RISK: Health, Safety & Environment (1990-2002)

Volume 4
Number 2 *RISK: Issues in Health & Safety*

Article 6

March 1993

Consensus Development at NIH: What Went Wrong

Itzhak Jacoby

Consensus Development at NIH: What Went Wrong?

Itzhak Jacoby*

Introduction

Shortly after the Science Court first received national attention, then-director of the National Institutes of Health (NIH) Dr. Donald F. Fredrickson conceived of a program of consensus development with it as a model.[1] Since then, the NIH program has sponsored more than 100 consensus development conferences to clarify issues involving application of medical technology to clinical practice. Having outlived the Science Court and other similar efforts, the NIH program represents the most visible federally-mediated medical technology assessment activity in existence. Its principles and procedures have been studied and emulated within the U.S. and by many other countries that support consensus development programs.[2] Nevertheless, weaknesses in the NIH program can be traced to early decisions not to adopt certain aspects of the original Science Court concept.

First, the NIH program has chosen not to use adversary procedures, particularly cross examination of expert witnesses, that could have assured a more orderly and thorough airing of the facts on both sides of the argument and produced stronger conclusions. Second, the NIH is both the sponsor and the recipient of the results of the consensus process, violating the Science Court's principle not to accept funding from an agency that is party to a policy dispute.[3] This principle was to

* Professor and Director, Division of Health Services Administration, Department of Preventive Medicine and Biometrics, Uniformed Services University of the Health Sciences. B.S. (Industrial Engineering), Israel Institute of Technology, M.S. and Ph.D., (Operations Research), Cornell University.

1 *See, e.g., Before the Health Subcomm. of the Senate Comm. on Labor and Public Welfare,* 94th Cong., 2d Sess. (1976). (testimony of Donald S. Fredrickson)

2 *See, e.g.,* J. Lomas et al., *The Role of Evidence in the Consensus Process: Results from A Canadian Consensus Exercise,* 259 JAMA 3001 (1988).

3 Task Force of the Presidential Advisory Group on Anticipated Advances in Science and Technology, *The Science Court Experiment: An Interim Report,* reprinted *infra,* at 179.

assure that considerations of the acceptability of the ruling to that agency would not contaminate the deliberations; that is, the participants should feel that there are "no strings attached." Third, in contrast to the primary aim of the Science Court model, the consensus process rarely debates a pending governmental policy decision in collaboration with another agency, even though doing so would help ensure the significance and timeliness of the results. NIH conferences instead focus on producing recommendations to guide decision making by practitioners, not external policy-makers, even though the guidance is often transformed de facto into policy by its application to reimbursement decisions. It is therefore somewhat ironic that the NIH selects topics and participants for its conferences without the broad participation of the very parties it aims to serve, contrary to what the Science Court architects prescribed. The NIH consensus development process therefore suffers from an insularity that raises questions about possible bias in its judgments and damages the effectiveness of its consensus program.

Evolution of the Process

The consensus development process matured in the 1980's, during an unprecedented proliferation of emerging medical technologies. There was an urgent need for knowledgeable, impartial scientists to scrutinize the safety and efficacy of new technologies and arrive at policy decisions regarding their readiness for adoption in clinical practice. Many of the emerging procedures promised great advances in diagnosis and treatment but were seen by some to remain too long in investigative circles. The consensus program therefore aimed initially to encourage the adoption of new procedures, provided they were found to be safe and effective and incorporated means for wide dissemination of the conference deliberations. As Fredrickson then informed the Senate Subcommittee on Health:[4]

> It seems clear that in the future, the NIH and the rest of the scientific community must assume more responsibility for the effect of research on the quality of the health care delivered. The need for accelerating the transfer of new technology across the "interface" between biomedical research and the health care community and systems is a major issue.

4 *Supra* note 1.

At the same time, concerns about the impact of expensive technologic innovations on escalating health care costs brought pressures from many sources, including Congress, to stem the flow of ineffective technologies and bring an end to outdated practices. Almost from the beginning, therefore, the program focused both on new technologies and reexamination of existing clinical practice; existing practices are considered when there is new scientific information concerning their use. Further, the evolution of the program was fostered by the enactment of prospective payment legislation for Medicare in 1983: For a technology to be reimbursable, it had to be established as safe and effective as well as cost-effective compared to alternatives. Submitting new technologies to careful assessment thus became a necessary component of the practice of medicine.[5]

After an initial five-year period of experimentation, NIH codified the format and procedure for its consensus conferences. These guidelines have remained stable over the past 10 years. Planning a conference takes from 12 to 15 months, during which the NIH selects a topic, forms a planning committee, invites participants, and analyzes supporting data. The planning committee, consisting of NIH staff, non-NIH experts in the field of discussion, and the prospective conference chair, defines the scope and focus of the conference by drafting specific questions related to the safety and efficacy of the technologies being assessed. The planning committee also selects the speakers and panelists. At the conference, speakers supply the evidence, which is evaluated by the consensus panel.

The panel responds to the questions posed by the planning committee in a consensus statement that reflects its assessment of the evidence. The 10 to 15 panel members include basic researchers, clinical practitioners, methodologists (epidemiologists and biostatisticians), and public representatives, none of whom are supposed to hold an advocacy position on the topic of the conference. Except for the conference chair, the groups — planning committee, speakers and panel — responsible for the final consensus product share no common membership.

The conferences begin with a day-and-a-half plenary session with public discussions among speakers, panelists, and members of the

[5] Fitzhugh Mullan & Itzhak Jacoby, *The Town Meeting for Technology: The Maturation of Consensus Conferences*, 254 JAMA 8 (1985).

audience. On the afternoon and evening of the second day, the panel drafts the consensus statement in executive session. The panel presents the draft statement to the audience on the morning of the third day and invites comment and discussion from the audience. The panel may later incorporate comments it receives into its final consensus statement. The conference concludes with a press conference and usually generates extensive coverage, including publication of the consensus statement in medical and specialty journals as well as summaries in the lay media.

The duration of deliberations, roles of participants, processes for selecting issues and participants and participants' roles all distinguish the NIH process from that proposed for the Science Court. These differences have had a major impact on the outcome of the NIH process.

Issue Selection

Both the Science Court and NIH consensus development models were designed to resolve a scientific controversy with policy implications. Both models therefore developed criteria for selecting from among numerous candidate issues. In the NIH process, topics proposed for evaluation involve a gap between knowledge and clinical practice. Relevance of issues is based upon their medical importance, which involves the number of people affected and cost implications. There must be a scientific debate about the use of the technology and a discrepancy between the available knowledge and its practical application in medicine. Most important, there must be an available base of scientific information about the technology to support an empirical discussion of its merits.

In the Science Court model, persons outside of government agencies would choose cases from among topics proposed by those agencies. This model prohibits, as previously mentioned, agencies from funding adjudication of their own policy issues. In the NIH program, the topic selection is performed internally. While the primary audience for the process is acknowledged by its own guidelines to be the community of medical practitioners, neither this community nor other organizations that depend upon the results of the consensus process for guidance are typically involved in topic selection. For example, the process rarely involves physician specialty societies or payer organizations which might help identify the most urgently needed consensus development issues. In fact, since NIH neither regulates science policy nor clinical

practice, the assessments being made more properly involve policy decisions of other agencies, like the Health Care Financing Administration, which reimburses providers and beneficiaries with government funds. As a result, NIH conferences often fail to match topic selection with the clinical community's need for guidance and the availability of evidence necessary for developing a consensus.[6]

The atypical, and controversial, 1983 conference on Liver Transplantation exemplifies the use of this process to arrive at a policy decision through debate on the evidence. The conference was stimulated by Medicare's need for guidance on coverage, assuring the timeliness of the assessment. The evidence was carefully analyzed and presented to the panelists well before the conference was held. Both sides of the argument were well represented, and the debate was focused. The conference attracted intense political and public attention. Medicare's decision to provide coverage for liver transplantation, which followed soon after the consensus statement was issued, can be linked directly to the recommendations of the conference. This felicitous joining of need and response has unfortunately been the exception.

Although the broad goal of the consensus development program is to facilitate the appropriate and timely application of biomedical research results to clinical practice, NIH tends to select topics circumscribed by the agency's biomedical research interests. For example, many of the conferences examine biomedical technologies like computerized axial tomographic (CAT) scanning, magnetic resonance imaging (MRI), endoscopy (the subject of two conferences), and the use of microprocessor-based "intelligent" machines in patient care. Such conferences have been criticized for not meriting the rigorous evaluation or expenditure associated with the program because they lack controversy and serve largely to promote the agency's desire to encourage the diffusion of emerging technology and justify further research.[7] These shortcomings contributed to producing the unremarkable findings of the 1984 conference on diagnostic ultrasound imaging in pregnancy. This conference can be contrasted with an earlier one on cesarean childbirth, sponsored with the active involvement of the

6 Jacqueline Kosecoff et al., *Effects of the National Institutes of Health Consensus Development Program on Physician Practice*, 258 JAMA 19 (1987).

7 Drummond Rennie, *Consensus Statements*, 304 NEW ENG. J. MED. 665 (1981).

clinical community, which tackled a controversial issue in the appropriate forum and produced valuable recommendations.

The program's unwillingness to assess the current state of clinical practice along with the current state of science when considering potential conference topics also contributes to poor timing of results. Since CAT scanning technology was already well diffused and its pattern of utilization established before the 1981 conference on CAT scanning was held, the results came too late to influence clinical practice. The outcome of consensus conferences on breast cancer held in 1979 and 1980, which considered uses for procedures — like radical mastectomy for management of local disease — that were already discredited by clinicians, also typified the problem of widespread preconference conformity to its recommendations.

As further indication of its detachment from policy concerns, NIH limits the scope of inquiry of its conferences to evaluation of safety and efficacy, even though these factors provide only part of the information needed by health care professionals, patients, third-party payers, and other decision makers. To be effective in improving health care practice, health technology assessments must also address relevant economic, social, and ethical consequences, such as cost, access, and quality of life. The results of the 1987 conference on MRI generated only limited interest because they focused on safety and efficacy instead of issues of cost-effectiveness that concerned policy-makers. Similarly, the 1984 conference on Limb-Sparing Treatment of Adult Soft-Tissue and Osteosarcomas lost much relevance by exploring issues related to preventing disease recurrence while resolutely avoiding discussion of an issue of great importance to patients — the impact of alternative treatment on the quality of life of amputees.

Selection of Advocates

In the Science Court model, issue selection was followed by identification of adversarial "case managers" for each side. This selection process would have involved broad advertisement for proposals to demonstrate the respondent's credentials to represent one side of the issue. The respondent was envisioned to be an interest group or any consortium of groups and individuals with the requisite expertise and constituency. The Science Court and/or the collaborating agency

were expected to participate in the selection of case managers.[8]

There is no comparable selection process or analogous case management function in the NIH consensus development program. The speakers invited to present at the consensus conferences come closest to functioning as the advocates of the Science Court. These speakers give expert testimony and may advocate strong and sometimes contrary points of view. The speakers, however, have no formal role to defend a given position in opposition to proponents of an opposing view. Nor do they question the statements of other speakers, as advocates of the Science Court concept would have done. Rather, the NIH program tends to invite speakers from the biomedical research community who present in a manner resembling a scientific meeting. Unfortunately, the program often fails to invite researchers and analysts with relevant data from other sources, such as clinical practice.

Selection of Judges and Referees

The Science Court method for selecting judges and referees included consultation with appropriate scientific societies and organizations to assure unusual scientific capability and no obvious connection to the disputed issue. As a further check, the case managers were to examine proposed names for prejudice. A referee, or chief judge advised by legal counsel, concerned with implementing procedures was also proposed to enable full control by the scientific community.

In the NIH process, the panel undertakes the role of the judges and the panel chair assumes the role of chief judge and referee. Panel members, although possessing collective expertise on the topic being discussed, are expected to arrive with an open mind and listen impartially to the scientific data presented by the speakers. These key appointments are made by the conference planning committee with minimal input from organizations or individuals interested in the consensus topic. In the absence of case managers or other systematic control of the presentation of evidence, the panel chair becomes a critical figure. The individual selected can control the objectivity of the process, questioning of speakers, comments from the audience, synthesis of data, and development of the consensus statement. The success of the process can therefore hinge on the prestige, knowledge, impartiality, leadership, and group process skills of the chair. This betrays several

8 *Interim Report, infra* at 182.

inadequacies of the consensus development process. In several conferences, the chairs were performing research on the topic being discussed or held an advocacy position on the issue and could not have been considered impartial.

Presentation of Facts, Challenges and Adversary Procedures

In the Science Court procedure, case managers individually were to reduce all aspects of the issue to statements of scientific fact which the referee or judges would have reviewed before allowing mutual scrutiny of the statements by case managers for the opposing view. The case managers either would have accepted or challenged one another's statements. Challenged statements first would have been subjected to a mediation process to produce consensus between the parties; if this fails, the statements would have become the focus of an adversary procedure. The intended challenge resolution procedure involved both oral and written presentations, in an effort to produce statements of the highest possible validity in the given time constraints. While some aspects of this process were negotiable, there was no question that the right of each case manager to cross examine the positions taken by his or her adversary would be preserved.

In the NIH consensus process, the speakers present in a collegial manner comparable to a scientific conference. There is little effort made to organize the data presented by the speakers or to align participants into opposing positions. In the 1985 meeting on adjuvant chemotherapy of breast cancer, the pro-chemotherapy viewpoint was well-represented while the opposing view was not. The 1984 conference on lowering blood cholesterol to prevent heart disease also promoted cholesterol lowering with inadequate representation of opposing views. The lack of confrontation damaged the credibility of the predictable recommendations. There is also no provision for speakers to challenge each other's findings or ask for clarification. Questions for the speakers come from the Panel or audience and vary in their degree of probing. Since there is never adequate time for audience members to comment or ask questions, the entire responsibility for evaluating and reconciling the sometimes conflicting data, as well as translating it into a consensus statement, falls to the panel. Depending upon the extent to which

speakers provide their presentations in advance or NIH staff provide the panel with a synthesis of the evidence before the meeting, much of this integration takes place at the conference.

There is an obvious risk that the speaker's manner of presentation or the availability of certain kinds of expertise on the panel will cause great variation in the results of the data evaluation. Also, the task may prove unworkable in the time available, causing the panel to compromise its evaluation of the evidence and make more subjective judgments. At best, this lack of precision in resolving evidentiary issues weakens the credibility of the statement. At worst, it may result in a statement which records the most commonly accepted issues — results on which achieving consensus is easy — while failing to include specific guidance on the controversial points the conference was held to resolve. The 1992 conference on laparoscopic cholecystectomy illustrates the latter outcome; the consensus statement endorses a procedure that is largely safe when performed by a competent surgeon but omits guidance to patients or policy-makers on selecting a competent surgeon during the current period of initial rapid diffusion. Nor did the statement address the complications associated with the procedure.

Strengthening the Consensus Development Process.

Comparing the Science Court model and the NIH consensus development process suggests several areas of potential improvement in the latter. The NIH program should consider broadening the focus of its consensus development process from practice guideline development and technology transfer to include specific federal issues of health policy. The program would also benefit from opening its issue selection process to members of the intended audience of the consensus findings, including national health authorities, industry, payers, and other agents to help ensure that consensus conferences concentrate on areas of practice most in need of change. These parties also should participate in identifying conference participants, particularly panelists, and contribute to the definition of questions to be answered by the conference.

The program should expand its scope beyond safety and efficacy to issues of cost and related economic concerns, quality of life, ethical and legal concerns, medical necessity, and others as appropriate to the topic being debated. For all these issues, special efforts need to be made to obtain the best information available and systematically synthesize and

order the data before the beginning of the consensus conference, borrowing from the goals of the Science Court process for presenting and challenging evidence. This synthesis could employ meta-analytic or decision theoretic modeling of available data.[9] The planning effort should result in an organized compilation of points to be addressed at the conference, far enough in advance of the conference to afford participants time for a thorough review. Such advance preparation would better equip panelists for their role in evaluating presentations and developing useful guidance. The sponsoring organization should provide for peer review of the consensus statement to ensure that the questions posed to the panel were adequately addressed and that the findings were reasonably supported by the evidence. These recommendations have been made to the NIH program by the Institute of Medicine and others in the past but have not been adopted.[10]

Conclusion

The time may have come to transfer some of the responsibility for health technology assessment to an agency with a broader role than the NIH in this area. One likely candidate would be the Agency for Health Care Policy and Research (AHCPR). Congress established the AHCPR in 1989 as a sister agency to the NIH in the Public Health Service. AHCPR's authorizing legislation directs the agency to "conduct and support specific assessments of health care technologies."[11] This legislation specifically charges the agency with considering not only safety, efficacy, and effectiveness, but, "as appropriate, cost-effectiveness, legal, social, and ethical implications, and appropriate uses of such technologies." Alternatively, other government or private organizations could fill this role; to a limited extent, some already do. Whether NIH expands its current narrowly focused technology assessment program alone or in collaboration with other organizations, or allows the baton to be passed, hopefully the principles of the Science Court experiment will influence future participants in this process to produce more effective guidance for health care decision making.

---

9 Itzhak Jacoby & Stephen G. Pauker, *Technology Assessment in Health Care: Group Process and Decision Theory*, 22 ISRAEL J. MED. SCI. 183 (1986).

10 INSTITUTE OF MEDICINE, CONSENSUS DEVELOPMENT AT THE NIH: IMPROVING THE PROGRAM (1990).

11 Omnibus Budget Reconciliation Act of 1989, P.L. 101-239.