Pew Memorial Trust Policy Synthesis: 7

Unnecessary Surgery

Lucian L. Leape

The extent of unnecessary surgery has been the object of considerable speculation and occasional wild accusation in recent years. Most evidence of the existence of unnecessary surgery, such as information from studies of geographic variations and the results of second surgical opinion programs, is circumstantial. However, results from the few studies that have measured unnecessary surgery directly indicate that for some highly controversial operations the fraction that are unwarranted could be as high as 30 percent. Most unnecessary surgery results from physician uncertainty about the effectiveness of an operation. Elimination of this uncertainty requires more efficient production and dissemination of scientific information about clinical effectiveness. In the absence of adequate data from scientific studies, the use of a consensus of expert opinion, disseminated by means of comprehensive practice guidelines, offers the best opportunity to identify and eliminate unnecessary surgery.

PURPOSE, RELEVANCE, AND AUDIENCE

In 1984, surgeons performed 25.6 million operations, an increase of 5.6 million since 1975 (U.S. Dept. of Commerce 1988). Allegations have been made that as many as 20 percent of operations are unnecessary. If this is true, or even close to true, unwarranted surgery represents a problem of staggering magnitude in terms of needless pain, suffering, and death, as well as a substantial waste of human and financial resources.

The purpose of this synthesis is to assess the extent of unnecessary surgery from evidence recorded in the literature and to make policy

---

This article was commissioned by the Foundation for Health Services Research and supported in part by a grant from the Pew Memorial Trust to the Foundation.

Address correspondence and requests for reprints to Lucian L. Leape, M.D., Adjunct Professor, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115.

recommendations for decreasing the amount of unnecessary surgery. I address the following questions:

1. What is unnecessary surgery?
2. What is the evidence for unnecessary surgery?
3. Why does unnecessary surgery occur?
4. How can unnecessary surgery be reduced?

Unnecessary surgery is an appropriate subject for review for several reasons. It has intrinsic interest for physicians, since it represents bad medicine. "Above all else, do no harm." No ethical surgeon wishes to perform an operation that will not help the patient, and certainly not one that may endanger the patient's life. Physicians have a responsibility to ensure that treatment is appropriate, and most are concerned that their patients perceive them as doing so. If inappropriate and ineffective operations can be identified, suitable measures should be taken to eliminate their use.

The public has an interest, piqued in no small measure by press reports of Medicaid fraud, "tonsil mills," and, in particular, the congressional hearings of 1976 that concluded that more than 2 million unnecessary operations were performed annually, with a toll of nearly 12,000 needless deaths (U.S. Congress 1976). It is in the public interest that surgical therapy be provided when it is appropriate and discouraged, or prevented, when it is not. Quality control mechanisms and reimbursement policies depend on these judgments. They should be made carefully.

Those responsible for the organization and provision of health care and those who pay for it also have a pecuniary interest. Surgery accounts for a substantial fraction of health care costs. If a significant number of operations currently being performed are not indicated, their elimination in the future could significantly reduce health care expenditures.

Since all operations have some element of risk, a useless or ineffective operation is also dangerous. The patient is placed at risk of life or injury without possibility of significant benefit—hence the moral fervor attached to the efforts to eliminate unnecessary surgery. Not only is it a waste of resources, it places the patient in jeopardy. If doctors knowingly engage in conduct that endangers patients' lives for no predictable benefit, then unnecessary surgery is indeed a moral issue of some gravity.

This synthesis should be of value to a number of interested parties. Directors of health care systems need to know which operations and services are effective and which are not, and thus are interested in the efficacy of methods (such as second surgical opinion programs) that

purport to determine that. Legislators need to know which methods improve the efficacy and efficiency of health care for the recipients of the major government-funded programs, Medicare and Medicaid, and legitimately look upon unnecessary surgery as an appropriate target for cost saving. Those who pay for health care, insurers, large companies, and the Health Care Financing Administration (HCFA), likewise have an interest in identifying measures that will decrease expenditures for ineffective care.

DEFINITION OF UNNECESSARY SURGERY

WHAT IS UNNECESSARY SURGERY?

While the term "unnecessary surgery" is widely used and commonly understood, it is ill defined. For example, if a person has an operation that fails to relieve the symptoms for which it was performed, that person might well conclude that the operation was unnecessary, regardless of the medical evidence of the value of the operation. Others confuse "unnecessary" with "discretionary" or "elective." Discretionary operations are those that do not affect physical function but are desired by the patient to improve his or her sense of well-being. Most cosmetic surgery is in this category. "Elective surgery" commonly refers to operations for conditions that pose no immediate threat to life or health and can, therefore, be scheduled at a time of convenience.

A common use of the term unnecessary surgery relates it to frequency of performance. If the number of operations being performed in an area, at a given hospital, or by an individual surgeon is significantly greater than the norm, those responsible may be accused of performing unnecessary surgery. In other cases, the charge of "unnecessary" reflects patient preferences. One person finds a given operation necessary, another does not.

Unfortunately, none of these considerations leads to a definition specific enough to permit collection and analysis of data to determine whether unnecessary surgery is in fact being performed. A more rigorous definition is needed. If it can also be simpler, all the better.

A logical place to start in the quest for a definition of unnecessary surgery is the definition of "necessary." According to Webster's Third New International Dictionary (1976), necessary is defined as something that "must be by reason of the nature of things," "cannot be otherwise," is "determined and fixed and inevitable," or "logically required."

But no operation is "inevitable," or "must be," or is "logically required" in all patients with any given indication. Determination of necessity requires consideration of all of the factors that enter into the decision to recommend an operation. In addition to the efficacy of the operation, these include the nature and stage of the disease, the benefit:risk ratio, the availability and efficacy of nonoperative treatment, and the presence of other disease conditions. While an operation may be "necessary" to save the patient's life, there are situations where such an operation would not be appropriate—as in a patient who is comatose and dying from disseminated cancer, for example.

Most importantly, the decision to have an operation is made by the patient in the context of his own values—how he weighs the perceived benefits of surgery against the costs. These costs are determined by his tolerance of risk, fear of surgery, tolerance of pain or disability, preferred lifestyle, requirements for peace of mind, and how he envisages living the rest of his life. In the last analysis, only the patient can decide if something is "necessary" for him. Clearly, this is a subjective judgment, one that cannot be readily quantified or defined in a manner that permits monitoring or evaluation.

Indeed, Pauly (1979) claims that because of the central role of patient values in the judgment of necessity, the medical profession cannot generate the necessary conceptual apparatus or information that would enable it to arrive at its definition. He defines an operation as necessary if it improves well-being and unnecessary if it makes the individual worse off. Only the patient can make that judgment, and he can do it properly only if he is fully informed. He must know the probable risks and benefits, and then express his own preferences.

Thus, "necessary" is a relative concept in medicine—it depends. In a given situation, an operation may be deemed to be *appropriate* for that patient in the sense that its benefits are generally perceived to exceed its risk; it is not possible, however, to say categorically that the operation is necessary. In apparently identical situations, two individuals may make opposite decisions about whether or not to have an operation. One deems it to be necessary, the other, unnecessary—for him.

If no operation is categorically "necessary," does it then follow that all operations are therefore "unnecessary"? Of course not. While Webster's first definition of unnecessary is "not necessary," the second definition, "useless," is more helpful. Although "necessary" in medicine is always situational, dependent both on the specifics of the clinical situation and on the individual patient's values, "useless" can be an absolute. If an operation is known to be ineffective, that is, if it does not accomplish its claimed objective, then it is useless. Since it will cause some pain

and expose the patient to some risk, its expected utility is negative, and therefore it will always be unnecessary for any patient. Other considerations, such as the stage of the disease, the benefit:risk ratio, the presence of other disease conditions, and patient preferences, are irrelevant. And, unlike "necessary," which depends to a significant degree on patient values, if "unnecessary" is defined as useless or ineffective, it is capable of reproducible and fairly objective determination.

An unnecessary operation, then, is one that is *useless*. On balance it does not benefit the patient. In other words, it does not do what it purports to do, or at the most, carries benefits so small that they are outweighed by the costs in terms of risks, morbidity, disability, and pain. The patient is not better off. This is the definition we will use.

METHODS

SUBJECTS

This synthesis focuses on three areas of study that have been cited as providing evidence of unnecessary surgery: (1) geographic variations in the rates of surgery, (2) second surgical opinion programs, and (3) appropriateness of indications. A sizable body of literature exists concerning these subjects, and information from these studies should be generalizable to the population as a whole. Indeed, many such generalizations and extrapolations have already been made. Inferences will also be drawn from the experience of the Peer Review Organization (PRO) and from comparison of surgical rates in health maintenance organizations with those in fee-for-service practice.

I will not review the evidence of unnecessary surgery from data that can neither be quantified nor generalized to the overall experience. Specifically, this report will not deal with

1. Anecdotal reports of cases of unnecessary surgery. While often dramatic, and in part the reason for public interest in the subject of unnecessary surgery, these types of case reports cannot be quantified in any meaningful way.
2. Reports of Medicaid fraud. It is difficult to evaluate the validity of indications for operation in these cases and impossible to derive any meaningful quantitative extrapolation. In addition, most fraud cases concern deceit in billing rather than performance of unnecessary surgery.

3. Hospital utilization review programs. Little information relating specifically to unnecessary surgery is available. Further, the heterogeneity of these programs and the variety of methods used for review make quantitative analysis and generalization exceedingly difficult. Utilization review was the subject of a previous synthesis (Payne 1987).
4. Malpractice suits and risk-management program information. These data have several limitations that make them unsuitable for estimating the extent of unnecessary surgery. First is the problem of the denominator. It is not known what fraction of cases are represented by those that come to legal action. Second, both sources concentrate on bad outcomes, technical errors, and other problems that do not directly reflect unnecessary surgery. Finally, even in cases of alleged inappropriate surgery, the results of malpractice trials reflect the judgments of lay juries, not a systematic evaluation of scientific evidence.

SEARCH STRATEGY

The reference search began with the standard citation indexes: MEDLINE, HEALTH, NTIS, SOCIAL SCIENCE, and GPOM. English language references from 1966 to 1988 were sought, for (1) operative surgery: standards, indications, second opinion, avoidable, appropriate, inappropriate, unnecessary, quality assurance, operative utilization, mortality, outcomes; (2) geographic variations; (3) surgical decision making; (4) technology assessment: standards, guidelines, assessment, criteria, consensus; (5) patient care management; (6) quality of health care; and (7) delivery of health care: appropriateness.

Additional references were obtained from bibliographies in reviews, from reference lists in the papers that were retrieved, and by consulting with experts. The *Federal Register* was consulted for information on hearings on unnecessary surgery and for standards. Finally, a computer search was conducted of related research by names of authors of several of the classical works.

SELECTION CRITERIA

Titles from the data base searches were examined for those that were possibly relevant to the topics. These references were then retrieved, and those that provided numerical data on surgical operations were analyzed. References from other sources were evaluated the same way.

CRITERIA OF RELEVANCE

Studies that met the entry criteria were evaluated for the validity of the methods used to determine unnecessary surgery. Four questions were asked:

1. Is the method that was used to determine that an operation was ineffective (useless) clearly described? What were the criteria?
2. Is the method of determining that a *given operation* met the criteria of ineffectiveness also clearly described?
3. Are the methods of gathering data described in sufficient detail to permit analysis of the validity of the results? This includes sample selection, methods of measurement, and identification of the control group or "denominator."
4. Are the results generalizable, either to the universe of patients with similar conditions, or to all candidates for surgery generally?

THE EVIDENCE FOR UNNECESSARY SURGERY

GEOGRAPHIC VARIATIONS

History

For several decades, students of the delivery of health care have been fascinated by the accumulating evidence of significant geographic variations in the use of medical services. Early studies reported differences among small areas (counties or hospital service areas), but recent studies have shown impressive differences between large areas (states or regions) as well (Chassin, Brook, et al. 1986b).

From the beginning, geographic variation has been regarded as evidence of unnecessary surgery. The early studies focused on surgical operations, and either stated or implied that the differences in use resulted from overuse in the high-rate areas (Lembcke 1952; Lewis 1969). Subsequent studies have found variations in nonsurgical services as well.

Wide variations have been found. In one of the earliest of the modern studies, Lewis compared utilization of six common surgical procedures by Blue Cross enrollees among 11 health planning regions in the state of Kansas in 1969 (Lewis 1969). Variations in use ranged from

Table 1: Geographic Variations in the Use of Selected Operations by County of Residence

	Operations per 10,000 Persons					
	Vermont 1969			*Maine 1973*		
Operation	*High*	*Avg.*	*Low*	*High*	*Avg.*	*Low*
Tonsillectomy	151	43	13	122	62	23
Appendectomy	32	18	10	22	17	11
Hemorrhoidectomy	10	6	2	19	7	3
Herniorrhaphy	48	41	29	60	45	35
Prostatectomy	38	20	11	40	25	18
Hysterectomy	60	30	20	93	59	39
Cholecystectomy	57	27	17	55	35	27

Source: Wennberg and Gittelsohn (1973), Vermont; and Wennberg and Gittelsohn (1975), Maine.

2.3 times for inguinal hernia to 3.8 times for appendectomy. He noted a correlation of surgical rates with the number of hospital beds and the number of surgeons.

Wennberg and Gittelsohn (1973) studied variation in performance of the same six operations plus some others among 13 hospital service areas in Vermont in 1973. Differences ranged from a factor of 1.7 for herniorrhaphy to 11.6 times for tonsillectomy. In a 1975 study, the same authors found similar variations (see Table 1) in the utilization of these same operations among 42 Health Service Areas in Maine (Wennberg and Gittelsohn 1975).

Significant variations have been found even when large areas (states or parts of large states) were used as the unit of analysis (Chassin, Brook, et al. 1986b). Studying a large number of medical and surgical procedures in Medicare Part B enrollees in 1981, they found, for example, that the rates of performance of hip arthroplasty varied by 11.4 times, carotid endarterectomy by 4 times, and herniorrhaphy by 1.4 times.

Cross-national comparisons also reveal significant differences. Hysterectomy is performed nearly 3 times as often in the United States as in England and Wales, and prostatectomy 2.5 times as frequently (McPherson, Wennberg, et al. 1982). Rates vary widely within other countries as well. Among 56 small rural areas in Manitoba, Roos and Roos (1982) reported a 2.7 overall variation in surgical rates, with a high of 4.2 times for cataract removal. Vayda found a more than fourfold variation in cesarean-section rates and more than ninefold variation for colectomy among counties in Ontario in 1977 (Vayda, Barnsley, et al. 1984). Wide variations in surgical utilization have been demonstrated in

the United Kingdom, Norway, and Canada (Cageorge, Roos, and Danzinger 1981; McPherson, Wennberg, et al. 1982; Vayda, Mindell, et al. 1982; Anderson and Lomas 1984).

Some have noted that much of the reported variation could result from statistical problems of sampling (Diehr 1984). However, consistency of variations over time in one region and the findings from multiple studies that certain operations display wide variations in use wherever studied provide abundant evidence that most variations are real (Gittelsohn and Wennberg 1977; Lewis 1969; Wennberg and Gittelsohn 1975, 1973).

The Significance of Geographic Variations

From the beginning, the causes of geographic variations have been an object of active speculation. A common assumption has been that variations indicate overutilization in high-use areas (Lewis 1969; Wennberg and Gittelsohn 1982). Clearly, the reverse hypothesis is equally plausible: differences could also result from underuse in the low-rate areas. In recent years an astonishing variety of hypotheses has been subjected to study in the search for evidence that geographic variations in the use of operations reflect unnecessary surgery.

The first question is whether geographic variations in surgical rates result from differences in the incidence of specific diseases. Most investigators have assumed that they do not, but the subject has not been studied adequately. Circumstantial evidence, such as extensive variations between adjacent regions with apparently similar populations, as well as the mobility and heterogeneity of the population in the United States, suggests that it is unlikely that variations are due to differences in disease incidence. Measures of health status have not shown differences between high- and low-use areas (Roos and Roos 1981, 1982; Wennberg and Gittelsohn 1975; Wennberg and Fowler 1977). In one of the few studies to assess the relationship between a specific disease and utilization, Roos et al. (1977) found no relationship between tonsillectomy rates and rates of respiratory infection.

We will consider briefly some of the other potential explanatory variables that have been investigated: (1) supply of beds, (2) supply of physicians, (3) socioeconomic and demographic factors, (4) patient characteristics, (5) health care system features, and (6) physician characteristics.

Supply of Beds. In 1969, Lewis linked variations in surgical utilization to availability of both beds and surgeons, likening the results to "a medical variation of Parkinson's Law: patient admissions for surgery

expand to fill beds, operating suites and surgeons' time" (p. 884). Using multiple regression, Lewis was able to explain 49–70 percent of the variance by availability of doctors, surgeons, or beds (Lewis 1969). Others have found a strong relationship between bed supply alone and surgical utilization (Stockwell and Vayda 1979; Wennberg and Gittelsohn 1973). However, Roos (1984) found no relationship between geographic variations in the performance of hysterectomy and the availability of hospital beds.

Hospitals add beds for many reasons. Sometimes it is because existing facilities are strained by increased physician activity. Often, however, there are political, status-related, or financial reasons for hospital expansion that do not reflect need. It is difficult to believe that many physicians recommend an operation simply because an empty bed is available—or that they fail to do so because a patient might have to wait a week to get into a hospital. However, easy availability of beds might well serve as an inducement for a physician to admit his patient to one hospital rather than another.

Supply of Physicians. Lewis (1969) found the availability of surgeons to be an even more powerful predictor of utilization than bed supply. In several of his regressions, the number of physicians and surgeons alone accounted for nearly 50 percent of the variation observed.

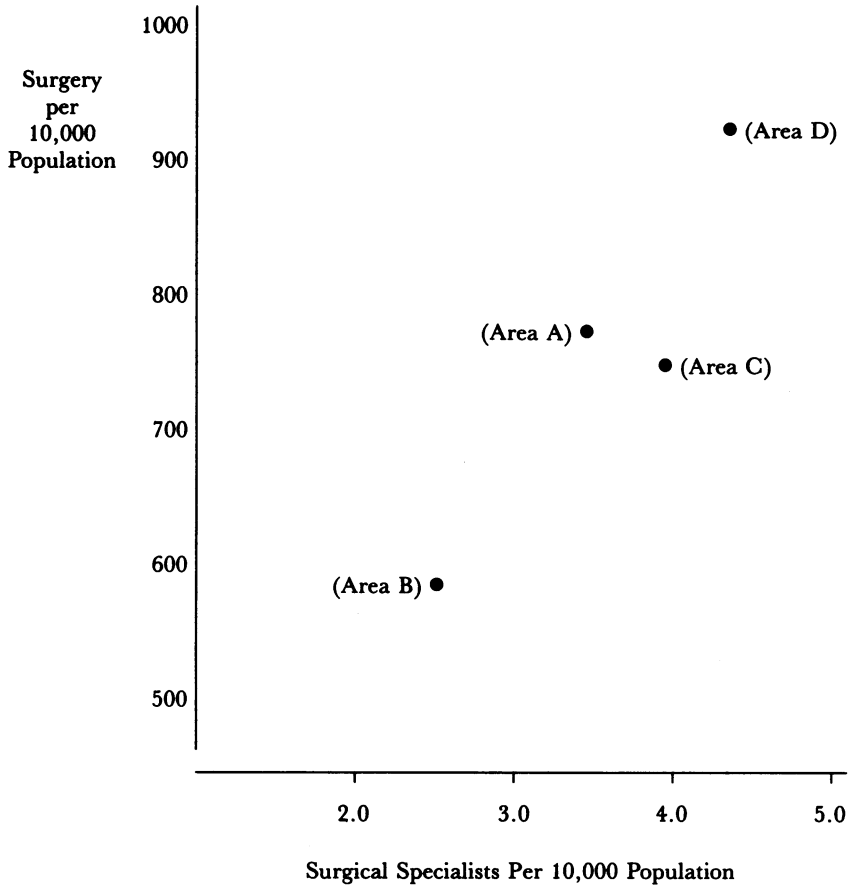
In 1970, Bunker observed that twice as many surgical procedures were performed per capita in the United States as in Great Britain, and that the United States also had twice as many surgeons (Bunker 1970). Similar findings have been noted when Canadian surgical experience has been compared to that of England and Wales (Vayda 1973).

The landmark Study of Surgical Services in the United States in 1975 reported an almost linear relationship between the number of surgical specialists and the number of operations performed per 10,000 population (American College of Surgeons and American Surgical Association 1975) (Figure 1). Wennberg's studies have consistently found a correlation between numbers of physicians, especially surgeons, and surgical utilization (Wennberg and Gittelsohn 1982).

Two studies from Manitoba refute these findings. For tonsillectomy and cholecystectomy the highest rates were found in regions with the lowest number of physicians doing the operations (Roos et al. 1977; Cageorge 1981).

Overall, the weight of evidence indicates that if one area has more surgeons than another, its citizens will have more operations. This is to be expected, for an underemployed surgeon has strong incentives to stimulate referrals. Despite the conflicting evidence of the ability of surgeons to induce demand for their services (Fuchs 1986; Wennberg,

Figure 1: Operations Performed and Supply of Surgical Specialists in Four Study Areas



Source: American College of Surgeons and American Surgical Association. *Surgery in the United States: A Summary Report of the Study on Surgical Services for the United States.* Chicago: American College of Surgeons and American Surgical Association, 1975, Table 11, Chapter IV.

Barnes, and Zuboff 1982; U.S. Dept. of Health and Human Services 1981), it is a common observation that a new surgeon in town can develop a practice that far exceeds any decrease in the patient loads of his competitors.

Socioeconomic and Demographic Factors. Age is a strong predictor of utilization of all forms of medical services (Hulka and Wheat 1985). Surgical rates rise with virtually every decade of life. Gender affects utilization: women undergo surgery more frequently than men. Income has a strong positive effect, especially for children and the aged (Bombardier, Fuchs, et al. 1977), but social background has not been found to be significant (Wennberg and Gittelsohn 1975). The effect of education level is bimodal: surgical rates are lower in those without a high school education and in those who have attended college (Bombardier, Fuchs, et al. 1977; Roos and Roos 1982).

While these socioeconomic and demographic factors have been shown to alter utilization patterns, none of them has been found to explain more than a small fraction of measured geographic variations (Hulka and Wheat 1985; Wennberg and Gittelsohn 1975; Roos and Roos 1982; Bombardier, Fuchs, et al. 1977).

Patient Characteristics. Variations in aggregate health status in a region do not seem to be a significant predictor of geographic variations. Nor are illness levels regularly higher in high-use areas (Wennberg and Gittelsohn 1975; Roos and Roos 1982; Wennberg and Fowler 1977). Medical sophistication does increase utilization. Operative rates among physicians and their spouses are 25–30 percent above average (Bunker and Brown 1974). For hysterectomy, the excess was 50 percent. But again, these differences have not been linked to geographic variations.

Health Care System. Geographic variations have not been shown to be related to the ability of the patient to find or get to a physician (Roos and Roos 1982; Wennberg and Fowler 1977). While at a national level the system of financing is correlated with utilization rates (e.g., Britain versus the United States), method of payment has not been linked to regional variations within the United States. Similarly, geographic variations have not been found to be related to participation in health maintenance organizations, despite the fact that HMOs have been shown to decrease significantly the use of medical care (Manning, Leibowitz, et al. 1984).

Physician Characteristics. The practice style or philosophy of the physicians in a region appears to be an important explanatory variable for geographic variations. Bunker (1970) noted long ago that U.S. surgeons were more aggressive than their overworked British counterparts. Vayda (1973) found similar differences in comparing Canadian surgeons to those in England and Wales.

Wennberg has coined the term "surgical signature" to refer to surgical decision patterns in small areas (Wennberg and Gittelsohn 1982). He considers differences in physician propensity to employ surgical treat-

ment the most important variable explaining regional variations. Such regional "surgical signatures" are specific for each operation, but remarkably constant over time. A key aspect is caseload. Individual surgeons' caseloads can have a significant effect on regional surgical rates, especially in very small areas where a few high-volume surgeons can markedly alter the local rate (Roos, Roos, and Henteleff 1977; Wennberg and Gittelsohn 1982).

If variation in practice style is the primary explanation for geographic variations, the question then becomes, Why do practice patterns of surgeons vary so much? Wennberg and others believe that patterns of practice vary because of the degree of uncertainty surrounding much of medical decision making.

Few laypeople are aware of the extent of uncertainty in the modern practice of medicine. The dramatic improvements in care that have resulted from the biomedical advances of the past 30 years have led many to think that the practice of medicine is the predictable application of science. In fact, the number and variety of diagnostic and treatment modalities has increased at a far more rapid pace than the assessment of their value. As a result, the degree of uncertainty about the effectiveness of therapy is greater than ever before (Eddy 1984; Barnes and Zuboff 1982; Wennberg 1984). Decision making under conditions of uncertainty results in variability of response.

If the uncertainty hypothesis is correct, variations will be greatest when the level of uncertainty is high, and least when the level of uncertainty is low. Such appears to be the case. Inguinal hernia and fractured hip are good examples of conditions with low levels of uncertainty regarding indications for surgery. In all geographic-variations studies, these two conditions show small degrees of variation (Wennberg 1986; Chassin, Brook, et al. 1986b). Conversely, the indications for carotid endarterectomy are highly controversial, and there are wide variations in its use (Chassin, Brook, et al. 1986b; Winslow, Solomon, et al. 1988).

Geographic Variations as Evidence of Unnecessary Surgery

While the relationship of these factors—bed supply, number of physicians, socioeconomic differences, patient characteristics, the nature of the health care delivery system, and the practice style of physicians—to geographic variations is interesting, none provides a direct answer to the question of whether geographic variations indicate unnecessary surgery. The criteria of relevance were not met by any of these studies.

Two types of investigations do give more useful information: stud-

ies of the effects of feedback and studies of the appropriateness of indications. Unfortunately, the results are conflicting.

Feedback. Wennberg showed that when practitioners were informed that their rates were substantially above state averages, rates subsequently dropped. In some cases the results were dramatic: a 50 percent decrease in the rate of hysterectomy and a 90 percent decrease in the rate of tonsillectomy (Wennberg 1984). The logical conclusion from this experience is that the surgeons recognized that some of the previous operations had been unnecessary.

Appropriateness. Two studies directly addressed the question of whether increased use of operations reflects increased inappropriate use. Roos et al. (1977) examined standards of practice in their assessment of geographic variations in tonsillectomy rates. Using as their standard of appropriateness the criterion of four or more episodes of tonsillitis, pharyngitis, or upper respiratory infection before operation, they found high percentages of inappropriate use, but no correlation of inappropriate use with rates.

Researchers at RAND/UCLA examined the presumption that high rates mean inappropriate use in a study of geographic variations among large areas. The appropriateness of highly detailed indications for three controversial operations and procedures was evaluated by a consensus method using a panel of experts. Nearly 5,000 hospital patient records were then abstracted in high-, medium-, and low-use regions for each of the three procedures, and appropriateness of indications was rated using the explicit expert criteria. While for some procedures the fraction of operations done for inappropriate indications was distressingly high (e.g., 32 percent for carotid endarterectomy), they found no significant difference between high- and low-use areas in the percentage of operations done for inappropriate reasons (Chassin, Kosecoff, et al. 1987).

Conclusions

Geographic variation studies provide provocative but only inferential evidence of unnecessary surgery.

1. If an operation is being performed ten times as frequently in one area as in another, it is reasonable to assume that the high utilization must represent some excess use. The observation that feedback of information about variations in surgical rates leads doctors to decrease use supports this conclusion. At the same time, it is probable that low rates indicate underuse.

2. The uncertainty hypothesis explanation of geographic variations also implies some degree of unnecessary surgery. Eventually, some of the controversial indications undoubtedly will be found to be inappropriate. Their current use thus represents unnecessary surgery.
3. Geographic variations are greater for controversial operations. Controversy is most likely when the evidence for effectiveness of an operation is weak. The finding of the RAND/UCLA appropriateness studies that the fraction of procedures performed for inappropriate or equivocal reasons is higher for procedures with greater variation is further evidence that large geographic variations are associated with a significant amount of unnecessary surgery.
4. The fact that rates of inappropriate or equivocal use were virtually identical in *both low- and high-use areas* in the RAND/UCLA study demonstrates that the *extent* of variations (e.g., 3:1 or 10:1) is not a direct measure of the amount of unnecessary surgery.

Thus, geographic variations in surgical rates seem to be indirect indicators of unnecessary surgery. A high degree of geographic variation is associated with a high fraction of inappropriate use, although the correlation is far from one-to-one. Geographic variations, therefore, are evidence of unnecessary surgery, but they do not provide the data to quantify it.

SECOND-OPINION PROGRAMS

Many third party payers provide for second surgical opinion (SSO) programs that encourage or require a subscriber to get a second opinion after an operation has been recommended by a surgeon. The second-opinion surgeon confirms, denies, or questions the diagnosis and the recommendation for the operation. The patient then decides whether or not to have the operation performed.

The feasibility of an SSO program was established by McCarthy at Cornell in a program developed for two unions in New York City in 1972 (McCarthy and Widmer 1974). He stated that the purpose of second-opinion programs is "to help the patient make a more informed decision regarding previously recommended surgery" (McCarthy and Finkel, 1978, 985). He felt that it would improve the quality of care through educating the consumer and would enable the patient to participate more fully in health care decision making.

Although his results have been used as evidence of unnecessary surgery, McCarthy was careful not to draw that conclusion, noting that the method was not designed to measure the reliability of judgments between physicians.

Second-opinion programs are applied only to elective surgery, which is defined as nonemergency operations and those that are not performed for life-threatening conditions. The third party pays for the consultation and any further studies it requires. In some programs, all elective surgery is covered; in others, SSO is required only for specified operations—those thought to be performed excessively. The most common operations in this category are hysterectomy, tonsillectomy, hemorrhoidectomy, removal of varicose veins, cholecystectomy, and operations on the knee, nasal septum, and intervertebral discs.

SSO programs may be voluntary or mandatory. In the latter, failure to obtain an SSO results in a penalty, most commonly denial of reimbursement of the surgeon for the operation. The participation rate in voluntary programs is typically low, 1 to 2 percent of those having operations, and thus their effect is minimal (U.S. Congress 1985). Programs also vary in specification of the second-opinion physician. In some, the plan provides recommendations from a panel of physicians; in others the patient is free to choose. Some require the second-opinion surgeon to be board certified, others do not. Some plans do not even require the second opinion to be given by a surgeon.

History of Second-Opinion Programs

McCarthy's pioneering study of two SSO programs, one voluntary and the other mandatory, was published in 1974. He demonstrated the feasibility of carrying out a second-opinion program, and established the ground rules that have been generally accepted by other programs. If the second-opinion surgeon agrees with the original recommendation, it is "confirmed." If he disagrees, it is "not confirmed." Nonconfirmation rates were 30.4 percent for the voluntary program and 17.6 percent for the mandatory one (McCarthy and Widmer 1974).

Congressional Hearings: "Unnecessary Surgery." Congressional hearings were held that same year on the issue of unnecessary surgery by the Subcommittee on Oversight and Investigations of the Committee on Interstate and Foreign Commerce, John E. Moss, Chairman. Although the committee received testimony from a number of sources, the McCarthy findings clearly created the strongest impact. Despite the investigator's caveats, the raw nonconfirmation rates were interpreted as "unnecessary surgery" rates, and the committee staff extrapolated the

numbers to calculate the national surgical experience. In the final report of the committee, "Cost and Quality of Health Care: Unnecessary Surgery," issued in January 1976, they estimated that in the preceding year 2.4 million unnecessary operations had been performed, with 11,900 deaths, at a cost of $3.9 billion (U.S. Congress 1976).

The American Medical Association (AMA) criticized the committee results vigorously, claiming that the estimates were erroneous because they were based on invalid data. They challenged both the estimate of the number of elective operations performed in the United States and the use of the McCarthy study figure of 17 percent to derive the national figures.

In subsequent hearings in the fall of 1977, the subcommittee reviewed the AMA rebuttal and heard testimony concerning a study the AMA had cited in support of its position. The study was performed by Ralph Emerson, M.D., in New York state, and reviewed indications for operations retrospectively, using preset criteria. He reported that only 1 percent of operations were found to be unjustified (Emerson 1976). The subcommittee staff discovered many discrepancies and methodological errors in the Emerson study, including, most importantly, widespread failure to apply standard criteria. A re-review of a subset of the Emerson study patients by an outside expert found a 16 percent unjustified rate. The subcommittee's rejection of the AMA position was scathing (U.S. Congress 1978).

Government Second-Opinion Programs. The most important result of the second round of subcommittee hearings was the strong recommendation that the Department of Health and Human Services (DHHS) promptly institute an SSO program for elective surgery funded by Medicare and Medicaid. DHHS responded quickly and later that year launched a national voluntary SSO program with great publicity and the establishment of a national hotline. Demonstration projects were established in New York and Michigan in which the Medicare copayment and deductibles were waived if beneficiaries sought second opinions. Medicaid agencies in seven states introduced mandatory SSO programs for certain operations, selected on the basis of volume, cost, and expected rates of nonconfirmation.

Commercial Insurance SSO Programs. Private-sector major insurers began to offer voluntary SSO programs to employers in the 1970s. Failure of these programs to be cost effective led to experimentation with mandatory programs. By 1984, 28 percent of the surveyed firms had a mandatory second-opinion program. In 1983, the Prudential Insurance Company estimated net savings of $6.95 for each dollar spent in its mandatory programs. Business coalitions also "jumped on the SSO

bandwagon," and Blue Cross/Blue Shield SSO plans expanded rapidly—from 10 in 1982 to 60 in 1985 (U.S. Congress 1985).

Further Congressional Action. In 1985, Congress again held hearings on unnecessary surgery, this time by the Senate Special Committee on Aging, chaired by Senator John Heinz. The Inspector General of DHHS reported that voluntary SSO programs were ineffective, but that mandatory programs in Medicaid in three states (Massachusetts, Michigan, and Wisconsin) had reduced utilization 20–35 percent, for a total estimated savings in those three states of $7.5 million (U.S. Congress 1985). He noted the great expansion of mandatory SSO programs in the private sector and in Medicaid, and recommended that a mandatory program be instituted by Medicare. The chairman was obviously of the same mind, and castigated the DHHS for "dragging its feet" in beginning mandatory programs.

HCFA did not support mandatory SSO, recommending instead strengthening of the voluntary program and reliance on PROs to enforce standards of appropriateness. The American College of Surgeons took a similar view, supporting voluntary programs, but opposing imposition of a mandatory one (American College of Surgeons 1982).

How SSO Works

Second-opinion programs are thought to reduce surgical rates by altering the behavior of both the patient and the surgeon.

Effect on Patients. The effect on the patient is thought to be most marked in those who are undecided or who have questions about the desirability of the proposed operation. For these patients, SSO offers time, information, and often reassurance. The almost inevitable delays imposed by selection and scheduling of the second opinion (average, one month) give the patient time to think over the situation, reevaluate his symptoms or disability, and weigh the pros and cons.

Additional information and the opportunity to hear a different point of view provided by the SSO enable the patient to make a more informed decision. For the 80–85 percent of patients in mandatory programs who receive a confirming SSO, the second opinion provides reassurance that the initial recommendation was valid. For the undecided patients in voluntary programs, the SSO can be reassuring with either result. For some, SSO has a "barrier" effect: the requirement that the patient must obtain a second opinion discourages him from undergoing the operation (Martin and Shwartz 1980). If the operation would have been beneficial, this barrier effect is deleterious to his health.

Effect on Physicians. Physicians respond to second-opinion require-

ments in several ways. Publicity about SSO and notification that it is required for reimbursement for certain operations suggest to primary care physicians that perhaps too many of these operations are being performed. They are likely to use more stringent criteria for deciding to refer for surgery. In addition, McCarthy and his colleagues have described a "sentinel" effect: a decrease in the number of operations recommended by surgeons after introduction of a second-opinion program (Grafe, McSherry, et al. 1978). Presumably, the possibility of having his opinion not confirmed makes the surgeon more conservative in his initial recommendation.

The net effect of mandatory SSO programs appears to be a decrease in the number of operations performed. Costs have also decreased. As a result, SSO programs are widely believed to be an efficient method of cost containment. They are also among the most appealing, because the decision to forgo surgery is made not by the doctor or the payer but by the one who bears the consequences—the patient.

The Experience with Second Surgical Opinion Programs

To determine whether surgical second-opinion programs indicate unnecessary surgery, studies of these programs must show that (1) operative rates have decreased, (2) the decrease has resulted from omission of unnecessary operations, and (3) their omission has caused no deleterious health effects. All of these assessments require that a control group be studied. To evaluate health effects it is necessary to measure the outcomes of both the surgically and nonsurgically treated patients over time. The requirements for an appropriate study, therefore, are:

1. Identification of a control group: patients who do not have access to an SSO program, but who are otherwise similar
2. Evaluation of health outcomes for both study and control patients, and for those who had surgery and those who did not
3. Follow-up of patients for a period long enough to discover long-term health effects of the decisions to operate or not—a minimum of two years.

None of the studies that have been uncovered in the literature meets all of these criteria. Accordingly, any conclusions about whether surgical second-opinion programs either measure or deter unnecessary surgery must be tentative. Assertions regarding the effectiveness of SSO programs as deterrents to unnecessary surgery are based on analysis of three

Table 2: Second Surgical Opinion Programs Nonconfirmation Rates (in Percent) by Program (Mandatory Programs)

*Program**	*Nonconfirmation Rate*	*Source*
Cornell-New York Hospital	18.7	McCarthy and Finkel (1980)
BC/BS Michigan Medicaid	11.0	McCarthy and Finkel (1980)
Michigan Medicaid	12.3	Roenigk & Bartlett (1982)
New Jersey Medicaid	11.5	New Jersey Dept. of Human Services (1984)
Wisconsin Medicaid	7.5	Wisconsin Dept. of Health and Social Services (1984)
Prudential	10.0	American Medical Association (1980)
Massachusetts CPES	13.3	Poggio et al. (1981)

*Adapted from Poggio (1985).

types of information: the rate of nonconfirmation of the initial opinion, the percentage of patients who do not have the operation, and changes in overall surgical rates after an SSO program is initiated.

Nonconfirmation Rates. Nonconfirmation rates in mandatory programs vary from 7.5 to 19 percent, (McCarthy and Finkel 1980; Poggio 1985) (Table 2). Combined nonconfirmation rates after a third opinion have been reported at 10 percent or less (Martin, Shwartz, et al. 1982). McCarthy's initial study rate of 17 percent was used for the congressional extrapolation that concluded that there were 2.4 million "unnecessary" operations in the United States in 1975 (McCarthy and Widmer 1974).

Several investigators have studied the effects of the Massachusetts Medicaid surgical second-opinion program (Gertman, Stackpole, et al. 1980; Martin, Shwartz, et al. 1980, 1982; Poggio 1985). Martin's detailed analysis of the confirmation pattern in this program showed a 14.5 percent initial nonconfirmation rate. Nearly half (44 percent) of the nonconfirmed patients sought a third opinion, which resulted in reversal of the nonconfirmation in 65 percent. Thus the final nonconfirmation rate was 10 percent. Overall, 72 percent of patients had the proposed surgery performed, 85 percent of those confirmed and 31 percent of those not confirmed. The estimated cost savings for the state was $856,500, a benefit:cost ratio of 3.5 (Martin, Shwartz, et al. 1982). The authors note that this estimate is based on untested assumptions about the percentage of patients who would have had the recommended operations in the absence of the second-opinion program. The significance of their results is further clouded by the coincident imposition of a 30 percent reduction in fees paid to doctors by Medicaid.

In 1981, McCarthy's group reported the results of a long-term (eight-year) study of second opinions in the Cornell study: 5,079 patients in voluntary programs and 6,799 in mandatory programs. The findings were remarkably similar to those of the 1974 study: nonconfirmation rates of 33.4 percent for voluntary programs and 18.7 percent for mandatory ones. One year after the initial recommendation for surgery, 45 percent of the nonconfirmed patients reported that they had received neither surgery nor further medical treatment of any kind. Although the health status of these patients was not assessed, this group was labeled potential "surplus surgery" (Finkel, Ruchlin, and Parsons 1981).

Not surprisingly, voluntary SSO programs have higher nonconfirmation rates than mandatory programs. Patients typically seek a second opinion voluntarily when they wish to avoid an operation, or when they perceive that either the indications for surgery or its proposed benefits are not clear-cut. This perception may accurately reflect the equivocal nature of the clinical situation or the patient's sense that the doctor's recommendation is inappropriate. Nonconfirmation rates in voluntary second-opinion programs are about 30 percent (McCarthy and Widmer 1974; Joffe 1980; Finkel, Ruchlin, and Parsons 1981).

Nonoperative Rates. Second opinions seem to carry more weight than the initial recommendation in most programs. The great majority of patients follow the advice of the second doctor, for or against surgery (Table 3). McCarthy reported that in the mandatory program 88 percent of those who had a confirmation second opinion ultimately had the operation, while 61 percent of those receiving a nonconfirmation decided against surgery (McCarthy and Finkel 1980). In the voluntary program, compliance was 70 percent for confirmation SSO and 83 percent for nonconfirmation opinions. Results in the mandatory Massachusetts Medicaid study were similar to McCarthy's mandatory program: compliance rates were 83–93 percent for confirmation SSO, and 61–63 percent for nonconfirmation SSO (Martin, Shwartz, et al. 1980).

Nonconfirmation and operative rates overestimate the effects of second-opinion programs, however. Many patients who receive a nonconfirming SSO would not have had surgery in the absence of a second opinion. Poggio et al. estimated this aspect in the Massachusetts Medicaid second-opinion program by interviewing a sample of participants to find out if the program affected their decision, either way, about whether to have an operation (Poggio et al. 1985). They found that only 4.1 percent of participants changed their minds as a result of the second opinion. While 2.9 percent of participants said they decided to forgo surgery as a result of the program, 1.2 percent changed their minds in favor of surgery because of the program. The net direct effect of the

Table 3: Second Surgical Opinion Programs Surgery Rates (in Percent) by Program (Mandatory Programs)

*Program**	*Confirmed Cases†*	*Nonconfirmed Cases†*	*Length of Follow-Up (Months)*	*Source*
Cornell-New York Hospital	87.7	38.6	12	McCarthy and Finkel (1980)
New Jersey Medicaid	67.8	33.8	2–16	New Jersey Dept. of Human Services (1984)
Massachusetts CPES				
Bay State	88.9	39.3	8–19	Martin, Schwartz, et al. (1980)
Central Mass.	91.6	37.3	3–14	Martin, Schwartz, et al. (1980)
Charles River	93.1	—	12	Poggio et al. (1981)
Pilgrim	88.2	—	12	Poggio et al. (1981)
Western Mass.	82.8	—	12	Poggio et al. (1981)

*Adapted from Poggio (1985).

†Rates shown are the proportion of patients undergoing surgery for their condition, including surgical procedures differing from the originally recommended procedure.

SSO, therefore, was a reduction of only 1.7 percent in the number undergoing an operation.

Since participation in voluntary programs rarely exceeds 2 percent, their effect on the number of operations performed is negligible. In 1983, the Inspector General of the DHHS concluded that they were ineffective (U.S. Congress 1985).

Overall Surgical Rates. The reduction in the overall rate of surgery as a result of an SSO program is usually much greater than can be accounted for by the nonconfirmation or nonoperative rates.

In a detailed study of the Massachusetts Medicaid second-opinion program, Poggio et al. examined the effects of adding or removing the requirement for a surgical second opinion for specific operations over a three-year period. Monthly surgical rates for each of the eight procedures for which SSO was required were analyzed for each of the five regions of the state, before and after institution of SSO. Using a pretest, posttest design, with comparisons before and after a reduction in Medicaid fees, they found reductions of 29–41 percent in monthly mean rates of surgery for the studied procedures. They estimated the total reduction in surgery at 24 percent (Poggio et al. 1985). Other programs have estimated reductions from 10 to 40 percent in the performance of

covered procedures (Finkel, Ruchlin, and Parsons 1981; U.S. Congress 1985).

Poggio's group noted that the indirect effect of the SSO programs was far greater than the direct effect, which only accounted for about 2 percent of the total 24 percent reduction in the rate of surgery.

Health Outcomes. The only study to look at health outcomes resulting from second-opinion programs was reported by Poggio et al., who interviewed 365 Medicaid second-opinion program participants in Massachusetts (Poggio et al. 1985). They found no significant deleterious effects as a result of the program, but noted that since relatively few patients changed their minds as a result of the SSO (2 percent), it would be difficult to detect health effects. They also pointed out that because only participants were studied, nothing is known about the impact on health outcomes of the indirect effect of the SSO program. Since reduction in overall surgery rates is the major effect of second-opinion programs, this information would be of some consequence.

Cost Containment. Cost containment has been the major motivation behind the initiation of most SSO programs. Since these considerations are tangential to the issue of necessity, I will not present a detailed analysis of the savings and cost effectiveness. A few comments are in order, however. Two types of financial analyses have been presented: cost effectiveness and total savings. Benefit:cost ratios from 1.1 to 22 have been reported, and annual savings for statewide Medicaid programs are typically estimated at several million dollars (U.S. Congress 1985).

These conclusions are not justified because there are no data on the costs to which the program results are being compared, that is, costs for comparable patients in the absence of a second-opinion program. Not only are control groups lacking, but since a significant fraction of patients do not follow the doctor's advice it is fallacious to calculate "savings" or cost:benefit ratios, as most have done, by adding up the cost of surgery for all with nonconfirmation SSOs. Finally, health costs further downstream have not been examined, either for SSO patients or for controls.

Second-Opinion Programs Do Not Identify Unnecessary Surgery

Do surgical second-opinion programs identify and prevent unnecessary surgery? Analysis of reported studies of surgical second-opinion programs suggests that we do not know. Both the design of the studies and the conceptual underpinnings of SSO programs have serious shortcomings.

Design Problems

Absence of Controls. No controlled study has been performed. That is, no study has been reported in which the experience of a group of patients in a second-opinion program is compared to that of a comparable group of patients who did not participate in an SSO program. In fact, it is not even known what percentage of patients who have an operation recommended without an SSO ultimately have the operation performed—within a month or a year, or ever.

The problem of lack of controls has been handled by investigators in three ways. The most common method is to do without, making the assumption that without an SSO everyone would follow the initial recommendation and undergo the operation (U.S. Congress 1976). All forgone operations associated with an SSO program are assumed to result from the program. This assumption ignores the well-known fact that many patients do not follow their doctor's advice. Even after two recommendations (a confirmed SSO), 12–15 percent of patients do not have the recommended operation (McCarthy and Finkel 1978).

The second method is a variant of the first. However, instead of using forgone operations as the measure of effect, the number of nonconfirmations alone is used (McCarthy and Widmer 1974). This approach ignores the evidence that some of these patients would not have had the operation anyway, even without a negative second opinion, and that some of them had the operation in spite of the negative second opinion.

The third way of dealing with the absence of appropriate controls has been to use a subset of the SSO patients. In calculating the cost effectiveness of a second-opinion program, Ruchlin compared the total costs of a group of nonconfirmed patients with an equal number of confirmed patients (Ruchlin, Finkel, and McCarthy 1982). Since these patients were not matched by diagnosis, age, severity, or comorbidity, selection bias may be considerable. This group also excludes others who were affected by the SSO program—those who did not get a second opinion but decided against surgery because of the requirement for a surgical second opinion.

In the absence of data from a control population, it is not possible to measure the effect of surgical second-opinion programs on either patient health or costs.

Paucity of Outcome Data. The answer to the question, Was the forgone operation truly unnecessary? requires follow-up information. What happens to those patients who choose not to have the operation? Do they have continuing symptoms or disability? Do they later require surgery

or develop complications related to delay of surgery? Of equal interest, what happens to those who did have surgery? If they remained symptomatic, were their operations "unnecessary"?

Few studies provide data regarding these questions. None assesses health status before the surgical decision was made. The follow-up study from McCarthy's program found that 50 percent of patients who elected not to have surgery (whether confirmed or not) had no treatment for at least a year (McCarthy and Finkel 1978). No assessment of their symptoms or disability was made, nor were they followed beyond one year. Poggio's study was limited to retrospective self-reported health status of a small sample, but found no effects (Poggio et al. 1985). Clearly a major limitation to assessing the medical impact of SSO is the absence of adequate follow-up health status information.

Conceptual Problems with SSO

Aside from the problems associated with study design, significant conceptual problems arise if one attempts to use information from surgical second-opinion programs to assess unnecessary surgery.

Criteria for Judgment. SSO is a form of implicit review. The consultant surgeon evaluates the patient's symptoms and findings using his own criteria based on knowledge and experience. No predetermined agreed-upon (explicit) standard is used for assessing the appropriateness of the indications for the operation. Thus, we do not know if the second opinion is any more valid than the first. There is no benchmark against which to measure whether an SSO does in fact identify and discourage unnecessary surgery.

Unnecessary versus Uncertain. For some operations, many indications are shrouded in uncertainty. Scientific evidence of efficacy is lacking, and experts disagree on the appropriateness of a number of indications. For other indications there may be a broad consensus that the operation is useless. Analyses of the results of SSO programs do not differentiate between the two. Nonconfirmations that result from differences of opinion are not distinguished from those that reflect a general consensus that an operation is ineffective.

Second Opinions by Peers. Although McCarthy's first SSO program utilized academic specialists to provide the second opinion, most of the programs subsequently developed by Medicaid and insurance companies have required only that the opinion be rendered by an equivalent specialist. In some cases, not even that stricture is applied, and nonsurgeons may give the second opinion. Much of the benefit of SSO as a deterrent to unnecessary surgery may be lost if peer second opinions are

Table 4: Effects of Reliability of First and Second Opinions for 100 Patients Recommended for Surgery

Results of the First Opinion		*Results of the Second Opinion*			
		20% Nonconfirmation Rate		*15% Nonconfirmation Rate*	
"True" Appropriateness	*Number of Patients*	*Confirms (80%)*	*Does Not Confirm (20%)*	*Confirms (85%)*	*Does Not Confirm (15%)*
Appropriate	75	60	15*	64	11*
Inappropriate	25	20*	5	21*	4
Totals	100	80	20	85	15

*Final recommendation (SSO) is inappropriate.

used. In simple terms, two wrong opinions do not make an operation appropriate. Indeed, if both opinions are subject to equal and independent error rates, probability theory suggests that the number of patients treated inappropriately will be *increased* by such a program, even if surgical rates drop.

- For some operations, the paucity of information on effectiveness is such that 20–30 percent of the indications for the procedure are, or ultimately will be found to be, inappropriate. Thus, many of the initial recommendations for surgery will be (unknowingly) inappropriate. Absent additional knowledge, the second-opinion surgeon will tend to agree with the first about 80 percent of the time (Table 4). Thus, he will confirm 80 percent of both the appropriate and the inappropriate original recommendations, rejecting 20 percent of each.
- The result is that of 100 patients initially recommended for surgery, 20 will not be confirmed, but 35 will have a recommendation for inappropriate treatment. Twenty will be confirmed for surgery who will not benefit and 15 will not be confirmed who would benefit. These 35 patients represent a 40 percent increase over the 25 percent inappropriate rate in the first recommendation alone. In practice, nonconfirmation rates are closer to 15 percent. Even this lower rate would result in inappropriate treatment for 32 percent, an increase of 28 percent.

Excessive Influence of the Second Opinion. Both in mandatory and voluntary programs, the vast majority of patients follow the advice of the second opinion whether it supports or contradicts the initial recommen-

dation. In McCarthy's follow-up study, patients decided against having surgery after receiving a nonconfirmation opinion 68 percent of the time in the mandatory program and 83 percent of the time in the voluntary program. While perhaps these results are to be expected when the second opinion is rendered by an academic expert, they would seem inappropriate when the second opinion is given by a peer. Yet, similar results are found in peer second-opinion programs. It is evident that when a difference of opinion exists, the chances are much greater than 50:50 that the patient will take the advice of the second consultant. The reason is probably related to reaction to the uncertainty engendered by conflicting advice. While a confirming opinion is reassuring to the patient, nonconfirmation is unsettling. The patient is most likely to resolve the dilemma by deciding against the operation. Why take a chance if the doctors are not sure?

False Negatives. While SSO programs theoretically could decrease unnecessary surgery if an expert panel were used for the second opinion, they do not address the problem of patients who receive an erroneously negative first opinion. These patients never "get into the system" to have a decision on surgery evaluated. This denial of potential surgical benefit can result from failure of the primary physician to refer the patient or from failure of the surgical consultant to recommend surgical treatment when it is appropriate. Both types of limitation of access may increase after institution of a second-opinion program—inappropriate consequences of the "sentinel" effect.

Conclusions

1. SSO programs appear to reduce the number of operations performed on participants. Even though control data are lacking, the fact that every study has shown reductions cannot be summarily dismissed. Decreases in overall surgical rates have been demonstrated for individual operations and for entire populations. Whether these reductions are appropriate is another matter.

2. By providing additional information, or at least a different point of view, SSO programs undoubtedly help many patients in decision making for or against surgery. For the undecided patient, one who is seriously questioning whether or not to have the recommended operation, the second opinion may tip the balance. Conversely, if the individual has his or her mind made up, the second opinion probably makes little difference.

The major value of voluntary programs may be for those who are undecided.

3. SSO programs may improve the process of decision making by the surgeon by means of the "sentinel" effect. Knowing that the recommendation will be reviewed causes a surgeon to be more cautious—more certain that the operation is truly indicated before recommending it. This increased caution seems particularly appropriate for elective operations where a number of nonbiologic factors enter into the decision-making process. It also seems appropriate for controversial operations or those that may be performed excessively.
4. Results of SSO programs do not define unnecessary surgery in an intellectually defensible way. In published studies, necessity has not been established—nor has it been refuted—for any of the indications for the operations. SSO programs only determine whether there is agreement between the consultant and the primary surgeon. Disagreement may reflect either a difference of opinion or the second surgeon's specific knowledge that the operation will be ineffective for this patient.
5. The reduction in surgical rates in most SSO programs may be indiscriminate. In programs where the consultant is a recognized expert, the SSO theoretically should identify some unnecessary operations, although this is unproved. In peer programs, such identification is less likely. The indirect effect, reduction in patients referred for surgery, is even less discriminating. Since SSO probably eliminates both needed and unneeded operations, it is a blunt instrument to use for quality control.

Because of these limitations, conclusions about the extent of unnecessary surgery in the United States cannot legitimately be derived from extrapolation of the results of surgical second-opinion programs.

CRITERIA STUDIES

The obvious way to assess the extent of unnecessary surgery is to measure it. If it is possible to define the conditions under which an operation is useless, those criteria can be applied to the evaluation of a series of surgical cases to determine which were not indicated and thus represent unnecessary surgery.

Despite the wealth of data from geographic variation studies and

second-opinion programs suggesting that unnecessary surgery is a significant problem, very few attempts have been made to measure it directly. In 1953, Doyle reported the results of a study of 6,248 hysterectomies of which 39 percent were found to be "unjustified" (Doyle 1953). There were few similar published reports in the ensuing 25 years, until publication of the results of the RAND/UCLA Health Services Utilization Study in 1987 (Chassin 1987; Winslow, Solomon, et al. 1988; Winslow, Kosecoff, et al. 1988).

Methodology

Measurement of unnecessary surgery is not simple. Definition, method of measurement, criteria, and sources of data all present specific obstacles.

Definition. While no one has used the definition "useless," the use of "inappropriate," as employed by the RAND/UCLA researchers, comes close. They defined "appropriate" to mean that the "expected health benefits (i.e., increased life expectancy, relief of pain, reduction in anxiety, improved functional capacity) exceeded the expected negative consequences (i.e., mortality, morbidity, anxiety of anticipating the procedure, pain produced by the procedure, time lost from work) by a sufficiently wide margin that the procedure was worth doing." (Park, Fink, et al. 1986, 767). "Inappropriate" is the converse: the expected benefits do not exceed the expected negative consequences. When terms such as "appropriate" or "justified" are used without definition, it is not possible to evaluate the validity of the results.

Methods. Donabedian and others have categorized quality evaluation methods as explicit, in which quality of care is appraised by use of predetermined criteria, or implicit, in which the evaluator (usually a physician) uses his own experience and judgment rather than specific criteria to determine appropriateness (Donabedian 1982). Most students of quality assessment believe that explicit evaluation is more objective and, therefore, is more likely to be reproducible than implicit evaluation.

Criteria. The validity of explicit review, however, depends heavily on the quality of the criteria used which, in turn, depends on the process that is used to derive them. While derivation of standards in surgical practice is not new, defining criteria of appropriateness in significant detail is. To permit a valid judgment of appropriateness, these criteria must be defined in a comprehensive fashion, embracing all of the critical variables that must be considered in the decision to operate. They must have sufficient detail to be clinically relevant, yet

manageable in their use. Most importantly, they must clearly distinguish between appropriate and inappropriate indications.

Most experts agree that a structured process that combines analysis of the scientific information in the literature with a group judgment by experts is preferable to reliance solely on published information or on informal expert consensus alone. Despite the increasing recognition in recent years of the importance of patient values in the medical decision-making process, methods have not been developed in which those values are directly incorporated into criteria. That is, patient utilities, if considered at all, are imputed by the medical experts who are judging the criteria. There is evidence that these judgments differ significantly from those that patients would actually make in practice (McNeil 1982). The process of developing suitable criteria is discussed in greater detail in the section titled, "Why Is There Unnecessary Surgery?"

Data. Finally, one must decide what source of information to use: the medical record, discharge abstracts, reimbursement claims, monitoring information, interviews with patients or doctors, or outcomes data of various sorts. Each has its advantages and limitations. A detailed evaluation of the validity of each type of source is beyond the scope of this synthesis, but it is essential to recognize that the source of data imposes significant limitations on the validity of criteria studies.

Findings

Six studies have been identified that have looked directly at unnecessary or inappropriate surgery. Table 5 summarizes the evaluation of these reports.

In one of the earliest studies of unnecessary surgery (actually a study of geographic variations), Roos used claims data to assess indications for tonsillectomy in 3,072 patients in Manitoba (Roos, Roos, and Henteleff 1977). The criteria for appropriateness were not met in 86 percent of patients. The accuracy of this finding may be questioned because of the general nature of the criteria used and the fact that they were derived solely from literature review. On the other hand, the criteria were minimal, so more detailed standards might result in an even higher level of inappropriateness.

The four RAND/UCLA studies applied a structured consensus method of deriving highly specific criteria of appropriateness to carotid endarterectomy (CE) (Merrick, Brook, et al. 1986; Chassin, Kosecoff, et al. 1987; Winslow, Solomon, et al. 1988) (Winslow and Chassin report the same data) and coronary artery bypass surgery (CABS)

Table 5: Criteria Studies

Characteristic	*Study*					
	Chassin, Kosecoff, et al. 1987; Winslow, Solomon, et al. (1988)	*Elliott et al. (1981)*	*Greenspan et al. (1988)*	*Merrick et al. (1986)*	*Roos et al. (1977)*	*Winslow, Kosecoff, et al. (1988)*
Number of operations	1302	4850	382	107	3072	386
Unambiguous definition of unnecessary	yes	yes	yes	yes	yes	yes
Source of information	R	R	R	R	C	R
Type of review	E	E/I	E	E	E	E
Type of criteria used	S	G	G	S	L	S
Findings—percent unnecessary	32	3	20	13	86	14
Methodologically valid	yes	no	yes	yes	+/-	yes
Evidence of unnecessary surgery	yes	no	yes	yes	yes	yes

R = patient's medical record.
C = payment claims data.
E = explicit review, reviewer uses predetermined criteria.
I = implicit review, professional uses his own judgment.
S = structured criteria development process involving critical analysis of the literature data and consensus of experts.
G = criteria developed by a group of experts.
L = criteria developed from review of the literature.

(Winslow, Kosecoff, et al. 1988). The criteria were used to evaluate the appropriateness of therapy as revealed by a detailed record abstraction process. Thirteen to 32 percent of CE were found to be done for inappropriate indications, and 14 percent of CABS were found to be unjustified. An additional 32 percent of CE and 30 percent of CABS were judged to be performed for equivocal indications—those on which the experts disagreed, or for which the evidence of effectiveness was felt to be inadequate. These studies have high methodological validity and provide unequivocal evidence of unnecessary surgery.

Greenspan evaluated insertion of cardiac pacemakers by means of an explicit review of medical records using predetermined criteria (Greenspan, Kay, et al. 1988). While these standards were not as detailed and sophisticated as those used in the RAND/UCLA studies,

they did discriminate between appropriate and inappropriate indications. Twenty percent of these procedures were judged to be unnecessary. This conclusion appears to be justified by the methods and evidence presented.

Elliott reported the results of a Professional Standards Review Organization-sponsored study of 13 surgical procedures using criteria developed by a medical committee (Elliott, Kahn, and Kaye 1981). Nurse reviewers used the criteria for screening, and care was considered to be appropriate if any of the criteria were met. Those that failed the screen were given a second implicit evaluation by a physician. Only 3 percent of the operations were found to be unjustified. The general nature of the criteria (e.g., appendectomy is justified if the patient has acute abdominal pain and a fever of 37.8°C or greater) and the use of implicit review are serious methodological shortcomings that render the conclusions of this study invalid.

Conclusions

Criteria studies provide the only direct and specific evidence about unnecessary surgery. Conclusions from the studies with sound methodology indicate that the extent of unnecessary surgery ranges from 13 to 32 percent. If our definition of unnecessary as "useless" or "ineffective" is broadened to include indications judged "equivocal" (i.e., lacking evidence of effectiveness), the percentage of unnecessary surgery increases to upwards of 64 percent. Before generalizations are drawn from these results, we must remember that the operations selected for study were those with considerable geographic variations in use, and about which substantial controversy exists regarding the appropriateness of their use. Thus, these figures are not representative of all of surgery. They furnish evidence, however, that for some operations the problem is far from trivial.

OTHER EVIDENCE OF UNNECESSARY SURGERY

Surgical Rates in Prepaid Group Practice versus Fee-for-Service Practice

It has been shown that enrollees in health maintenance organizations (HMOs) use fewer medical services (Manning, Leibowitz, et al. 1984), and there is some evidence that this reduction in utilization is not associated with impaired health outcomes (Ware, Brook, et al. 1986). Are reductions in surgical rates in prepaid groups evidence of unnecessary surgery in the fee-for-service sector?

Data from the federal employees health benefits programs show substantially lower surgical rates in prepaid health plans than among patients enrolled in Blue Shield plans (Dept. of Health, Education and Welfare 1971). Overall, the number of operations for Blue Shield enrollees was 2.2 times the number for members of the health maintenance organizations. While differences were slight for some operations (e.g., appendectomy and cholecystectomy), for controversial operations such as tonsillectomy the ratio was 2.9:1.

In the only reported exception to the finding of lower surgical rates in prepaid practice, Perkoff found no differences in surgical rates between a prepaid group practice (PGP) and fee-for-service (FFS) practice (Perkoff, Ballinger, et al. 1975). Similar rates were found even for controversial operations such as hysterectomy and tonsillectomy. However, the small enrollments (approximately 1,000 patients in each group) and small number of surgical cases (139 in each group) limit the validity of this study.

LoGerfo found higher rates of surgery for low-income persons in Seattle in fee-for-service practice compared to prepaid practice, but concluded that differences in rates could not be attributed solely to unnecessary surgery in the FFS (LoGerfo, Efird, et al. 1979). While the overall surgical rate in FFS patients was 3.8 times that in the PGP, significant differences persisted even when the analysis was limited to patients whose operations were judged to be "necessary, appropriate or justified." In fact, the ratios of FFS rates to rates for PGP patients in these "appropriate" cases were 6.8:1 for hysterectomy and 2.8:1 for tonsillectomy.

While his conclusion—that all of the difference was not due to unnecessary surgery—seems justified, reanalysis of his data also reveals that 85 percent of the difference in the rate of tonsillectomy *was* accounted for by unjustified use and 26 percent of the difference in hysterectomy rates was accounted for by unjustified use. Although the methods used to render judgments of "justified" or "appropriate" are open to question, this is the leading bit of evidence linking higher surgical rates in FFS to unnecessary surgery.

Peer Review Organizations

In 1984, Congress, in an effort to ensure the quality of care for Medicare patients, as well as to contain expenditures, established the Peer Review Organization(s) (PRO), successors to the Professional Standards Review Organization(s) (PSRO). These independent contractors are charged with improving quality of care and reducing utili-

zation of all health services for Medicare patients, particularly surgery. The quality control efforts have focused on unnecessary admissions and unnecessary surgery.

Targets for unnecessary surgery were developed from published reports on overuse, such as geographic variations data and results of second-opinion programs. When a surgeon proposes one of these operations, approval must be obtained before the patient can be admitted to the hospital or have the operation performed. Approval is based on the results of a screening process using locally derived criteria followed by implicit physician review of any that fail to meet the screening criteria.

The criteria used for review are similar to those used for hospital quality assurance efforts and those that were developed by the specialty societies in the 1970s for the PSRO. These criteria are characterized by their brevity and their orientation toward establishing the *presence of disease*, not whether the proposed treatment is appropriate for the individual patient. Specifically, they do not consider severity of disease, comorbidity, alternative forms of treatment, patient risk, or outcome probabilities.

The general nature of the PRO screening criteria limits their usefulness in identifying unnecessary surgery. Not surprisingly, the aggregate denial rate for the PROs nationwide is only 2.3 percent (Webber 1988).

CONCLUSIONS FROM AVAILABLE DATA

What can we conclude from the available evidence about the extent of unnecessary surgery in the United States?

Geographic variations are indicators of unnecessary surgery, but do not measure it. They largely result from differences in practice style, which in turn reflect the uncertainty of much of medical decision making and the lack of consensus, even among experts, on the appropriate uses of many procedures. Since variations are greater when uncertainty is high, geographic variations provide only circumstantial evidence of unnecessary surgery, but do not provide quantifiable data.

Second-opinion programs also provide only inferential evidence of unnecessary surgery, despite the fact that their nonconfirmation rates are widely quoted as evidence of unnecessary surgery. Even more clearly than geographic variations, the experience with second-opinion programs reveals the effects of uncertainty and lack of consensus in medical practice. The absence of controls and outcomes data for second-opinion programs makes it impossible to draw conclusions from them about unnecessary surgery.

Criteria studies, of which there have been very few, furnish the only specific evidence of unnecessary surgery. While the operations were selected for study specifically because of a perception of widespread inappropriate use, the magnitude of inappropriate care revealed is disturbing, particularly if the fraction of patients who received care for indications of dubious value ("equivocal") is added to that of those who received care judged inappropriate.

In summary, it is not possible to determine the extent of unnecessary surgery from the available data nor to estimate the magnitude of the problem with any degree of accuracy. But it is evident that the amount is not trivial. Available evidence suggests that of highly controversial operations 30 percent or more are performed for clearly inappropriate reasons. If debatable indications are included, more than half would be judged unnecessary. Several of these operations, such as carotid endarterectomy and coronary artery bypass graft surgery, are among the most frequently performed operations. Elimination of those that are unnecessary would result in significant savings of lives and resources.

WHY IS THERE UNNECESSARY SURGERY?

Why do surgeons perform unnecessary operations? Do they do so deliberately or do they act in ignorance? While there may be a few unscrupulous individuals who deliberately perform unnecessary surgery on unsuspecting victims out of greed, the number is surely small. Few surgeons do operations they *know* are unnecessary. Unnecessary surgery is more likely the unfortunate by-product of decision making under conditions of uncertainty, in which a number of factors, personal, social, and economic, influence the decision to recommend an operation.

UNCERTAINTY

As Eddy (1984) has noted, "uncertainty creeps into medical practice through every pore" (p. 75). Accuracy of diagnosis, validity of laboratory results, and effectiveness of treatment are often incompletely known. The physician frequently makes decisions on the basis of inadequate evidence. Uncertainty has a profound effect on the nature of medical practice, producing significant variations in physicians' practice styles. As we have seen, variations in practice style are probably the most important determinants of geographic variations in use of

operations and of nonconfirmations in second-opinion programs. Differences in practice style result from a complex array of factors influencing physician behavior (Eisenberg 1985; Greer 1987, 1988).

SOCIAL FACTORS

A major determinant of practice style is local community consensus. Studies of adoption of new technologies reveal a complex social mechanism whereby physicians incorporate new treatments into practice. As Greer has described the process, typically a local "innovator" introduces a new technology, but other physicians are slow to accept it until a consensus of local peers develops. This consensus is usually preceded by the acceptance of the new modality by the local opinion leader (Greer 1988).

Of particular interest to those seeking to alter this process is the observation that most new product or procedure information is disseminated among physicians by word of mouth and in person. Community physicians distrust the scientific literature and the motives of the scientists. In addition, research findings are often not presented in a way that can be applied easily in the clinical setting. Thus, physicians may not accept new ideas until they have had the opportunity to evaluate them by discussion with experts or peers. Their appraisal of new treatments is guided by the verbal opinions of those they respect.

Clearly, an advantage of this conservative approach is that it acts as a safeguard against too rapid introduction of a treatment that might turn out to be ineffective or even hazardous. The major disadvantage is that even when there is dissemination of scientific evidence of effectiveness of a new treatment, it still may not be accepted.

The method by which outdated or ineffective treatments are *abandoned* is not well understood. In the absence of a highly publicized, definitive study demonstrating that a practice is ineffective, there is little pressure for developing a consensus for abandonment. There is usually no professional or financial reason to give up a long-accepted treatment, and there is seldom an outspoken idea champion to urge its discontinuance. The opinion leader is usually slow to take a highly visible stand. As a result, ineffective treatments or operations are given up much more slowly than they are adopted. The process of abandonment is spotty; some physicians will be found adhering to a practice long after others have abandoned it.

INDIVIDUAL FACTORS

Practice style is also influenced by a host of individual factors that affect physicians in unique ways. These include training and tradition, state of knowledge, personal characteristics, and motivation.

Training and Tradition

Surgeons tend to practice in the way they were taught in residency. They may follow the precepts of a distinguished and revered teacher long after new, more effective practices have been accepted by their colleagues. This behavior is particularly likely to persist if it is successful. If the patients seem to do well and everyone is satisfied, why change? Such behavior is not without merit. Although it may perpetuate some outmoded therapies, it also can prevent adoption of new treatments while their effectiveness is uncertain.

Knowledge

Sometimes physicians do not know what is known. They have not assimilated and used available information that would improve the quality of their decision making. They do not keep up. Doctors as a group are hardworking people, but they are often preoccupied with patient commitments. They often find it difficult to find the time to keep up with the literature.

A more compelling reason why doctors do not keep up is that they *cannot* keep up. The technological explosion of the past 25 years has produced a deluge of information, typically fragmented, unconnected, and difficult to evaluate. Research findings are seldom presented in a form that makes them useful to the practicing clinician. The physician is presented with a variety of treatment options. Because of conflicting and confusing information in the literature it is impossible for an individual physician to evaluate all possible forms of therapy for a given condition (Eddy 1984).

This is both a quantitative problem—the volume alone is overwhelming—and a qualitative one: sorting out the useful information from the plethora of irrelevancy. Despite the abundance of medical journals, or perhaps because of it, there is poor dissemination of clinically relevant information in a form that is useful to the practicing physician.

Personal Characteristics

Age, experience, personality characteristics, and specialty influence practice patterns, such as the ordering and use of tests (Eisenberg and Nicklin 1981). Family practitioners and internists have approaches to patient care that differ from a surgeon's approach. Individual physicians may be more or less risk averse, and this influences their decisions to recommend for or against an operation.

Motivation

A number of factors motivate physician behavior and thus influence practice style. Self-image as a professional is clearly a major one. This includes the doctor's sense of responsibility to the patient and his personal and professional standards. Physicians are also motivated by their preferences for a particular practice style or for certain kinds of patients (Eisenberg 1985; Wennberg and Gittelsohn 1982).

Much has been written in recent years about economic motivation. While characterization of physicians in private practice as "income maximizers" is clearly overly simplistic, in a society where income determines both status and standard of living, maximization of income is surely rational behavior. The evidence that patients in fee-for-service systems have more operations overall than those in prepaid plans has been widely accepted as evidence that economic motivation leads to increased provision of services.

The important question is whether motivation to increase income leads physicians to perform unnecessary surgery. It is probable that at the margin it does. When the patient's situation is clear-cut, it is unlikely that a surgeon would recommend a useless operation in order to increase his income. But many decisions in medicine are problematic, with no clear-cut right or wrong answer. In these uncertain situations the fact that provision of a service will enhance income may influence a physicians' decisions. The proliferation of endoscopic procedures in recent years suggests that this phenomenon is not confined to those who wield the scalpel.

PATIENT FACTORS

The essence of professionalism has been described as the ability to place the interests of the client above one's own. Most observers believe that most physicians do this most of the time. That is, physicians' practice patterns are driven more by a desire to act on behalf of their patients than by their own self-interest (Eisenberg 1985). Unfortu-

nately, this agency role often exacerbates the problems that are generated by professional uncertainty. The average patient prefers action to inaction. Lacking information on probabilities and risks, the physician as patient agent is more inclined to treat than not to treat, particularly when that is what the patient desires.

While scientific decision making involves assessment of *probabilities*, in the face of inadequate data, patients and physicians often think in terms of *possibilities*. The patient wants something done about a problem, particularly if he perceives that inaction poses a significant risk or continued discomfort. An operation might help. It is worth a try—especially if the risk of surgery does not appear too great.

Carotid endarterectomy is a good case in point. The objective of this operation is to prevent a stroke. A patient who has had a transient ischemic attack (TIA) or "little stroke" is justifiably terrified at the possibility of a major stroke. If carotid endarterectomy *might possibly* prevent it, the patient is strongly motivated to have the operation. Similarly, the surgeon is inclined to recommend the procedure if he believes there is a chance it might work. In an action-oriented society where patients expect cures, the recommendation seems preferable to doing nothing.

POLICY IMPLICATIONS

Whether from response to uncertainty, lack of knowledge, failure to change old habits, or response to perceived patient needs, performance of unnecessary surgery results, in the last analysis, from inadequate information. It results from the inadequate production, evaluation, dissemination, and use of information. If the effectiveness of an operation were established, and widely known, the opportunities for misuse would be drastically diminished. To prevent unnecessary surgery it is necessary, first, to define it, to clearly delineate the circumstances under which an operation is not effective. Then it is necessary to disseminate that information in a manner that will induce the appropriate change in surgical practice.

Policymakers who would eliminate unnecessary surgery must therefore concentrate on development and dissemination of methods of determining effectiveness. Ideally, all decisions to recommend surgery should be based on scientific evidence of efficacy. The "gold standard" for assessing efficacy is the randomized clinical trial (RCT). For a number of reasons, not the least of which is their enormous expense, RCTs have been performed for only selected indications for a few operations. Lacking this kind of information, it is necessary to fall back

on analysis of available data from other types of clinical research and outcomes information. These data can be combined with opinions of experts to produce state-of-practice criteria of appropriateness. Such criteria are far more likely to be valid than ad hoc decisions by individual practitioners. They can be used to develop practice guidelines or standards. An analysis of efforts to develop criteria of clinical effectiveness is presented in the next section.

DEVELOPING MEANINGFUL PRACTICE GUIDELINES

Practice guidelines[1] are standardized specifications for care, either for the use of a particular procedure or for the management of a specific clinical problem. To be clinically useful, guidelines must be specified in sufficient detail to distinguish what is from what is not appropriate care. Ideally, guidelines are derived from evidence of effectiveness.

Various types of guidelines have been used by physicians for years: protocols for diagnosis and management of many forms of cancer, recommendations for the use of screening tests such as mammography, protocols for trauma care, and manuals for diagnosis and treatment. In the performance of many operations the surgeon follows highly specific technical guidelines. Guidelines to be used for assessing the quality of care, however, need to be much more comprehensive and detailed, and they must provide specific recommendations.

Guidelines developed for the RAND/UCLA Health Services Utilization Study are a good example of useful guidelines. The criteria for the use of coronary angiography specify which indications would be appropriate, inappropriate, or equivocal. One of these criteria follows.

> Coronary angiography is indicated in a patient with chronic stable angina (without strong contraindications to CABG surgery) in whom angina occurs with mild exertion (Class III or IV) and who has received maximal medical management and (has) a very positive exercise ECG, a negative exercise thallium scan, and a positive exercise MUGA. (Chassin, Kosecoff, et al. 1986a, 84)

Despite the apparent need for a more thorough and sophisticated method for defining quality of care, the medical profession has been slow to develop and use practice guidelines. Most doctors do not perceive a need for detailed guidelines or standards. They are not dissatisfied with the quality of care they provide, and from the practitioner's perspective individualized decision making is more satisfying and seems more professional than use of standardized criteria or algo-

rithms. Many are skeptical about the capacity of existing methodology to arrive at adequate and useful definitions of quality of care. They point to the lack of professional agreement on many aspects of care.

As a result, most guidelines are superficial and too general to be meaningful. They have been developed in response to external requirements, such as from the PSRO or Joint Commission on the Accreditation of Healthcare Organizations (JCAHO), and often they have been oriented more to cost containment than to improving quality of care. Until recently, except for the American College of Physicians, professional organizations have been reluctant to take responsibility for developing guidelines. In addition to the foregoing reasons, they have had concerns about potential liability and misuse. Within the last few years, the American College of Cardiology, the American Heart Association, the American Society for Gastrointestinal Endoscopy, the American College of Obstetrics and Gynecology, and the American Society of Anesthesiologists have begun to develop practice guidelines.

TYPES OF OPERATIONS THAT NEED GUIDELINES

Guidelines are not needed for all operations. Their major value is for procedures that are controversial and have a significant effect on the patient or the health care system. Guidelines are appropriate for

1. Operations that have unusual potential to harm the patient, for example, those with a significant risk of complications or loss of life,
2. Operations that involve extensive use of resources, both human and financial; examples: organ transplantation, open heart surgery, and hip replacement,
3. Operations that are controversial or in part outdated, or
4. Operations suspected of overuse or inappropriate use, particularly if they are performed frequently, for example, pacemaker insertion and cataract surgery.

Developing guidelines for a relatively small number of operations could have a major effect on the quality of surgical care, for just a few diagnoses account for a major share of operations performed. For example, it is estimated that ten operations and procedures account for 50 percent of surgical charges for Medicare patients.

CHARACTERISTICS OF USEFUL GUIDELINES

Guidelines will be more effective in improving care if they are perceived by decision makers to be of high quality. Listed are the characteristics of good guidelines.

1. They must be *comprehensive*, including all likely indications for the use of the operation.
2. They must be *specific*, clearly describing the exact conditions for which the operation is recommended.
3. They should describe in meaningful *detail* the distinguishing features that separate one indication from another.
4. They should clearly indicate the circumstances under which an operation is *appropriate* and *inappropriate*.
5. They must be *inclusive* of all major relevant additional factors to be taken into consideration in the decision to recommend an operation.
6. They must be *manageable*. Despite the above requirements, the number and complexity of indications must not be so great that the guidelines are difficult to use. Guidelines must be presented in a form and in language that make them easy to understand and to implement.

EXPERIENCE WITH GUIDELINE DEVELOPMENT

A number of professional groups and payers have constructed practice guidelines in the past few years. Guidelines developed by eight widely different groups are compared in Table 6 to the criteria just listed.

Specialty Societies Guidelines

In recent years, a number of medical specialties have taken on the responsibility to set practice guidelines for procedures in their specialty. Most of these have been detailed lists of tests or procedures to be considered for a particular patient. Other specialty guidelines are scientific papers in which an expert focuses on the background, indications, and limitations of a procedure or test. The common shortcoming of these types of guidelines is that the important clinical variables are not specified in enough detail to readily discriminate between care that is appropriate and that which is not.

An exception has been the guidelines generated by the Task Force on Assessment of Diagnostic and Therapeutic Cardiovascular Procedures of the American College of Cardiology and the American Heart

Table 6: Evaluation of Existing Practice Guidelines

Criterion	*Specialty Societies*	*CEAP*	*ACC/ AHA*	*DATTA*	*JCAHO*	*NIH*	*RAND*	*Payer*
Guidelines								
Comprehensive—include all likely indications	NO	NO	YES	NO	NO	NO	YES	NO
Specific—describe exact conditions	NO	NO	YES	NO	NO	NO	YES	NO
Detailed—describe distinguishing features	NO	NO	YES	NO	NO	NO	YES	NO
Distinguish appropriate from inappropriate	NO	NO	YES	NO	NO	NO	YES	NO
Inclusive—consider all relevant factors	NO	YES	YES	NO	NO	NO	YES	NO
Manageable—not too complex to be usable	YES	NO	+/-	YES	YES	YES	+/-	YES

Association (1987). The Guidelines for Coronary Angiography are highly specific and place each indication into one of three classes: Class I, where there is "general agreement that coronary angiography is justified"; Class II, where there is "divergence of opinion"; and Class III, where there is "general agreement that angiography is not ordinarily justified."

These guidelines specify the details of the clinical and laboratory findings that discriminate among candidates in a meaningful way. They are specific, comprehensive, and detailed, indicating clearly when the procedure is appropriate and when it is inappropriate.

The Clinical Efficacy Assessment Project (CEAP)

CEAP began in 1976 as the Medical Necessity Project, a joint undertaking of the American College of Physicians (ACP) and the Blue Cross and Blue Shield (BC/BS) organization. Its purpose was "to identify outmoded tests thereby eliminating reimbursements for useless medical procedures." Since 1981, it has been carried out by the ACP single handedly for the purpose of elevating the standards of medical practice. A long list of tests and procedures used in internal medicine has been evaluated (but no surgical operations). The results recently have been accepted by national BC/BS as guidelines, with recommendation for adoption by local BC/BS organizations.

The major shortcoming of many of the CEAP guidelines is their lack of specificity. While recommendations are comprehensive in

scope, they are not stated in terms that permit clear discrimination between the specific clinical differences that determine whether a test or procedure is or is not appropriate for a given individual.

Diagnostic and Therapeutic Technology Assessment Project of the American Medical Association (DATTA)

This American Medical Association (AMA) project polls a large number of specialists (15–120) for their opinions on the safety and effectiveness of new technologies. The questions posed are very general, and no attempt is made to be comprehensive in scope or specific in description. Respondents are asked to classify only safety or effectiveness in a single, general clinical situation.

Joint Commission on Accreditation of Healthcare Organizations

The Joint Commission (JCAHO) requires hospital surgical departments to develop and apply standards for indications for operations. These are derived locally by each individual hospital. Most are neither comprehensive nor specific enough to discriminate among indications except at a general level. Consequently, physicians usually find them of little relevance or value except to identify truly egregious abuse.

Consensus Development Project of the National Institutes of Health (NIH)

Sponsored by the Office of Medical Applications of Research of NIH, the Consensus Conferences bring together a prestigious panel of bioscientists and nonmedical experts to hear evidence and testimony in open forum concerning the efficacy of new technologies. The panel then adjourns to write its draft report in private. The draft report is considered during a second hearing for commentary, and the panel then writes a final revised report.

Although discussions are erudite and often fairly thorough, recommendations are general and are not framed in a format that permits discrimination among similar candidates for a procedure. No attempt is made to be all-inclusive in considering indications for the procedure.

RAND/UCLA Health Services Utilization Study

As part of the Health Services Utilization study of the extent and causes of geographic variations in the use of services, the RAND/UCLA Corporation developed a modified Delphi and interactive group technique for obtaining expert consensus on the appropriateness

of indications for six operations and procedures. Highly specific and mutually exclusive indications were derived after comprehensive literature review and consultations with experts. These indications were then rated for appropriateness by panels that were carefully balanced with medical and surgical specialists and generalists (Park, Fink, et al. 1986).

The RAND/UCLA guidelines are the most comprehensive and the most specific that have been developed. The factors on which a judgment of appropriate or inappropriate is based are inclusive and specified in detail. The resulting guidelines clearly distinguish between appropriate and inappropriate indications, as well as an intermediate "indeterminate" category. The use of RAND/UCLA guidelines in practice has not been evaluated.

Third Party Payer Guidelines

Payers have long found it necessary to have criteria for payment. For the most part, these do not constitute practice guidelines as we have defined them. Since the purpose is to reduce expenditures, they typically consist merely of a determination that a procedure is not effective and, therefore, that it will not be paid for. These guidelines are usually simple, specific, and easily understood, but they do not specify in detail which indications for a procedure are considered appropriate and which are inappropriate. Associated conditions and risk factors are also rarely part of the definitions.

PROCESS FOR DEVELOPING GUIDELINES

The quality of the process for developing guidelines is critical, for it determines both their usefulness and the likelihood of their acceptance by the profession. Credibility is enhanced if guidelines are developed through a rigorous, structured process that synthesizes the information in the scientific literature, extends it through the knowledge of expert physicians, and expresses the information in statements that are specific, precise, and comprehensive.

Where possible, guidelines should be based on scientific evidence of effectiveness of the operation. The ideal standard is the randomized clinical trial. Unfortunately, there are few operations for which such trials have been conducted. In the absence of hard scientific data investigators have turned to less precise methods, such as structured synthesis of evidence using meta-analysis or decision analysis, and use of expert opinion. Practice guidelines thus typically represent a marriage of evidence, experience, and opinion. The art lies in getting the "best"

opinions. Various group judgment methods have been used to accomplish this.

A suitable method of synthesizing evidence and extending it by expert clinical opinion for the development of practice guidelines would include the following characteristics:

1. The available scientific data in the literature are critically analyzed and synthesized to provide usable evidence of effectiveness.
2. The method for selecting experts for consensus panels minimizes bias and ensures appropriate representation.
3. The panel process maximizes exchange of ideas and minimizes dominance by individuals.
4. The panel process includes explicit consideration by panelists of outcome probabilities and risks when making their judgments of appropriateness.
5. The results of the process meet the criteria for good guidelines, that is, they are comprehensive, specific, detailed, and inclusive; they distinguish between appropriate and inappropriate; and they are not unduly complex.
6. The results have validity when compared with other evidence.
7. The results are reproducible. Duplicate panels achieve similar results.
8. The method is acceptable to all interested parties—physicians, patients, regulatory authorities, hospitals, and payers.
9. The method is practical. It can be replicated for a large number of operations at reasonable expense.

Of the methods that have been developed to date, none meets all of these criteria. The CEAP, American College of Cardiology/American Heart Association (ACC/AHA), and RAND/UCLA techniques contain many of the essential ingredients, however, and have a number of features in common. Whether they can be transformed into guidelines that are easily used in practice remains to be seen. The methodology of developing guidelines is now sufficiently advanced, however, to make it clear that reasonable additional effort can lead to development of a process to produce clinically relevant criteria.

WHAT CAN BE DONE ABOUT UNNECESSARY SURGERY?

Unnecessary surgery is a problem of unknown dimension but undeniable significance. While it is impossible to estimate the extent of unnecessary surgery in the United States with any precision, we have seen that for some controversial procedures 30 percent or more may be performed for inappropriate reasons. Elimination of these operations would result in significant savings of lives and resources.

How can the amount of unnecessary surgery be reduced? The issue boils down to how to identify potential unnecessary surgery and how to prevent it. Identification requires the use of guidelines specific enough to distinguish useless operations from those of value. Prevention requires that the guidelines be followed as standards by physicians when surgery is recommended. Control of unnecessary surgery, then, depends on the development and effective use of guidelines.

Four questions in the use of guidelines are of major importance to policymakers:

1. How should guidelines be developed?
2. Who should be responsible for developing guidelines?
3. Who should pay for the development of guidelines?
4. How should guidelines be used?

METHODOLOGICAL ISSUES IN GUIDELINE DEVELOPMENT

Significant progress has been made in the development of guideline methodology within the past few years. While the methods used by the ACP, the ACC/AHA, and the RAND/UCLA study meet many of the requirements for credibility, acceptability, and feasibility, a number of methodological issues remain to be addressed before these or other methods are put to use:

1. What are the best methods for evaluating and synthesizing the information in the scientific literature? Is meta-analysis required?
2. What factors should be included in the determination of appropriateness of an indication for a service or procedure?
3. If the following are included, how is each best evaluated?
   - Effectiveness
   - Safety

- Alternative therapies
- Risk/benefit
- Patient preferences.

4. Is a consensus method the most appropriate way to summarize and codify expert opinion?
   - What are the pros and cons of face-to-face, Delphi, voting, and survey methods?
   - How reliable (reproducible) are consensus methods?
   - Who should be on the panels? (specialists who perform the procedure, referring specialists, generalists, laymen, administrators, payers?)
   - How do you minimize bias, dominance, the "halo" effect in the panel process?

5. How should consensus methods be validated?
   - Face and content validity?
   - Specialist review, comparison to implicit reviews?

6. Is there an effective process for implementation, periodic review, and revision of the guidelines?

WHO SHOULD BE RESPONSIBLE FOR DEVELOPING GUIDELINES?

Professional Societies

Are professional societies the logical groups to develop guidelines? On the surface, the answer would seem to be yes. The development and use of guidelines is an appropriate function for professional societies, and the surest way for a society to retain control over its most important *raison d'être*, maintenance of high standards of care. The American College of Physicians, the American Heart Association, and the American College of Cardiology have already taken the lead in developing useful guidelines. The recent interest of the American Medical Association in fostering the development of practice guidelines is a welcome advance.

Some have questioned whether specialty organizations are sufficiently free from bias and self-interest to be objective. If these objections can be overcome by inclusion of specialists from other disciplines on the expert panels, should professional society standard-setting activities be encouraged?

Academic Centers

The lack of agreement on methodology and the past reluctance of professional societies to take responsibility for generating guidelines has led some academic medical centers to play a role in the development of the appropriate methodology for deriving guidelines. These institutions are able to focus considerable intellectual and clinical resources, and are less open to charges of specialty professional or financial self-interest. They, too, represent the taking of responsibility by the profession itself. Can their efforts be coordinated toward the development of an acceptable methodology? If so, how should it be supported?

Government

Should government be responsible for developing guidelines for practice? While governmental agencies could more readily command the necessary financial resources than other interested parties, there are important reasons why it would not be desirable for them to develop guidelines. First, it would give government an unprecedented level of control over the nature of medical practice. It is unlikely that this would be tolerated by either the profession or the public. Second, the pressure on governmental agencies, particularly HCFA, to control costs would lead doctors and the public to question whether the purpose of the guidelines was quality control or reduction in utilization. Finally, there are concerns that the rigidity of the government's bureaucratic structure is such that revising and updating guidelines might be a difficult undertaking, inhibiting practice and progress.

At present, the methodology for developing guidelines and their actual generation is proceeding irregularly and in an uncoordinated fashion. Should the government facilitate or stimulate this process? Should one of its agencies, such as HCFA or the National Center for Health Services Research set standards for the development process or its results? If so, the obvious mechanism is through funding.

FUNDING THE DEVELOPMENT OF GUIDELINES

Developing the methodology for deriving guidelines is a time-consuming and expensive task, as is the actual production of the guidelines once the methods have been agreed on. It is unlikely that either professional societies or academic medical centers can provide the necessary resources.

On the other hand, the task is manageable. Application of guide-

lines to a relatively small number of operations could have a major effect on unnecessary surgery. The first targets of a standard-setting process should be those frequently performed operations for which there is a lack of consensus regarding their appropriateness. These would include: cesarean section, hysterectomy, tonsillectomy, hip replacement, prostatectomy, disk surgery, joint operations, coronary bypass surgery, cataract surgery, and carotid endarterectomy.

If such an effort is to occur, it will require support from private foundations and the government. Guidelines are a classic "public good," one from which everyone should benefit. Thus, support of their development is a logical government function. As the largest payer for medical care and as guardian of the public trust, the government (HCFA) also has a responsibility to ensure that its funds are efficiently spent. Elimination of ineffective care is clearly an appealing way to save health care dollars.

HOW SHOULD GUIDELINES BE USED?

While development of meaningful guidelines will permit identification of unnecessary surgery, little will be gained unless the guidelines are also used to dissuade performance of the useless operations they identify. To abolish unnecessary surgery it is necessary to change physician behavior, motivating doctors to abandon operations that are useless. There are several ways in which this can be done.

Education and Feedback

If physicians are furnished with credible information on effectiveness by means of carefully crafted authoritative guidelines, will unnecessary surgery disappear? Wennberg's experience with feedback suggests it will (Wennberg 1984). Others have shown that local educational efforts and dissemination of information can bring about changes in practice patterns (Griner 1979; Berwick and Coltin 1986). Unfortunately, in the absence of continuing incentives, such changes are often transient (Eisenberg 1977). In any educational system, support of the leadership and continuing effort are required for long-term success. For example, the NIH consensus panels have not had such interest and support, and have been found to be ineffective in altering physician practice patterns (Kanouse, Winkler, et al. 1987). On balance, the evidence from voluntary efforts to change physician behavior is not reassuring.

Local Standards, Locally Enforced

Can hospitals incorporate guidelines into their quality-assurance mechanisms in a way that changes practice patterns? There is little evidence that current quality-assurance activities either identify or significantly reduce unnecessary surgery. In part, this may be because most utilization review and precertification programs rely on implicit review or use locally developed explicit criteria that are not specific enough. No hospital has the resources necessary to develop detailed guidelines for more than a few, if any, procedures. Since quality-assurance programs consume a tremendous amount of human and financial capital and also disrupt the medical decision-making process, it is appropriate to ask whether they are worth what they cost.

If hospital programs are to be relied on as serious guardians of the quality of care, they require significant modification. One way in which they might be made effective is by adoption of explicit review and by use of meaningful guidelines from a credible national source. Whether hospitals have the will and the motivation to do this is open to question. The JCAHO could play a catalytic role in the process. If it were to insist on the use of appropriate guidelines as standards for the performance of operations, hospitals could become the effective agents of quality control in surgery.

National Guidelines as Standards for Payment

The most efficient incentive for change in physician behavior is reimbursement policy. Virtually all payers have lists of operations that they will not pay for—largely outdated and discredited procedures, about which there is wide consensus. Would payers also use more detailed and sophisticated guidelines that distinguish not among operations, but among specific indications for a single operation? Almost certainly. If the guidelines were authoritative and accepted as reasonable by the profession, they would be welcomed by payers as a rational and desirable means of reducing expenses without lowering the quality of care.

Would the government use guidelines for determining eligibility for payment under Medicare and Medicaid? Again, the answer is almost certainly yes. William Roper, head of HCFA in the Reagan administration, launched an "effectiveness initiative" pledged to identify which services work and which do not, with the declared intent to pay only for the former (Roper 1988). Detailed, authoritative guidelines acceptable to the profession are the ideal way to accomplish this objective.

Peer Review Organizations are well aware of the limitations of locally derived standards and are searching for more effective methods of quality control. Suitably derived guidelines that specify in detail which indications for surgery are inappropriate might be welcomed by the PROs. Use of nationally recognized and credible guidelines would be more effective than local standards in identifying unnecessary surgery, and they would be more acceptable to physicians. They also would undoubtedly be more economical since the review process could be simplified.

CONCLUSION

However they are used, well-conceived and practical practice guidelines have the potential for improving the quality of medical decision making. They offer the best opportunity for eliminating unnecessary surgery.

ACKNOWLEDGMENTS

The author gratefully acknowledges the reviews and critiques of this synthesis by Robert H. Brook, M.D., Deputy Director, Health Sciences Program, the RAND Corporation; Dennis O'Leary, M.D., President, Joint Commission on the Accreditation of Healthcare Organizations; and Donald M. Berwick, M.D., Vice-President for Quality Care Management, Harvard Community Health Plan. While each made valuable suggestions and criticisms, the responsibility for the final product is, of course, solely mine. I also want to thank the Pew Memorial Trust for its sponsorship of the synthesis series, which served as the motivation for this review.

NOTE

1. Much of this section has been excerpted and adapted from the author's contribution to the chapter "Practice Guidelines" in the 1989 "Annual Report of the Physician Payment Review Commission." Washington, DC: Government Printing Office, 1989.

REFERENCES

American College of Cardiology/American Heart Association Task Force on Assessment of Diagnostic and Therapeutic Cardiovascular Procedures. "Guidelines for Coronary Angiography." *Journal of the American College of Cardiology* 10, no. 4 (1987):935-50.

American College of Surgeons and American Surgical Association. *Surgery in the United States: A Summary Report of the Study on Surgical Services for the United States*. Chicago: American College of Surgeons and American Surgical Association, 1975.

American College of Surgeons. *Second Surgical Opinion Programs: A Review and Progress Report*. Chicago: American College of Surgeons, 1982.

American Medical Association. *Status Report on Second Surgical Opinion Programs*. Chicago: AMA Division of Health Policy and Program Evaluation, Dept. of Health Care Financing and Organization, 1980.

Anderson, G. M., and J. Lomas. "Determinants of the Increasing Cesarean Birth Rate." *New England Journal of Medicine* 311, no. 14 (1984):887-92.

Berwick, D. M., and K. L. Coltin. "Feedback Reduces Test Use in a Health Maintenance Organization." *Journal of the American Medical Association* 255, no. 11 (1986):1450-54.

Bombardier, C., V. R. Fuchs, et al. "Socioeconomic Factors Affecting the Utilization of Surgical Operations." *New England Journal of Medicine* 297, no. 13 (1977):699-705.

Bunker, J. P. "Surgical Manpower. A Comparison of Operations and Surgeons in the United States and in England and Wales." *New England Journal of Medicine* 282, no. 3 (1970):135-44.

Bunker, J. P., and B. W. Brown. "The Physician-Patient as an Informed Consumer of Surgical Services." *New England Journal of Medicine* 290, no. 19 (1974):1051-55.

Cageorge, S. M., L. L. Roos, and R. Danzinger. "Gallbladder Operations: A Population-Based Analysis." *Medical Care* 19, no. 5 (1981):510-25.

Chassin, M. R., J. Kosecoff, et al. "Does Inappropriate Use Explain Geographic Variations in the Use of Health Care Services?" *Journal of the American Medical Association* 258, no. 18 (1987):2533-37.

_____. *Indications for Selected Medical and Surgical Procedures—A Literature Review, and Ratings of Appropriateness: Coronary Angiography*. Publication R-3204/1. Santa Monica, CA: The RAND Corporation, 1986a.

Chassin, M. R., R. H. Brook, et al. "Variations in the Use of Medical and Surgical Services by the Medicare Population." *New England Journal of Medicine* 314, no. 5 (1986b):285-90.

Diehr, P. "Small Area Statistics: Large Statistical Problems." *American Journal of Public Health* 74, no. 4 (1984):313-14.

Donabedian, A. *The Criteria and Standards of Quality*. Ann Arbor, MI: Health Administration Press, 1982.

Doyle, J. C. "Unnecessary Hysterectomies." *Journal of the American Medical Association* 151, no. 5 (1953):360-65.

Eddy, D. M. "Variations in Physician Practice: The Role of Uncertainty." *Health Affairs* 3, no. 2 (1984):74-89.

Eisenberg, J. M. "An Educational Program to Modify Laboratory Use by House Staff." *Journal of Medical Education* 52 (1977):578-81.

_____. "Physician Utilization: The State of Research about Physicians' Practice Patterns." *Medical Care* 23, no. 5 (1985):461-83.

Eisenberg, J. M., and D. Nicklin. "Use of Diagnostic Services by Physicians in Community Practice." *Medical Care* 19, no. 3 (1981):297.

Elliott, R. V., K. A. Kahn, and R. Kaye. "Physicians Measure Up." *Journal of the American Medical Association* 245, no. 6 (1981):595-600.

Emerson, R. "Unjustified Surgery: Fact or Myth?" *New York State Journal of Medicine* 76, no. 3 (1976):454-60.

Finkel, M. L., H. S. Ruchlin, and S. K. Parsons. *Eight Years' Experience with a Second Opinion Elective Surgery*. Baltimore, MD: HCFA Office of Research, Demographics and Statistics, 1981.

Fuchs, V. R. *The Health Economy*. Cambridge, MA: The Harvard University Press, 1986.

Gertman, P. M., D. A. Stackpole, et al. "Second Opinions for Elective Surgery." *New England Journal of Medicine* 302, no. 21 (1980):1169-74.

Gittelsohn, A., and J. E. Wennberg. "The Authors Respond." *Journal of the Maine Medical Association* 68 (February 1977):53-57.

Grafe, W. R., C. K. McSherry, et al. "The Elective Surgery Second Opinion Program." *Annals of Surgery* 188, no. 3 (1978):323-30.

Greenspan, A. M., H. R. Kay, et al. "Incidence of Unwarranted Implantation of Permanent Cardiac Pacemakers in a Large Medical Population." *New England Journal of Medicine* 318, no. 3 (1988):158-63.

Greer, A. L. "The State of the Art Versus the State of the Science." *International Journal of Technical Assessment in Health Care* 4, no. 1 (1988):5-25.

_____. "The Two Cultures of Biomedicine: Can There be Consensus?" *Journal of the American Medical Association* 258, no. 19 (1987):2739-40.

Griner, P. F. "Use of Laboratory Tests in a Teaching Hospital: Long-Term Trends—Reductions in Use and Relative Cost." *Annals of Internal Medicine* 90, no. 2 (1979):243.

Hulka, B. S., and J. R. Wheat. "Patterns of Utilization: The Patient Perspective." *Medical Care* 23, no. 5 (1985):438-60.

Joffe, J. "Evaluating a Voluntary Second Surgical Opinion Program." *Evaluation and Health Professions* 3, no. 4 (1980):421-33.

Kanouse, D. E., J. D. Winkler, et al. "The NIH Consensus Conferences." Publication No. R-3060. Santa Monica, CA: The RAND Corporation, 1987.

Lembcke, P. A. "Measuring the Quality of Medical Care through Vital Statistics Based on Hospital Service Areas: 1. Comparative Study of Appendectomy Rates." *American Journal of Public Health* 42, no. 3 (1952): 276-86.

Lewis, C. E. "Variations in the Incidence of Surgery." *New England Journal of Medicine* 281, no. 16 (1969):880-84.

LoGerfo, J. P., R. A. Efird, et al. "Rates of Surgical Care in Prepaid Group Practices and the Independent Setting." *Medical Care* 17, no. 1 (1979): 1-10.

Manning, W. G., A. Leibowitz, et al. "A Controlled Trial of the Effect of a Prepaid Group Practice on Use of Services." *New England Journal of Medicine* 310, no. 23 (1984):1505-10.

Martin, S. G., M. Shwartz, et al. "Impact of a Mandatory Second-Opinion Program on Medicaid Surgery Rates." *Medical Care* 20, no. 1 (1982): 21-45.

_____. *The Effect of a Mandatory Second-Opinion Program on Medicaid.* Washington, DC: Health Care Financing Administration, 1980.

McCarthy, E. G., and G. W. Widmer. "Effects of Screening by Consultants on Recommended Elective Surgical Procedures." *New England Journal of Medicine* 291, no. 25 (1974):1331-35.

McCarthy, E. G., and M. L. Finkel. "Second Opinion Elective Surgery Programs: Outcome Status Over Time." *Medical Care* 16, no. 12 (1978): 984-94.

_____. "Second Consultant Opinion for Elective Orthopedic Surgery." *American Journal of Public Health* 71, no. 11 (1981):1233.

_____. "Surgical Utilization in the U.S.A." *Medical Care* 18, no. 9 (1980): 883-92.

McCarthy, E. G., M. L. Finkel, and H. S. Ruchlin. "Second Opinions on Elective Surgery." *The Lancet* no. 8234 (20 June 1981):1352-53.

McNeil, B., and S. G. Pauker. "On the Elicitation of Preferences for Alternative Therapies." *New England Journal of Medicine* 306, no. 21 (1982): 1259-62.

McPherson, K., J. E. Wennberg, et al. "Small-Area Variations in the Use of Common Surgical Procedures: An International Comparison of New England, England, and Norway." *New England Journal of Medicine* 307, no. 21 (1982):1310-14.

Merrick, N. J., R. H. Brook, et al. "Use of Carotid Endarterectomy in Five California Veterans Administration Medical Centers." *Journal of the American Medical Association* 256, no. 18 (1986):2531-35.

New Jersey Dept. of Human Services. *Evaluation of Medicaid Second Opinion Program.* Trenton, NJ: New Jersey Dept. of Human Services, 1984.

Park, R. E., A. Fink, et al. "Physician Rating of Appropriate Indications for Six Medical and Surgical Procedures." *American Journal of Public Health* 76, no. 7 (1986):766-72.

Pauly, M. V. "What is Unnecessary Surgery?" *Milbank Memorial Fund Quarterly* 57, no. 1 (1979):95-117.

Payne, S. M. C. "Identifying and Managing Inappropriate Hospital Utilization." *Health Services Research* 22, no. 5 (1987):709-69.

Perkoff, G. R., W. F. Ballinger, et al. "Lack of Effect of an Experimental Prepaid Group Practice on Utilization of Surgical Care." *Surgery* 77, no. 5 (1975):619-23.

Poggio, E., et al. *Second Surgical Opinion Programs: Analysis of Public Policy Options.* Washington, DC: National Technical Information Service, March 1985.

Poggio, E., et al. *Second Surgical Opinion Programs: An Investigation of Mandatory and Voluntary Alternatives.* Cambridge, HA: Abt Associates Inc., September 1981.

Roenigk, D., and L. Bartlett. *Controlling Medicaid Costs: Second Surgical Opinion Programs.* Washington, DC: National Governors' Association, Center for Policy Research, November 1982.

Roos, N. P. "Hysterectomy: Variations in Rates Across Small Areas and Across Physicians' Practices." *American Journal of Public Health* 74, no. 4 (1984):327-34.

Roos, N. P., and L. L. Roos. "High and Low Surgical Rates: Risk Factors for Area Residents." *American Journal of Public Health* 71, no. 6 (1981): 591-600.

_____. "Surgical Rate Variations: Do They Reflect the Health or Socioeconomic Characteristics of the Population?" *Medical Care* 20, no. 9 (1982): 945-58.

Roos, N. P., L. L. Roos, and P. D. Henteleff. "Elective Surgical Rates—Do High Rates Mean Lower Standards?" *New England Journal of Medicine* 297, no. 7 (1977):360-65.

Roper, W. L., W. Winkenwerder, et al. "Effectiveness in Health Care." *Journal of the American Medical Association* 319, no. 18 (1988):1197-1202.

Ruchlin, H. S., M. L. Finkel, and E. G. McCarthy. "The Efficacy of Second-Opinion Consultation Programs: A Cost-Benefit Perspective." *Medical Care* 20, no. 1 (1982):3-20.

Stockwell, H., and E. Vayda. "Variations in Surgery in Ontario." *Medical Care* 17, no. 4 (1979):390-96.

U.S. Congress. House. Subcommittee on Oversight and Investigations. *Cost and Quality of Health Care: Unnecessary Surgery*. Washington, DC: Government Printing Office, 1976.

_____. *Quality of Surgical Care*. Washington, DC: Government Printing Office, 1978.

U.S. Congress. Senate. Committee on Aging. *Unnecessary Surgery: Double Jeopardy for Older Americans*. Washington, DC: Government Printing Office, 1985.

U.S. Dept. of Commerce. *Statistical Abstract of the United States*. Washington, DC: Government Printing Office, 1988, 100.

U.S. Dept. of Health and Human Services. Health Care Financing Administration. *Physician-Induced Demand for Surgical Operations*. Report prepared by J. B. Mitchell and J. Cromwell. Washington, DC: Government Printing Office, 1981.

U.S. Dept. of Health, Education and Welfare. "The Federal Employees Health Benefits Program—Enrollment and Utilization of Health Services, 1961-1968." Washington, DC: Government Printing Office, 1971.

Vayda, E. "A Comparison of Surgical Rates in Canada and in England and Wales." *Manpower* 289, no. 23 (1973):1224-29.

Vayda, E., J. M. Barnsley, et al. "Five-Year Study of Surgical Rates in Ontario's Counties." *Canadian Medical Association Journal* 131 (15 July 1984): 111-15.

Vayda, E., W. R. Mindell, et al. "Measuring Surgical Decision-Making with Hypothetical Cases." *Canadian Medical Association Journal* 127 (15 August 1982):287-90.

Ware, J. E., R. H. Brook, et al. "Comparison of Health Outcomes at a Health Maintenance Organization with Those of Fee-For-Service Care." *Lancet* 1, no. 8488 (1986):1017-22.

Webber, A., personal communication, 1988.

Webster's 3rd New International Dictionary. Springfield, MA: G & C Merriam Co., 1976.

Wennberg, J. E. "Dealing with Medical Practice Variations: A Proposal for Action." *Health Affairs* 3, no. 2 (1984):6-31.

_____. "Which Rate is Right?" *New England Journal of Medicine* 314, no. 5 (1986):310-11.

Wennberg, J. E., and A. Gittelsohn. "Health Care Delivery in Maine I: Patterns of Use of Common Surgical Procedures." *Journal of the Maine Medical Association* 66, no. 5 (1975):123-49.

_____. "Small Area Variations in Health Care Delivery." *Science* 142, no. 4117 (1973):1102-08.

_____. "Variations in Medical Care among Small Areas." *Scientific American* 246, no. 4 (1982):120-34.

Wennberg, J. E., and F. J. Fowler. "A Test of Consumer Contribution to Small Area Variations in Health Care Delivery." *Journal of the Maine Medical Association* 68 (August 1977):275-79.

Wennberg, J. E., B. A. Barnes, and M. Zuboff. "Professional Uncertainty and the Problem of Supplier-Induced Demand." *Social Science Medicine* 16, no. 7 (1982):811-24.

Winslow, C. M., D. H. Solomon, et al. "The Appropriateness of Carotid Endarterectomy." *New England Journal of Medicine* 318, no. 12 (1988): 721-27.

Winslow, C. M., J. B. Kosecoff, et al. "The Appropriateness of Performing Coronary Artery Bypass Surgery." *Journal of the American Medical Association* 260, no. 4 (1988):505-09.

Wisconsin Dept. of Health and Social Services. *Surgery Rates under the Medicaid Second Surgical Opinion Program.* Madison, WI: Bureau of Evaluation, Division of Policy and Budget, March 1984.