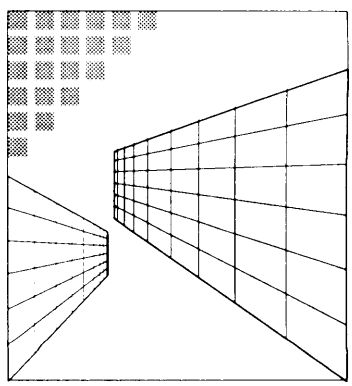
Behind the Search for Evidence 2

The present interest in evaluating the worth of health technologies and clinical practices owes much to two men whose works were separated by time and geography. One, Earnest A. Codman, was a Boston surgeon practicing in the early 1900s. He believed adamantly that the path to improvement in medical care depended on documenting the outcomes of patients who had been treated. Codman's call for detailed public reports of these outcomes, including long-term followup assessments of patients, faced strong opposition in his own day and was never adopted on a large scale (562). Nonetheless, his work set the stage for modern day efforts to focus on comparative patient health outcomes as a basis for improving the quality and effectiveness of care.

Six decades later and a continent away, Archie Cochrane, a physician and epidemiologist. changed the way researchers, poli - cymakers, and clinicians viewed medical care with the 1972 publication of his book *Effectiveness and Efficiency: Random Reflections on Health Services (* 130). In it, he argued that to provide the best health care at a given level of national health expenditures, society first must improve the effectiveness (eliminating ineffective care) and efficiency ("the optimum use of personnel and materials") of the health care system (130).



Cochrane pointed out that a major cause of ineffective care was that too much medical decisionmaking was based on poor evidence—''expert opinion" or, at best. observational studies that could not adequately differentiate effective from ineffective (or harmful) medical care. He advocated an emphasis on randomized controlled trials to evaluate medical interventions. Most importantly, however, he stressed that more valid information on the effectiveness and cost-effectiveness of health care interventions

BOX 2-1: The Tragedy of DES

Diethylstilbestrol (DES) is a dramatic example of a drug that became widely used in clinical practice before it was found to have major adverse effects An estimated three million American women took DES between 1948 and 1970 (728).

DES became popular in the early 1950s, after the publication of several studies that suggested that it was efficacious In treating placental insufficiency, a condition that often causes stillbirths (1 54, 279,595,653,707a,927). None of these studies were randomized controlled trials and none used double-blinding. Five other contemporary studies that did use double-blinding failed to show that DES improved pregnancy outcomes (1 45,1 75,233,642,838). Nonetheless, individual cases of women who had had previous stillbirths, and who were finally able to have children after taking DES, provided physicians with anecdotal evidence supporting the management of high-risk pregnancies using this drug Despite the absence of reliable evidence supporting the use of DES, U.S. clinicians began to prescribe the drug widely

In 1970, two researchers published a paper that reported a number of cases of a rare cancer in daughters of women who took DES (338) A second paper published the following year found maternal usage of DES to be strongly associated with the development of tumors in young women (340). In 1971, the FDA announced that DES was contraindicated for use in pregnant women By this time, however, several million men and women had already been exposed to DES in utero Numerous studies have since identified a range of adverse effects, including increased incidence of certain rare cancers in DES children, reproductive system anomalies in both sexes, and an increased incidence of negative pregnancy outcomes for DES-exposed women (61 ,62,337,339,685,886).

The tragedy of DES is not only that the drug proved to be harmful to the children of women who took it, but it was never really shown to be effective even for the condition for which it was so enthusiastically prescribed. Ironically, a reanalyses of data used in one of the first studies that purported to support the use of DES found the drug to be associated with an increase in "miscarriages, 'premature' deliveries and neonatal deaths" (77)

SOURCE Adapted from PA Goldstein H S Sacks, and T C Chalmers Hormone Administration for the Maintenance of Pregnancy, I Chalmers M Enkin, M Keirse (eds) *Effective Care in Pregnancy and Childbirth, Vol 7 Pregnancy* (New York NY Oxford University Press, 1992)

was crucial to improving both the quality and the efficiency of medical care.

THE NEED FOR EVIDENCE

The basic foundation of the evaluation of a health technology (or any health care intervention) is information about its efficacy and safety: whether, under at least some conditions, the technology provides a health benefit that outweighs any attendant risks (779). The evaluation of efficacy and safety is far from a theoretical concern. Experience with technologies such as diethylstilbestrol (DES), a cancer-causing drug prescribed to millions of pregnant women in the 1950s and 1960s, has taught that even the most enthusiastically adopted technologies can be not only ineffective but lethal (box 2-1).

The federal government has long had a role in evaluating the efficacy and safety of certain categories of medical technologies. Within the Department of Health and Human Services, for example. the National Institutes of Health (NIH) conducts and sponsors both basic biomedical research and clinical trials to test some of the most

promising technologies developed by its scientists. The Food and Drug Administration (FDA) regulates drugs, biologics, and medical devices, requiring manufacturers to provide evidence of safety and efficacy before their products can be marketed.[1] Other departments such as the Department of Veterans Affairs (VA) and the Department of Defense often sponsor both the development and testing of technologies intended to improve the health of the population in their charge.

For all of this regulation and testing, however, society's understanding of the full effects of most of the health technologies it uses is remarkably small. This state of affairs has four causes.

First, much of what medical care has to offer was part of customary practice before rigorous testing for efficacy became common. Randomized, controlled trials to demonstrate the efficacy of interventions have been openly advocated only since the 1940s, and they have been used widely only since the 1970s (784). Yet drugs to treat open-angle glaucoma, for example, have been prescribed since the 1800s (400). The first randomized controlled trials of the effect of a drug in preventing vision loss due to open-angle glaucoma were not undertaken until the 1980s (227, 416).[2]

Second, a high proportion of newly introduced technologies, even today, are not required to show rigorous evidence of efficacy before they are adopted. Only the most novel medical devices, for example, are subject to individual scrutiny and approval by the FDA before they can be marketed (370). Therapies such as psychological counseling and surgical procedures are subject to no regulatory requirements regarding efficacy at all (except to the extent that they involve drugs or devices that are regulated). Promising new procedures thus are often widely publicized and adopted by physicians and patients without undergoing any formal evaluation (box 2-2). The longstanding estimate that only about 10 to 20 percent of procedures have ever been formally evaluated for safety and efficacy (924) remains a rule of thumb (e.g., see reference 208).

The third reason is that a technology, once introduced, is frequently used in circumstances that are quite different from those in which it was first shown to be efficacious. The effects of the technologies under the new conditions can be very different as well. Drugs tested and approved for use for one type of cancer, for example, are frequently used to treat other cancers as well (881). Neither providers nor patients can be certain that a treatment used for a new population or in a new setting will actually have the same risks and benefits as those shown in the initial efficacy studies.

And fourth, as meager as society's knowledge of the health effects of many medical technologies is, our knowledge of their economic and social effects pales by comparison. In 1982, the Office of Technology Assessment (OTA) concluded that "No class of technologies is adequately evaluated for either cost-effectiveness or social and ethical implications" (783). Recent observers have suggested that this is still the case (606).

Thus, the deficits in evidence regarding the value of existing health care interventions are substantial. Nonetheless, in the two decades since the publication of Cochrane's seminal work, the

[1] FDA regulations actually specify that the agency consider safety and "effectiveness," but FDA's interpretation of "effectiveness" is more akin to "efficacy" as used in this report. The kinds of requirements that medical products must meet to satisfy this standard depend on the type of product. Drugs, and some medical devices considered to present a high possible risk of harm to users, must meet the most stringent requirements. Medical devices in the lowest risk category are required to demonstrate only such features as whether the manufacturer of the device met standards for good manufacturing practices.

[2] New antiglaucoma drugs seeking FDA approval (e.g., topical timolol) have had to undergo rigorous testing for some time, but such drugs have had to show only that they could reduce intraocular pressure. Although high intraocular pressure is strongly associated with open-angle glaucoma, until recently no rigorous studies had actually investigated whether reducing intraocular pressure through drug therapy protected patients from losing vision (786a). The National Eye Institute is currently funding a large multicenter trial to examine more specifically which patients with slightly raised intraocular pressures would benefit from the preventive application of antiglaucoma drugs (853).

BOX 2-2: Promising but Untested: Examples of Two Proposed New Surgical Therapies

Surgical innovations are especially likely to enter mainstream medical practice without ever being exposed to formal testing The lack of tradition among surgeons in testing new therapies through randomized trials, and the perceived difficulty in conducting such trials, may explain some of this phenomenon In addition, however, new surgical therapies often are incremental, have theoretical appeal, and are not subject to regulatory oversight. All of these characteristics make surgical Improvements difficult to Identify and study before they diffuse into the health care system,

A recent example of a surgical innovation is a technique to improve lung functioning in emphysema patients (9) Emphysema is a potentially fatal disease in which extensive damage to lung tissue (usually as a consequence of smoking) Impairs respiratory functioning,

The new technique revolves the surgical removal of 20 to 30 percent of a patient's lung A similar technique was Introduced in the 1950s but was rejected by the medical establishment on the grounds that the removal of lung tissue to treat symptoms (i e., shortness of breath) caused by tissue damage could only have a negative impact on patients. However, the newly refined procedure has been tried in 20 patients, all of whom have reportedly shown functional improvements as a result of surgery No randomized studies have been performed to confirm that the apparent short-term improvements are real, and the long-term effects of the procedure remain unknown, If the technique captures the interest of physicians and patients it may never undergo further testing before being adopted into clinical practice, since it is not subject to the safety or efficacy standards of any regulatory body

Another example of the kind of innovation that may never undergo rigorous evaluation is a potential new surgical procedure to preserve the salivary glands of head or neck cancer patients, These glands are often destroyed during radiation therapy (80), To avoid such damage, a researcher at the Tufts University School of Dental Medicine has proposed transplanting the glands temporarily to the patient's abdomen After the last radiation treatment, the glands could be re-transplanted into the mouth

So far, the procedure has been attempted only in animals The biggest challenge facing scientists is making sure that the glands can survive long enough in their temporary location to enable a full cycle of cancer therapy to be completed However, this problem may soon be solved, Because this new procedure has considerable theoretical appeal, it could well become an accepted strategy in cancer care based primarily on a demonstration of its feasibility

SOURCE Off Ice of Technology Assessment 1954 based on sources as shown Full citations are at the end of the report

movement to improve the assessment of the health, economic, and social effects of health care technologies has increasingly, though erratically, gained momentum. One result of this movement has been the growing accumulation of "research-based evidence" (705). That evidence, in turn, can be used to support judgments about the value of the myriad components of health care: "evidence-based medicine" (3 15).

A FRAMEWORK FOR EVALUATION

Improving medical care through increased knowledge about what works, and the applicat ion of that

knowledge, is a powerful concept. As support for the concept has increased, however, the language describing it has become increasingly muddled.

One of the most common phrases used to describe this effort is "outcomes research." The term originally arose to describe the line of health services research that has emphasized how little is often known about the effectiveness and outcomes of care that patients receive. This line of research, described in more detail below, ultimately led to the federal government's medical effectiveness initiative and the creation of the Agency for Health Care Policy and Research (AHCPR) to carry out this effort (Public Law 101 -239). The term has come to be used so sweepingly, however, that it has become problematic. For example, it is now often used synonymously with "outcomes-based management," a technique through which purchasers and providers hope to be able to manage the quality and cost of care provided to patients. This technique uses information on the outcomes of patients treated by a particular provider, or enrolled in a particular health plan, to stimulate actions that will improve care (box 2-3). The phrase "outcomes research" is rarely used in this report.

The convergence of terms has led to confusion among policy makers and the public alike between activities to improve the quality of care and those primarily aimed at identifying and improving its effectiveness. Although the concepts of quality and effectiveness are closely related—both are aimed at making health care "work" better—they are not identical. Activities to improve quality generally focus on improving the process by which an activity is performed, or the capabilities of those performing it, in order to improve outcomes. In contrast, research to investigate effectiveness focuses on what outcomes are associated with a given technology (or clinical management strategy, or any other health care intervention), and whether and under what circumstances that technology is better than alternatives. The relative effectiveness of a technology does indeed depend in part on how well providers perform it. Policy interventions to address problems in the quality of care, however, may be different from those interventions that address the overall effectiveness of care. The focus of this report is on the latter.

In this report, the phrase "effectiveness research" describes the category of research efforts aimed at identifying effective care and developing and refining methods to support the identification of effective care. The concept of effectiveness includes both whether the technology has a given effect and whether the technology is more effective than alternatives.

It is sometimes useful to make a conceptual distinction between efficacy and effectiveness. One generally wants to know whether a technology works at least under ideal circumstances (efficacy) before applying it more broadly (effectiveness).[3] In reality, however, the distinction between efficacy and effectiveness is often fuzzy. If the patient population in an initial efficacy study is sufficiently broad, for example, the study results may be credible evidence of effectiveness more generally. Conversely, a demonstration that a technology is generally effective in one population (e.g., women) does not necessarily imply effectiveness in a differently defined population (e.g., all adults).

Cost-effectiveness analyses are an increasingly common step in evaluating medical care. They use the results of effectiveness research, in conjunction with detailed cost information, as part of a structured, comparative evaluation of the relative costs and effects of two or more health care interventions.

Information on effectiveness and on cost-effectiveness, in turn, can form the basis of a health technology assessment: an analysis of a technology-related issue conducted for the purpose of

[3] For a more detailed discussion of usage of the terms "efficacy" and "effectiveness," *see the* OTA report, *Assessing the Efficacy and Safety of Medical Technologies (779).*

BOX 2-3: Using Patient Outcomes in Health Care Management

Stimulated in part by research emphasizing the final health outcomes of patients as an endpoint for assessing care, health care payers and providers have become increasingly interested in "outcomes-based management " In this case, data on patient outcomes is used as a way to permit payers, providers, or patients themselves to make choices or implement programs that are hoped to improve the quality and cost of care

Integral to many of these efforts is some form of "report card, " a profile of data on the outcomes of patients treated by particular hospitals or physicians, or enrolled in particular health insurance plans. Among the measures of quality commonly found in report cards are mortality, rehospitalization, length of stay, childhood immunization rates, and cancer screening rates.

States and private organizations have been particularly active in embracing the use of report cards as an approach to quality monitoring and quality improvement. In some cases, the dissemination of cost and outcomes Information has been mandated by state governments (e.g , in Illinois, Missouri, and Pennsylvania) In 1988, New York State began collecting cardiac surgery outcomes data intended only for use by hospitals and physicians but was later forced to make the data public as the result of a lawsuit (721).

However, many providers have independently initiated report card programs to market health plans and as a means of identifying aspects of clinical management that deserve closer scrutiny Examples of such private sector activity include United HealthCare Corp , a large managed care network that has used quality indicators since 1991 (334,449,941); and the Cleveland Health Quality Choice Project, which in 1993 released its first assessment of the quality and efficiency of 31 participating hospitals in northeastern Ohio (364). Interested persons and organizations can purchase the project's report cards for a fee. The Maryland Quality Indicator Project, which was initiated in 1985 by the Maryland Hospital Association, now covers over 600 participating hospitals Among the 15 quality indicators measured quarterly are hospital-acquired infections, Cesarean sections, and unplanned readmission The data allow participating hospitals to compare themselves with their peers and decide what, if any, action to take in response to their results (271).

While the diversity in approaches to quality assurance indicates that such projects have a promising future in many environments, the variability has also meant that the field of quality measurement has remained largely unstandardized, confounding purchasers' ability to make meaningful comparisons among competing insurers or hospitals. Two recent nationwide projects have be-

providing input to a policy decision. In this latter case, the policy decision itself has ramifications for clinical decisionmaking.

Thus, the findings from effectiveness research may be applied directly by the practitioner and the patient to improve clinical decisionmaking. Alternatively, information on effectiveness may form part of the evidence base for more detailed analyses that incorporate information on costs and on other important social considerations. In the latter case, information on a technology's effectiveness affects clinical decisions and patient outcomes indirectly, by way of their incorporation into cost-effectiveness analyses, technology assessments, and policy decisions (figure 2-1).

Clinical practice guidelines created by expert groups lie in an intermediate area in this framework. They are sometimes treated as an extension

BOX 2-3 continued: Using Patient Outcomes in Health Care Management

gun to address this problem The most prominent quality initative is the development of a prototypical standard report card by a nonprofit organization called the National Committee for Quality Assurance (NCQA) Among the 21 managed care organizations participating in the effort are Kaiser Permanence and U S Healthcare, Inc (48), United HealthCare Corp (described above) is also participating in the effort, Its own quality indicators are compatible with those of the NCQA Initiative Indicators planned for the standardized report card include childhood vaccination rates, breast and cervical cancer screening rates, and hospitalization rates for pediatric asthma cases (941). A preliminary version of the report card IS projected for completion by the end of 1994 (48)

Second, the Joint Commission on the Accreditation of Healthcare Organizations (JCAHO) announced, in 1993, the introduction of the first two sets of quality indicators for a program in which hospitals could participate on a voluntary basis; participation in the Indicator Monitoring System IS to become compulsory in 1996 (429) The data collected by JCAHO from Individual hospitals wiII be translated into scores (based on compliance with recognized standards of care, such as clinical guidelines) in 50 areas. The scores will be available for use by the hospitals themselves as well as by consumers (553)

It is Important to note that, while most private sector quality initiatives have focused on producing report cards that may be used by employers, the legislative language of several health reform proposals implies that the explicit audience for quality assessments should shift to the individual, who will be choosing coverage from a selection of plans made available through a purchasing cooperative (S 1757, H.R 3222, S 1770),

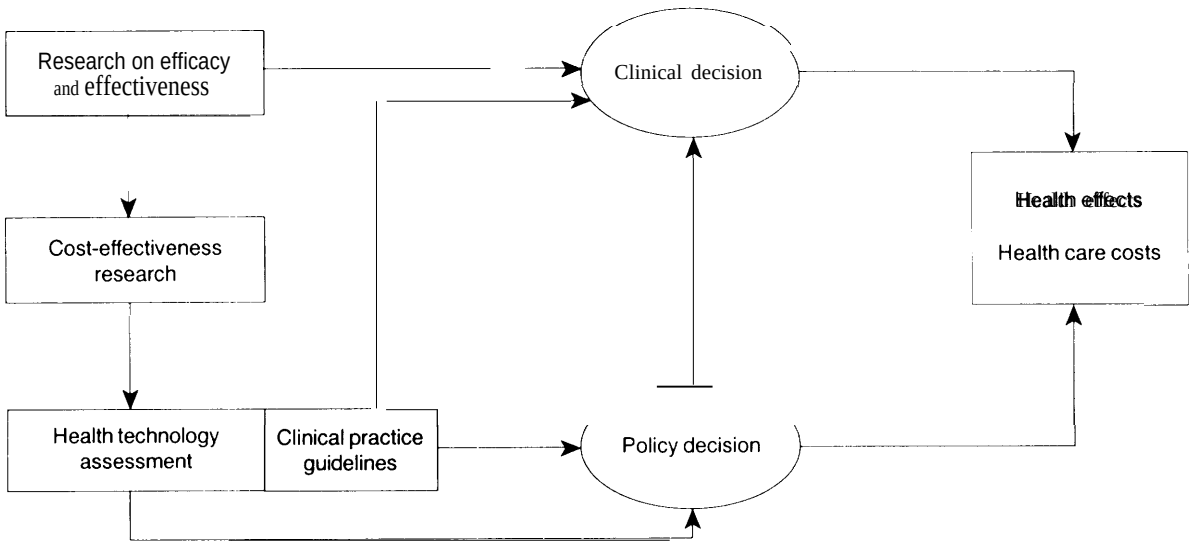
The switch from the employer to the Individual as the unit that generates and controls demand for health care coverage raises additional issues to confront in the development of report cards Whether the level of interest in using report cards on the part of individual consumers will be comparable to that of employers is still unclear (553) Some observers have argued that outcomes-based report cards wiII not be easily interpreted by consumers unfamiliar with medical issues (566) The extent to which other factors, such as personal relationships with physicians and the recommendations of peers, may compete with or outweigh the value of report cards in individual decision-making is also unknown (553)

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

of effectiveness research, assimilating existing research and adding educated opinions where research results are lacking or controversial. And a major purpose of most guidelines is to inform an individual clinical decision directly. At the same time, guidelines also often provide a basis for broad clinical policies (e.g., in a medical practice) and insurance coverage policies, and they certainly can include the consideration and analysis of information other than effectiveness information (e.g., costs).

In this report, clinical practice guidelines that address medical technologies and practices, and that are created through a structured format of synthesis and analysis, are considered a special and particularly relevant category of health technology assessments. Many of the clinical guidelines discussed here are undertaken in order to guide the formation of a clinical policy rather than a purchase decision or insurance coverage policy, and information on effectiveness is often (he predominant concern. but this report

FIGURE 2-1: A Framework for Evaluating Health Care Technologies



SOURCE Office of Technology Assessment 1994

does not consider such efforts to be part of effectiveness research itself.

THE SHAPING OF EFFECTIVENESS RESEARCH

The prime contributor to the current enthusiasm for effectiveness research, and for the use of particular tools and methods in that research, derives from fertile field of health services research. This field first became a recognized discipline in the late 1960s, as it brought together people from diverse social science and clinical backgrounds with interests in untangling the underlying factors affecting the patterns, quality, and cost of health care.

Much of this research comprised studies that investigated relationships within the health system as a whole. Relationships between peoples' access to care and health status, between trends in health care services and trends in health care costs, and other subjects relating to the cost, quality, and accessibility of care are longstanding areas of health services research. One segment of research into the patterns and quality of care, however, developed lines of inquiry that began to focus on the patient-level consequences of clinical care. This line of research, which received its impetus from intriguing findings about variations in clinical practice across geographic areas, led to a number of different research efforts examining the appropriateness and outcomes of patient care, and it ultimately resulted in the federal government's medical effectiveness research initiative.

Geographic Variation in Medical Practice

Research into geographic patterns of care was one of the earliest areas of health services research. Variations in the rate at which patients use medical services, and the rate at which physicians perform them, have been an intriguing topic of health care research for decades. A seminal study by Glover in the 1930s showed that the percentage of British schoolchildren who had undergone tonsillectomies varied more than tenfold across areas of England and Wales (285).

Studies of surgical procedures in the United States and Canada in the 1970s and early 1980s documented similarly large differences across small geographic areas. Hysterectomy rates var-

BOX 2-4: Variation Between Two Cities: The Case of New Haven and Boston

In a frequently cited 1987 article, Wennberg and colleagues [Illuminated a dramatic example of geographic variation in the utilization of Inpatient care (91 7) New Haven, Connecticut and Boston, Massachusetts are demographically similar cities in which most hospital care IS provided by academic health centers However, 1982 per capita expenditures for inpatient care in Boston were roughly twice those of New Haven ($451 vs $889), making the Boston community one of the biggest consumers of health care services and New Haven one of the smallest in New England The authors estimated that about 80 percent of the increased utilization in Boston was attributable to higher hospital admission rates (as opposed to greater lengths of stay) A look at rate differences for specific procedures and operations found ratios of utilization to favor New Haven in some cases and Boston in others

These observations led the authors to ask whether hospital services were being rationed in New Haven or over-utilized in Boston An assessment of hospital resources in the two cities showed that Boston residents were allocated 55 percent more beds per capita than residents of New Haven The researchers wondered whether New Haven suffered from a shortage of beds, which would force doctors to consciously control rates of hospital admission However, based on conversations with physicians, as well as the observation that New Haven hospital bed occupancy rates averaged only 85 percent, Wennberg and colleagues concluded that this was not the answer

Wennberg has suggested that Boston's higher hospital admission rates, and thus its higher health care expenditures, might be the result of a need to cover the cost of maintaining additional resources in the form of beds, personnel, and equipment (91 2). In 1989, 38 hospital beds were available per 1,000 Bostonians, the statistic for New Haven residents was 26 per 1,000. Because there IS no evidence that the additional expenditures are linked with better outcomes for patients, the Boston-New Haven example has been evoked often to support the view that a significant fraction of health care spending may be wasted on unnecessary care.

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

ied five-fold across hospital service areas within the province of Manitoba, Canada, for example (646). Tonsillectomy rates in Vermont varied from three to 15 tonsillectomies per 1,000 residents (91 4). Similar variations across small areas in other states were found for procedures such as hernia repairs, appendectomies. and Cesarean sections (39,456,9 15). Considerable variations occurred across states as well. Rhode Island physicians, for example, performed twice as many hysterectomies and prostatectomies per thousand population as did physicians in Maine (39).

Variations were not limited to surgical procedures. Rates of hospitalization for back injury varied tenfold across Maine (909), and physicians in Boston hospitalized adult patients with medical problems at one and a half times the rate of physicians in New Haven (917) (box 2-4). Rates of use of both medical and surgical procedures by Medicare beneficiaries varied tremendously both within and across states (11 7).

Along with the accumulating evidence of great variations in the use of medical services was an increasing awareness through the 1980s that these variations were not necessarily associated with discernible differences in the need for health care among different populations. Areas showed great differences in practice variation even when they

had populations that were similar in both their demographics and their measurable rates of morbidity, such as days in bed due to disability (649, 908,909,917).

Outcomes of Patient Care

A critical question raised by the research on practice variations was whether these differences in medical practice were associated with corresponding differences inpatient outcomes (916). In one of the earliest studies examining this question, Daniels and Schroeder found no relationship between physicians frequency of laboratory test use and the degree to which their hypertensive patients' blood pressures were under control (149). Other studies suggested that for some surgical services, having surgery was not associated with a decreased risk of death (646,915).

Where differences in medical practice could not be linked to differences in underlying health needs or to differences in health outcomes, researchers theorized that the rate of procedures in the high-rate areas could be lowered, and costs reduced, while maintaining good patient outcomes. Conversely, where different rates were associated with different outcomes, then overall patient outcomes and the quality of care could theoretically be improved by moving practice towards the best-outcome rate. In either case, physician preferences and uncertainty appeared to be a major determinant of the procedure rate in any given community (196,9 11). If these factors were indeed at the root of practice variation, then the tantalizing possibility arose that many instances of medical intervention might be avoided, and better health outcomes achieved, simply by more agreement on the best course of care.

The Study of Outcomes of Prostate Disease

In the early 1980s, Wennberg and colleagues began to focus specifically on the study of the treatment of benign prostatic hyperplasia (BPH), a noncancerous enlargement of the prostate gland that is a very common condition in older men. Prostatectomy is one treatment for BPH, and the proportion of men undergoing the procedure by age 85 varied from as low as 10 percent to as high as 50 percent in neighboring communities (911).

Discussions with physicians confirmed the existence of two differing views within the physician community. One view held that prostatectomy should be performed as early as possible after diagnosis. Doing so, these physicians maintained. would avoid the development of later symptoms, and the need for surgery when the patient was older and the procedure riskier. The opposing view held that surgery did not improve overall life expectancy and should be reserved for patients with symptoms. When the researchers examined the literature and insurance claims data on treatments for BPH and compared them with mortality rates, they concluded that surgery did not improve life expectancy and might actually decrease it (911).

Besides creating an interdisciplinary approach to research that focused on patient outcomes. "outcomes research" as defined by the activities of the BPH project had three characteristics that shaped the directions of subsequent effectiveness research.

First, the project drew attention to the differences between outcomes predicted by clinicians, for various alternative therapies, based on their knowledge and experience, and the outcomes experienced by patients, as documented in the data. Neither existing literature nor expert opinion on prostatectomy suggested as high a mortality rate as that found for the patients represented in the claims database used by the BPH research team (918).

Second, it made great use of insurance claims data as a basis for assessing the actual outcomes associated with particular therapies in practice. The identification of greatly varying rates of prostatectomy through insurance claims data was responsible, in part, for the decision to focus on BPH as a condition to study (248). The BPH study also found higher reoperation and mortality rates associated with an increasingly popular, less invasive form of prostatectomy (transurethral prostatectomy, or TURP), based on an analysis of claims data for the procedure (920). Those results gained wide publicity as an unexpected finding that

demonstrated the benefits of this observational data-based approach to documenting outcomes. Although this association was later shown to be due at least in part to patient selection bias (physicians tended to refer higher risk patients for the less invasive surgery and reserve open surgery for lower risk patients) (248), the initial finding nonetheless helped promote the use of claims data as a method for studying patient outcomes.

Third, the BPH outcomes project placed a heavy emphasis on understanding patient preferences and measuring patients' self-reported symptoms and quality of life. Given that therapy for BPH is targeted at reducing or preventing symptoms rather than improving lifespan, and given the lack of clear objective benefit, based on existing studies, of one therapy over another, BPH researchers concluded that patient preferences should be a major component of the decision to select a particular mode of therapy (911).

Relationships Between Volume and Outcomes of Care

A separate cadre of researchers homed in on another aspect of variation in medical practice: the relationship between the volume of a procedure done in a hospital. or by an individual physician. and the outcomes of the patients who underwent that procedure. The common theme of this body of literature is that there is often a correlation between the number of procedures performed by a provider and the outcomes of care (generally measured by mortality rates).

Luft and colleagues published a landmark study in 1979, suggesting that, at least for some procedures, higher hospital surgical volumes were associated with better outcomes for patients (476). They compared mortality rates with surgical volume of 12 procedures for nearly 1,500 hospitals during 1974 and 1975 and found that for open-heart surgery, vascular surgery, transurethral resection of the prostate. and coronary artery bypass graft (CABG) surgery, high-volume hospitals (defined as hospitals that performed a given procedure 200 or more times annually) had mortality rates 25 to 41 percent below their low-volume counterparts. For four other procedures, researchers found that the volume-outcome curve flattened out at a much lower annual volume threshold (10 to 50 procedures per year as opposed to 200). In two cases. no volume-outcome relationship was observed.

Subsequent studies confirmed the finding that hospitals with more experience in a procedure-i.e., higher volumes of surgery—had significantly lower rates of in-hospital mortality (256.257, 426,693). Few equivalent volume-outcomes studies on medical conditions have been performed, although two studies of AIDS treatments found that patients with AIDS fared substantially better at hospitals serving large numbers of patients with AIDS cases, compared with their counterparts at low-volume hospitals (47,732).

Not all studies investigating possible volume-outcome relationships have found them. A 1987 review of the literature regarding this relationship for hospitals found that unlike most studies of other procedures, studies of treatment for femur fracture and for stomach operations tended not to support the "greater volume-better outcomes" hypothesis (477).

The research on volume-outcomes relationships emphasized the usefulness in health research of ultimate measures of health outcomes, such as mortality, rather than intermediate endpoints with less clear functional implications. Although the exact nature of the relationship between volume and outcomes remains murky. the research overall has tended to reinforce the idea that simultaneously reducing costs (through improved efficiency at high-volume institutions) while improving the quality of care (through better care outcomes) is an achievable goal.

The Medical Outcomes Study

The theme of improving measurement of patient outcomes gained substantial support from an entirely separate and ambitious research initiative. the Medical Outcomes Study (MOS), which began in 1986. The goal of the MOS was to follow the health care received by a large group of participants in order to answer outstanding questions

about the relationships between the structure and process of care and the health outcomes associated with that care (746). To do so, the MOS researchers collected cross-sectional (i.e., one-time) data on over 22,000 participants. In addition, the researchers identified a subset of over 2,000 patients who had at least one of five conditions (hypertension, diabetes, acute myocardial infarction, congestive heart failure, and depression) and began collecting detailed longitudinal data on their care. Data collection on these patients was still ongoing as of 1993 (745,746).

To assess the outcomes of care on patients, researchers used information from clinical examinations and from the patients' medical records (746). In addition, the researchers developed and tested at length a set of general health surveys, administered to patients, to assess the patients' own perceptions of their functioning and general well-being.[4]

The MOS made two crucial contributions that helped give focus to effectiveness research efforts. The first was its substantial investment in developing and validating general health measurement instruments, particularly the 36-question version, the "SF-36," to measure self-assessed patient functioning and wellbeing at any point in time (513). The second contribution was to link patient characteristics and particular components of care with care outcomes (725). Researchers have found, for example, that the negative effects of depression are additive for patients who are depressed in addition to having other chronic health problems (907).

Appropriateness of Care

Even when an intervention is generally effective, or effective under particular circumstances, it may sometimes be applied to patients for whom it is inappropriate. The research on variations in medical practice led directly to another question: Does the greater inappropriate use of procedures in high-use areas explain the geographic differences in rates of use?

There has long been evidence that some inappropriate medical practice does occur (242). A very convincing study done in the 1970s, for example, documented the inappropriate use of tetracycline, an antibiotic, among young children in Tennessee's Medicaid program (626). Complications related to the use of tetracycline in this age group had long been noted, and by the 1970s there were numerous alternative drugs. In January 1975, the American Academy of Pediatrics officially stated that there were "few if any reasons for using tetracycline drugs in children less than 8 years old." Despite the uniform agreement in the official medical community regarding tetracycline's inappropriateness for children in this age group, Ray and colleagues found that the drug had been prescribed for over 4,000 young children over a two-year period (626).

Researchers at RAND approached the question of appropriateness of care by focusing on specific procedures that are both costly and shown to vary across geographic areas (11 8). Initially, they chose six procedures to study:

1. coronary angiography (a diagnostic imaging procedure for heart disease),
2. coronary artery bypass graft surgery (a major surgical treatment for heart disease),
3. cholecystectomy (surgical treatment for gallstones),
4. diagnostic gastrointestinal endoscopy (a procedure to diagnose disorders of the digestive tract),
5. colonoscopy (a diagnostic procedure to detect disorders of the lower intestine), and
6. carotid endarterectomy (a surgical procedure performed in persons considered to be at very high risk of stroke).

A major obstacle to overcome was defining "appropriate" uses of these procedures. Unlike the tetracycline study, which had the advantage of an

[4] In an interesting example of the accumulative properties of health services research, the foundation for the health surveys was an assessment measure from a previous major federally funded research effort, the RAND Health Insurance Experiment (86).

BOX 2-5: Defining "Appropriate"

To define "appropriate'" indications for the procedures they studies, RAND researchers convened expert panels that reviewed the indications discussed in the literature, and in their own experiences, and arrived at group ratings of the appropriateness of each indication (see chapter 7) The panels used a rating scale of 1 through 9, with 1 representing extremely inappropriate and 9 representing extremely appropriate "Appropriate" was defined to mean that the expected health benefit exceeded the expected negative consequences "by a sufficiently large margin that the procedure was worth doing " "Inappropriate" meant that the negative consequences outweighed the health benefits Panelists were instructed not to consider financial costs

The researchers suggested a final, simpler spilt to categorize the ratings three categories of Inappropriate, " "appropriate, " and "equivocal " The definition of the last category was particularly Interesting, because it included both indications for which the panel agreed that the Indication was neither clearly appropriate nor clearly Inappropriate, and Indications for which there was substantial disagreement among panelists regarding appropriateness

SOURCE Off Ice of Technology Assessment 1994 See chapter 7 and appendix C text for more detailed discussion and reference sources

unambiguous measure of appropriateness in the statement of a major medical association, there was no universally acknowledged consensus about what constituted appropriate use. The issue was not that these procedures (e. g., bypass surgery) were themselves inappropriate, but that some of the reasons for doing them—the medical indications—were not appropriate. To define '*appropriate" reasons for performing the six procedures, researchers at RAND assembled "expert panels'* to rate the various identified medical indications for each procedure (box 2-5).

The results of applying appropriateness ratings to explain geographic variations in medical practice have been somewhat surprising. In the first study on this topic. researchers examined the reasons for performing three of the six procedures (carotid arterectomy, coronary angiography, and gastrointestinal endoscopy) in five sites across the country (11 8). The rates at which each of the three procedures were performed varied considerably across sites (in the case of carotid endarterectomy, they varied by almost a factor of four). The proportion of procedures performed "inappropriately" according to RAND criteria, however, was surprisingly consistent across sites (between 29 and 40 percent for carotid endarterectomy and between 15 and 19 percent for the other two procedures). Overall, there was an association between higher rates of use of a procedure and a higher proportion of inappropriately performed procedures, but the association was surprisingly small (1 18).

To test the possibility y that the use of large areas for comparison might have masked variations that would be apparent if smaller areas were contrasted, the researchers repeated the process in 23 counties in a single state (447). Both the rates of procedures and the percentage of procedures rated appropriate varied enormously across these small areas (table 2-1). Nonetheless, the association between the two measures was remarkable for its near absence (447). Although this study has been criticized as inadequate to test its hypothesis properly (152), its findings were so remarkable that they are hard to dismiss out of hand.

Using the RAND appropriateness criteria, the same group of researchers have documented significant proportions of inappropriately performed procedures in several patient populations (525a, 938,939). In a literature review that included these and other investigations into inappropriate care.

TABLE 2-1: Variation and Appropriateness of Three Procedures Across Small Areas

Procedures	Rate of use per 10,000 Medicare enrollees	Percent of procedures judged appropriate
Coronary angiography	13-158	8% - 75%
Carotid endarterectomy	5-41	O% - 67%
Upper gastrointestinal tract endoscopy	42-164	07. - 25%

SOURCE Based on data from L L Leape, R E Park, D H Solomon, et al "Does Inappropriate Use Explain Small-Area Variations in the Use of Health Care Services')" *Journal of the American Medical Association 263(5) 669-672, 1990*

the reviewers found documentation of inappropriate use ranging from 3 to 75 percent for procedures, 6 to 80 percent for hospital use and office visits, and 3 to 90 percent for drug use (83). They found evidence of underuse as well as overuse, although the latter was much more prevalent in the literature. They concluded by speculating that:

> ... as much as one-fifth to one-quarter of acute hospital services or procedures were felt to be used for equivocal or inappropriate reasons, and two-fifths to one-half of the medications studied were overused in outpatients (83).

International comparisons suggest that even countries with much lower overall rates of procedures than the United States have a substantial proportion of procedures that are performed for inappropriate reasons. Physicians in the United Kingdom, for example, perform coronary angiography and coronary bypass surgery much less frequently than do U.S. physicians (54). As expected, in a study comparing the appropriateness of indications for these two procedures in the two countries, researchers found that U.K. physicians rated a higher proportion of indications to be inappropriate for both procedures than did U.S. physicians (54,85) Nonetheless, the proportion of these procedures deemed inappropriate even by the U.S. physician panel was a substantial 17 percent (54).

Overall, the findings of appropriateness research have tended to support the belief that some portions of medical care can be eliminated while actually improving the quality and effectiveness of care provided. That belief may be somewhat overstated. The main message from the RAND review of appropriateness studies—that up to one-fourth of procedures and up to one-half of medications are prescribed for reasons that are inappropriate or equivocal—may imply more "wasted" care than is the case. The selection of technologies that have been studied, for example, may be biased if researchers have tended to study a particular technology or service precisely because inappropriate use was suspected.[5] In addition, the reviewers' generalization of appropriate use combined equivocal with inappropriate care. Recent studies suggest that the category of equivocal care is sometimes much larger than the category of care that is clearly inappropriate (54,343,446).

Nonetheless, appropriateness research has certainly documented that a significant amount of dubiously useful care is being provided. This research has also helped highlight the degree of professional uncertainty and disagreement that remains in the appropriate indications for performing many high-cost procedures. But the findings of this research also suggest that areas with high

[5] Indeed, the one study reviewed by the RAND researchers that looked at a broader set of 12 procedures found a much lower rate of overuse (3 percent) than any of the studies looking at overuse of a single procedure (the lowest rate of inappropriateness found in any of these studies was 13 percent).

rates of particular procedures do not necessarily have a higher proportion of inappropriate procedures, In fact, areas with low rates may not perform et-tough appropriate procedures.

The research addressing the question of the appropriate use of particular medical technologies has diverged somewhat from the line of research that makes up most of the federal medical effectiveness program. Unlike the work on patient care outcomes, the extensive RAND work on appropriateness of care has focused more on the pragmatic demand for information that can lead to immediate, relatively unambiguous decisions.

The ability to label some care as "inappropriate" is potentially useful to policy makers interested in taking immediate action to reduce some proportion of health care costs through the elimination of "wasteful" services. The attractiveness of the RAND approach is apparent in the fact that private sector payers and providers are expressing an interest in linking medical practice guidelines and payment to conclusions about appropriateness based on this approach. The limited assessment of this approach outside of the small group of researchers who developed it, however, has led some observers to criticize the adequacy of its evaluation (605). (Chapter 7 and appendix C of this report discuss the process used in the RAND approach in more detail.)

The Federal Medical Treatment Effectiveness Program

The different lines of inquiry into the variation and outcomes of current medical care practices began coalescing into a program identity in the late 1980s. Encouraged by the progress of research into the outcomes of treatments for prostate disease. Congress in 1987 ordered the National Center for Health Services Research (NCHSR)[6] to establish an "outcomes research program" to expand this approach to understanding medical care. NCHSR solicited applications for the first outcome research team program grants in 1988.

In the same year, William Roper, then administrator of the Health Care Financing Administration (HCFA), and several of his colleagues issued a call for "effectiveness research" (651). Roper's focus was on the effectiveness of medical care provided to elderly and disabled individuals covered by Medicare. He proposed to examine the outcomes of medical procedures and other care by making use of the rich resources that were the Medicare administrative databases (65 1).

The creation of the Agency for Health Care Policy and Research (AHCPR) by congressional mandate in 1989 eclipsed Roper's plans for a HCFA research initiative. A major part of AHCPR's role was to be the focal point for federally supported effectiveness research. To carry out this role, AHCPR established its Medical Treatment Effectiveness Program (MEDTEP), which subsumed both the HCFA initiative and the previous NCHSR outcomes research program. Research into practice variation and documenting outcomes of current medical practice continued to be part of the research portfolio.

Although "effectiveness research," as defined earlier in this chapter, could cover a very diverse set of research activities, the characteristics of the federal government effectiveness initiative have been shaped by its roots in research on practice variation and the measurement of health outcomes. (Research on the appropriateness of care, as carried out at RAND, has been carried out separately from the federal initiative.) The outstanding characteristics of the federal endeavor based at AHCPR are:

1. It is focused primarily on the evaluation of existing technologies and medical practice patterns, rather than on the evaluation of new interventions.

[6] NCHSR went through several name changes between its inception in 1968 and its replacement by the Agency for Health Care Policy and Research in 1989. In 1987 its formal name was the National Center for Health Services Research and Health Care Technology Assessment, but for the sake of simplicity the shorter title is used here and elsewhere in this report.

2. It has emphasized the need for research whose results will be widely applicable, including populations and settings that have often been underrepresented in efficacy studies. These include elderly populations, women, minorities, persons with disabilities or multiple health problems, and treatment settings such as physicians and health facilities that are not affiliated with teaching institutions.
3. It has stressed the use of outcome measures that assess factors that affect patients directly (e.g., physical and social functioning and pain), rather than only intermediate clinical measures (e.g., laboratory test scores).
4. It has included the substantial use of tools other than prospective, randomized clinical trials. In particular, it has historically placed a particular emphasis on analysis of large administrative databases. It has not absolutely excluded the use of randomized and other controlled clinical studies, but much of the impetus for the field has come from the expectation that for existing medical technologies, nonclinical research methods are both cheaper and more efficient.

EXPECTATIONS IN THE CONTEXT OF NATIONAL HEALTH REFORM

Effectiveness research stresses that medical practice varies for reasons unassociated with demographics and health needs, and that much current medical care is performed for inappropriate or at least equivocal reasons. If this is true, and if the most effective practices can be identified, described, and disseminated, then it might indeed be possible to raise the quality of health care while reducing its costs. This is the basic assumption that underlies many of the expectations of effectiveness research. It is also the assumption that led the federal government to invest substantially in the creation and dissemination of clinical practice guidelines, which would assemble the evidence and describe the best course of clinical care for the medical conditions they addressed.

The assumption found a ready audience in public policy makers, embroiled in the search for palatable solutions to the conundrum that is American health care. Since the early 1980s, researchers and commentators have promoted the idea that pursuing research into the effectiveness, cost-effectiveness, and broader effects of health care would be a small investment yielding a major improvement in both the quality and the cost of care (8,791a,908a,934). The message was clearly heard by members of Congress. At a Senate hearing in 1988, the opening statements of the Senators reflected a confidence that health services researchers would be able to define appropriate care in order to offer substantial cost savings and high quality, focusing on the advantages for Medicare beneficiaries (792). It was against the background of these expectations that AHCPR was created in 1989 to provide focused federal support for effectiveness research and clinical practice guideline development.

Since the establishment of AHCPR, the rhetoric emphasizing the cost-containment benefits of these activities has faded somewhat. Cautious notes have been sounded by reports from the Institute of Medicine and the Physician Payment Review Commission, which backed the idea of federally supported guidelines but questioned whether AHCPR's guidelines effort would necessarily lead to cost savings (376,607). Recently, the administrator of AHCPR has asserted bluntly that "outcomes research is not a cost cutting exercise" (494).

Nonetheless, with the prospect of national health reform on the horizon, effectiveness research, guidelines development, and other activities that involve the evaluation of clinical practices continue to play a part in policy makers' hopes for improving the health care system. President Clinton's health proposal, for example, included specific provisions to encourage "effectiveness research," "quality and outcomes research," and the "development and dissemination of guidelines." According to the proposal, this research would "increase the cost-effectiveness, appropriateness and quality of care" in the health care system (S. 1757).

Of particularly widespread interest in health reform proposals is the idea of "scorecards" or "per-

BOX 2-6: Examples of State Legislative Activities Regarding Clinical Practice Guidelines

- **Maine** has an ongoing five-year demonstration project that permits the use of guidelines as a defense in malpractice cases (453)
- **Vermont** has a similar law that calls for "recommendations for the development of standards of care and practice guidelines, ' which could be used as a defense in malpractice suits (Vermont Law Sec 1, 18 V.S.A Part 9, Ch 221)
- **Maryland** has established an Advisory Committee on Practice Parameters to oversee the design of guidelines whose content are to be based on effectiveness research and physician consensus
- In **Minnesota,** the Health Right Act of 1992 included the adoption of practice parameters as a means of assuring quality in health care Here, too, guidelines may be used as a defense in malpractice suits, and the fiscal expectations for guidelines are eloquently demonstrated by the fact that the Minnesota Department of Health listed the provision for practice parameters under the heading of cost containment, as a measure "to avoid unnecessary and Ineffective treatment and services (533)

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

formance indicators," which rate providers, insurers, or health care plans according to their performance along several criteria. These can include mortality rates, costs, rates of specific procedures, or rates of hospitalization for "preventable" diseases. A number of payers and providers already have systems in place to monitor performance indicators (see box 2-3). They see these systems as strategies to eliminate costly or inadequate physicians from the payment rolls, improve hospital quality of care over time, and assure premium-paying employers that their health care dollars arc being well spent. In this case. the results of effectiveness research and clinical practice guidelines can become the benchmarks against which providers are rated.

Clinical practice guidelines have also become a basis for policy makers' hopes of reducing malpractice insurance costs and physicians' use of defensive medicine (e.g.. H.R. 101), especially at the state level (box 2-6).[7] Perhaps most importantly, guidelines and data on effectiveness have also been proposed as the basis for defining health insurance benefits.

In perhaps the best known example, the State of Oregon, in 1989, officially proposed prioritizing health care services for its Medicaid beneficiaries according to such factors as the relative effectiveness of the services (722). Although in the end evidence on effectiveness played a relatively minor role in the prioritization process (788,794). the process shaped the discussion about the place of information on effectiveness, cost-effectiveness. and quality of life in health insurance coverage. More recently, legislation introduced in Oregon would require that medical guidelines be part of the basis for prioritizing services under the state's Medicaid demonstration program (Oregon Senate Bill 757. 1993).

[7] "Defensive medicine" occurs when doctors order tests, procedures or visits, or avoid high-risk patients or procedures, primarily (but not necessarily solely) to reduce their exposure to malpractice liability (790). For a detailed discussion of this topic, see the OTA report, *Defensive Medicine and Medical Malpractice* (790).

BOX 2-7: One Proposed Model for Basing Health Insurance Benefits on Clinical Practice Guidelines

In one model of how a benefits package might be based on clinical practice guidelines, Hadorn has proposed the development of a comprehensive set of "necessary care guidelines, " which would collectively represent a basic benefits package (318,320), He defines "necessary" as "reasonably well demonstrated to provide significant health benefits, " one step beyond appropriateness (320,403). Under this model, necessary care guidelines, resembling utilization review criteria in format, would be developed by expert panels and presented for public debate at hearings modeled after the '(science court experience" and the NIH Consensus Development Conferences (31 8,319)

Hadorn's proposal hinges on the ability to incorporate into the benefits development process an "objective standard of proof" that would consider health care needs as well as costs, thereby constructing a mechanism to judge a given type of care on the "net health benefits" that the population could expect to gain from it (319), Gwen that the goal of this model is to provide comprehensive coverage while cutting costs, the major assumption is that it would be unnecessary to make decisions based on cost alone (i e rationing) because "the volume of services excluded from coverage using a standard of proof approach would entail substantial cost reduction in and of Itself"' (319)

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

Other public and private policy makers have also begun to experiment with the use of clinical practice guidelines in defining or modifying health insurance benefits. In Canada, a preliminary agreement with the British Columbia Medical Association stipulates that patients who seek services outside the parameters of practice guidelines (now being developed by the province's Medical Services Commission) will not be covered by Canada's national health insurance (528). In the United States, Blue Cross and Blue Shield of Illinois implemented, on January 1, 1994, a policy requiring physicians to comply with practice guidelines. Participating specialists in the Illinois Blues' Managed Care Network Preferred, servicing over 100,000 enrollees, must follow guidelines covering 14 procedures or treatments, including bypass surgery, cholecystectomies, and blood transfusions. Except for guidelines on cancer care, which were developed by the insurance company, the practice parameters were produced by various specialty societies. The new policy was met with opposition from the American Medical Association, which argued this mechanism made physicians who participated in the medical plan subject to guidelines that they had no opportunity to help develop or modify (74).

The Clinton Administration's proposal for national health care reform also incorporated effectiveness and cost-effectiveness research results into its proposed benefits plan, at least for preventive services (S. 1757). Both this and alternative proposals that involve the establishment of a national board that would set benefits clearly envision that such a board would use the results of effectiveness and cost-effectiveness research and of clinical practice guidelines and other technology assessments in their decisionmaking (e.g., S. 1757, S. 1579, H.R. 3222). Some researchers have taken the concept a step further and proposed an insurance benefits model in which a battery of guidelines would themselves comprise a benefits package (box 2-7). In California, policy makers have considered using guidelines to create a benefits package for the state insurance plan for public employees (99).

If data on effectiveness and formally structured clinical practice guidelines are one of the bases for health insurance benefits under health reform,

then the validity and reliability of those inputs are clearly of considerable interest. Even in the absence of a benefits package that relied heavily on research-based evidence and guidelines, any reform proposal that relies on the expansion of "managed care" has a stake in the validity and impact of these activities. They represent some of the tools by which the managers in managed care organizations can hope to achieve high-quality, better-cost care. If these tools are inadequate, the assumption that managed care can solve many of America's health care problems would bear serious scrutiny.

CONCLUSIONS

Much, if not most, of existing medical technology and practice has been inadequately evaluated, even with regard to its effectiveness in improving peoples' health. Nonetheless, for all this dearth of information, society has gradually amassed a number of tools to evaluate the health, economic, and social effects of technologies (366,783), and the applications of those tools to the crucial questions of health care are slowly growing.

Research to address the deficit in evidence regarding current medical care has developed separately from the traditional clinical trials research community, influencing the kinds of tools it has applied. The research evaluating existing clinical practices has also tended to emphasize that considerable variations exist in how medical care is practiced; that considerable disagreement exists among clinicians regarding the circumstances under which particular treatments are appropriate; and that the health outcomes valued by patients are often not the same as those measured by researchers and clinicians. In the process, effectiveness research has created expectations among policy makers that further investments in this line of research. coupled with the aggressive development and promotion of clinical practice guidelines, can make great strides in eliminating ineffective care, improving the overall health of the population, and even reducing health care costs.

Despite the optimism prompted by early effectiveness research, there were and are still a number of ambiguities about the kind of change that can be expected. The research on appropriateness, for example, has found that higher rates of use of a procedure are not equal to higher levels of inappropriate care. Nor does current research necessarily support the idea that the source of variations in clinical practice is individual provider uncertainty that can be abolished by presenting that practitioner with good information or guidelines. Rather, research suggests that uncertainty lies in disagreements among physicians (459), with individual physicians possibly quite confident in their own opinions. Indeed, Chassin (11 5) theorizes that the main reason behind practice variation is the number of physicians who are "enthusiasts" for particular procedures or care processes. If this is true, there may be disagreement but not individual uncertainty, implying a more difficult job for federally sponsored activities whose ultimate goal is to affect clinical practice by improving outcomes, reducing costs, or both.

Implicit assumptions about the impact that these activities will have underlie a number of different aspects of proposals currently being discussed in the context of national health reform. It will affect, for example, the extent to which policymakers can depend on the idea that basing health benefits on guidelines and effectiveness information is feasible. likely to result in changes in clinical practice, and likely to help restrain system costs.

Moreover, the findings of effectiveness research and practice guidelines are a crucial underpinning of performance indicators, which are based on the idea that there is a proven standard of preferred practice to which a provider should adhere. Health reform proposals that emphasize a large role for other consumer and provider information, or for managed care providers, contain implicit assumptions that evidence regarding the

effectiveness and value of medical technologies and practices is sufficiently available, valid, and convincing that it will enable these players to improve their health care outcomes and costs.

As implemented in the federal government's medical effectiveness initiative in 1989, and in the charge to AHCPR, "effectiveness research" emphasized particular qualities and approaches toresearch. Those qualities (e.g., an emphasis on existing technologies and broad populations) and approaches (e.g., large database analyses) were emphasized in response to perceived deficits in the contemporary research agenda. However, "effectiveness research" includes a wider variety of potential activities than those emphasized in the first few years of AHCPR's existence.

Examining the federal government's current investment in activities that evaluate the effectiveness and value of medical technologies and practices in detail, and examining the extent to which expectations for that investment are well founded, is the focus of this report. The remainder of this report assesses the validity, potential usefulness, and efficiency of Federal activities regarding effectiveness research (chapters 3 and 4), cost-effectiveness research (chapter 5), health technology assessment generally (chapter 6), and clinical practice guidelines specifically (chapter 7). Finally (chapter 8), this report examines the ways in which these activities, and particularly clinical practice guidelines, are most likely to have an impact on clinical practice.