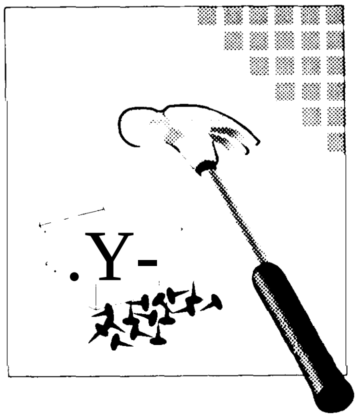
Tools for Effectiveness Research 3

The strength and believability of evidence on the effectiveness of health technologies rest largely on the underlying methods used to generate it. The purpose of this chapter is to describe the basic methods employed to generate evidence, emphasizing those techniques that have evolved recently, are particularly appropriate to broad research on the effectiveness of care, or have seen especially heavy use in effectiveness research as carried out thus far in the United States. [1]

The validity of the underlying methods being applied is a matter for particular interest in the area of research on the effects of health technologies, because making heavier use of certain techniques was an explicit component of the federal government effectiveness initiative. The law enacting the Agency for Health Care Policy and Research (AHCPR) specifically encouraged the use of particular research methods, such as large administrative database analysis (Public Law 101 -239). In addition, the increasingly intense interest in whether specific medical interventions are worth doing has stimulated research activity in areas ranging from the measurement of people's preferences for various health outcomes to the statistical synthesis of the results of pre-existing studies.

Legislation to encourage effectiveness research has not only encouraged certain research methods but also a particular organizational structure for applying them. This approach centers on Patient Outcomes Research Teams (PORTs)-groups of research-



[1] More detailed discussions of the applications. advantages, and limitations of some of these techniques are contained in a separately published set of background papers associated with this report (see appendix A).

ers with diverse backgrounds who join together to conduct research on a particular medical condition. This chapter concludes by examining the contributions of these teams and their implications for the future of effectiveness research.

TECHNIQUES TO EVALUATE HEALTH EFFECTS

Tools for generating evidence regarding the effectiveness of health technologies fall into three broad categories:

- *basic tools* for measuring health status and health outcomes;
- *primary studies,* such as clinical trials and administrative database analyses; and
- *secondary techniques* to synthesize the results of the primary studies in order to generate new insights or more powerful conclusions.

Basic Tools for Measuring Health Outcomes

Assessing the effectiveness of a medical technology (or any health care intervention) requires an evaluation of whether the health-related outcomes resulting from the use of that intervention are better than would have been expected without it. Such an evaluation requires measuring both what those outcomes are and what they would have been without the technology.

The simplest outcome to measure is death. For some conditions, it is probably the most important outcome as well. But most of the conditions that cause people to seek medical care, such as back pain and bronchitis (795), are not characterized by high fatality rates. Furthermore, even for conditions that are often fatal (e.g., cancer), improvements in the quality of life people have before death is a major goal of treatment (820).

An interest in measuring people's health status more directly has led to the development of tools to assess how patients feel and how well they can function. At the most basic level, existing tools differ according to two attributes: whether they depend on patients' own responses or the observations of others: and whether they are condition-specific or generic measures of health.

Some of the oldest health measurement instruments, such as the Karnofsky Index for patients with cancer (developed in the 1940s) and various Activities of Daily Living scales (developed in the 1950s), are still used today (75.161 ,503). What most of these measures have in common is that the assessment of the patient's health is usually performed by someone who observes the patient, often a clinician.

Recentl y, however, there has been an explosion of research interest in measures that incorporate the patient's own self-assessment. In particular, the past decade has seen increasing interest in the use of measures of self-assessed health status that might be applied across a wide variety of health conditions to evaluate the effects of health care technologies (6,5 19).

One reason for the surge in interest is the discovery that clinical markers of health often correlate very poorly with the patient's perception of his or her health status. Perhaps the best documented example of this phenomenon is the evaluation of the health status and progress of persons with benign prostatic hyperplasia (BPH), a noncancerous enlargement of the prostate gland. BPH is very common in elderly men and often results in a narrowing of the urethra (the urinary conduit), producing troublesome symptoms such as frequent urination and difficulty starting urination. To evaluate patients with BPH, urologists traditionally have used a measure of the amount of urine left in the bladder after voiding. They have also used a measure of the rate of urine flow to assess obstruction, and they have measured the size of the prostate through palpation and imaging. None of these measures. however, correlates well with how patients experience symptoms, or even with the frequency of their symptoms (2,25,43, 557).

A second reason for the interest in self-reported measures of health is the increasing evidence that health professionals are often not good proxies for their patients when it comes to reporting symptoms and health experiences (63,587). Studies comparing self-reports with reports from proxies suggest that the less observable a characteristic is (e.g., personal values about health care), the less likely it is that others can report on that characteristic accurately (227,482,655,773).

Interest in developing generic measures of self-assessed health status has derived in part from the desire to assess changes in a person's well-being when that person has multiple health conditions, and treatment for one condition can affect others (263). Generic measures also enable researchers to avoid reinventing new measures for every health condition. One widely used instrument developed for the Medical Outcomes Study, for example, has since been used in studies of such varying conditions as diabetes and knee replacement surgery (41 2,558).

A third reason for interest in generic measures of health is the desire to make comparisons across different conditions and treatments for the purposes of health policy and resource allocation decisions (59 1). This use of health status measures has been an especially strong incentive for the development of measures whose results can be summarized in a single number and incorporated into cost-effectiveness analyses (see chapter 5).

Measuring Health-Related Quality of Life

Most instruments used to measure health-related quality of life take the form of questionnaires that ask about at least four different dimensions of this attribute (503,726). These are:

1. *Functional ability.* This component relates to what people can do, without regard to their resources or the actual demands on them. Physical abilities included in a questionnaire might include climbing stairs, or being able to read a newspaper or hold a pen.
2. *Perceived health.* Worry about one's health and satisfaction with one's health are commonly measured aspects of self-perceived health. Or, a question may simply ask people to rate how healthy they think they are.
3. *Psychological well-being.* This component focuses on the extent to which people see themselves as distressed (e.g., depressed or anxious). It is intended to be broader than specific measures of mental health. although it is related.
4. *Role functioning.* Questions regarding role function ask about individuals' work, their resources, and what they ordinarily expect themselves to do on a day-to-day basis (e.g., care for one self and family, visit friends), These questions help accommodate the fact that the same condition can have very different effects on people. A knee injury that severely limits the normal activities of a professional athlete, for instance. may be much less limiting to a professional editor, even though both of them have the same absolute functional abilities.

For many of the generic health status measurement instruments, results are summarized by describing the results for each of the dimensions the instrument measures. For example, a conclusion might be that a patient improved in physical function but was unchanged with regard to role functioning or perceived health. The "SF-36" and the Sickness Impact Profile (box 3-1) are probably the best known U.S. examples of generic instruments measuring patients' self-assessed health.

When the purpose of the measurement is to make comparisons across conditions. however, researchers instead sometimes use an instrument that produces a summary value for quality of life-one that combines results for the different dimensions measured and presents them as a single number. To come up with such a summary value, scores for the individual dimensions must be combined. usually by assigning weights to the individual scores and adding these weighted scores. The weights, which are intended to represent the relative importance of the different aspects of health being measured, might be derived from statistical models or average ratings of health care workers, patients, or the general public. The Quality of Well Being (QWB) Scale (see box 3-1) is among the best known examples of instruments that produce a single quantitative score of health-related quality of life.

BOX 3-1: Examples of Instruments To Measure Health-Related Quality of Life

Many different quality-of-life instruments exist (75,503), and the emphasis on development and use of particular measures varies among different countries. The Health Utilities Index, developed by Statistics Canada, is being used in Ontario, Canada as a general population health status measure as well as a clinical and policy tool (263). The Nottingham Health Profile has been used widely in the United Kingdom (360), and the EuroQol index has been used in a 14-country study in Europe (228) Three multidimensional measurement instruments have been particularly widely used in the United States for studies of health outcomes and effectiveness: the Sickness Impact Profile (SIP), the Medical Outcomes Study 36-item short-form health survey (SF-36), and the Quality of Well-Being (QWB) Scale

The SIP was developed in the 1970s to create a comprehensive instrument to measure the impact of sickness on people (49). Containing 136 questions that measure health in 12 different areas, it is considered one of the most comprehensive measures of health. Portions of the SIP have been used in studies of patients with conditions as diverse as pneumonia and chronic pain (328,396) and by some of the federally funded Patient Outcomes Research Teams (PORTS) (263).

The SF-36 comprises 36 questions about 8 different aspects of health-related quality of life (892) Its great strength is parsimony, it is fairly brief to administer while capturing most of the in-reformation obtained from much longer surveys Like the SIP, the SF-36 has been well-studied, and the instrument (or portions of it) has been applied by the PORTS and in other research on a wide variety of conditions (263)

The QWB differs from the SIP and the SF-36 in that it was specifically designed to produce a single score that represents an Individual's reduction from perfect health (414). This instrument consists of a list of questions that ask the respondent to report opinions or experiences regarding various symptoms (e g , headaches), diagnoses (e.g., blindness), and activity limitations (e. g , being unable to drive a car) Respondents' answers are individually weighted according to the relative importance of those problems (based on pre-existing preference weights derived from surveys conducted by these researchers) and then totaled to produce the overall score The QWB has been used both in clinical studies of outcomes (413) and for health policy purposes, in the development of Oregon's prioritized list of Medicaid benefits (788)

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

Applications anti Limitations

Disease-specific health measurement tools are standbys of health research, both because of their sensitivity to the nuances of the health condition of interest and because they are often designed to be simple and inexpensive to administer (590). New disease- and condition-specific tools continue to be developed and validated, and many emphasize patients' self-assessments (42,103,486, 656).

Generic tools such as the SIP and SF-36 have the great advantage of enabling standardization and comparability of results across multiple conditions studied (590). Brief versions of such generic measures offer the possibility of much more detailed monitoring and comparisons of the health status of specific populations (e.g., enrollees in particular medical practice or health insurance plans) than is now possible (50). The greater use of generic measures in clinical trials could add to the understanding of the relative benefits of competing medical technologies and enhance clinicians' and patients' abilities to make informed decisions about treatment choices (316,847,848, 852).

There is growing agreement that generic measures of self-reported health status and quality of life can be reliable and valid for both health status monitoring and for comparing the outcomes of specific therapies. There is also growing experience in aulgmenting them with disease-specific questions to make them more sensitive. Current research is focused on: which measures are best for which applications: how and when to use disease-specific measures, or adapt general measures for specific diseases: how answers to these questionnaires might differ across specific subpopulations: and how to minimize the number of questions that must be asked while still capturing the essence (263,420,590).

There is some debate, however, about how best to encourage clinical researchers to incorporate quality-of-life measures into trials. Staff at the National Institutes of Health point out that adding this component can increase the cost and complexity of trials, and that trial researchers may resist incorporating it (849). AHCPR officials and advisors, on their part, express frustration at the sense that this component is frequent] y considered an external add-on, with experts in the topic consulted well after a trial has been designed, rather that including it as an intrinsic part of a trial (127,816).

Some of the difficulty in getting quality-of-life measures incorporated more extensively into clinical trials may derive from the fact that the superiority of these measures over existing measures of health outcomes is not clear to the trial researchers. Head-to-head comparisons of existing trial outcome measures with generic quality-of-life measures might be required to demonstrate the superiority (or lack of it) of the generic measures.

The most controversial area regarding the use of quality-of-life measures is the application of the findings from quality-of-life surveys to health policy and resource allocation decisions. Unlike applications to clinical outcomes studies, this use of health measurement tools essentially assumes that the average scores from instruments such as the QWB represent the value that society as a whole places on different levels of health. This issue is discussed in more detail in chapter 5.

Primary Studies To Evaluate Health Effects

Epidemiological studies to observe and compare the health outcomes of patients are the backbone of medical science. They can be roughly divided into two categories: observational studies, in which the actual experiences of the groups being compared are simply observed: and experimental studies, in which the experiences of the groups are intentionally influenced by the study.

Observational studies are the traditional source of information on suggestive associations in epidemiology. The recent reports of a series of similar cases of fatal and near-fatal illness among Native Americans in the rural southwest, for example. has suggested the introduction of a new infectious disease (832). Case-control and cohort study designs are types of observational studies commonly used to make direct comparisons where experimental designs are infeasible.[2]

In *experimental studies,* study participants are randomly allocated among treatment and control groups. Random allocation is intended to ensure that all comparison groups are reasonably similar not only with regard to known characteristics but also any characteristics that are unknown but

[2] In case-control studies, a group of individuals with the characteristic of interest (cases) are compared with individuals without that characteristic (controls) regarding their previous exposure to some factor. In cohort studies, individuals are classified according to their exposure or nonexposure to a disease or intervention and followed for\\ ard to track the outcomes.

might influence the outcome. Differences in the outcomes of the groups thus can be attributed to differences in the treatment, with a level of confidence that can be described statistically. In general, the larger the groups, the greater the level of confidence that an observed effect truly exists and was not merely due to chance.

Where the effect of an intervention is large and immediate, evidence based on the findings of nonrandomized observational studies is often enough to draw a confident conclusion that the effect is at least real. '*Slam-bang" technologies such as blood transfusions and antibiotics were convincing because the outcomes after the interventions were so dramatic compared with the expected course of the conditions they were used to treat.

But most modern medical advances are incremental rather than revolutionary. They are aimed at such improvements as reducing disease complications in people with diabetes (162), or slowing the decline in the cognitive functioning of people with Alzheimer's disease (153,230). Conversely, the predicted benefit of a new technology (e.g., immunotherapy treatment for women with recurrent spontaneous abortion) sometimes turns out to be illusory (267). In these circumstances, the ability to reliably distinguish real but modest effects from no effects through carefully performed studies is crucial to the credibility of the conclusion.

Randomized studies maximize internal validity—the certainty that (he treatment actually caused the effect. Because they are specifically designed to disprove the null hypothesis (i.e., that the treatment has no effect), they err on the side of finding no apparent effect even where a very small one actually exists. In contrast, nonrandomized study designs tend to favor the treatment being tested (784). Where both randomized and nonrandomized control studies of a new therapy have been performed, the nonrandomized studies generally (although not always) find the new therapy to be much more beneficial than do the randomized ones (136,529,669,950).

Thus, the validity of nonrandomized studies that conclude that a particular technology is beneficial is often suspect, because of the known bias in favor of finding a beneficial effect, and because it is often impossible to assess the extent to which the groups being compared were actually similar. This problem is especially acute in "case series" studies, where the "control" is how well patients have done in the past, or how well the case patient was expected to do in the absence of the new treatment. The hope generated from apparently positive results inferred from case series can make the inability of later randomized studies to show any effect especially disappointing (box 3-2).

Randomized study designs are unquestionably superior in being able to link cause and effect. No other tool offers the ability to exclude extraneous explanations with such confidence. Although for ethical and logistical reasons a randomized controlled trial (RCT) is not always possible, it is nonetheless well establishcd as the method of choice (88,390,784).

Despite their advantages. the RCT study design is frequently criticized as a basis for drawing conclusions about the effectiveness of medical technologies, Some of the major criticisms are:

- The applications of most technologies have never been tested in RCTs (779). Therefore, if decisions are needed now, other evidence must be used.
- Randomized trials of accepted technologies are difficult to conduct and may be unethical, because many physicians and patients already believe these technologies to be effective.[s]
- Some particular types of interventions. such as psychotherapy and new surgical techniques, have posed challenges to randomized study designs (see e.g., reference 730). Innovations in these areas e.g., laparoscopic surgery) often

[3] Peto has argued that it is equally unethical to prescribe untested therapies to patients even if these therapies are common (604).

BOX 3-2: Lorenzo's Oil

The story of "Lorenzo's oil," a combination of two fatty acids purified from olive oil, has been widely publicized, particularly with the release of a 1993 film dramatizing the efforts of Lorenzo Odone's parents to find a cure for their son's illness It was the Odones who first hypothesized that the mixture might be a therapy for adrenoleukodystrophy (ALD), a rare disease that causes the degeneration of myelln, the protective covering of nerve fibers Although there is also an adult form of ALD, the disease usually strikes boys between the ages of 5 and 10, resulting in death wlthln a few years of onset

In 1984, Augusto and Mlchaela Odone resorted to studying ALD themselves after doctors told them that Lorenzo's illness was untreatable. After extensive research, the Odones tried treating their son with what has become known as Lorenzo's 011, they became convinced that the mixture not only halted the progress of ALD but also caused a partial reversal in Lorenzo's condition

The Odones challenged the medical community to validate their accomplishment with formal clinical trials Despite high hopes for Lorenzos oil, it has not been proven to be effective once the identifying neurologic symptoms of ALD appear in boys (640,641) Also, a recent clinical trial to test Lorenzo's oil for sufferers of the adult form of ALD failed to yield evidence that it was effective (29) Although physicians have been prescribing Lorenzo's 011 for several years in hopes that it wilI hold off the disease in young boys who have not yet developed the symptoms of ALD (577), its therapeutic value is now thought to be much *more* limited than first suggested (640)

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

gain acceptance before they can be identified and studied by those outside the immediate practitioner community.

- The expense and administrative difficulties of establishing and running randomized trials, and the delay before answers are available, makes it impractical to conduct RCTs on every use of every technology.
- Trials are frequently too small to detect any but the largest effects, rendering a finding of "no effect found" difficult to interpret. Although the U.S. research establishment, and particularly the Department of Veterans Affairs (VA) and NIH, has considerable experience in collaborative efforts to conduct very large clinical trials, such trials to date have generally also been expensive.
- Strict trial protocols intended to ensure that any effect found can be attributed to the treatment being studied have often limited the generalizability of RCTs. In the past, for example, most trials of therapies for acute myocardial infarction (heart attack) excluded elderly persons in order to avoid any confounding due to the comorbidities that many elderly persons have (373). Women of childbearing age have also systematically been excluded from many clinical trials on the grounds that some women might be pregnant and the technologies being tested might prove harmful to the fetus (524). The consequence is that the results of many trials cannot be applied with confidence to women or to elderly persons.
- The fact that clinical trials are often conducted in teaching hospitals, by specialists, on highly selected patients according to strict protocols. makes their conclusions suspect when applied by community physicians to their patients in other settings. The surgeon performing a procedure in a clinical trial, for example, may be much more skilled than other surgeons who will later carry out the procedure (335). Pa-

tients who participate in trials are often more motivated or less sick than patients not participating in the trials (434).

A recently completed RCT of intensive insulin therapy for persons with diabetes exemplifies some of these criticisms. The trial successfully confirmed that intensive therapy yielded benefits beyond those of standard therapy (162). However, the intensive therapy regimen requires constant attention and commitment by the patient and has some risks; in the trial, patients were highly motivated, received much clinical attention, and were not representative of the general diabetic population. Although this very expensive trial certainly established that intensive therapy was more efficacious than standard therapy, clinicians are expressing doubts that its findings have much practical use or are attainable under ordinary circumstances (64,619).

As a consequence of these perceived barriers to randomized trials, and the fact that many medical innovations are not subject to regulatory review of their evidence of effectiveness, most interventions never undergo evaluation with RCTs (779). Furthermore, those that do often have not been evaluated with respect to the full range of patients and practitioners that use them. Two responses to this situation have emerged. One response, which has been emphasized by the federal government effectiveness research program, relies on innovative ways to assemble and study the observational data that currently exist in administrative health databases. The other response relies on innovative ways of applying the basic principles of RCTs to make them more adaptable to community settings or to a wider range of interventions. These innovations are described below.

Database Studies

Disillusioned with the lack of useful, relevant information existing for many medical technologies, many health policy makers and researchers embraced ideas for enhanced research use of large health care administrative databases in the late 1980s (247,295,373). Medicare and other health insurers keep computerized records of patient claims, which include such information as patient age, sex, diagnosis, procedures performed, and the charges billed for those services. Advocates of the greater exploitation of these administrative databases as research resources pointed out that observational data from these sources have a number of advantages compared with the collection of data through RCTs. These include:

- the large numbers of patients represented in the data;
- the fact that the data represent ordinary medical practice, rather than very selective patient populations or settings;
- the immediate availability of the data;
- the ability to track patients' health experiences back over time;
- the unobtrusiveness of data collection; and
- the expectation that analyzing existing data should be much cheaper than planning and implementing entire new trials.

On the other hand, researchers such as Byar have criticized this research technique, on the grounds that if the purpose is to compare medical technologies, administrative databases—like all observational data sources--contain biases that often render the results invalid (94,95,299,747, 893). The heavy role these databases play as sources of information in effectiveness research warrants a detailed examination of their uses.

Descriptive uses of administrative databases

The data from insurance claims and hospital discharge databases have long been used to describe various aspects of health care. Medicare claims data, for example, have been used widely to provide estimates of the costs of health care for elderly persons and to examine the characteristics of persons who incur high costs (26,954). Analyses of the direct medical costs of specific illnesses have also drawn heavily on information from administrative data (35,47 1,639).

The use of administrative data by researchers in the 1970s and early 1980s to describe tremendous variations in medical practice patterns across different areas and populations focused attention on the potential power of this tool. The use of admin -

istrative databases for documenting variations in medical practice continues to draw considerable interest, because of the implications of these variations. Administrative data from several sources, for example. have been used in studies that found that African Americans undergo coronary artery bypass surgery at lower rates than do white Americans (287,921,931). Such findings can stimulate a search for ways to improve access to services for particular populations. The Institute of Medicine (IOM) suggests that the existence of documented practice variation is an important criterion for the selection of medical technologies to assess, because it implies that uncertainty or disagreement exists in the field (377). The IOM has also suggested that the existence of such uncertainty improves the likelihood that an assessment can affect clinical practice, although this assertion is open to challenge (see chapters 7 and 8).

Administrative databases have also been useful for describing trends in the use of individual medical technologies of interest. Some studies have used this tool to monitor trends in a technology's applications over time, such as the increasing use of total hip replacement surgery (647). Researchers have also used administrative data to describe changes in the treatment of prostate cancer, and to examine whether guidelines intended to affect treatment patterns were associated with any changes (690).

In a related application, administrative data have proven useful in describing the relationships between new and existing technologies. One recent study, for example, found that a new technique to improve blood flow to the legs—peripheral artery angioplasty—was associated with an increase rather than a decrease in peripheral artery bypass surgery, an older technique with a similar purpose (767).

Examining the health outcomes (e.g., mortality rates) associated with the use of particular procedures has been one of the most publicized uses of administrative databases. Studies of mortality following transurethral prostatectomy and carotid endarterectomy (650,939) have been quoted widely. Published studies using administrative data have also examined the impact of specific procedures on outcomes such as rehospitalization (24) and reoperation (648).

Descriptive studies making use of administrative databases encounter a number of generic problems related to the data sources. One of the most pervasive issues is whether the numerically encoded diagnoses and procedures that appear in administrative databases accurately represent the real circumstances involving that patient. Some well-recognized problems include inaccurately assigned codes, particularly when coding accuracy does not affect payment (158.249,358,365): incomplete codes, particularly for patients with multiple diagnoses and procedures (5 14); and difficulty ascertaining whether a coded condition was actually a pre-existing condition or a consequence of treatment (929).

A second generic issue for studies using administrative data is the actual difference between the population represented in the database and the population of interest. For example, a study describing rates of a procedure among veterans that used administrative data from the Department of Veterans Affairs health care system might underestimate procedure use, since these data would not capture procedures performed in non-VA hospitals.

Describing practice pattern variation, trends in the use of particular technologies, and health outcomes associated with particular technologies and patterns of care is relatively straightforward. The validity of the description depends largely on the extent to which the database examined. and the analysis of it, was appropriate to the question. The researcher must be confident, for example, t hat the database actually represents the entire population of interest, and that the occasions where the technology was applied are reasonably complete and accurately recorded.

But when the descriptive information derived from the database is used to suggest associations between trends and events, new issues arise. Like any such conclusions based on observational data, these arc always subject to a healthy dose of skepticism, because the researcher can never know that the event of interest+. g., a change in technology or the release of a guideline—was the sole, or even a major, cause of the observed trend. Confidence in the conclusion can be strengthened by showing that no other known relevant events occurred, but it rests on negative evidence.

The major caveat that suggestive associations are only that—suggestive—is especially true for one current use of administrative databases: their use in documenting and contrasting health outcomes resulting from a particular procedure or medical intervention. The fundamental task of documenting health outcomes is a purely descriptive one. Once documented, however, those outcomes are frequently used, either implicitly or explicitly, to compare the outcomes—and, by inference, the relative effectiveness—of alternative medical technologies.

Comparative uses of administrative databases

One potential comparative use of administrative databases would be to contrast the outcomes reported in a clinical trial during the investigation phase of a technology, and the outcomes that occur when that technology is in general use. Such comparisons might illuminate differences in the efficacy of a technology under strict conditions and the outcomes borne out in widespread use. It is widely believed that there are substantial differences between outcomes in randomized studies and outcomes in general practice; however, it is surprisingly difficult to find specific documented cases.

A second potential application is to compare the database-derived outcomes of apparently similar patients undergoing alternative treatments. This application, however, raises more serious issues.

As with any observational study, the valid it y of comparative results derived from information in large administrative databases rests heavily on the degree to which the populations being compared are truly equivalent in all relevant respects. Unlike randomized experiments, however, it cannot be assumed that if the groups are large enough, any material characteristics-e. g., those risk factors that make someone more or less likely to do well after a procedure—are likely to be evenly divided between groups.

As Byar observed, “in medicine, the doctor chooses the therapy precisely in order to affect the outcome” (97). Patients’ medical and other characteristics are generally expected to differ among groups receiving different therapies. To exclude these differences as reasons for different outcomes among patients receiving different therapies, researchers may examine the data to see if they can detect known risk factors. In their analysis, the researchers then “adjust” the results to account for the different distribution of these risk factors across the study populations. The degree to which the populations being compared are equivalent, and the analytic results valid, thus depends heavily on whether the researchers know all of the risk factors that might affect the results and can identify them accurately in the data.

Where identified differences in the outcomes of apparently comparable groups are very large, the differences are probably real, although the real differences may not actually be as large as the apparent ones. Probably the most striking example of this to date is a recent finding regarding outcomes in patients who have undergone cataract surgery. As a part of that procedure, some patients also undergo posterior capsulotomy, an additional optional procedure sometimes done to prevent the future development of certain visual problems. (Patients who do not undergo capsulotomy at the time of cataract surgery but who later develop the visual problems can have the procedure at that later time.) In their examination of the outcomes associated with cataract surgery, database researchers found that the rate of retinal detachment—a rare but severe complication that can lead to blindness in the eye—was over three times higher in patients who had undergone the additional procedure (39 1). Because the difference was so great, even after accounting for possible differences between patients selected for the adjunct procedure and other patients, the finding was a credible one.

But other outcomes comparisons have proved misleading. An early finding based on database analyses, for example, was that the mortality rates of men who had undergone traditional open surgery for prostate disease were lower than those of men who had undergone transurethral prostatectomy (650). Until then, the transurethral procedure had been considered the safer and less invasive of the two alternatives. Subsequent research confirmed that the less invasive procedure was associated with higher mortality even after adjusting for population risk factors as represented in the database (140). But this research, a more detailed review of the actual medical charts, also revealed that the patients chosen for the transurethral procedure were sicker than those chosen for the open procedure. Thus, patients undergoing the less invasive procedure probably had a higher mortality rate afterwards because physicians tended to refer sicker patients to the procedure that they perceived to be less risky. This tendency could not be identified in the original claims database, making inferences about relative effectiveness of the two procedures based solely on claims data misleading.

Unfortunately, adjusting the data to account for measures of illness severity that are represented in standard administrative databases can sometimes actually make things worse. In one study, researchers adjusted the data to account for differences in a number of secondary conditions across patients, and they discovered that analyses of the adjusted data actually suggested that patients with serious illnesses had *higher* survival rates than patients without those illnesses. At the time, the researchers speculated that the reason for this anomalous finding was the limitations to the number of diagnosis codes that could be entered on a discharge abstract (leading certain diagnoses that are often secondary to be mentioned only when there were no more important diagnoses to be coded) (395). But a later study found that the same kinds of anomalous findings occurred even when there were no restrictions on the number of codes that could be included (363).

To test more directly whether it is possible to use administrative data to make valid comparisons among technologies, several researchers have compared the results of observational studies, including administrative database analyses, with the results of clinical trials.

A recent example that is still the subject of some controversy was a study of the drug lidocaine, in which researchers hypothesized that the drug would help prevent deaths in people with myocardial infarction. An observational study found that prophylactic administration of lidocaine was beneficial in this population (356). Yet subsequent randomized trials, and a meta-analysis synthesizing those results, were unable to find any effect (346). This example is particularly interesting because the researchers in the observational study were especially careful to use stringent entry criteria, a well-defined endpoint, and adjustment for differences in risk of the endpoint. Although the trials, and even the meta-analysis, were not powerful enough to detect small differences in the subcategory of deaths most likely to be preventable with lidocaine, this comparison certainly raises the question of whether even rigorously conducted observational studies can be relied on to give valid answers.

Three important examples exist of observational studies whose results were confirmed in randomized trials. One study compared the results of a clinical trial of tonsillectomy in which children were randomized with the results of a study of children whose parents refused to participate in the trial, but who were followed observationally. The results of the randomized and nonrandomized portions of the study were indistinguishable (584).

A second study compared two technologies for coronary artery disease (coronary artery bypass grafting and medical management) using information in the Duke Database for Cardiovascular Disease. It identified people in this database who would have been eligible for each of the previous randomized trials of this topic, and predicted what the survival for these patients would have been if they had all undergone medical management, and if they had all undergone surgery. These survival curves were then compared with the actual survival of participants in each arm of the trials. The differences between the trial results and the database analysis results were remarkably small and within the range that would have been expected simply due to random variation (349).

The third study examined the effectiveness of beta-blocking drugs after heart attack, based on observational data on patients in a particular hospital. The researchers compared observational data with results from a specific randomized trial of the drug. The results were in agreement, not only in the direction of benefit (the drugs were found effective) but in the approximate magnitude of the benefit (357).

Thus, it is clear that it is possible to obtain valid results from observational database analyses comparing technologies. It is also clear that it is possible to get invalid results. One interpretation of the lidocaine study is that "sometimes nonrandomized studies will tell you the right answer, sometimes the wrong answer, and there is no way to tell the difference without an RCT to determine the 'true' answer" (929). Unfortunately, unlike randomized trials, multiple database analyses with similar results do not necessarily raise the level of confidence that the answer is the true one, because the same unknown bias—an unknown but important risk factor, for example—may pervade all the analyses.

The finding that differences in outcomes between two groups are large lends more validity to the findings of a database analysis, because unless unknown biases are also very large it is likely that the direction, if not the magnitude, of the finding is correct. For studies in which the expected differences in outcomes are smaller, the validity of this technique for making direct comparisons is much more questionable.

There are some factors that seem to increase the chance that the results of a comparative database analysis under these circumstances will be valid. They include:

1. Detailed and pretested knowledge of the **risk factors relevant to predicting an outcome.** Hlatky and colleagues, who performed the study of observationally based versus experimentally based results in coronary artery disease management, point out that the salient risk factors for death due to coronary artery disease are well studied (929). Where modelers can predict risk of death better than most clinicians, the effect of patient selection bias in who received what therapy becomes less important.
2. Access **to sufficient clinical data in the database to detect risk factors such as secondary chronic illnesses.** The analysis of prostate surgery based solely on claims data would have been problematic even if all relevant risk factors had been known, because those factors were not adequately represented in the data.
3. **Great care in designing the database study—i.e.,** ensuring that the populations, encounters, and procedures being measured actually represent what the researchers want to measure. This factor by no means assures validity, as the lidocaine example shows, but it is difficult to believe that a study could be valid without it.

These factors exist together for relatively few conditions. In addition, these factors together still cannot guarantee validity, although they increase its likelihood. Some researchers have pointed out that even in administrative databases that are supplemented with added clinical data, it is difficult to answer questions that were not formulated carefully before data collection began (929).

Currently, considerable effort is being made to address the second of the three factors above: the need for richer databases. New directions include combining and augmenting existing databases to produce much richer sources of information. The Health Care Financing Administration (HCFA) and the National Cancer Institute, for example, are collaborating in an effort to merge Medicare claims data with cancer registry data (849). HCFA is also pilot-testing a study that will augment existing Medicare claims data with survey data on health statutes of beneficiaries (766). Combining Medicare data with data from other payers (e.g., the VA. private insurers) and with research-related data, to gain a more complete record of beneficiaries health care experience, is also an area of interest (799).

Another use of administrative and other descriptive databases, with broad potential application, is the use of a database as a sampling frame from which to draw patients for a prospective study. Medicare databases. for example, include data on nearly all elderly individuals, making a random sample drawn from it a good representation of persons in this category. Administrative data can also include information that can be used to focus the selection of individuals for a study. A study of the quality of medical care after the implementation of Medicare's prospective payment system for hospitals, for instance, used claims data to identify patients with target conditions discharged from hospitals before and after the payment system was put in place (406). Hospitals' individual computerized billing records have been used to identify appropriate persons for studies of specific conditions (702). Large administrative databases make particularly useful sampling frames for case-control studies in which cases are rare and difficult to identify through other means (396,397).

Finally, one of the most important contributions of analyses of large administrative databases may be to illuminate uncertainty and provide a focus for discussion of its resolution. This may have been the central benefit of the prostatectomy finding. Despite the fact that the actual comparative outcomes of the two prostate procedures were misleading, they highlighted the degree of uncertainty in the field. Indeed. to some degree they created it by convincing practitioners that the presumed benefits of the less invasive procedure were not obvious, and by pointing out the degree of variation in practice patterns for the two procedures.

Innovations in Randomized Trials: Large, Simple Trials

One experimental technique successfully applied to overcome some of the problems of traditional randomized trials is the large, simple randomized trial. The fundamental characteristics of such a trial are:

1. it enrolls a very large number of patients. enabling it to detect even very small differences between treated groups with confidence; and
2. it is very simple in design. requiring data collection on only a few significant endpoints (604,95 1),

One of the best known trials ever conducted was essentially a large, simple trial, although it was not labeled as such. Forty years ago, the National Foundation for Infantile Paralysis recruited a team of physicians and public health researchers to mount a huge trial testing the efficacy of the Salk polio vaccine (266). In the spring and summer of 1954, the vaccine was administered to over 200,000 U.S. schoolchildren, with an additional 200,000 receiving a placebo injection. The outcome measured was simply the rate of hospitaliza - tions for polio in the test areas, Over the course of only a few months. the trial demonstrated the effectiveness of the vaccine in preventing a serious and disabling disease.

In the case of the polio vaccine, the size and simplicity of the trial design were to a great extent dictated by the urgency of the public health problem. Nonetheless, the trial remains a convincing demonstration of the potential power of the large, simple trial technique. It implemented many of the principles of this technique that have only more recently been formally described.

The modern prototype of the large, simple trial was the original ISIS (International Study of Infarct Survival) project, the first of a series of collaborative trials testing therapies to treat acute myocardial infarction (heart attack). The first ISIS trial, ISIS- 1, began in 1981 with the goal of examining the effects on mortality of the intrave-

BOX 3-3: The Thrombolytic Trials: GISSI, ISIS, and GUSTO[1]

Following the success of ISIS-1 in determining the effectiveness of atenolol administered after myocardial infarction, cardiovascular researchers turned the spotlight on thrombolytic drugs, a major area of controversy in the field One thrombolytic drug, streptokinase, has existed for years, but until the mid-1980s it saw relatively little use The 20 randomized trials that had previously examined the efficacy of this drug had shown conflicting results, due to the fact that the effect of the drug was modest (a reduction in death from heart attack of about 10 to 30 percent) and the individual trials were fairly small, enrolling only a few hundred patients.

The first of the thrombolytic trials, GISSI-1, compared streptokinase with “usual treatment. ” It recruited over 11,000 myocardial infarction patients in Italy and included as participating centers nearly 90 percent of that country’s coronary care units. It ultimately documented a reduction in mortality of about 18 percent associated with the use of the drug (309).

Following the GISSI trial, ISIS-2 again used the large, simple trial design, with the participation of hospitals in 16 countries, to examine the relative effects on survival of aspirin, streptokinase, and a combination of both The trial enrolled over 17,000 patients over a three-year period and demonstrated an incremental improvement when both drugs were used in combination (385).

The emergence of new, expensive, bioengineered thrombolytic drugs on the market led both the **ISIS** and GISSI collaborative groups to conduct additional trials in the second half of the 1980s In ISIS-3, over 41,000 patients in 17 countries were randomized to a head-to-head comparison of three different thrombolytic drugs streptokinase, TPA, and APSAC.[2] In addition, half the patients

[1] The GISSI and ISIS collaborations have conducted multiple trials, which are distinguished by abbreviations and numbers (e g , GISSI-2 is the second trial conducted by the GISSI collaborators)

[2] TPA and APSAC are relatively new bioengineered drugs

nous infusion of atenolol (a beta-blocking drug) immediately after the hospitalization of a patient with suspected myocardial infarction (384). Previous evidence on beta-blocker drugs suggested a net benefit but was not conclusive. Although the expected effect of the treatment was small—perhaps a 10-or 20-percent reduction in mortality-it would be important if documented. The total number of deaths from myocardial infarction is large, so even a small reduction in mortality rates would translate into many lives saved.

To detect such a modest effect, however, the researchers calculated that they might need to enroll up to 20,000 patients (384). They formed a network of 245 hospitals in 14 countries to enroll patients. To encourage the participation of these centers in the trial, the trial organizers specifically designed the trial procedures to include very simple entry criteria, treatments, and follow-up. Entry into the trial was based on only a few specific patient characteristics, and randomization occurred over the telephone. Outcome measures were primarily in-hospital and post-hospital mortality. Ultimately, over 16,000 patients were enrolled, and the trial did indeed detect a statistical y significant reduction of about 15 percent in deaths among patients treated with the drug (384).

The success of ISIS-1 led to the use of this basic trial design in a series of additional collaborative trials examining the effectiveness of thrombolytic drugs, administered shortly after the onset of myocardial infarction to break up the blood clot (thrombus) blocking blood flow to the heart (box 3-3). The purpose of these trials was to establish,

BOX 3-3 continued: The Thrombolytic Trials: GISSI, ISIS, and GUSTO[1]

were assigned to receive the anticoagulant drug heparin, while the other half received a placebo (in addition to the thrombolytic drug they received) (386), Simultaneously, GISSI-2 compared streptokinase and TPA, again with a secondary test on half of all patients with heparin as adjunct therapy (310) Both trials showed that although TPA reduced later heart attacks more than streptokinase it also resulted in an Increased risk of stroke Ieading to an insignificant difference in overall mortality between two drugs of greatly dlffering costs

The GUSTO trial was undertaken after both ISIS-3 and GISSI-2 failed to show any slgnificant benefits from TPA over streptokinase TPA advocates hypothesized that the lack of apparent effect was due to the fact that TPA had not been administered in the most effective fashion, rapidly and in conjunction with Intravenous heparin (the previous trials had used subcutaneous heparin) (314) GUSTO did show an advantage to TPA under these circumstances (314), although the trial IS still being debated in the clinical community (450,637)

ISIS-4 and GISSI-3, both completed in 1993, tested additional promising treatments for acute myocardial infarction oral nitrate, oral converting enzyme inhibitors, and Intravenous magnesium, alone and in combination They found that although enzyme inhibitors dld lower the mortality rate from heart attack, the other therapies had no clear effect (either on the overall population or on subpopulations of women and elderly persons) (31 1,381)

KEY APSAC anisoylated plasminogen-activator complex GISSI - The Gruppo Italiano per 10 Studio della Streptochinasi Nell Infarto Miocardio GUSTO Global Utilization of Streptokinase and Tissue Plasminogen Activator for Occluded Coronary Arteries ISIS International Stucy of Infarct Survival TPA tissue-type plasminogen activator

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

first. whether such drugs do in fact reduce deaths from heart attacks; second, whether one thrombolytic drug is more effective than another: and third, whether the administrate ion of adjunct drugs such as heparin (an anticoagulant) improves the effects of thrombolytic drugs.

Implications and limitations

The ISIS trials have provided solid support for two guiding principles of large, simple trials. First, they demonstrated that modest but clinical] y important treatment effects could indeed be detected with confidence if sample size was sufficiently largc. In the case of thrombolytic therapy for myocardial infarction, for example, even a randomized trial of 2,000 patients would not have been sufficient to reliably detect a 20-percent reduction in short-term mortality. This consideration becomes increasingly important as adjunctive therapies need to be tested—for example, heparin in addition to a thrombolytic drug, Each incremental advance in treatment can be expected to have only a modest absolute impact on an already reduced mortality rate. Head-to-head comparisons of treatments for their comparative efficacy and differences in side effects also require especially large sample sizes (92).

Second, the ISIS trials demonstrated (hat highly simplified procedures and data requirements can induce many health care providers to participate in the trial, enabling the enrollment of the large number of patients needed. The broad trial participation by providers and broad patient entry criteria, in turn, ensured that the results of these trials would have direct relevance to the broad spectrum of acute heart attack patients seen in ordinary clinical practice.

Yusuf and colleagues have argued that simplifying the trial design is not merely a poor, second-best solution where very large sample sizes are needed. They maintain that the most "important" effective treatments are those that are broadly applicable and practical (95 1). Such treatments are preferable to equivalently effective treatments that are highly complex. Because the most widely practicable treatments are often those that are fairly simple, the clinical trial protocols testing these treatments can be relatively simple, too. And simple trial protocols can be implemented without undue burden on community hospitals and practitioners, the very settings where most broadly applicable treatments take place (951).

If data collection is limited to major endpoints—those most likely to directly affect physicians' and patients' decisions regarding treatment—the trial protocol can be kept simple (95 1). In the case of thrombolytic therapy, the use of a surrogate endpoint-e. g., the destruction of the clot blocking blood flow to the heart—is actually a poor indicator of the likelihood of short-term mortality, which is the endpoint that is probably the most important to patients and their physicians.

Although patients' characteristics span a wide spectrum, and those characteristics can affect the outcome of treatment, treatment effects are nonetheless expected to be largely in the same general direction. The magnitude of the effect maybe different in patients with different characteristics, but in general all groups of patients would still be expected to have a reduction in mortality (or whatever change in major outcome is being measured) (95 1). Therefore, entry criteria in a very large trial measuring a few major endpoints can be very broad without sacrificing the validity of the results. An additional strength of trials using broad entry criteria is that their findings are of direct relevance to the broad spectrum of patients seen in ordinary clinical practice.

A very important practical strength of large, simple trials is that, because the trial protocol is kept very simple, the per-patient cost of conducting the trial can be kept relatively low. For example, in a relatively traditional clinical trial begun in 1977, the Beta-Blocker Heart Attack Trial, researchers randomized 3,800 patients at a cost of $20 million (71), or an inflation-adjusted per-patient cost of over $11,000 (92). In contrast, an ongoing trial testing the drug digitalis among patients with congestive heart failure, which employs a simplified protocol, has a total budget of $16 million and has randomized 7,790 patients, for an inflation-adjusted cost of slightly over $2,000 per patient **(92). A** large, simple trial of aspirin and beta-carotene in healthy men, the Physicians' Health Study, has had annual costs of approximately $80 per participant for the first five years of the trial (92).

A major potential limitation of large, simple trials is the other face of one of its defining characteristics: the need for simplicity in design. Corresponding to the need for simplicity is the need to collect data on only a very few critical patient characteristics and outcomes, such as mortality. The minimal data collection burden is what enables a large trial to be conducted at low cost and in community settings, but it also means that the richness of detail provided by many traditional trials is lacking. Although proponents of large, simple trials argue that most of these details are of far less importance and are therefore unnecessary to collect anyway, many U.S. researchers may be uncomfortable with their absence. A simple trial, for example, would not collect detailed information on quality of life effects, or background biochemical information from detailed and intensive laboratory tests that could be used in other aspects of research into the mechanisms of disease.

A less discussed but equally important limitation of large, simple trials in the United States is that they depend on a committed infrastructure of community health care providers, many of whom at present have had little experience in participating in clinical research. Although one of the guiding principles of large, simple trials is to minimize the number of physician encounters and tests related solely to the research protocol so that the research and data collection burden on providers is kept light, investigators at each center nonetheless must be organized and trained. Furthermore, the providers must be reassured that patient care costs associated with the research will be covered by payers. If participating providers cannot be assured of recouping these costs, either from private health insurers or from the research sponsors, they are unlikely to stay committed to the project.

One of the benefits of creating a broad communit y research infrastructure is that it can be used repeatedly in future studies. Both the ISIS and GISSI collaborative groups have been able to perform repeated studies building off of their initial network of participating centers.

Current applications

Despite their ability to address some of the criticisms of traditional RCTs and their impressive track record in the area of treatments for heart disease, large, simple trials are still used relatively infrequently in the United States. One of the areas where they might find potential application is in primary care and preventive services, where an intervention often must be applied to a very large population in order to see the ultimate effects on major morbidity and mortality. In fact, one of the few prominent U.S. examples of a large simple trial is the Physicians' Health Study, which is examining the preventive effects of taking low-dose aspirin and beta-carotene regularly (box 3-4).

One recently begun treatment trial that draws on some of the principles of large, simple trials is testing the use of digitalis to treat congestive heart failure, a condition in which the contractions of the heart become progressively weaker. Digitalis is a commonly prescribed and longstanding drug used to treat this condition, but the small trials conducted on this therapy have yielded inconsistent results on mortality. A large multicenter trial with a relatively simple trial protocol began in 1991 with funding from the National Heart, Lung, and Blood Institute and should provide definitive findings on the net mortality effect of digitalis (92).

Several investigators have suggested that identification of the most useful treatments for AIDS might be efficiently achieved through the conduct of large, simple trials, because most therapies are likely to have modest rather than overwhelming effects (96,98,22 1,223,707). As with the incremental addition of new therapies to treat myocardial infarction, each new therapy must show equal or greater effectiveness in relation to an expanding array of standard therapies (92). And unlike the case with many other conditions. an established network of community physicians willing to participate in research studies of AIDS treatments already exists (22 1,223).

To address the need for more extensive data on some aspects of the treatment tested, some researchers have suggested that selected participating sites (e.g., academic clinical centers) could augment the basic data collection with additional, more detailed data gathering. This strategy was used successfully in a trial of the effectiveness of routine fetal movement counting in pregnancy. which randomized over 68,000 women in a variant of the large, simple trial design (293). Although the primary outcome in the trial was a simple one—perinatal mortality—researchers also gathered more detailed data (e.g., on psychosocial effects) from a subset of the women participating in the trial. A hybrid approach such as this might be particularly appropriate for AIDS treatment, where the rapid development of new experimental treatments means there is frequently much less long-term experience with a drug's toxicity or other effects than is often the case with agents being tested in large, simple trials (92).

Researchers have also proposed that factorial designs might be productive in large trials of AIDS treatments; for example, one arm of the trial could compare two antiretroviral drugs, while another compares treatments to prevent occurrence of opportunistic infections (92). Other uses of large, simple trials in AIDS treatments might be to test the effects of different dosages of particular drugs (221) and to compare new antiretroviral drugs against existing therapies (707).

Even trials with very large sample sizes do not always provide unambiguous answers when the differences being measured are very small. The question of the relative effectiveness of TPA and

BOX 3-4: The Physicians' Health Study

The Physicians' Health Study, an ongoing randomized, double-blind, placebo-controlled trial, began in 1982. The goal of the trial was to test simultaneously the effect of low-dose aspirin on cardiovascular disease and the effect of beta-carotene on cancer risk, among a population of apparently healthy U.S. male physicians. Trial participants were randomly assigned to one of four treatment groups aspirin alone, beta-carotene alone, both active agents, or both placebos.

The entire study—including dispensing study medications and collecting annual followup data on compliance and health outcomes—was conducted by mail. Physicians were chosen as study participants because they were presumed to be accurate reporters of their own health, and they were presumed more likely than other populations to comply with the necessary daily pill-taking regimen for an extended period of time. Self-reported compliance was tested in a subset of participants, and self-reported outcomes were confirmed against medical records.

The initial assembly of the study population was more involved that of many large, simple trials in disease treatment However, once participants were randomized, the trial procedures and followup in the study were highly streamlined The annual cost per participant is about $80 per year. After 10 years, compliance rates are over 80 percent, morbidity followup is over 95 percent, and no participants have been lost to mortality surveillance.

In 1988, the aspirin component of the trial was terminated prematurely due to the emergence of a statistically extreme 44 percent reduction in the risk of first myocardial infarction among those assigned to aspirin. At that time, there were insufficient numbers of total cardiovascular deaths—the trials' primary endpoint—to permit reliable assessment of aspirin's effect on this outcome, but the extremely low mortality rate precluded informative results until at least the year 2000 There were also insufficient numbers of strokes to permit reliable assessment of aspirin's effect on this outcome. Although aspirin may prevent strokes that result from blockage of arteries to the brain, because this drug increases the tendency to bleed, it is possible that its use increases the risk of the much less common, but clinically more severe, strokes that result from bleeding into the brain These questions are currently being addressed in the Women's Health Study (below).

One way in which the Physicians' Health Study diverges from the principles of large, simple trials as articulated by Yusuf et al. (951) is in its relatively homogeneous study population (male physicians aged 40 through 84). This homogeneity reduces generalizability (e. g., to women). In this case, homogeneity was purposefully selected to ensure valid results within an acceptable time frame, and the direction of effect is expected to apply to a more general population even if the exact balance of benefits and risks differs.

Low-dose aspirin in women is being tested in a separate trial, the Women's Health Study, in which approximately 40,000 female health professionals are being enrolled to evaluate the balance of risks and benefits of low-dose aspirin, beta-carotene, and vitamin E in cardiovascular disease and cancer.

SOURCE Adapted from J E Buring, M A Jonas, and C H Hennekens,"Large and Simple Randomized Trials Strengths, Successes and Limitations," paper prepared under contract to the Off Ice of Technology Assessment, forthcoming 1994 streptokinase, two thrombolytic drugs, is still being energetically debated in the U.S. academic community despite the findings of several very large trials comparing the two drugs (see box 3-3). Controversies such as this one suggest not that RCTs are inadequate, but that some differences may be so small that factors other than relative average effectiveness must be the basis for choosing between two interventions.

Firms Trials

Another intriguing application of the randomized trial design is the firms trial, in which patients are randomized among entire clinics or other institutional settings (105,541 ,564,863). Neuhauser (563) describes firms research as resting on three underlying basic concepts:

1. **Parallel providers of care.** Patients are assigned to one of several providers, who may be anything from a single physician to an entire hospital. Neuhauser points out that systematic assignment to parallel providers is not unique to firms research: existing examples include medical societies' referring patients inquiring about physicians to the next physician on their lists (to equalize referrals), and Boston City Hospital's historical assignment of new patients in rotation to the Boston University, Harvard, or Tufts teaching services (323).
2. Ongoing random assignment of patients to these parallel providers. At one hospital in Cleveland, this literally means the assignment of patients by randomly generated numbers, to ensure that the assignments are fairly distributed. Similarly, all new staff and attending (i.e., patients' personal) physicians are also randomly assigned. Once assigned, however, staff remain with their team to permit patient/provider continuity.
3, Continuous evaluation and improvement. "A change can be made in the way one provider or firm provides care, leaving the other as is. Differences in care can be observed. If the change is favorable, then it can be implemented by all firms. This becomes the new platform of care, and the next change can be started in the same way" (563). Because the firms are parallel in structure, and the patients are randomly assigned, any change in outcome can be reasonably attributed to the change in care, rather than to other patient or provider characteristics.

The idea of firms trials was first implemented at the Cleveland Metropolitan General Hospital, which began randomizing patients to care settings in 1980 (133,891). The hospital set up an experimental clinic to which relevant patients were randomly assigned when they came to the hospital for their outpatient care. Patients not referred to the experimental clinic received their care in an equivalent clinic providing usual care (564). Each clinic operated as an independent firm-hence the name of the technique.

Advantages

The power of firms trials is that not only changes in specific therapies but changes in the processes of care can be evaluated, taking advantage of all of the design strengths and statistical validity of randomized controlled trials. The technique is particularly amenable to studies of educational interventions and health delivery changes, and in fact many of its applications have been in those two areas.

Two major advantages of firms trials are that the basic structure is always in place, with random patient assignment happening continuously; and that the intervention itself is carried out as part of a patient's ordinary care. The consequence of these two features is that the incremental costs of conducting a study of a particular intervention are very low. Researchers cite the cost of one randomized trial testing a change in computer-based feedback to house staff, for example, as less than $1,000 (342,561). Neuhauser points out, rather colorfully, that the cost of writing up the study was the largest component cost of conducting it (563).

Limitations

A disadvantage of the firms approach is that hospitals or physician practices with relatively few patients cannot realistically maintain truly parallel providers and still assign enough patients to each provider to permit statistically valid conclusions. Neuhauser notes, for example, that existing firms systems are mostly in general internal medicine, and that there are fewer pediatric care settings with enough patients to have parallel providers (563). Similarly, trials within a single institution must generally focus on common conditions or processes in order to keep sample sizes large enough for valid results (105).

Firms trials must also grapple with methodological issues that affect the validity of their results. A crucial concern is the potential for the clinics, presumed to be equivalent due to similar structure and ongoing randomization of patients and physicians, to become less equivalent over time. Problems could occur if, for example, patients in one clinic had a higher drop-out rate than another over time, or if staff had differences in expertise that was relevant to the topic of the trial (105,155).

As they are other clinical studies, cross-firm *contamination* and the *Hawthorne effect* are ongoing methodological issues in firms trials (105). Staff from the clinic in which the intervention is being introduced, for example, might discuss it with staff from other clinics, leading to changed behavior in the "control" clinics as well. Or, the simple fact that staff in one clinic know that an experiment is ongoing may lead them to change their behavior in ways that affect the results.

Certain kinds of trials are not well suited to firms research, at least as it is currently carried out. A clinical trial that requires the presence of a highly specialized physician, for example, would be ill-suited for this design (105).

The most significant limitation of the firms method. however, is probably in the initial difficulties of implementing a system of parallel providers with random physician assignment. Doing so requires, for example, that a randomization procedure be designed and taught to staff (559). It can also require substantial changes in the duties of individual physicians and nurses (194,559), which might often encounter considerable organizational resistance or require adaptations of the firms design. In one hospital, for example, the administrators discovered that it was not possible to randomize individual private physicians to firms without also assigning their partners or covering groups to the same firm. This hospital also found that the number of patients admitted overnight could vary substantially among firms as a normal consequence of random assignment, resulting in resident physicians' complaints about unequal workload (194).

Applications

The number of health providers with established firms research systems is still very small, although it is no longer limited to only one or two unusual institutions. A significant newcomer, for instance, is the Wade Park Veterans Affairs Medical Center in Cleveland (439). Firms systems are in place in at least eight other hospitals, including one other VA hospital and an army medical center (564). The technique has not yet spread outside of academic medical centers, however. probably in part due to the need for ongoing expertise in such fields as biostatistics and epidemiology in order to carry out research (761).

Applications of the firms trial research design to date have been on quite diverse topics, with educational and service delivery topics prevalent. Examples included research on colorectal screening performed by nurse clinicians (104); counseling patients to quit smoking (134); alcohol dependency counseling (286): and the influence of new physician staffing patterns (704).

Future applications for firms research might include research into the effects (on, e.g., costs, patient health outcome, and patient and physician satisfaction) of implementing new clinical practice guidelines. Indeed, some previous trials have been conducted on subjects that examine specifically the effects of interventions to change physician behavior, Hershey and colleagues used the firms system to study the effects of computerized reminders to clinicians on practice change (341). Researchers at the Reganstrief Health Center in Indiana have conducted a series of studies that investigated the effects of various interventions on physicians' test-ordering behavior. They studied, independently, the effects of displaying to the physician a patient's prior test results; the probability that the test would be abnormal; and the patient charges for each test ordered, and discovered that each intervention resulted in a reduction in the number of tests ordered (761).

The general concept successfully tested in firms trials—that prospective randomization based on units larger than the individual can produce valid results—is still relatively rare but has considerable potential. Medical practices, health care plans, communities, or other units are all possibilities. Bakketeig, for example, has suggested testing the effects of different ways of providing prenatal care by using geographical areas as the units of randomization (36). Such trials might be more difficult to carry out in heterogeneous countries such as the United States, but the suggestion serves as a reminder that additional innovations in the use of firms and other variations of controlled clinical research methods might be rewarding.

Secondary Techniques To Synthesize Results

Despite its drawbacks, the medical literature is nothing if not voluminous. As Glass so succinctly stated (about the social science literature) in 1976:

> We face an abundance of information. Our problem is to find the knowledge in the information. We need methods for the order] y summarization of studies so that knowledge can be extracted from the myriad individual researches (28 l).

This section describes two different methods for synthesizing information, each with a different purpose. The first, meta-analysis, is aimed at synthesizing research results in order to draw more powerful and confident conclusions about the state of the world they describe. In its purest form it is a straightforward research tool, but it is also being used as a way of drawing together information for decisionmakers. The second tool, decision analysis, is expressly oriented to the purpose of organizing existing information and assumptions for decisionmaking.

Meta-Analysis and Other Systematic Reviews

The traditional method of synthesizing the results of previous research on a topic is the research review, a discussion and analysis of work to date on the topic of interest. The need for reviews in order to make sense of existing research is great enough to support entire periodicals that publish nothing else (e.g., the publications of Annual Reviews, Inc.).

Despite science's reliance on reviews to synthesize pre-existing results, the traditional narrative review often suffers from a number of weaknesses. Reviewers often do not define clearly the methods they used to identify and select information, they often review the information haphazardly, and they rarely assess the quality of data systematically (550). The consequence is that two researchers reviewing the same topic, and even the same group of studies, can come to diametrically opposed conclusions (457). The burgeoning liter--ature and conflicting reviews have led to increasing use of more systematic reviews of the literature, using structured methods to reduce the opportunities for bias (583). A type of systematic review that has received particular attention recently is meta-analysis—a structured review that incorporates statistical methods to combine the results of the individual primary studies (220,437, 860).

The idea of combining study results quantitatively dates from 1904, when Pearson summarized the relation between inoculation against enteric fever and mortality by calculating the average correlation between those variables across five communities (593). Meta-analysis as a formal discipline, however, arose out of work on the social sciences literature in the 1970s (281,458).

The essential characteristics of a meta-analytic review are that it is systematic and quantitative (473). Meta-analysis requires that the analyst undertake a formal, explicit consideration of what literature will be represented in the review. In addition, the analyst does a quantitative reanalysis of the relevant results of those studies (box 3-5).

BOX 3-5: The Steps of a Meta-Analysis

A meta-analysis is a systematic process (190,213,436,668) that involves nine steps:

1, Defining the research question, The analyst specifies the treatment under investigation, its alternative, the outcome, the study populations, and the quantitative effect measure of interest

2. Defining the admissibility criteria for studies, Examples of possible criteria are that for a study to be considered relevant, it must: be blinded, compare the treatment with a placebo, include elderly persons as study subjects, be written in English, and present results in such a way that the effect measure of interest can be calculated,

3. Searching for relevant studies. This step usually involves a computerized literature search, supplemented by perusing the reference lists of identified articles, abstracts from conferences, and any other informally identified sources.

4 Reviewing the retrieved studies for admissibility, The analyst reviews the identified articles to see if they meet admission criteria, abstracts relevant information, and if necessary re-expresses study results in a standard fashion for subsequent statistical analysis,

5 Assessing the quality of the admissible articles. Objective methods for assessing study quality are frequently used, and published criteria exist (90, 112, 159). Subjective criteria and criteria specially tailored to the research question under investigation have also been employed (51 ,472). Study quality criteria might include, for example, whether the investigators in the study knew which patients received treatment and which received placebo; whether the presentation of data was appropriate, and whether the statistical analyses were appropriate,

6. Correcting for probable bias. If a treatment effect observed in a given study is not an accurate measure of the true treatment effect, the study is biased Certain study designs are associated with known biases, for example, studies in which the investigators know which patients got which

(The broader term, “systematic reviews,” includes meta-analyses, but it also includes reviews that undergo the same process without the quantitative step.)

In most meta-analyses, the quantitative reanalysis involves recalculating individual study results so that the treatment effects are all portrayed in a consistent manner. If some results are portrayed in the original articles as differences (between outcomes of treatment and control groups) and other results portrayed as ratios, for example, the analyst might recalculate them so they are all portrayed as ratios. In addition, the analyst must calculate, for each treatment effect, the precision with which that effect was measured in that study (i.e., the variance around the treatment effect). In general, smaller studies will have larger variances—less precision—than larger studies, because there is a greater chance that random variation is responsible for the observed effect in a small study. The analyst weights each study result according to its precision and then combines all the results in a single calculation to assess the overall treatment effect implied from the studies as a group. Because the meta-analysis of results includes many studies, it has more precision than any individual studies. Thus, a meta-analysis can increase the confidence that a real effect does (or does not) exist, even when individual studies differ in whether they find an effect.

Rather than combining the results of individual studies, some analysts actually combine the raw

BOX 3-5 continued: The Steps of a Meta-Analysis

treatment tend to find a larger treatment effect than do studies in which the Investigators were blinded to the subjects treatment (1 36). Where the size of the likely bias is predictable, the observed treatment effect can be adjusted to account for this bias (209,303,882) Bias correction is often a skipped step, however, because there is often no basis for estimating the likely size of the bias, or even whether it exists

7 Analyzing the data The data analysis step is the one in which the results of the individual component studies are actually reanalyzed. Often the Individual results are displayed graphically A study with a large confidence Iinterval, displayed as a long bar, represents less certainty about the result Combining the results of all the individual studies, weighted by the degree of certainty of each result, gives a new result with very high confidence (i.e , a small confidence Interval) Individual studies may find no effect, but collectively the meta-analysis has the power to detect even very small effects with much greater certainty As part of the data analysis, the meta-analyst also conducts sensitivity analyses (e g , to show whether the results of the meta-analysis depend heavily on one or two particular studies)

8 Assessing publication bias The process of peer review and journal publication can winnow out studies that are considered to be less interesting simply because they found no effect, biasing the published literature in favor of studies that found effects The analyst may take steps to account for this

9 Interpreting the results As with other studies, the final step of a meta-analysis is for the analyst to interpret the results so that their generalizability and Implications for practitioners and researchers are clear

SOURCE Adapted from M P Longnecker Tools for Health Technology Assessment Meta-Analysis paper prepared under contract fo Off Ice of Technology Assessment forthcoming 1994

data from the studies. In such a "pooled analysis." the data are treated as if they are all from a single very large trial, rather than from many independent trials. Pooled analyses and meta-analyses usually give similar results. Pooled analyses facilitate the analysis of subgroups of patients, but they often require the cooperation of many scientists in order to obtain the raw data (473).

Well-done recta-analyses seem to be reasonably well established as reliable and valid. In one assessment of the reliability of this technique, Chalmers and colleagues investigated 20 replicated meta-analyses (111). They found that the differences in meta-analyses of the same research question were "almost always of degree rather than direction" (111). A similar study of meta-analysis reliability by Henry and Wilson (336) found similar results. It also found the recent meta-analyses it assessed to be generally more reliable than the older analyses studied by Chalmers et al.

In addition to comparing the results of meta-analyses with each other. both groups of researchers also compared the results found through meta-analyses with the results of single, large randomized clinical trials. Chalmers and colleagues found agreement between meta-analyses and later large trials for just one of three meta-analyses studied, while the one comparison conducted by Henry and Wilson found that the recta-analysis and the clinical trial results agreed (113,336). Other instances of agreement between individual meta-analyses and clinical trials have also been found (442), although no comprehensive comparative survey of the field has been attempted.

There have, of course, been instances of disagreement among meta-analyses as well. Several researchers who have conducted meta-analyses of the literature on interventions to reduce cholesterol, for example, have concluded that lowering cholesterol does not result in lower mortality rates overall (150,15 1,544,621). In contrast, Law and colleagues concluded from their own meta-analysis that reducing cholesterol levels reduces the risk of ischemic heart disease and does not raise the risk of death from any other cause except stroke, a risk outweighed by the reduction in heart disease deaths (443). Law and his collaborators point out that the different conclusions derive in part from the different outcomes examined (all-cause mortality vs. disease-specific mortality) and differences in the trials selected for analysis.

Issues

The growing acceptance of meta-analysis notwithstanding, meta-analytic results can be controversial (487,889). Three issues are especially hotly debated in the field: the combinability of results from the studies used in a meta-analysis, the importance of procedures to account for publication bias, and the protocol followed by the analyst.

▮ **Combinability.** The justification of meta-analysis is based on the assumption that the component studies all address similar research questions. If the populations, the treatment, the study design, and the outcomes measured in each study are sufficiently similar, then the meta-analysis is analogous to a multicenter clinical trial. Differences in the treatment effect across the component studies—the "centers" of the meta-analysis "trial"-can be presumed largely due to chance (473).

As the component studies of the meta-analysis become less similar, the appropriateness of their joint analysis becomes a matter of judgment and is thus subject to debate. Many of the criticisms of meta-analysis revolve around this specific aspect of the technique (69,251.274,352). For example, one meta-analysis of nonmedical treatments for chronic pain calculated the average effect of one treatment on several different kinds of pain. Some studies included in the meta-analysis measured headache pain, while others measured cancer-related pain. Critics of this meta-analysis charged that the treatment effect might have been very different for headache and cancer pain (352). If this were the case, summarizing across the different types of pain might have obscured the true treatment effects in these different groups.

An equally controversial issue in combinability is whether it is appropriate to combine the results of nonrandomized studies. While evidence from good randomized clinical trials is widely accepted as valid, the validity of results from nonrandomized trials is less clear, and these results are excluded from many meta-analyses (602). Some researchers define meta-analysis to include only analyses of randomized studies (91,928).

For many research questions, however, only data from observational studies are available (46). Dickersin and Berlin (170) point out that a meta-analysis of such studies should be as acceptable as are the studies themselves. The crucial point is that the meta-analysis cannot entirely overcome the deficiencies of the studies on which it draws: if the studies are biased, the results of the meta-analysis will probably be biased, too.

Separating analyses of studies based on study design may be one way of detecting and reducing potential bias. In a meta-analysis of alcohol consumption in relation to risk of breast cancer, the estimate of the effect of alcohol derived from the combined case-control studies was larger than the estimate derived from the combined follow-up studies (472). The authors kept the analyses of the two types of studies separate and argued that, for various reasons, the result based on the combined follow-up studies was the more like] y to represent an unbiased result.

A third issue in combinability of studies arises when the treatment effect found in the component studies varies markedly among studies. Summarizing a single treatment effect across studies under these circumstances is commonly done, but when and how to do it are subjects of debate among researchers (303,603).

A common quantitative method for combining study results is the "random effects model." in which the calculated summary treatment effect is assumed to be an estimate of the average treatment effect in the universe of hypothetical studies with differing treatment effects. The component studies in the meta-analysis are assumed to be a sample from this universe. However, some analysts prefer the "fixed effects model," which assumes that there is a single "real" treatment effect that the different component studies are all attempting to estimate, with varying degrees of success.

In practice. the two methods give similar results when the results of the component studies of a meta-analysis are not too variable. When the results of the component studies do vary substantially, the "'fixed effects" model gives heavy weight to the largest studies, while the "random effects" model gives a result somewhat closer to a simple average (473).

▮ **Publication bias.** *Publication bias* refers to the well-documented fact that studies that get published differ from studies that do not. in ways that are not just related to the quality of the study. Several researchers have shown. for example, that studies with statistically significant results are more likely to be published than other studies (46, 170, 193a). Results perceived as important are also disproportionately likely to be published (172,193a).

One of the characteristics that sets meta-analyses and other systematic reviews apart from traditional narrative literature reviews is the use of explicit rules for including studies in the analysis, and researchers in the field of meta-analysis have carried on a longstanding debate about how to prevent, or adjust for, publication bias. A number of formal statistical methods to detect and assess the extent of publication bias in a meta-analysis have been proposed, but as yet there is no widespread agreement on their use (473).

Some researchers suggest that the solution to this problem is to include all relevant unpublished studies, as well as the published ones, in the meta-analysis (269,952). Most analysts agree that when unpublished studies can be obtained, they should be assessed along with published ones (143). Differences in the results of published and unpublished studies can be assessed by presenting the results of the meta-analysis with and without the unpublished studies (143). Unless registries of all studies undertaken in a given field exist, however. including all unpublished studies may be impractical or impossible (171,890,952).

▮ **Meta-analytic protocol.** In addition to procedures for summarizing treatment effects and for accounting for publication bias, meta-analysis researchers debate a number of other aspects of the meta-analytic process.

Chalmers, for example, argues that the evaluation of studies to be included in the meta-analysis should be blinded (107). He follows a protocol in which the names of the authors, the actual results of the studies, and other study characteristics that might bias the reviewer are hidden during the study selection process. In addition, he recommends that two people independently evaluate the quality of the studies in a meta-analysis (107). Most researchers agree that these procedures should improve the quality of the meta-analysis (473). They come at considerable cost in reviewer time, however, and the degree to which they improve the quality of the analysis has not been shown. Hence, they are often not followed.

Considerable debate also surrounds the issue of how best to judge the quality of the individual studies considered for inclusion in the meta-analysis. One possible option, for example, is to assign each study a numerical score according to how well it meets each of a number of specified indicators of presumed quality (159). Low-scoring studies could be excluded, given a lower weight in the analysis, or analyzed as subgroups. Alternatively, Rubin has suggested that characteristics of component studies be analyzed in relation to the treatment effect, to see if particular characteristics (e,g., study design) strongly affect the result of thc meta-analysis (662). There is no uniform protocol or agreement among analysts regarding the approach to follow. There is agreement, however, that explicit attention to study quality is important (473).

Finally, some researchers specifically advocate a Bayesian approach to meta-analysis (210,288). This approach explicitly incorporates the analyst's own presumptions about the likelihood of certain things, such as whether a particular study to be included might be biased. Its potential advantages include statistical results that are easier to interpret than those of traditional meta-analy - sis, and greater flexibility in combining different types of information in the meta-analysis. Its disadvantages include the need for special software to perform the analyses, the greater susceptibility of the results to debate (because the analyst's assumptions are fundamental components of the analysis), and the fact that even fewer people understand Bayesian methods than understand traditional meta-analysis (473).

Applications

Meta-analysis is unquestionably gaining in popularity, application, and influence, especially in medicine and public health. The number of published meta-analyses on health topics, and articles about meta-analysis, has grown from fewer than 100 in the entire decade prior to 1987 (171) to over 200 in 1989 and well over 300 in 1991 alone (473). Topics range from the usefulness of prophylactic antibiotics for children with recurrent ear infections (933) to the effect of garlic on cholesterol levels (895).

Evidence from meta-analyses has been used to support a number of the federal government's clinical practice guidelines. Because it not only synthesizes existing information but adds value to it, by a more robust estimate of whether a given health care intervention is effective, meta-analysis has become a standard input to the Agency for Health Care Policy and Research's guidelines effort (80 1,802,8 10). The U.S. Preventive Services Task Force also considers meta-analyses as evidence for its recommendations (87 1). According to a member of the Task Force, a meta-analysis is given the same grade of evidence as the grade that would have been applied to its component studies (868). The U.S. Food and Drug Administration allows the results of meta-analyses to help support new drug applications (25 1).

The U.S. General Accounting Office has proposed that meta-analyses be conducted that combine results from randomized clinical trials with those from analyses of large administrative and other databases (882). The purpose of such "cross-design syntheses" is to enable statements about a treatment's effect in the general population (that represented in the database) to be made, while grounding the certainty that the treatment is efficacious in the randomized trial data. Since the essence of this method is a meta-analysis that combines randomized with observational data, it is likely to be controversial, and its validity may be difficult to establish. The feasibility y of the technique is currently being tested by GAO researchers (700).

Recent research suggests that while individual trial populations may differ from the population at large, pooling the results from many trials may give a more representative finding. Klawansky and colleagues examined age-specific survival rates in four clinical trials of breast cancer patients and compared them with U.S. cancer registry data (430). They found wide variability in survival rates across trials, suggesting that individual trials did indeed vary from each other and the general population of breast cancer patients. When the results of the trials were pooled, however, the overall survival rates were quite similar to average survival rates for those age groups in the registry. Thus, the problem of nonrepresentative trial populations may be lessened if the results of multiple trials are combined.

In summary, meta-analysis' applications in areas where multiple randomized trials exist are considerable. Under these circumstances, the technique permits a statement about two treatments' relative effects to be made with considerably more certainty than is possible from the individual trials, and it is a useful tool in assessing effectiveness. It can enable more robust estimates not only of the efficacy of an intervention in a broader population than is enrolled in any one trial, but also on particular subgroups of special interest (e. g., elderly persons) to see if there are differences in effectiveness for those subgroups. Its major limitations are, first, that relatively few researchers are trained in the technique; and second. that the reliability and validity of a meta-analysis—indeed. the ability to do one at all—are limited by the studies that exist for it to draw upon.

Decision Analysis

Decision analysis, a technique for guiding rational decisionmaking under uncertainty (620), has been rapidly gaining in its application in health care effectiveness research and technology assessment. 1(is not a new field, nor is it historically associated with health care, but its applications in this area are spreading rapidly, and some of the implications of those applications have considerable public policy consequences.

The essence of a decision analysis is the systematic. schematic presentation and examination of all of the relevant information for a decision, the points at which decisions or uncertain events occur, and the relative preferences the decisionmaker would have for the array of various possible outcomes of the decision. A simple decision analysis is frequently depicted as a decision '-tree," which branches at points of decision (e.g., surgery vs. no surgery) or uncertainty (e.g., getting a post-surgical infection vs. no infection). (See box 7-4, p. 162, for an example.) The decision analyst records, at each appropriate branch, the best estimate of the probabilities that various outcomes might occur and what those outcomes are.

The use of decision analysis to improve medical decisionmaking was proposed by Lusted in 1971 (478). One of its most familiar (although not necessarily most frequent) health care applications has been decisions about the best course of treatment for a particular individual patient, In this context. decision analysis is primarily a way of laying out the options available to a physician or patient and organizing the information relevant to those options in a way that helps the individual make the decision. It serves as much as a discussion tool as a decision tool.

A physician, for example. can discuss with a patient that person's "preferences" for various possible outcomes of treatment. The patient in this example might assign death a "preference" weight of O, permanent disability a weight of 80, and eventual full health a weight of 100. A decision tree can then be drawn that included the various treatment options and the chances, under each option, that each of those three outcomes would occur. The physician and patient then can multiply the probabilities by the outcomes and arrive at a number representing the net "desirability" of choosing each option.

Applied to decisions for or by groups, some of the characteristics of decision analysis have additional implications. Matchar (496) argues that the greatest benefit of decision analysis as a tool to aid in expert group decisionmaking is its ability to be "a language for the representation of difficult decisions." Unlike many more complex models that can be used to aid decisions (e.g., detailed computer-based simulation models), in a simple decision analysis the information and assumptions can be laid out in a way that makes them easily comprehensible to all members of a group. The group can then discuss the assumptions and the factual information and save their most heated discussions for discussing the importance of relative outcomes, rather than on what assumptions are implicit in the model.

In addition to providing a framework and language for discussion among individuals in a group, decision analysis helps make clear what important information is missing (496). By testing the sensitivity of the results of the analysis to different estimates of preferences for specific outcomes, the group can examine the range of potential implications of its decisions.

The critical controversy over decision analysis, however, relates not to its use in organizing information but to its explicit incorporation of preferences. Calculating which decision path is preferred requires that the decision analyst assign to each possible path not only the outcome of that

BOX 3-6: The Creation and Evolution of the Patient Outcomes Research Teams

The concept of multidisciplinary research teams to study the outcomes of ordinary patient care pre-dates the establishment of the Agency for Health Care Policy and Research (AHCPR). The research program supporting such teams (originally labeled "Patient Outcome Assessment Research Program" grants) was funded by the National Center for Health Services Research (NCHSR) under its Outcomes Research Program The program formally began soliciting grant applications for research teams in 1988 (839a)

The assessment teams to be funded under the program were modeled on the original prostate disease outcome research team (see chapter 2 text) Each team was to focus on a particular medical condition, They were to be composed of 5 to 7 full-time-equivalent professionals and were required to include persons with expertise in at least nine specified subject areas:

- clinical competence in the study subject,
- epidemiology,
- biostatistics,
- research design,
- economics,
- decision analysis,
- survey research,
- data management, and
- research synthesis and meta-analysis (839a).

The assessment teams were also given a very specific charge as to how they should go about their research efforts (839a). They were to:

- conduct literature reviews and research syntheses of the condition,
- use existing "routine" databases to develop hypotheses about practice variation,

path but also the relative preference for each outcome. Incorporation of these preferences, or "utilities," is part of the reason that decision analysis can be a powerful tool. But unlike the example with the individual physician and patient making a decision discussed above, assigning utilities to outcomes in the context of a decision with group-level application requires two crucial assumptions. First, it requires that the preferences being included in the analysis are the correct ones for the decision-e. g., the preferences of the group that the decision will affect. This requirement is simple to state but can be—and has been—the source of considerable disagreement in practice. Second, for the assigned utilities to be valid in the context of the decision they must be truly valid measures of the real preferences for that outcome. This topic is an area of intense empirical research and theoretical debate (see chapter 5).

Thus, decision analysis, although it depends on existing information, has several uses in clinical evaluation. It can be used in both research and clinical practice to display outcome probabilities and preferences of individual patients. It can be used in cost-effectiveness analysis, where cost as well as health outcomes are incorporated into the analysis. And it can be used in clinical practice

BOX 3-6 continued: The Creation and Evolution of the Patient Outcomes Research Teams

- m develop more extensive data sets to examine these hypotheses, includlig the use of primary data gathering through interwews and surveys,
- based on this information, design “carefully-focused epidemiologic or experimental clinical trials that NCHSR WiII consider conducting, ”
- disseminate research findings to physicians, and
- evaluate the impact of the research and dissemination on physician behavior and practice patterns

The first four teams, funded in the fall of 1989, addressed cataracts, myocardial infarction, prostate disease, and back pain (797)

AHCPR, which replaced NCHSR in 1989, funded more outcome research teams in the following years These “Patient Outcomes Research Teams” (PORTS) were to become the centerpiece of the Federal government’s effectiveness initiative As with the pre-AHCPR teams, PORTS were required to conduct literature reviews and synthesis, analyze practice variations and associated patient outcomes, using available data augmented by primary data collection where desired, disseminate research fundings, and evaluate the effects of dissemination (797) By October 1992, a total of 14 PORTS (including the first four) were receiving AHCPR funding (817) No new awards were made in 1993

A recent program announcement re-inviting applications for new PORTS relaxed substantially these methodological requirements placed on the first set of PORTS Teams are still to be interdisciplinary and focus on a specific condition or problem, but they are given more leeway to define for themselves the methods they choose to address the issue (81 1) Six new “PORT-II ‘ grants were awarded under the revised program in 1994

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

guidelines development to structure expert group dicisionmaking. The lattcr two uses are discussed further in chapters 5 and 7.

APPLYING THE TOOLS: THE PATlENT OUTCOMES RESEARCH TEAMS

The PORTS are the “showcase investment" of the federal government’s effort to apply the tools of effectiveness research (800). The government’s goal for them was explicit and ambitious:

> The goals of a PORT project are to identify and analyze the outcomes and costs of current alternative practice patterns in order to determine the best treatment strategy and to develop and test methods for reducing inappropriate variations (797).

The characteristics and methods of PORTS were closely prescribed, by statute and by the terms of the requests for grant applications put forth by AHCPR. In structure. PORTS were to be multidisciplinary, multi site. large-scale, and long-term (800) (box 3-6). All were required, regardless of their particular topic and thrust, to include four components: a comprehensive literature review and synthesis (e.g.. a recta-analysis); an analysis of variations in medical practice and associated patient outcomes (using claims and other sources of data); dissemination of findings about effectivc care: and an evaluation of the effects of dissemination (“to demonstrate methods that encourage voluntary change in provider behavior”). The effectiveness of dissemination was to be judged “in terms of reduced variation in practice patterns, more appropriate use of health care resources, and improvements in patient outcomes” (797).

Cross-cutting methodological issues faced by the PORTS are discussed in periodic meetings of interPORT work groups. These groups grew out of a meeting held shortly after the first four PORT grants were awarded in 1989 (485). They offer chances for PORT investigators to explore common issues and problems and to consult with additional experts about those issues. AHCPR provides formal support for the six groups. which are on the topics of:

- literature review and meta-analysis,
- use of claims data,
- decision modeling,
- outcomes assessment (e.g., measuring quality of life),
- cost of care, and
- dissemination of findings (485).

In addition to their roles in research and information dissemination:

> the clinical recommendations developed by PORTS [were] intended to be a primary source of scientific information for use by independent expert panels in the eventual development of practice guidelines (800).

The agency has several times deliberately assigned the same medical condition to both a PORT and a guidelines panel. This dual attention is in part the result of the priority AHCPR staff have placed for both activities on high-frequency procedures and conditions that have correspondingly high costs to the Medicare program. It also has allowed the guidelines panel to take advantage of previous or concurrent work done by PORT teams. The cataract panel, for example, relied extensively on the review performed by investigators on the cataract PORT team (724a). In that case, the principal investigator of the PORT was also the consulting methodologist to the guideline panel.

In another case, one of the investigators of the prostate disease PORT was actually appointed a member of the guideline panel on the same topic. The influence of the PORT’s work is evident in the emphasis the practice guideline ultimately put on eliciting patient preferences as a crucial determinant of the most effective and appropriate treatment (819).

Although a number of PORTS are much too new to be expected to have any findings yet, it is not too soon for preliminary judgments about what can be expected from this centerpiece of federal effectiveness research. Of the 14 PORTS ongoing as of early 1994, four were in the fifth and final year of their grants (table 3-1). Another seven were in their fourth year, the year they were to begin disseminating the results of their research. The contributions of the PORTS thus far can be judged on three grounds:

1. For the PORTS nearing completion, have the original goals of these projects been met?
2. Aside from those goals, have the PORTS contributed new insights, knowledge, or evidence regarding the effectiveness and cost-effectiveness of medical interventions?
3. Has the work of the PORTS contributed to the infrastructure of health research in other ways (e.g., through advances in methodological techniques)?

Contributions

The PORTS have developed topic-specific expertise in great detail, using the talents of investigators with diverse backgrounds and training. Many of the methodological developments described earlier in this chapter have been in part the contributions of PORTS, particularly in the areas of meta-analysis, administrative database analysis, and the application of measures of patient functioning and quality of life. These contributions are illustrated by some of the specific output of the initial four PORTS (which have had the most time

TABLE 3-1: Current and Planned Patient Outcomes Research Team Projects (PORTS) as of July 1994

Start date	End date	Topic
9/89	8194	Back Pain Outcome Assessment Team *Uhversity of Washington, Seattle, WA*
9/89	8/94	Consequences of Variation in Treatment for Acute Myocardial Infarctlon (AM I) *Harvard Medical School, Boston, MA*
9/89	9/94	Variations in Cataract Management Patient and Economic Outcomes *Johns Hopkins University Bait/more, MD*
9/89	8/94	Assessing Therapies for Benign Prostatic Hypertrophy and Localized Prostate Cancer *Dartmouth College, Hanover, NH*
4/90	3/95	Assessing and Improving Outcomes Total Knee Replacements *Indiana University Indianapolis, IN*
6/90	9/95	Variations in the Management and Outcomes of Diabetes New *England Medical Center, Boston, MA*
7/90	8/95	Outcome Assessment Program in Ischemic Heart Disease *Duke University Durham, NC*
8/90	8/95	Outcome Assessment in Patients with Biliary Tract Disease *University of Pennsylvania, Philadelphia, PA*
9/90	9/95	Analysis of Practices Hip Fracture Repair and Osteoarthritis *University of Mary/and, Baltlmore, MD*
9/90	9/95	Assessment of the Variations and Outcomes of Pneumonia *University of Pittsburgh, Pittsburgh, PA*
9/90	9/95	Variations in Management of Childbirth and Patient Outcomes *The Rand Corporation, Santa Monica, CA*
8/91	8/96	Secondary and Tertiary Prevention of Stroke *Duke Universlty Medical Center, Durham, NC*
9/92	9/97	Schizophrenia Patient Outcomes Research Team *University of Maryland, Baltimore, MD*
9/92	9/97	Patient Outcomes Research Team (PORT) on Low Birthweight in Minority and High-Risk Women *University of Alabama, Blrmingham, AL.*
9/94	8/99	PORT-II For Prostate Disease *Massachusetts General Hospital, Boston, MA*
9/94	9/99	Cure, Costs and Outcomes of Local Breast Cancer *Georgetown University Washington, DC*
8/94	7/99	Cardiac Arrhythmia PORT *Stanford University Palo Alto, CA*
7/94	6/97	Homemade Cereal-Based Dehydration Therapy *Health and Hospitals of the City of Boston, Boston, MA*
7/94	6/99	Dialysis Care: Choices, Outcomes, Costs and Tradeoffs *Johns Hopkins University Baltimore, MD*
9/94	8/98	Value of Medical Testing Prior to Cataract Surgery *Johns Hopkins University Baltimore, MD*

SOURCE U S Department of Health and Human Services, Public Health Service Agency for Health Care Policy and Research Rockville MD 1994 to obtain results) and, to a lesser extent, by more recent PORTS.

Prostate Disease

The PORT on prostate disease is in some ways the most defensible one on which to base conclusions about the contributions of this organizational form of research, because the research team itself actually predated the formation of AHCPR. It was the prototype for the PORT concept, and it has actually had several additional years to carry out its line of research. **The clearest and most widely acknowledged contribution of this team has been its investigation into the role of patient preferences and functional outcomes in treatment decisions for prostate disease (both BPH and prostate cancer).** Several of the insights into the importance of patients' reports discussed above, for example, are based on research by the prostate PORT. Among its specific contributions are:

- highlighting disagreements among physicians in treatment for prostate disease, and identifying the discrepancies between the great increases over time in the number of prostatectomies performed, and the lack of evidence that this treatment was more effective than alternatives (253,480,896,91 1);
- demonstrating the importance of patient self-assessments and preferences in determining the appropriate treatments for BPH and prostate cancer, and the discrepancies between patients' reports, physicians' assessments, and outcomes of treatments reported in the literature for these diseases (42,253,264); and
- convincing both the clinical research and the practicing urology communities that good clinical studies comparing alternative intervention strategies for BPH and prostate cancer are needed.

Back Pain

The prime success of the back pain PORT has been to demonstrate, repeatedly and convincingly, that a major reason for great variation in treatments for back pain is the utter lack of evidence that any one treatment is more effective than any other. In one study, for example, the researchers identified a sevenfold variation in the rate of cervical spine surgery to treat neck pain among counties in the state of Washington. The authors pointed out that this large variability in practice is not at all surprising in light of the lack of clinical evidence that might support any unified approach to the treatment of this problem. The abysmal state of the literature on both the diagnosis and treatment of back pain, and the great need for good studies, is a major theme in a number of publications by the investigators in this PORT (350,769,770).

In other contributions, an interesting physician survey conducted by the back pain PORT demonstrated great variation in the diagnostic tests used for low back pain and showed that the physicians' specialty (e.g., neurology, rheumatology) is strongly associated with the type of diagnostic test ordered (11 9). PORT researchers have also worked with Maine physicians to conduct a prospective study examining the outcomes of disk herniation and stenosis (821).

Acute Myocardial infarction (AMI]

In stark contrast to the back pain PORT, the AMI PORT focused its investigations in an area in which claims data were relatively plentiful and in which data from high-quality comparative studies already existed. **The major contributions of this PORT were its various examinations of the concordance between the evidence regarding effective interventions that exists, and the extent to which those interventions are applied in practice.**

Some of the most powerful findings of this PORT came from its meta-analytic studies and the contrast between treatments shown to be effective based on meta-analyses and their acceptance in the medical community (27,442). The concept and insights possible from the technique of cumulative meta-analysis were a clear contribution of this PORT to the methodological development of the field. Analyses of claims and other administrative databases also proved illuminating; they documented great variations in the rate with which generally effective interventions are performed across gender, age, and racial subgroups (33.34,772). Although these analyses could not fully identify the reasons for these differences, they raised clear questions about whether the processes by which treatment decisions are made are fully equitable.

In conjunction with database analyses that link differences in particular treatments with mortality outcomes. AM I PORT researchers have applied some novel statistical techniques (hierarchical modeling and instrumental variable analyses) (5 15). These techniques have not yet been applied, evaluated, and critiqued by peers in detail, however, so their full contribution towards drawing conclusions about the comparative effectiveness of different technologies cannot yet be assessed.

Cataracts

The PORT examining the effects of cataract surgery in Medicare patients, like the AMI PORT, had the advantage of being able to identify relevant patients and procedures in claims data with fair accuracy. Researchers examined mortality outcomes of elder-l y cataract patients overall (734) and more specific clinical outcomes associated with particular types of procedures (101 ,392, 393). For the most part. this research confirmed previous studies and estimates of complications and outcomes. Other relevant contributions of this PORT have been estimates of the costs of the episode of care surrounding cataract surgery, and a measure of vision function for cataract patients (821).

The chief success of the cataract PORT was its finding, based on claims data analysis, that a particular adjunct procedure (laser capsulotomy) maybe associated with a greatly elevated risk of retinal detachment (391). This complication is a severe one, and although the absolute risk found in the study is small, if confirmed it would imply that performing capsulotomy as a preventive procedure is not necessarily a good idea. The finding is notable because the rarity of the complication would make it very difficult to detect in a clinical setting, and because the magnitude of the increased risk makes it very difficult to dismiss the finding out of hand as an artifact of the method.

Although this is probably the most direct and credible finding of comparative safety and effectiveness based on claims data analysis (from the PORTS or other research), it has not gone unchallenged. Among the criticisms, for example. is the fact that the data do not permit researchers to identify whether the eye suffering retintal detachment was actually the eye that underwent the procedure in question. For this and other reasons, many ophthalmologists apparently do not find the results convincing (724). AHCPR is currently funding a case-control study, conducted by the same researchers, to confirm the results of the claims data analysis (724,821).

Other PORTS

Although the longest of the remaining PORTS have been in existence for only four years, several have reported findings.

- The stroke PORT has reported differences among racial groups in the receipt of technologies to diagnose and treat the disease (575). It has also done extensive work examining the factors that predict outcomes after stroke (511) and the usefulness of decision models in helping expert panels rate the appropriateness of indications for carotid endarterectomy (a major surgical procedure sometimes performed to prevent strokes) (576).
- The PORT studying knee replacement surgery has reviewed the rating systems used in assessing outcomes, with the goal of helping develop and encourage more consistent and valid methods of assessing patients' levels of improvement after surgery (179). They have also confirmed that most patients do consider themselves better off after surgery, are conducting a cohort study of surgery for arthritis of the knee, and are examining the comparative outcomes of surgery in different subpopulations of elderly patients undergoing the procedure (764).
- Members of the pneumonia PORT conducted a prospective follow-up study in which they doc-

umented substantial variations in lengths of hospital stay for pneumonia patients, particularly in low-risk cases (238). PORT researchers have also developed a pneumonia-specific prognostic index that they believe could be a useful tool for clinicians (239).

Limitations and Frustrations

Despite their several notable successes, the PORTS have suffered equally notable disappointments. Most of these are directly related to the limitations of the methods they employed for the objectives they were ostensibly trying to address.

First. **the PORTS have been largely unsuccessful at identifying the most effective treatments among the alternative treatment patterns in existence,** one of the fundamental stated goals for this research investment. The closest successes have probably been the retinal detachment finding of the cataract PORT, the findings regarding the importance of patient preferences in determining treatment appropriateness in the prostate PORT, and the possibility that the techniques of the AMI PORT might produce some credible findings on relative effectiveness of treatments. Nor do there appear to be critical findings on relative effectiveness on the immediate horizon from any of the newer PORTS. It appears to be rare that data from existing claims and other databases, even augmented ones, are sufficiently clear and show differences of a sufficiently large magnitude to be useful in drawing conclusions about relative effectiveness.

Second, **no particular research method has proven universally fruitful; the mandate to use particular research methods has frequently led to inefficient or unproductive lines of research for individual PORTS.** Exhaustive reviews of the literature, for example, have proven a very expensive undertaking for some PORTS. with relatively little to show except to document the poor quality of existing evidence (485,807). Similarly, analyzing variations and outcomes from claims data did not prove particularly useful or productive in some PORTS. In the hip fracture PORT, for example, claims data were not very useful, since nearly all patients with fractured hips are hospitalized, and since when hip surgery (or resurgery) is performed it is not possible to tell which hip was the subject of the operation (807).

Third, **the ability of the PORTS to undertake active dissemination of their findings, and to evaluate their effects on clinical practice patterns, has been very limited.** The first and most obvious reason for this is that the PORTS have offered very little in the way of delineation between appropriate and inappropriate and ineffective practices. Most of the contributions of the PORTS have dealt with insights that can improve the processes of care-e. g., through accommodating more explicitly patients' preferences and leading providers to question the equity of patients' referral to specific treatments-rather than insights into which practices lead to better health outcomes overall.

A second reason is that disseminating information, and setting up a process for evaluating the effects of that dissemination, is a very different activity than the initial research, requiring both new planning and new skills. Overall, the PORT investigators have been dissatisfied (and rightly so) with their ability to perform this function within the constraints of their five-year grants (807). Only the back pain and prostate disease PORTS appear to have fully operative dissemination and evaluation programs, and neither has yet formally presented any results from these studies.

Fourth, **perhaps the most glaring failure of the federal government's investment in the PORTS has been the inability to follow up the detailed examinations of the poor and conflicting evidence justifying current alternative practices with primary research to resolve the questions.** Both the prostate disease and the back pain PORTS, for example, identified specific questions about treatment effectiveness that could only be answered in prospective studies (9 18). For the most part, however, these studies have not been forthcoming. Exceptions include cohort studies by the back pain and knee PORTS, the case-control study now underway to confirm the findings of the cataract PORT, and two prostate disease studies that examine some aspects of the questions raised by the prostate PORT. In contrast to these small exceptions. the uncertainties demonstrated by these PORTS in current practice was enormous. evidence to resolve them was lacking, and explicit attempts by the prostate PORT to initiate a trial directly testing specific questions it raised were unsuccessful. This issue is discussed in greater detail in chapter 4.

Future Plans

Although there has been no formal agency assessment of AHCPR's effectiveness research program in general, or the PORTS in particular, the agency has engaged in some introspective discussion about this approach. AHCPR held a small conference in 1993, at which agency staff, PORT investigators, and other attendees discussed some of the lessons of the PORTS and promising directions for future research teams (807). Among the conclusions of participants were:

- The approaches used by the PORTS had not been universally successful, particularly the overemphasis on claims data analysis and exhaustive literature reviews. More future emphasis on more flexibility in methods, and more small prospective studies (within the limitations of AHCPR's resources) was warranted,
- Evaluating the dissemination of their findings was not something most teams could accomplish within their time. expertise, and financial constraints.
- There was considerable merit to having interdisciplinary research teams gain indepth knowledge and expertise in a particular clinical condition, and for the more successful of the PORTS there might be merit in maintaining these centers of expertise.

AHCPR did not fund any new PORTS in the fall of 1993. Instead. the agency released a new request for application for future PORTS that was great] y changed from the initial PORT solicitation five years earlier and incorporated many of the sentiments expressed at the conference. Among the most notable differences, the new grant announcement stressed the following:

- Conditions affecting mainly children. youth, or nonelderly adults would be given as much priority as conditions affecting the predominantly elderly Medicare population.
- Investigators were encouraged to “design new research strategies, to use new combinations of methods. or to tailor existing methods” in order to obtain evidence for the comparative effectiveness of clinical interventions. Experimental and quasi-experimental research designs were explicitly mentioned.
- The use of secondary sources of data, such as claims data, was not given prominence, and suggestions for the use of administrative data were the more modest possibilities of using them “in identifying cases and controls, estimating costs, or measuring selected outcome s.”
- Researchers were specifically instructed to include women and minorities in study populations.
- There was no mandate that the research teams disseminate their findings and study the effects of that dissemination on changes in clinical practice (811).

The first six new PORT grants (PORT-II) began in the summer of 1994 (table 3-1). Some of these new PORTS follow up, with prospective studies, questions raised by the initial PORTS (300,81 7).

CONCLUSIONS

The ability of patients, providers, and payers to get valid and reliable information on which health care technologies work best, for whom, and under what circumstances, has always been limited. AHCPR was created in 1989 in part to fulfill this need. The establishment of that agency marked not only a commitment to effectiveness research but also an emphasis on particular facets of that research. Thanks in part to the stimulus provided by the federal government’s emphasis on increasing that understanding, the tools now available to enhance our understanding are many and are continuing to be developed and refined. Their applications are likewise growing. In its level of sophistication. the science of evaluating the comparative effectiveness of existing health care interventions has passed from infancy to somewhere in early childhood.

The focus on evaluating the outcomes of health interventions that most matter to patients, and the refinement of tools to achieve this aim, is one of the major contributions of effectiveness research as it has been carried out thus far through the federal effectiveness research initiative. The current debates over which health survey questions and instruments to use for this purpose have not been resolved. Global measures enhance comparability of results across studies, while disease-specific measures offer more opportunity for relevant detail. Brief measures are simpler to use and might enable information on patient functioning and qua] it y of life to be incorporated in studies more widely, while longer measures offer more ability to be sensitive to specific problems. Areas still in need of attention are:

- Continued methodological research into different measures, and different applications of those measures, to understand more fully the advantages and drawbacks of each.
- Development of common measures for the sake of enabling more valid comparisons across studies of the same disease.
- Better collaboration between quality-of-life researchers and researchers conducting comparative clinical studies, so that study results can be more meaningful to more patients. AHCPR and NIH both clearly have much to contribute, yet cross-fertilization between the agencies on this topic has been limited. Many institutes seem to have relatively little interest in the methodological work done at AHCPR; and where there is interest, AHCPR seems to perceive it as interest in that agency's resources rather than real interest in intellectual collaboration.

The analysis of large administrative databases—a tool deliberately emphasized in the federal effectiveness initiative and the mission of the PORTs—has proved quite useful for several specific purposes. Among its important contributions have been its uses in:

- highlighting variations in medical practices and paving the way for serious discussion about the reasons for these differences, including future prospective studies;
- identifying appropriate candidates for primary studies;
- highlighting the differences between medical practices shown to be effective and their use in particular populations of patients, as demonstrated by the research involving data on AMI patients;
- identifying rare adverse events; and
- enriching clinical with administrative data, which offers possibilities for much richer descriptive information on the experiences of patients with particular conditions and undergoing particular treatments.

In contrast, **the notion that the analysis of large administrative databases could address the need for information on the comparative effectiveness of alternative treatments has proved misguided. No** clear, wholly credible finding about the direct effectiveness of one medical practice over another has been derived directly from this research method thus far. Even the finding of the cataract PORT regarding the risk of retinal detachment has not been entirely convincing to clinicians. Other research suggests that there are areas where credible, or at least highly suggestive, findings might be forthcoming, but administrative databases themselves are not the most productive means for determining the comparative effectiveness of most medical technologies and services. Focusing on this research method as a relatively simple, inexpensive first-line tool for answering comparative questions is unwarranted.

Prospective comparative studies, and particularly randomized controlled trials, have been underused in the government's effectiveness initiative. Although they are often considered to be tools applied to medical technologies at an early stage, variations of the randomized clinical trial design can and have been applied to compare two or more existing interventions, and to include broadly representative populations. One of the main contributions of administrative database analysis (o effectiveness research, in fact, has been to highlight uncertainties and—even more importantly-create an environment in which patients and clinicians alike can agree that comparative trials are needed.

An aspect of comparative effectiveness trials largely uncommented on, in either the literature or the health policy debate, is the relationship between effectiveness trials conducted within a committed infrastructure and the goals of continuous quality improvement—a topic very much the subject of current discussion. This relationship is particularly marked in the GISSI large, simple trials. which included most of the coronary care units in Italy. As the trials were completed, units could incorporate the findings, and new trials were begun to achieve the next level of quality improvement. Questions of generalizability of findings were almost irrelevant. since most units and patients participated. Firms trials have accomplished this objective on an institute-specific basis; as an intervention proved effective, it was adopted by the other firms in the institutions and became the new level against which future improvements would be measured.

Large, simple trials seem a particular y promising tool for comparing the effectiveness of some interventions in ordinary practice, where large sample sizes might be necessary but provider participation and funding may require that protocols be kept very simple. The major drawback of this simplicity is that it conflicts with the need to better measure patient-centered outcomes and preferences. Further innovations in trial design might overcome this and other drawbacks+. g., through “nesting” a smaller trial with more detailed data collection in a larger, simpler trial. At least one successful example of such a “-nested” trial in primary care already exists (293). **In general, the experiences of effectiveness research thus far suggest that it is not the rejection of randomized controlled trials but innovations in the design of clinical trials, and greater incorporation of RCTs into ordinary practice, that is needed to improve the level of knowledge about the comparative effectiveness of existing medical interventions.**

The refinement and greater application of meta-analysis and other systematic reviews of the literature is a useful contribution of the effectiveness initiative. The experience of the PORTS suggests that while an insistence on exhaustive literature collection can be both inefficient and unnecessary, systematic reviews nonetheless have been important in highlighting both research and practice deficiencies. When good studies do exist, meta-analysis can also often derive more powerful and convincing statements from the findings of previous research. The greater use of systematic reviews could reduce unnecessary and duplicative research, enable important information already in the literature to gain broader exposure, and reduce inconsistencies among literature reviews. Although the application of meta-analysis to nonrandomized studies has limitations, it also has promise. It would behoove clinicians and health policy makers alike to learn to be able to judge the quality of a meta-analysis at a basic level, and to be able to interpret its results.

Decision analysis is a tool for organizing existing information, incorporating information on effectiveness and outcomes from pre-existing studies and structuring it to help clarify the choices to be made. Like meta-analysis, it can also help point to needed areas of research, where information to make a decision is especially poor and especially important. The power of decision analysis derives from its ability to structure the information needed to make a decision and assess consequences. Distinguishing among the quality of different kinds of studies and other information used in the analysis, however, is the responsibility of the analyst—a feature the users of decision analyses must bear in mind when using decision analysis to compare the outcomes of different alternatives.

The PORTS have been a successful testing ground for developing and applying many of the tools of effectiveness research. They have especially excelled at raising the level of discussion about what is known, and what is not, about the effectiveness of treating particular diseases. They have also contributed to an improved set of measures for assessing the outcomes of therapies for problems such as prostate disease and knee conditions.

In so doing, the PORTS have created a fertile environment for new research on existing medical technologies and services. Attempts to generate new evidence regarding effectiveness using the tools they have emphasized in the past, however, have met with only rare success and point to the limitations of a research model that, at least until now, has emphasized secondary research methods and the use of existing rather than newly generated data.

The inability of the federal government effectiveness research efforts to follow up the questions highlighted by PORT research with comparative clinical trials is one of the signal failures of those efforts. Although the next round of PORTS may include a better balance of research methods, including more comparative prospective studies, these will still be constrained by numbers and resources in the questions they can address. The implications of this and some of the other issues raised by effectiveness research thus far are discussed in the next chapter.