5. Synthesis, Cost-Effectiveness Analysis, and Decisionmaking

Truth is rarely pure, **and** never simple.

—Oscar *Wilde*

Contents

List of Tables

5. Synthesis, Cost-Effectiveness Analysis, and Decisionmaking

INTRODUCTION

This chapter is concerned with synthesizing research information and economic data in a systematic fashion so that results are useful for policy. This is not an easy task. There is an overwhelming amount of information on medical technologies (basic research, applied research, cost/outcome data), and policy analysts are faced with the problem of making sense of this information. This chapter describes and critiques methods that are intended to assist policy analysts in weighing information from diverse sources. Much of the discussion is based on material presented in appendix C and previous OTA reports (270,271).

The first section discusses methods for synthesizing research results from multiple studies which are generally concerned with safety and efficacy. Next, economic concepts are included in a presentation of the principles of cost-effectiveness analysis (CEA). This discussion is followed by a presentation of methods for soliciting group opinions and a brief discussion of quantitative decisionmaking techniques.

SYNTHESIZING RESEARCH RESULTS

Individual research studies, in themselves, do not constitute a technology assessment. An assessment must consider the evidence from a set of studies, evidence concerning social effects as well as safety and efficacy. Typically, the first step in conducting a medical technology assessment is a review and synthesis of available research evidence. Although issues pertaining to synthesis have been neglected in the health care literature (266,398), consideration of these issues is essential if the range of information available about particular medical technologies is to be utilized. Problems encountered in using traditional synthesis procedures such as the literature review are described below. Also discussed are various formal procedures to systematically synthesize research results.

The Literature Review: Unstructured Synthesis

The traditional approach to synthesizing research is the literature review. Typically, a reviewer selects a set of studies believed to be most relevant and summarizes the evidence. There are a number of problems in relying on literature reviews. First, such reviews tend to be subjective. Second, methodological problems in individual studies are often ignored, distorted, or obscured. Finally, there is the problem of timeliness: the reviews must be available prior to the dissemination of ineffective or unsafe treatments.

A number of biases may affect reviews of the research literature. The understandable enthusiasm which investigators have for any new treatment that potentially improves care can lead to errors. This problem may be somewhat diminished when reviews are prepared by groups (e. g., National Institutes of Health (NIH) panels) rather than by individual physicians or researchers,

Aspects of the problem of investigator bias are illustrated by the controversy over electronic fetal monitoring (EFM). Different perspectives apparently cause reviewers of the literature to attend to different aspects of the evidence (see app. C). The paucity of randomized clinical trials (RCTs) generally available (4oo) presents another impor-

98-144 O - 82 - 5

tant obstacle to synthesis, because well-controlled research methods are probably the best mechanism to guard against investigator bias. When such methods are employed, a reviewer can use methodological arguments to discuss disagreements between studies.

However, even when well-controlled trials are available, they may not fully answer questions about a technology. Tonsillectomy is a case in point. There exists a substantial literature on the safety and efficacy of tonsillectomies, including clinical trials and other types of research. According to Cochrane (6o), none of the three different clinical trials on tonsillectomies conducted in England during the 1960's resolved the policy controversy over the appropriate use of tonsillectomy. The available RCTs, he noted, had two methodological problems: 1) the treatment was compared with no or inadequate treatment (instead of an alternate treatment); and 2) the patients' parents were not blind to the conditions of the experiments, so those whose children were on the waiting list may have exaggerated their children's symptoms.

Structured Synthesis Procedures

Because of the problems inherent in literature reviews, efforts have been made to develop more systematic procedures to integrate and interpret sets of research evidence. These range from elementary classification procedures to sophisticated statistical techniques. A description of such techniques is presented below. Additional material is presented in appendix C.

Classification or Voting Method

A simple structured synthesis technique involves organizing a body of literature according to a prespecified set of criteria and is actually a classification procedure ***(226)*** Sometimes called the "voting method, " this synthesis technique involves selecting a particular sample of evaluative studies of a technology, coding some aspect of the design and/or conceptual framework, classifying observed outcomes as to whether they are favorable, neutral, or unfavorable (i.e., "taking a vote"), and then constructing tables of research findings. This method is frequently used to demonstrate the differences obtained by various methodological approaches used to assess the same technology.

The value of the voting method lies in the precise identification of the type of studies to be sampled, and the coding of substantive and methodological aspects of the studies according to clearly defined procedures. More widespread use of this classification technique could probably aid in determining the specific patient populations and/or conditions that can be effectively treated by a medical technology (i.e., to establish external validity). This technique helps to avoid the problems of more traditional literature reviews, noted above, which selectively describe research evidence and which neglect consideration of methodological strengths and weaknesses.

Krol ***(216),*** however, cites three problems with the method: sample size, effect size, and Simpson's paradox. The first problem is that large studies are likely to produce statistically more significant results than studies with small numbers of subjects. Thus, for example, a study finding of "no results" or "no difference" between treatment and control conditions may be correlated with small sample size. The second problem is that the voting method does not take effect size into account, i.e., small, marginal effects are not distinguished from large effects. Simpson's paradox is a more subtle statistical point. It is possible, under certain conditions, to reach conclusions by aggregating data from all studies that are different from the conclusions reached by counting each study separately. A fourth problem with the voting method is that it may oversimplify the results of studies and cause reviewers to overlook subtle interactions among variables ***(226).***

Meta-Analysis

A rigorous statistical approach to research synthesis is a sophisticated quantitative synthesis technique called meta-analysis (166,165). This technique uses the actual results of studies and permits the determination across a set of studies of the magnitude of treatment impact. Meta-analyses are useful in assessing treatments for which a large number of studies are available and findings across studies seem to have great variability.

As used by Glass (165), such analyses require that both comparison and treatment groups be available and that the original research reports contain appropriate statistical information such as the group mean and standard deviation. Effect sizes (ES) are calculated by determining the difference between the mean of the treatment group (T) and the mean of the comparison group (C), divided by the standard deviation of the comparison group (SD). Thus,

$$ES = \frac{T - C}{SD}$$

This procedure converts the average effect of each outcome measure into a common scale (i. e., standard deviations) that can be compared to *re*sults of other studies. If a treatment has no effect, then the effect size would be zero; if the treatment is effective (i. e., better than the current alternative), then the effect size is positive; and if the treatment is inefficacious, the effect score is negative. If some assumptions are made about the skewness of experimental and control group scores within each study and the distribution of effect sizes across a large number of studies (i.e., that they are normally distributed), effect sizes can be converted into percentile ranks and inferences can be made about the overall effects of a medical technology.

One recent example of the use of meta-analysis is Smith, Glass, and Miller's ***(353,354)*** review of the outcome literature on psychotherapy. These investigators searched the published literature and included within their analysis all available control group studies of the effectiveness of any form of psychotherapy. * For each study, the investigators coded an extensive number of variables, including methodological criteria such as the nature of patient assignment to condition (e.g., random v. matching), experimental mortality, and other threats to internal validity. Effect size scores were calculated for each principal outcome. A code was also developed for validity of the outcome measures.

'Drug studies were analyzed separately, and studies that did not involve the use of professional therapists (operationally defined as psychologists, psychiatrists, and social workers) were eliminated from the analysis.

Smith, Glass, and Miller's (354) findings indicated that, on average, the difference between scores of the groups receiving psychotherapy and the control groups was 0.85 standard deviation units. Assuming the normal distribution of effect size scores, this average standard score can be translated to indicate that a typical person who receives psychotherapy is better off than 80 percent of the persons who do not. The investigators also performed a number of analyses to determine whether the methodology of the outcomes study affected results and whether different therapies (or other factors) were differentially efficacious.

The work of Smith and colleagues (353) has been criticized, because it "lumps together" a large number of what some consider incomparable treatments and outcomes (e.g., 137). It should be noted, however, that the strength of the effect size technique is that it provides a common metric that permits analysis of these differences (methodological and substantive). Smith, et al. 's (354), classification variables for each study were fairly comprehensive and yielded a systematic comparison of studies on the basis of their conceptual and methodological designs. What is problematic, however, is that the findings are heavily dependent on a number of decisions that are not always made explicit. These include criteria for selecting literature and criteria for selecting variables. It is not possible to ascertain biases resulting from the investigators' sampling decision and whether only certain types of studies, therapies, or variables are assessed using control group designs ***(273).***

Other Synthesis Procedures

A number of other methods exist for statistically combining the results of independent studies (see 69,292,324). The effect size method described above actually incorporates several procedures. The most important of these is the comparison of treatments to detect interactions between the characteristics of a study and outcome (i.e., external validity). As noted in the discussion of the classification method, some of these procedures can be employed when effect scores are not computed.

Additional statistical methods can be used to combine probability values from various studies and to adjust outcome scores according to the rele-

vance of the data. Rosenthal (324) describes a number of procedures for combining probabilities. These range from adding observed probability levels across different studies to adding weighted standardized scores. They also include the testing of mean probability values. Use of such procedures indicates whether significant effects are obtained across a set of studies. The problem in using probability values is that the number of subjects per study influences the statistical power to detect whether significant overall differences are present.

An interesting method for synthesizing the results of experiments done with human and animal species according to the relevance of the data has been proposed by Du Mouchel and Harris (131). This method involves the sophisticated application of a statistical theorem (Bayes' theorem) to provide a quantitative prediction of data relevance. Du Mouchel and Harris use the procedure to provide estimated carcinogenic risk of various substances derived from the results of a series of epidemiological studies.

Advantages of Structured Synthesis Procedures

A number of advantages result from the use of formal procedures for data synthesis ***(292).*** One benefit is that formal syntheses help to identify contradictions in the literature by systematically organizing studies according to specified classification factors. Thus, differential outcomes can be segregated according to treatment chacteristics and/or methodological approaches.

The use of effect size scores offers a second benefit. Such scores provide insight as to the worth of a treatment and provide a benchmark for later research. Thus, for example, an analysis of ***23*** controlled studies of patient education programs by Posavac ***(294)*** found a ***0.75*** average effect size. According to Posavac, this should provide a standard against which new patient education programs can be assessed. A finding that the effect size of a new program is only ***0.20*** (providing that similar dependent measures are employed), would probably indicate that the program was not particularly effective, at least for the problem or population for whom it had been designed.

Another advantage of quantitative synthesis methods is that they serve to control for some statistical conclusion validity problems (e.g., power) that some commentators have reported as severe in the medical literature (e.g., 141,172,337). The widespread use of meta-analysis and other quantitative approaches to research synthesis would likely improve statistical reporting practices by calling attention to investigators' use of data. Furthermore, most quantitative analyses of multiple studies would compensate for errors in analyses such as the use of multiple-independent inferential tests without appropriate error rate control or incorrect inferences because of a lack of power. Although errors in data collection would not be corrected by quantitative synthesis methods, the systematic analysis of multiple studies should render the effect of such errors less consequential. The attention to systematic considerations of the "weight" of evidence across research studies should have a generally salutary effect.

These synthesis procedures seem most appropriate for evaluating more mature medical technologies about which there has accumulated a considerable body of research. Often however, they may be applicable to less developed technologies. In some cases, where only meager evidence is available from a small set of studies, the effectiveness of a medical technology maybe suggested by a review of specific components from some other portion of the literature.

COST-EFFECTIVENESS ANALYSIS

CEA can be thought of as an aid to synthesis of both the health effects and the economic effects of a medical technology. CEA was itself the topic of a recent OTA assessment (270,271,272, 273,274). OTA's findings from that study are summarized here.

The value of CEA lies more in the process of performing the analysis than in any numerical results. There are a number of reasons for this, among the most important of which are CEA'S inability to adequately address ethical issues and the uncertainty of many of the key variables re-

quired for the analysis. Addressing uncertainty is a topic addressed later in this chapter.

A second major finding from the OTA study was that there is no one "correct" way to do an analysis. Not only does each analysis differ in terms of which benefits and which costs must be considered, but each analysis differs in terms of how the benefits and costs are valued. A driving force behind these variations are the social/ethical concerns mentioned earlier.

OTA suggested that the most appropriate approach to any assessment is to perform it in an open forum so that assumptions and underlying values can be challenged; to identify, measure, and, to the extent possible, value all relevant benefits/effects and costs; and to present the results of the analysis as an "array" of benefits/effects and costs rather than forcing them into some aggregate single measure. By arraying effects in a systematic fashion, one can place the appropriate relative emphasis on given effects whether they are quantifiable or not. This technique is designed to make more explicit the health, economic, and social consequences of any decision.

In suggesting the "array" method, OTA recognized that CEA is a decision-assisting technique, rather than a decisionmaking one. In some instances, however, a cost per aggregated effect may be possible, appropriate, and quite acceptable. A case in point is OTA's own cost-effectiveness study on pneumococcal vaccine, which calculated a ratio of $1,000 per quality-adjusted year of life saved for the elderly (282). In that case, the study was performed under public scrutiny, and the analysis and assumptions were subjected to extensive outside review.

Finally, although OTA concluded that there was no single "correct" methodology for conducting CEA, it did find general agreement on 10 principles of analysis. Those principles, including a short explanation of each, are reproduced below.

1. Define Problem.—The problem should be clearly and explicitly defined and the relationship to health outcome or status should be stated.
2. *State Objectives.* —The objectives of the technology being assessed should be explicitly stated, and the analysis should address the degree to which the objectives are (expected to be) met.
3. *Identify Alternatives.* —Alternative means (technologies) to accomplish the objectives should be identified and subjected to analysis. When slightly different outcomes are involved, the effect this difference will have on the analysis should be examined.
4. *Analyze Benefits/Effects.* —All foreseeable benefits/effects (positive and negative outcomes) should be identified, and when possible, should be measured. Also, when possible, and if agreement can be reached, it may be helpful to value all benefits in common terms in order to make comparisons easier.
5. Analyze *Costs.* —All expected costs should be identified, and when possible, should be measured and valued in dollars.
6. *Differentiate Perspective of Analysis.* —When private or program benefits and costs differ from social benefits and costs (and if a private or program perspective is appropriate for the analysis), the differences should be identified.
7. *Perform Discounting.* —All future costs and benefits should be discounted to their present value.
8. *Analyze Uncertainties,* —Sensitivity analysis should be conducted. Key variables should be analyzed to determine the importance of their uncertainty to the results of the analysis. A range of possible values for each variable should be examined for effects on results.
9. *Address Ethical Issues.* —Ethical issues should be identified, discussed, and placed in appropriate perspective relative to the rest of the analysis and the objectives of the technology.
10. Discuss *Results. —The* results of the analysis should be discussed in terms of validity, sensitivity to changes in assumptions, and implications for policy *or* decisionmaking.

GROUP DECISION METHODS

Although the application of formal quantitative procedures for the integration of data from individual studies and the use of quantitative decision-assisting techniques should improve the process of technology assessment, such procedures cannot, by themselves, resolve policy controversies. These procedures cannot go beyond the data available on a particular problem, nor can they substitute for informed judgment or include societal values (see ch. 4).

A frequently used method for soliciting group opinions is the unstructured conference at which a given topic (or topics) is discussed. There are many conference formats, ranging from formal to informal presentations with or without questions and answers. Sometimes prepared critiques or presentations are employed; other times a debate format is used.

Another informal group opinion technique is the advisory panel approach such as that used by many Government agencies (including OTA). Sometimes, the panel is required to endorse a given study. In such cases, studies are often modified until most members are satisfied with the major points. Often a single member or small number of members of the panel can exercise de facto veto authority over certain elements of the study; other times a minority is included in the final report.

Four formal methods for resolving conflicts across research studies and for developing assessments of particular technologies are described below: 1) the Delphi technique; 2) the nominal group technique; 3) a new group opinion process, referred to as consensus development, sponsored by NIH; and 4) a computerized knowledge base being developed by the National Library of Medicine (NLM). The Delphi and nominal group techniques are based on behavioral science principles (163), the goal being to aid groups composed of individuals with different information and perspectives to develop group judgments that best take account of the positions of the individual members. The discussion below illustrates both the potential and limitations of these methods in synthesizing technology assessment information.

Delphi Technique

Delphi is probably the oldest structured model for involving groups in decisionmaking processes and has been used widely in health care ***(78).*** The Delphi technique uses a series of questionnaires (or individual interviews), each followed by anonymous feedback summarizing all the participants' responses. Although Delphi was originally developed by the Rand Corp. to synthesize expert opinions on national defense problems, it has since been extended to medical problems ***(232,246,250, 318,336).***

A unique feature of the Delphi technique is that persons selected to participate in the process generally have no direct contact with one another. Instead, participants are provided with a summary of the questionnaire responses, usually by mail. Personality or status variables thus have little chance to influence participants' opinions, as they might in face-to-face meetings. By using anonymous feedback, each participant has an equal chance of influencing other participants (41). The technique also provides a framework within which to approach a problem in a focused manner. Finally, and perhaps most importantly, the Delphi technique provides a limited time frame in which to achieve consensus (41,135). There are a fixed number of iterations, usually three, in the questionnaire-feedback process. (For a discussion of a modified Delphi technique, see ref. ***293.)***

Nominal Group Technique

Although there are several variations of this technique, typically, all participants are seated at a common table and asked to write their views on each of a number of issues posed by the leader of the meeting. Delbecq and colleagues (82) call this the "nominal" group technique, because the individuals at the table (at the outset) are a group in name only. Each person's view is recorded on a separate card, and talking is prohibited. The (silent) presence of others while writing the cards is supposed to stimulate participants to perform better. The cards are collected, and their contents

(but not the authors' names) are listed for all to see. Subsequent discussions focus on the ideas that have been proposed, not on who did the proposing.

Consensus Development

In response to congressional pressure to assist in the transfer of technology, NIH initiated its consensus development program in 1977. Its goal is to bring together various concerned parties—physicians, consumers, bioethicists, etc.—to seek consensus on the safety, efficacy, and appropriate conditions for use of various medical technologies. Judgments about the technology under consideration are intended to be based on the available scientific evidence. At conferences jointly sponsored with the National Center for Health Care Technology (NCHCT), information about the technology's social, ethical, economic, and legal impacts was included, as well. * The consensus development process is designed to produce a written document, called a "consensus statement, " that can be accepted by clinicians, researchers, and the public. The statement is supposed to identify both what is known and not known about the technology ***(287).*** (A list of NIH consensus development meetings from September 1977 through October ***1982*** appears in table 1.)

The consensus development conferences are coordinated by NIH's Office for Medical Applications of Research (OMAR). The topics are selected by the relevant institutes, but OMAR helps to make the final decision in cooperation with the bureaus, institutes, and divisions of NIH about the suitability of the topic, panel composition, and the proposed format for a consensus exercise. Adversary groups and task forces have been almost entirely abandoned. Moreover, the questions that have been posed to the conferences have been addressed strictly to those issues on which there is enough factual evidence to reach agreement. Unlike the techniques discussed above, consensus conferences have no particular theoretical bases for their format.

A panel of experts is selected by NIH to hear presentations by the leading medical researchers addressing a prespecified set of questions about the technology. These presentations, usually summarizing the latest research findings, are made over a 2-day period during which both panelists and audience members discuss the research findings. On the evening of the second day, the panel is requested to draft a statement responding to the questions. Usually, the panel deliberates through much of the night, often writing four or more drafts of its consensus statement. In some rare cases, minority reports are developed to indicate disagreement with the majority recommendations. The next morning, the consensus statement is read to the audience for their comments and criticisms. The conference concludes with a press conference. After the panel disperses, it sets about the final task of revising the statement. The statements from consensus development conferences are widely disseminated to thousands of organizations and individuals by NIH and by publication arrangements with leading medical journals such as the New *England Journal of Medicine.*

Knowledge Base Development

In response to the often overwhelming number of articles and other information concerning a particular topic, NLM's Lister Hill National Center for Biomedical Communications developed a unique system for soliciting expert opinion. NLM's system, termed the "Hepatitis Knowledge Base, " is intended to function as a continually updated, computerized body of knowledge concerning viral hepatitis (129). Although this system's topic is a disease rather than a technology, its approach to soliciting and maintaining expert opinion could also be applied to the latter.

Information for the Hepatitis Knowledge Base was originally assembled by a small body of experts who reviewed and combined the critical information from an identified set of ***40*** important review articles on the subject. Any inconsistency in the information was addressed by these experts; they either reached a consensus or noted that there was an unresolvable conflict.

The resulting base of "knowledge" was computerized and made immediately accessible to each expert within this group. Subsequent information which appears in the literature is reviewed

● NCHCT, which was established in 1978, was not funded in 1982.

Table I.—NIH Consensus Development Meetings, September 1977 Through October 1982

Sponsors	Title	Dates held
NCI	Breast Cancer screening	Sept. 14-15, 1977
NCI	Educational Needs of Physicians and the Public Regarding Asbestos Exposure	May 22, 1978
NIDR	Dental Implants Benefit and Risk	June 13-14, 1978
NCI	Mass Screening for Colo-Rectal Cancer	June 26-28, 1978
NIA	Treatable Brain Diseases in the Elderly	July 10-11, 1978
NINCDS	Indications of Tonsillectomy and Adenoidectomy: Phase I	July 20, 1978
NIAID	Availability of Insect Sting Kits to Non physicians	Sept. 14, 1978
NCI	Mass Screening for Lung Cancer	Sept. 18-20, 1978
NIGMS	Supportive Therapy in Burn Care	Nov. 10-11, 1978
NIAMDD	Surgical Treatment of Morbid Obesity	Dec. 4-5, 1978
Interagency Committee on New Therapies for Pain and Discomfort (Organizer)	Pain, Discomfort, and Humanitarian Care	Feb. 16, 1979
NICHD	Antenatal Diagnosis	Mar. 5-7, 1979
NHLBI	Transfusion Therapy in Pregnant Sickle Ceil Disease Patients	Apr. 23-24, 1979
NHLBI	Improving Clinical and Consumer Use of Blood Pressure Measuring Devices	Apr. 26-27, 1979
NCI	The Treatment of Primary Breast Cancer: Management of Local Disease	June 5, 1979
NCI	Steroid Receptors in Breast Cancer	June 27-29, 1979
NEI	Intraocular Lens Implantation[a]	Sept. 10-11, 1979
NIA	Estrogen Use and Post-Menopausal Women	Sept. 13-14, 1979
NIAID	Amantadine: Does It Have a Role in the Prevention and Treatment of Influenza?	Oct. 15-16, 1979
DRS	The Use of Microprocessor-Based "intelligent" Machines in Patient Care	Oct. 17-19, 1979
NIDR	Removal of Third Molars	Nov. 28-30, 1979
NHLBI	Thrombolytic Therapy in Thrombosis	Apr. 10-11, 1980
NINCDS	Febrile Seizures	May 19-21, 1980
NCI	Adjuvant Chemotherapy of Breast Cancer	July 14-16, 1980
NCI, NIA, NICHD[a]	Cervical Cancer Screening: The Pap Smear	July 23-25, 1980
NIAMDD[a]	Endoscopy in Upper GI Bleeding	Aug. 20-22, 1980
NICHD[a]	Childbirth by Cesarean Delivery	Sept. 22-23, 1980
NCI	CEA and Immunodiagnoses	Sept. 29-Oct. 1, 1980
NHLBI[a]	Coronary Artery Bypass Surgery: Scientific and Clinical Aspects	Dec. 3-5, 1980
NINCDS	The Diagnosis and Treatment of Reye's Syndrome	Mar. 2-4, 1981
NINCDS	CT Scanning of the Brain	Nov. *4-6, 1981*
NIAID	Defined Diets and Childhood Hyperactivity	Jan. 13-15, 1982
NIAID	Total Hip Joint Replacement	Mar. 1-3, 1982
NIAID	Immunology—The Bee Sting	Oct. 6-8. 1982[b]

[a] Jointly sponsored with the National Center for Health Care Technology.
[b] Planned date.

SOURCE: National Institutes of Health, Office for Medical Applications of Research, 1982.

individually by the panel, consensus is sought, and the knowledge base is updated. The panel can caucus through a technique known as "computer conferencing" without being together geographically or temporally. As time permits, each member can access the terminal o enter his or her comments and to review the comments of others.

To date, the Hepatitis Knowledge Base system for reaching consensus on a timely, complex biomedical topic is a research effort only. It has not been evaluated for either clinical utility or cost effectiveness.

An important outgrowth of the Hepatitis Knowledge Base research effort is the develop-

ment of the Knowledge Base Research Program. This experimental program, which currently includes viral hepatitis, peptic ulcer, and human genetics, may contribute to more effective access to and use of available biomedical information in solving the daily problems of diagnosis, prognosis, and treatment of illness. The program is exploring ways by which medical computer scientists and medical subject matter experts can select and organize relevant and accurate information from the biomedical literature.

Advantages and Disadvantages of Formal Group Process Methods

The methods discussed above are the better known and developed methods designed to facilitate group decision processes and make them more efficient. These methods are also intended to produce more reliable and valid information than an unstructured conference. Evidence of effectiveness, however, is somewhat contradictory for the Delphi and nominal group techniques, and sparse for the newer methods such as NIH's consensus development conferences and NLM's knowledge base.

The relative strengths and weaknesses of both the Delphi and nominal group process methods have been summarized by Delbecq, et al. ***(82).*** They believe that these methods are superior to simple, unstructured group interaction providing for a higher level of independent thinking both in terms of quantity and quality, especially with respect to specificity. Delphi seems to be particularly relevant for generating predictive information ***(78)*** when data are poor and for resolving highly controversial issues likely to be distorted when participants interact with one another.

Nevertheless, in comparison to formal decision analysis, the Delphi technique has been criticized as being little better than the "seat-of-the-pants" methods currently employed by policymakers and as being a method which bases "knowledge" on an informal set of opinions (332). Others (10) maintain that it is subject to the same total error as most predictions. The process is also time- and group-dependent, because the results are based on information available to a specific group of experts at a specific point in time. As a consequence, the process should be repeated as data change with time. It also appears less well suited as a process for resolving minimally controversial issues (318) or for synthesizing the state of the art in a given field (163).

A recent study compared Delphi with the nominal group process technique ***(368).*** Physicians were randomly assigned to one of three Delphi or nominal group technique panels to develop procedures for handling four hypothetical emergency medical services cases. In order to determine the reliability of the decisions, panelists were contacted individually ***6*** months later and asked to cast an anonymous vote on the procedures originally discussed. The degree of consensus achieved was the same for both techniques. The most striking finding, however, concerned the reliability of decisions over time. There were "very extensive" changes in the nominal group technique vote ***6*** months later, suggesting that it is "a less than reliable technique for reaching a consensus. " Although the physicians reported that they liked the nominal group technique much more than Delphi, group norms and pressures were developed using the nominal group technique that produced unstable or false consensus.

Virtually no critiques of NIH consensus conferences have been published to date. However, a major study is currently being funded by NIH to evaluate its process. A recent Institute of Medicine (IOM) publication notes that consensus conference results were reported to be particularly useful in health planning, quality assurance, and setting reimbursement ***(197).*** Although no current evidence is available regarding the usefulness and impact of the conference results, NIH is planning to fund an ambitious impact study. Wagner suggested that the success of any consensus format is dependent on the following considerations ***(197):***

1. the composition of the evaluation panel, especially participation by epidemiologists and biostatisticians;
2. the amount, quality, and comprehensiveness of available data;
3. the duration of the process (sufficient time for participants to synthesize information is critical; several meetings held during a longer

period are preferable to an intensive, one-time meeting); and
4. the resources for support.

Although highly structured, the NIH consensus format, is not designed, as the Delphi and nominal group techniques are, to limit group dynamic problems such as potential dominance by selected individuals or groups within the panel. From a methodological perspective, two aspects of the NIH consensus development process are of concern: its sensitivity to the limitations of the research evidence and the extent to which it considers a comprehensive and systematic review of the research literature. However, NIH is aware that its search for consensus resolution must be confined within the limits of the expertise and evidence assembled. In order to broaden the expertise and evidence to the maximum extent practicable, NIH strives for diverse, open, and balanced representation and participation at its consensus conferences. It also performs an exhaustive review of the research literature to compile relevant background evidence. A more complete discussion of NIH consensus meetings is presented in appendix C.

Relatively less is known about the value of the knowledge base approach. Nevertheless, the concept seems extremely interesting and potentially quite helpful to researchers and clinicians alike. One major concern is the cost of such a system.

DECISIONMAKING UNDER UNCERTAINTY

Elements of uncertainty abound throughout the process of assessment. Over the years, various techniques have been developed to assist in making rational decisions under such conditions. The discussion here is intended to introduce the reader to the more useful techniques for evaluating uncertainty within the field of medical technology assessment. It is not an exhaustive description of such methods, nor is it intended to provide the reader with a complete understanding of any one technique.

It is important to distinguish between two types of uncertainty: that which reflects the presence of random events and that which reflects a basic lack of knowledge. Random events occur according to a known probability distribution. For example, the flipping of a coin will result in heads roughly half of the time and tails the other half. Thus, although the result of a single flip of a coin remains uncertain, the probability distribution of heads and tails is known. By contrast, when uncertainty reflects a lack of knowledge, one not only does not know the probabilities of various outcomes' occurrence, one may not even know which types of outcomes can occur. Thus, administering a brand-new chemotherapeutic agent to a terminal cancer patient may have any of several therapeutic and toxic effects; conceivably, the latter might include side effects never previously recognized in cancer chemotherapy.

Uncertainty due to random events can be adequately handled by a variety of techniques. A common approach is statistical confidence limits. Another common method is decision analysis. * Possible courses of action (outcomes, etc.) are diagramed on a "decision tree. " Branches in the diagram are associated with known or imputed probabilities and "payoffs. " As a result, an outcome value (positive and negative) is associated with each pathway. Thus, analysts can trace plausible paths to determine the probability and expected value of each final outcome. (For a discussion of decision analysis in clinical decisionmaking, see ref. ***386.)***

In addition to decision analysis, a variety of computer simulation techniques allow analysts to model real-world phenomena and estimate their consequences over hypothetical periods of time. By manipulating all such models until outcomes mirror empirical findings, analysts may be able to acquire valuable insight into real-world processes. The potential usefulness of modeling techniques is great, but analysts and policymakers should always retain an awareness of the influence of underlying assumptions. Technical sophistication can mask tenuous assumptions, particularly for those individuals who lack familiarity with the analytical approaches.

*Decision analysis is also helpful for uncertainty due to lack of knowledge.

One general approach to handling uncertainty is called sensitivity analysis, which is a conceptually simple but powerful tool with which to address both random and nonrandom events. Actually a series of techniques, sensitivity analysis can test whether variations in assumptions affect the qualitative conclusions of an analysis. It is particularly *useful* in CEA. For instance, if an analyst assumes a discount rate of 2 percent and concludes that the program in question is desirable (i.e., its benefits exceed its costs), he or she can try discount rates of **O** to 4 percent to determine whether the program's basic desirability is a function of, or is sensitive to, the discount rate. Thus, sensitivity analysis can shed light on the importance of certain assumptions, especially as to whether an analysis is meaningful despite the presence of uncertainty.

Sensitivity analysis can produce four important results. First, it can demonstrate that a conclusion of a study substantially depends on a particular assumption, thereby suggesting that the overall analysis cannot be viewed as "definitive. " Second, it can demonstrate that a conclusion is "robust" with respect to a particular assumption (i.e., that violation of the assumption does not significantly affect the conclusion) and, hence, that the tenuousness of the assumption is not a source of concern. Third, it can establish a minimum or maximum value which a variable must have for a program to appear worthwhile. Finally, sensitivity analysis can identify issues (uncertainties) deserving of research attention.

CONCLUSION

The techniques for synthesizing research, CEA, soliciting group opinions, and decisionmaking all have the common goal of making sense out of a body of information.

In technology assessment, the objective of synthesizing the information gained from multiple research studies is to understand the cause and effect relationships from the use of a technology. As chapter 3 explained, research studies, no matter what type of design, never provide perfect information and often provide only insights into relationships. Most of the synthesis techniques described in this chapter were developed to provide analyst/researchers with systematic means to make those relationships clear. These techniques are primarily concerned with safety, efficacy, and effectiveness.

CEA can be regarded as adding at least one more dimension to the synthesizing techniques described above. Traditionally, CEA has been used to balance the safety /efficacy/effectiveness of a technology with its economic effects. However, a recent OTA report, The Implications *of* ***Cost-Effectiveness Analysis of Medical Technology (270),*** described how social values can also be incorporated into the analytical framework. The 10 principles of analysis, described by OTA, are intended to make CEA more policy-relevant than most applications of the technique are. But, as OTA concluded, CEA is a *decision-assistin* $_g$ rather than a decisionmaking tool.

A systematic approach to decisionmaking is to solicit group opinions, This approach can be used to synthesize diverse individual and group values with clinical research findings and economic effects. Several of the group process techniques described in this chapter were developed and refined by the social science community to maximize the benefits and minimize the liabilities of group dynamics. Evidence indicates that these formal techniques produce more and better information than unstructured conference-type sessions. To date, however, there have been few broad applications of these techniques.

One purpose of ***assessing*** medical technologies, is to produce information needed by policy makers. Policies that affect the development, diffusion, and use of medical technologies—in particular, drugs, medical devices, and medical and surgical procedures—are described in the next chapter.