considering the costs and benefits associated with each outcome (box 7-4).

(Algorithms are a related framework that are sometimes used to assist the clinical decision maker actually using a guideline. In contrast to decision analysis, algorithms prescribe, "Given this. do that” (490) (see box 7-1).)

Although they are theoretically attractive and can be very useful, decision models also have limits. Generally, the systems are complex and time consuming. Specification and structuring of the problem, obtaining the values for the data inputs, and computation of the primary and subsidiary analyses require substantial expertise in clinical

BOX 7-4: Steps in Using Decision Support Systems

- Identify the problem in terms of the clinical presentation, population, time frame, and perspective (e g patient, payer, provider, society) The perspective of the model affects the costs and values assigned to events and outcomes and thus strongly influences the results.
- Structure the problems and explicitly describe the underlying logic and reasoning. Alternate courses of actions and their consequences must be specified. Outcomes of interest might include physiological parameters, such as mortality/survival and complications, and physical, social, and psychological, cognitive, role, social, and other functional measures Broader measures that incorporate patient values and preferences and costs, such as quality adjusted life years, cost-effectiveness, and cost-benefit also should be considered. Based on the outlined problem structure, the probability of occurrence of events and outcomes must be obtained from objective, published, peer-reviewed scientific literature, but may be based on expert judgment when other data are unavailable. The model is often expressed as a decision tree with branches representing different outcomes



- Select preferred options considering the expected value of each alternative strategy conduct sensitivity analysis when data underlying the decision model are uncertain to assess the likely range of values associated with options.

SOURCE J S Schwartz "Decision Support Systems and Their Potential Contribution to Consensus Development paper presented at "Workshop To Improve Group Judgment for Medical Practice and Technology Assessment, " sponsored by the Institute of Medicine, Division of Health Care Services, Council on Health Care Technology, Washington, DC, May 15-16, 1990

medicine, epidemiology, biostatistics, economics, psychology, and decision sciences. Many times, data required to model decisions are not available. Also, models only inform decisions; they are not definitive.

There have not been many scientific evaluations of the impact of using decision analysis in group judgments. The limited experience that does exist suggests that for decision support systerns to work, the group must be receptive to the concept. The technique is unfamiliar to many, so it must be taught to potential users (681).

Decisionmaking Rules and Procedures

A number of methods to combine the opinions of individuals in a decisionmaking group are avail-

able, some of them employing sophisticated mathematics and weighting schemes. These different methods may give widely different answers for certain questions (942).

Some guideline panels require the consensus of group members. while others allow for a range of dissenting opinion. Consensus does not necessarily mean unanimous agreement. In fact, it can be taken to mean group solidarity in sentiment and belief. a general agreement, or the judgment arrived at by most of those concerned **(855,899).** Most of the guideline processes reviewed for this report use informal consensus methods to arrive at recommendations. Groups consider evidence and usually iron out differences in roundtable discussions, but sometimes vote when there is disagreement. Few groups require unanimous approval of the guideline, and somc have established mechanisms to include dissenting opinions into guideline reports (e.g., ACP, HCHP).

Some observers suggest that requiring unanimity may result in recommendations that represent the 'lowest common denominator" of opinion. Instead, levels of agreement or disagreement can be establishcd according to votes taken during the group process. This provides a mechanism to voice disagreement without endangering the overall group process (123,51 ()). Voting can be either anonymous or public. If anonymous, those who are in the minority have some protection from undue pressure to change their position. Public votes may allow the group to focus on the problems that remain to be resolved (or that cannot be resolved) and force dissenters to defend their positions. Voting can be done On a simple yes/no basis or on a scale that reflects the level of agreement or disagreement. Using a scale allows panelists more latitude in expressing their opinion and can be used to qualify any recommendations according to strength of opinion (5 10). Dissenting opinions can also be included in the final report (123).

The RAND process allows participants to rate appropriateness indications anonymously. Some fear that the RAND process may lead to conclusions that diverge from the medical literature because of the nature of group process. For example, the reduction in disagreement over Delphi iterations could be the result of well-known psychological pressures toward conformity in groups, or a methodological artifact resulting from statistical regression to the mean (949).

However, some evidence suggests that conflict resolution in groups is determined more by the availability of research evidence than by the personalities and predilections of panel members (469). There arc limits to the extent to which agreement can be reached when good evidence is lacking. Nearly three-quarters of conflicts were resolved during a consensus process when good data were available, while only about one-quarter of conflicts were resolved when good data were not available (469).

The outcome of group processes can be enhanced if sources of disagreement are identified and discussed (5 10). It is informative, for example, to know whether disagreement stems from some panelists' concerns about a poor health outcome or from perceived unfavorable patient attitudes toward the intervention under consideration.

Not surprising] y, how agreement is defined can greatly affect a panel level of agreement. RAND assessments of appropriateness for coronary angiography, for example. ranged from 31 to 63 percent depending on how agreement was defined, and whether some panelists opinions that represent extremes were discarded in the final judgment (586).[9,10]

[9] The purpose of discarding opinion outliers is to better represent the group view. It is not necessarily the case, however, that the outlier is wrong (684).

[10] There was agreement for 31 percent of coronary angiography cases when all nine of the ratings fell within any three-point range. There was agreement for 63 percent of the cases when after discarding one extreme high and one extreme low rating, the remaining seven ratings all fell within any three-point range (586).

The Basis for Guideline Recommendations

Professional judgment is used throughout the guideline development process—from reviewing and interpreting key evidence to discussing personal opinions and experience and formulating recommendations. There is, however, great variation in the extent to which expert opinion or judgment is used as the basis of guideline recommendations. At the extreme is NCI, which has recently decided against issuing recommendations at all. Panels apply their judgment in evaluating and summarizing the available literature, but conclusions are limited to scientific statements that do not explicitly promote particular clinical policies.

The prevention guidelines issued by the U.S. Preventive Services Task Force in 1989 set a benchmark in the use of evidence to support guideline recommendations (871). Unlike any previous U.S. guideline efforts, the task force prefaced their work with the development of an explicit approach to selecting and evaluating the existing literature, as described above (see box 7-2). Further, their guidelines graded the strength of each recommendation according to the strength of the evidence supporting it. The USPSTF was the first major officially sanctioned group to produce practice guidelines linked directly to evidence, and its efforts were fundamental in establishing their practicality and acceptance. ACP's CEAP program also produces evidence-based guidelines.

Like the Preventive Services Task Force guidelines, AHCPR's guidelines effort was established with the intent of applying an explicit, systematic approach to the selection and evaluation of evidence regarding the effectiveness of managing a spectrum of health conditions (812). Unlike the USPSTF effort, however, AHCPR has a separate guideline panel for each condition selected for assessment, and different panels have interpreted and carried out this task in different ways (678). Where there has not been a strong evidence base on a topic, panels have sometimes made recommendations primarily on the basis of clinical opinion. The panel for the AHCPR guideline on pressure ulcers conducted a systematic review of the literature, found few quality studies, and so based guideline recommendations on the expert opinion of the panel.

The NIH Consensus Development Conference Statements are based on the consensus of opinion of a group of nationally recognized experts who consider evidence presented to them over the course of a three-day meeting. The panelists review clinical evidence and use their best judgment to make recommendations. There is no attempt to link recommendations explicitly to a particular source (either the oral presentations or published literature). Similarly, physicians rating the appropriateness of interventions at RAND use a combination of scientific evidence and expert judgment. NHLBI's cholesterol guidelines rely heavily on indirect laboratory evidence and expert opinion in addition to considering evidence from clinical studies.

Those supporting a role for expert opinion in forming guideline recommendations argue that where good evidence does not exist, the best judgment of experts is at least better than no guidance for clinicians at all. Advocates of this position also point out that there are many circumstances where decisions have to be made in the absence of good evidence. For example, HCFA has had to make many decisions regarding coverage of specific technologies under Medicare on the basis of expert opinion alone (188). In contrast, those supporting a strong evidence-based approach argue that where good information is lacking, there is no sound basis for creating or promoting a position. Without the benefit of strong scientific evidence, groups of experts can come up with very different recommendations, making the guidelines unreliable, One researcher, for example, identified 21 guidelines on asthma management which varied greatly in content (277).11 Indeed, guide-

[11] The guidelines were identified through the National Library of Medicine's MEDLINE® database covering the period 1989 to mid-1993.

lines can differ even when some evidence exists. Past experience with guidelines+. g., guidelines on screening for colorectal cancer (box 7-5)-confirms that conflicting recommendations frequently occur.

For some interventions, it may not be possible to produce good evidence quickly. For many preventive interventions for chronic disease, for example, the length of time between intervention and potential outcome is so long as to make randomized controlled trial results unavailable to policymakers for many years. Studies of how best to diagnose and treat some acute conditions, however, can often be conducted relatively quick] y. In a recent example. decision rules for the use of radiography in acute ankle injuries were developed and then assessed in a controlled trial. Investigators were able to link implementation of the decision rules to a decrease in use of ankle radiography, waiting times, and costs without patient dissatisfaction or missed fractures (502,723).

Even when good evidence from randomized controlled trials is available, clinician judgment is needed to interpret the results. For example, often trials may be confined to a limited group of patients (e.g., middle-aged males) and clinicians need to assess whether the trial results could safely be generalized to other groups of patients (elderly people, women, etc.) (378). And just because relevant evidence from clinical trials is available does not mean that the interpretation of results is always clear cut. There are many instances of contradictory results from multiple randomized clinical trials on the same topic. Sometimes these differences reflect deficiencies in design (e.g., they were too small), but in other cases conclusions from well-designed trials on the same topic differ because of variations in patients or therapies across trials (e.g., trial entry criteria or treatment of placebo groups may differ somewhat across trials) (355). Thus, even when several trials are available. there is often room for variations in opinion on the interpretation of evidence. For example, vigorous debate over whether mammography is indicated for women under age 50 followed the recent pub] i cation of results of a large randomized controlled trial designed to assess the effectiveness of mammography in reducing breast cancer mortality (37, 109,255,696).

Notwithstanding the fact that interpreting even good evidence is itself a matter of judgment, studies of group processes show that agreement among panel members is easier to reach when good evidence is available (469). Panel ratings of appropriateness of indications for Cesarean birth were much more likely to be in agreement for indications for which there was good evidence (e.g., randomized controlled trials or other prospective studies) than for those for which evidence was lacking or of poor quality (469). Likewise. groups are more able to make clear and precise recommendations when good evidence is available and agreed upon (469).

There are limits to the extent to which agreement can be reached when good evidence is not available. For example. when good evidence on the appropriateness of cardiovascular procedures was available, physicians in the United Kingdom and in the United States generally agreed, but the U.K. panel produced consistently lower ratings for many indications when evidence was unavailable. The U.K. panelists also seemed to require a higher standard of scientific evidence than did their U.S. counterparts (85).

How clinicians draw conclusions about appropriateness in the absence of evidence is very unclear, In one study, physicians rating the appropriateness of endarterectomy appeared to base their judgments mom on a patient's risk status prior to surgery than on their assessment of how the procedure would change outcome (i.e., probability of death or stroke). For example, even though a panel of expert clinicians assessed six indications for endarterectomy as inappropriate, six of eight panelists believed the procedure reduced the likelihood of adverse outcomes for these indications (500). This seeming contradiction is perhaps explained by the apparent importance of surgical risk in rating appropriateness. Patients at high surgical risk were often assessed as inappropriate candidates for the procedure in question.

BOX 7-5: Conflicts and Issues in Federally Sponsored Guidelines for Colorectal Cancer Screening

Several federal agencies sponsor the development of clinical practice guidelines for preventive services, and two have produced guidelines relating to colorectal cancer (CRC) screening The recommendations in these two guidelines differ from each other and serve as an illustration of the potential for contradiction among multiple guidelines. They also differ from the privately sponsored guidelines issued by philanthropic groups and by physician-specialty societies The various groups Issuing guidelines on this topic, and some of the differences among them, include

- the U S Preventive Services Task Force (USPSTF), which has declined to recommend either for or against periodic screening with either fecal occult blood testing (FOBT) or sigmoidoscopy in average-risk individuals,
- the National Cancer Institute (NCI), which has recommended an annual FOBT and a sigmoidoscopy every three to five years starting at age 50, with no suggested age at which to discontinue screening,[1]
- The American College of Physicians, whose recommendation is similar to that of NCI,
- the American Cancer Society which likewise recommends frequent FOBT and sigmoidoscopy but in addition recommends an annual digital rectal exam after age 40 and two initial sigmoidoscopies one year apart at age 50, and
- the American Society for Gastrointestinal Endoscopy and the American Gastroenterological Association, which endorse both FOBT and sigmoidoscopy screening beginning at age 50, but have not provided a recommended frequency.

The differences among groups in recommendations regarding CRC screening for average-risk people reflect two facts First, the evidence on the effectiveness of specific technologies is inadequate in many areas Second, the criteria (either implicit or explicit) for judging the evidence that does exist differ among expert groups.

At issue is whether a screening test for CRC must be shown to reduce CRC incidence or mortality in order to be considered effective, or whether demonstrating a shift in the distribution of

[1] NC I recently ceased issuing or endorsing recommendations for cancer screening and limits Its statements to reviews and interpretations of the evidence (see chapter text and appendix C)

even though clinicians, when asked to assign outcome probabilities to such patients, indicate that they would likely improve with surgery (500).

Organization and Administration of Guideline Activities

Almost all federal guideline activities described here are sponsored by the government rather than being developed internally. This distancing from the sponsoring agency provides a measure of independence. The guidelines of AHCPR and NIH, for example, do not need to be formally approved by the sponsoring agency. [12] The sponsoring agencies do, however, play important roles throughout the guidelines process, for example, in selecting topics and panelists, specifying methodology, and providing administrative and technical support to the panel. CDC guidelines published in *Morbidity*

[12] AHCPR reserves the right to publish AHCPR-sponsored guidelines and seeks both agency and departmental clearance (53).

BOX 7-5 continued: Conflicts and Issues in Federally Sponsored Guidelines for Colorectal Cancer Screening

detected cancers to earlier stages IS sufficient for considering a screening regimen effective Those who require direct evidence that CRC screening wiII reduce the Incidence of, or mortality from, CRC have found the existing evidence inadequate, The critics also point out that screening and diagnostic followup have medical risks and high costs Others focus on the heavy burden of illness and death brought about by CRC and conclude that even indirect evidence that screening may alter the course of disease in a substantial proportion of people screened cannot be ignored

The controversy around guidelines for CRC screening is likely to continue for some time The USPSTF IS currently updating its recommendations, including those for CRC screening At the same time, the Agency for Health Care Policy and Research (AHCPR) has recently begun sponsoring the development of its own guidelines on the topic The agency awarded the contract to develop the guideline to the American Gastroenterological Association (AGA) There is no formal mechanism for coordinating the USPSTF and AHCPR efforts, and history suggests that there IS considerable potential for conflicting recommendations between the two forthcoming sets of recommendations

The potential for conflicting recommendations is heightened by the fact that the panels creating the new CRC guidelines are likely to be quite different and to operate in different ways The USPSTF includes no gastroenterological specialists on its panel, and as described in the text it follows a rigorously structured process of considering evidence and developing recommendations In contrast, the AHCPR-sponsored panel is not required to follow any equivalent development process under its contract The contract does specify some of the procedures the AGA must follow in appointing panelists (e g , panel members must represent consumers as well as a variety of health care professionals, and AHCPR wiII review the proposed panelists for "potential conflicts of interest, but the contractor has considerable leeway in deciding exactly who will be on the panel Panel composition and the contract award may be an issue in future debates about the panel's recommendations, particularly if those recommendations differ from the recommendations of other publicly and privately sponsored groups

SOURCES Office of Technology Assessment *Costs and Effectiveness of Colorectal Cancer Screening in the Elderly* (Washington DC U S Government Printing Off Ice September 1990) U S Department of Health and Human Services Public Health Service Agency for Health Care Policy and Research contract no 282-94-2023 awarded to the American Gastroenterological Association May 25 1994

and Mortality Weekly Report, however, are ap - proved by CDC and represent government policy (751).

Some guideline panels are, in effect, standing committees that assume a long-term commitment to a particular topic. The USPSTF, NCI, and the CDC's ACIP panels follow this format. NHLBI activities are overseen by standing "Coordinating Committees." Each committee is charged with staying abreast of scientific developments and monitoring health education needs in its particular area. The committee can initiate a variety of activities from creating health education brochures to establishing panels to develop clinical guidelines. Standing committees have the advantagc of being able to keep abreast of the 1iterature on a given topic after a guideline has been published to decide when the guideline needs to be updated. The NCI has a formal process to continually monitor and update the information statements on the PDQ

computerized database. This medium has advantages over guidelines that may become out-of-date soon after they are published.

Methodological Research

Several groups have cited a need for further rew arch on the processes that underlie guideline development (37 1,376,607), but relatively few studies have judged the quality of guideline development processes. of the studies available, none compare the relative merits of one method to another.

Instead. research to date has focused on individual approaches—in particular, the RAND appropriateness method. As discussed above, research from this source on panel methods has illuminated the importance of such characteristics as panel membership. de fin it ion of agreement, and availability of evidence on the reliability of guideline results. RAND researchers have also found individual physicians able to be consistent across time in their recommendations. Physicians who rated appropriateness were able to reliability reproduce their ratings six to eight months after the completion of the RAND process (526).

In another interesting experiment, performed at HCHP, three panels of primary care internists were provided with identical literature summaries on the management of two common clinical problems: acute sinusitis and dyspepsia. Each panel used formal group processes to create clinical algorithms. Five of the six algorithms produced by the panels were similar, but one was substantially different. The authors concluded that "even with optima] literature support and a standardized consensus process, physician consensus groups may still produce guidelines that vary due to differences in interpretation of evidence and physician experience." Evidence available to the panels included a wide range of studies, with varying degrees of epidemiologic rigor and some conflicting results. The topic about which the panels' disagreement was greatest was not addressed in the available literature (594).

Guideline development methods used by the federal government have not been formally assessed for reliability y or validit y. The NIH Consensus Development Conferences have been evaluated only for their effects on physician practices. CDC has created a database of CDC-developed guidelines and developed resource materials relating to decision and cost-effectiveness analyses but has not evaluated the methods of developing guidelines themselves.

AHCPR has begun assessing its process of developing guidelines, and it is sponsoring a study on optimal methods for prioritizing guideline topics (813). Investigators at RAND are evaluating differences between their appropriateness rating method and the guideline methods of AHCPR (689). In addition. RAND researchers are investigating (403 ,823):

- the use of meta-analysis in the literature review,
- the effect on appropriateness judgment of having the panel consider probabilities and utilities explicit] y,
- the effects of alternativc methods of panel composition and function,
- methods to evaluate service underuse,
- the reliability y and validity of panel ratings, and
- the relationship between patient outcomes and inappropriateness ratings (81 ,373).

CONCLUSIONS

The Link Between Methods and Recommendations

The methods used to develop clinical practice guidelines might be a relatively uninteresting topic if the guidelines issued by the various organizations were consistent and uniformly accepted as valid. As previous examples demonstrate, however. they are not. Recommendations by various groups conflict with each other, and the recommendations of one group on a topic often are not considered valid or acceptable by others. Differences among guidelines can cause confusion and may undermine the basic credibility of guidelines themselves (205).

Although the focus of various federal guideline efforts vary somewhat from each other, there is no overall federal guideline strategy or coordination,

and current efforts are fragmented. For example, AHCPR has issued guidelines on topics also covered in the National Institutes of Health's Consensus Development Program (e.g., urinary incontinence, pain management). The U.S. Preventivc Services Task Force, sponsored by the Office of Disease Prevention and Health Promotion. and the Centers for Disease Control and Prevention have both issued guidelines on immunization. Recommendations regarding cholesterol screening are issued by both the USPSTF and the National Heart, Lung, and Blood Institute. Government-sponsored efforts sometimes also overlap with private sector activities. The American Academy of Pediatrics, for example, also issues immunization recommendations, as does the ACP.

The diversity of organizations producing guidelines and the methods they use-even within single agencies, such as NIH—suggests that **the potential for unnecessary duplication and contradiction between guidelines, and inefficient cross-agency use of resources needed to produce guidelines, is high. Furthermore, recommendations of guidelines available on the same topic sometimes differ markedly.**

Guideline methods vary considerably across federal agencies, yet there have been few efforts to compare them and identify the relative strengths of competing approaches.

The limited research on guideline development processes generally suggests that the availability of strong, high-quality evidence improves the likelihood that panels of experts will agree on practice recommendations. Group composition and aspects of group process become increasingly important determinants of guideline recommendations as the availability of evidence declines.

How evidence is considered, how group discussions are managed, and how agreement is defined also appear to affect the decisions that groups make. Consistency of methods appears to improve the reproducibility of guidelines, but where evidence is lacking the differing judgments of panel members place limits on the ability to produce similar guidelines even when similar methods are used. In general, formal group process techniques seem to improve group performance, but this has not yet been verified in the context of clinical guideline development. More research is needed to identify the factors that affect group judgments when evidence is lacking.

The composition of a guideline panel appears to affect the scope of a guideline, the kinds of issues considered, the way in which panel members consider and weigh different types of evidence. guideline recommendations. and the credibility of the guideline. There are probably tradeoffs in the effects of panel composition regarding such issues as credibility by physicians vs. considering topics such as cost and patient preferences, but the implications of these tradeoffs have not been adequately explored.

Evidence-based clinical practice guidelines have proved workable and politically acceptable, **In fact, the theoretical strength of such guidelines is so compelling that it calls into question the usefulness of federally sponsored guidelines not based on an explicit review of evidence that considers, in some explicit and systematic fashion, the strength of that evidence.** Guidelines with less evidence basis may be justified for some purposes (e.g., guidance on the use of very new technologies), but those purposes should be carefully thought out.

The identification of outstanding clinically relevant questions for research is an important contribution of guidelines, and such recommendations could be highlighted to a greater extent.

Developing Methods To Prioritize Guideline Topics

Guidelines are most likely to be influential when sound evidence supports certain clinical practices, but clinicians are not following those practices because they lack information or are uncertain. **Priorities for guidelines could be set according to formal reviews of available guideline-related evidence** *and* **analyses of clinical practice.** If guideline topics are selected solely on the basis of practice variation (or some other indicator of prac -

titioner uncertainty), there will likely be many topics selected that have an insufficient pool of evidence with which to develop a clear and specific guideline. It is difficult for groups to formulate specific guidelines useful to practitioners when evidence is not available and without good evidence, panel judgments seems to be vulnerable to significant bias.

Most groups developing guidelines select a topic and then proceed to assess the available evidence. Criteria could be established to help determine when sufficient data are available to develop a guideline (377). Such criteria are generally not being used now. Some suggest that the 1984 NIH recommendations on lowering blood cholesterol may have been issued too early, before adequate information was available on which to make recommendations. Evaluation of the state of evidence prior to guideline development will be easier with the establishment of a NIH clinical trial database (see chapter 4).

If the intent of a guideline is to inform and possibly change physician behavior, data could be collected prior to guideline development to evaluate the state of medical practice, and clinician knowledge, attitudes, and beliefs. Such data could be used to assess whether guidance is needed, and if so, if it is likely to be adhered to by the targeted population. Had the state of current practice been assessed, an NIH Consensus Development Conference on treatment of primary breast cancer might not have been held. A major recommendation of the conference—that few Halsted radical mastectomies be done—was found in a subsequent review of medical practice to be moot-the procedure was being performed very infrequently (372,41 1).

Behavioral science techniques such as focus groups and surveys might be helpful in understanding the sources of variation in practice. Both focus groups and surveys have been conducted by the NIH Consensus Development Program, but they are used to evaluate the impact of their program and not to identify topics. The AMA DATTA process assesses expert physicians' opinions regarding the safety and effectiveness of technologies, which could provide useful information to guideline developers. The AMA survey of physicians regarding the use of Teflon™ injections to treat incontinence would have informed the AHCPR guideline on urinary incontinence (the AMA survey was conducted after the AHCPR guideline was published). Most physicians polled in 1992 by the AMA felt that Teflon injections were effective in certain circumstances, but most questioned the safety of the technique (407). (The FDA has approved the use of Teflon preparations for injection into vocal cords to treat paralytic dysphonia. The use of this paste for the treatment of urinary incontinence is considered to be an "off-label" indication.) The 1992 AHCPR guidelines assess the effectiveness of the procedure (it is listed as a treatment option) but not its safety, which seems to be of concern to clinicians (531).

Research Needs

It is important to establish which processes produce the most valid. reliable, and usable guidelines. At present the various guidelines approaches vary markedly in terms of resource use, yet there is no clear indication as to whether one method produces a guideline that is any better than another. It may be that some processes are particularly appropriate to certain purposes or under certain circumstances, but at present there is little evidence on which to tailor guideline efforts.

Some federal guideline efforts have no published description of their process, making it difficult to judge the basis and soundness of the guideline or to compare different approaches. The methods used to make guideline recommendations need to be explicit if any evaluation comparing them is to be made. Aspects of the process that need to be fully described within the guideline include how the topic was selected, how relevant evidence was identified and considered, how panel members were identified and selected, and how evidence and expert opinion were used in making recommendations.

A number of methods are used to make clinical practice recommendations. but there is insuffi-

cient research with which to judge what method works best (246). Guideline development panels arc now often quite small (generally 10 to 20 members), but take on variety of tasks. They need to conduct literature reviews, organizc and synthesis evidence, evaluate safety and effectiveness of alternative interventions. consider cost-effectiveness and othcr health policy considerations, and identify areas of needed research within the guideline topic. In some ways current guideline panels are inefficient. Given the range of responsi - bilities of a panel, it is difficult to provide the needed expertise on panels of limited size. The process is sometimes very slow and experience gained is often lost after the panel completes the guideline.

One alternative model to test would be to create standing expert teams to support guideline panels. Such teams could experiment and develop methods for activities that all guideline panels must do: literature review; assess current practices: consider the health policy implications of the guideline such as the cost impact of guidelines and the cost-effectiveness of alternative clinical strategies.

Existing group processes should be further developed and tested and contrasted with one another. Formal group processes seem to enhance group performance, and some cvidencc suggests that guideline panelists prefer structured approaches when the topic is controversial (689). In the area of guidelines, the RAND adaptation of the Delphi group process method is a tested applicable formal group process model. This model structures group interaction and formalizes the elicitation and combination of' member opinion, but the foundation of that opinion is not specified. There is no way to know the extent to which assessments are based on opinion or evidence. The RAND method includes a review of relevant literature, but appropriateness ratings are not linked to the evidence. One way to strengthen their approach would be to identify the source of each rater opinion. This would be difficult at present, given the large number of indications that need to be rated. How'ever. most indications that are identidied are theoretical and are rarely or never seen in practice, [13] One could therefore consider asking panels to provide a richer set of judgments on fewer indications.

Federal agencies are in a unique position to be able to assemble resources needed for guideline development. In addition to funding clinically relevant research to serve as the basis for guidelines, agencies could develop tools potentially useful to guideline developers. These might include:

- identification of areas of clinical uncertainty and its sources—national databases can be used to identify practice variation, and national clinician surveys could be conducted to assess sources of variation:
- methodologic reviews of the literature on topics of likely interest—teams of methodologists could review literature identified through systematic searches;
- development of methods to incorporate cost assessments and patient preferences into practice guidelines; and
- listings and evaluations of clinical trials.

The planned creation of two clinical trial databases. if successful, will facilitate the identification of evidence for guidelines panels. All published randomized controlled trials will be identifiable through a database being developed through the efforts of the National Library of Medicine and the Cochrane Collaboration (see chapter 4). In a separate effort, NIH is creating a database of NIH-sponsored ongoing trials related to women's health, and the possibility of an all-NIH database is also under consideration (chapter 4).

[13] As many as 60 percent of indications for angiography, 79 percent of indications for endoscopy, and 68 percent of indications for endarterectomy were never seen in reviews of patients' charts (586).