Control over the Utilization of Medical Services

BRUCE STUART

RONALD STOCKTON

During recent years, the health care industry has been characterized by rapid increases in the volume of services delivered. This escalation is in part unjustified by medical need, and has produced a variety of efforts on the part of payers and providers to restrict overuse. In this article the authors consider the issues and problems involved in the control of medical utilization. Five categories of control are considered in detail: supply limitations, financial disincentives, authorization requirements, review mechanisms, and legal action. The article suggests that the success or failure of these various control mechanisms hinges upon four factors: whose use is being regulated, who performs the control activities, whether the attempted control involves a judgment as to the appropriateness of treatment, and whether the attempt to control occurs before, after, or during treatment. It is concluded that most current forms of utilization control suffer from ambiguity of purpose, organizational inefficiency, and undesirable side effects. The authors offer several proposals to correct these shortcomings, but conclude that the only long-range solution to overutilization lies in a more integrated approach to medical resource allocation and a consequent change in the structure of provider and user incentives.

Introduction

Notwithstanding current Federal price controls, inflation in the medical sector is still very much an issue of public interest. In fact, concern over the *price* of medical services has tended to draw attention away from an equally important determinant of growth in health care expenditures, namely, *utilization*. The public is well aware that medical prices have risen dramatically over the past ten years but is less conscious of the fact that approximately half the total increase in medical expenditures during the 1960s was the result of increased utilization and not price inflation (Klarman et al., 1970). Normally, a rise in the use of service does not elicit the

need should be satisfied, for clearly there is a point at which the benefit of providing an additional medical service is more than offset by the sacrifice in other goods that must be made in order to purchase the service. Some forms of utilization constraints (such as benefit exclusions under a health insurance policy) may thus be viewed as allocative devices designed to hold in check the expenditures for less essential types of medical care. Whether such constraints are properly applied only to services reflecting a low economic priority is another matter. But whatever the judgment, it is clear that the appropriateness of medical services may be viewed from an economic as well as a medical perspective.

Classification of Utilization Controls

Unfortunately, it is nearly impossible to separate those utilization controls whose primary effect is to reduce medically unnecessary care from others designed to limit services with a low economic priority. For example, the most prevalent type of utilization control practiced in American hospitals takes the form of hospital-based utilization review (UR) committee decisions. As mandated under Medicare and most Blue Cross plans, the UR committees have the responsibility of reviewing individual cases on the basis of "medical need." But the criteria of "need" are usually defined so that reimbursement will be denied both for services clearly unnecessary from a medical point of view (for example, tonsillectomies in the absence of any current or past history of infection or inflammation) *and* for services that are not covered under the insurance contract (for example, hospitalization for certain diagnostic tests which could be performed on an outpatient basis). As demonstrated by these two examples, economic as well as medical concerns are involved in the same UR mechanism.

If the dichotomy of medical and economic goals represents an inadequate schema for classifying utilization controls, at least the controls may be distinguished according to the means employed in reaching these goals. Three considerations are useful in this regard. First, are the controls applied before, during, or after the actual treatment? Second, do the controls incorporate decision rules which require a judgment as to the appropriateness of treatment? And third, who applies the controls and to whom are they applied?

ceived under public programs and a variety of ex post facto review mechanisms (utilization review, peer review, professional society review, medical audit, etc.). In general, these judgment-related controls are among the most controversial found in the medical system. For example, it is possible that mechanisms like surgery authorizations or recertification of care requirements may work to the detriment of the patient if the physician is careless in his appraisal of need. On a different plane, because ex post facto analysis of care represents a potential threat to health professionals whose competence might be called into question, one can postulate a situation in which peer review would function primarily to cover up all but the most egregious cases of "inappropriate" care. This is not to suggest that either of these situations will necessarily arise, but rather to indicate the potential importance of a third characteristic of utilization control.

Any effort to limit the usage of medical services, be it of a judgmental or nonjudgmental nature, may be relatively effective or ineffective, according to who is in a position to control the regulatory mechanism. Conceptually, at least, utilization control may be exercised either by users, providers, third-party payers, or the courts. Needless to say, certain of these groups are in a better position to regulate or resist regulation than others. To take just one example, physicians can regulate the supply of services by raising fees, refusing to make house calls, and reducing accessibility during nonoffice hours. Users, on the other hand, have little control over physicians since their major weapon—the power to take their business elsewhere—is reduced by unposted fee schedules, closed practices, and the urgency of need. Such structural inequalities in the market for medical services are pervasive, but as yet there is little direct evidence as to what effect they actually have on the functioning of utilization controls.

The characteristics of utilization control are summarized in the following chart under the five general categories of supply limitations, financial disincentives, authorization requirements, review mechanisms, and legal action. The remaining sections of the paper describe briefly the historical development of these methods of limiting usage and then concentrate attention on the specific types of control employed in actual practice. Finally, alternatives to current practice are examined in the light of probable developments in national health policy over the next decade.

The Origin and Rationale of Utilization Control

Aside from the constraints imposed by the threat of malpractice litigation, formalized utilization controls may be traced back to 1918, when the American College of Surgeons (ACS) first initiated its hospital approval program. As the forerunner of hospital accreditation, the ACS program required that hospitals meet minimum standards for the provision of quality care to each patient. To receive approval, the hospital had to show that review and analysis of certain services (limited initially to surgery and obstetrics) were being conducted on a regular basis. While it is impossible to separate the impact of these review requirements from other aspects of the ACS program, one may infer their significance from the fact that review mechanisms were eventually required in all hospital service departments for ACS approval (and were later required for approval by the Joint Commission on Hospital Accreditation, which assumed primary responsibility for hospital accreditation in 1952).

Interest in the review of hospital services led directly, but gradually, to the development of professional service accounting. As first practiced, this accounting method involved little more than the use of simple statistical summaries to record the number of patients treated, the length of stay, death rates, autopsy rates, etc. In the early 1950s, the Southwestern Michigan Hospital Council expanded the concept to include the use of data from patient medical records (Slee, 1968). By 1955, methods for handling data had advanced to the degree that several organizations (the American College of Physicians, the ACS, and the American Hospital Association) joined with the Council to form the nonprofit Commission on Professional and Hospital Activities. This commission organized the Professional Activity Study (PAS) and made the data it collected available to all member hospitals.

Interest in utilization review was stimulated by independent studies throughout the 1950s and 1960s which showed an alarming overutilization of hospital beds. One such study, conducted in Michigan, indicated that in 1962, 6.8 percent of the patient-days in the state's hospitals were unnecessary (Fitzpatrick et al., 1962). An analysis undertaken in Massachusetts determined that between four and eight percent of all hospital admissions involved cases that could have been satisfactorily treated elsewhere (Anderson and Sheatsley, 1967). A study in Indiana concluded that 20 percent of

low economic priority, the "success" of utilization control remains a hotly debated issue.

Control Through Supply Limitations

Perhaps the surest way to guarantee that a particular medical service will not be overutilized is to limit the physical availability of the service. As mentioned previously, almost any advance in medical technology is subject to restricted access because of limitations in supply. But limitations also affect utilization of traditional forms of medical treatment. These limitations run the gamut from policies which reduce the provision of marginal or clearly unnecessary services to actions which cut off the supply of care altogether.

At the one extreme, it might be noted that, in 1969, 134 counties in the United States had no practicing physicians (American Medical Association, 1970). While this situation obviously does not reflect any conscious attempt by American doctors to control "excess" utilization, it does indicate that a geographic maldistribution of physicians may have the effect of severely limiting the availability of services, particularly in rural areas. On a broader plane is the question of whether or not this country suffers from a general shortage of physicians. There are many views regarding the existence of a shortage (Fein, 1967; Hansen, 1970; Holtmann, 1965; Lynch, 1972; Ginzberg, 1966; Harrison and Nash, 1972), but there is a general consensus that, even if one does exist, it can be traced to professional concern over economic security rather than to any attempt to limit utilization per se (Burrow, 1966; Rayak, 1967; Kessel, 1970).

On the other hand, there are historical examples in which supply restrictions have been explicitly designed to reduce utilization. One such case was the decision by the British National Health Service (NHS) to authorize virtually no new hospital construction in the years following World War II, despite a rapidly increasing demand. The important point, however, is that if restrictions on supply do create shortages, the result will be the same whether the decision to limit usage is conscious or not; that is to say, either the price of the service will rise enough to clear the market or an increase in waiting times and difficulty in making appointments will serve to ration the scarce resource. In the case of the British NHS,

conditions of excess demand, there is some evidence that they may be useful in cases of excess capacity. Under "normal" business conditions, of course, the existence of excess capacity leads to a decline in profit rates which in turn is a signal to investors to find other outlets for their funds; new capital becomes scarce, and the industry goes through a period of disinvestment (i.e., depreciation exceeds net replacement of capital). Assuming that demand and supply are functionally independent, market forces will eventually eliminate excess capacity. But if demand and supply are *not* independently determined, such an adjustment process will either be hampered or cancelled out altogether. A situation akin to this arises whenever "supply creates its own demand." Although, technically speaking, excess capacity cannot exist if demand is totally dependent upon supply, the adequacy of capacity may be viewed in terms of some criterion other than demand (need, for example). It is on this basis that a number of studies have purported to show that excess capacity is a problem in the medical sector, particularly for hospitals ["a built bed is an occupied bed" (Roemer and Shain, 1959a; U.S. Department of Health, Education, and Welfare, 1968)] and for certain medical specialties such as general surgery (Fuchs, 1969; Bunker, 1970; Maloney, 1970; Owens, 1970; Hughes et al., 1972). It is worth noting, without attempting to support such claims, that some twenty states have considered hospital overbedding to be a serious enough problem to enact certificate-of-need laws which require prior approval by the state before new construction or facility additions may be initiated (Ingbar, 1972). The 1972 amendments to the Social Security Act extended this power to the Department of Health, Education, and Welfare (HEW) by authorizing the Secretary to deny Medicare and Medicaid payments for capital expenditures which are not in accord with area-wide planning board decisions regarding construction and expansion. Much less has been done in the field of manpower, although it has been suggested that licensing statutes be redefined on the basis of medical specialty. If, as has also been suggested, specialty licensing were combined with the power to establish regional quotas, then licensing boards would be able to prevent oversupply in professional specialty areas. There is precedent for such action. In Britain, for example, certain overdoctored areas are closed to new practices (Great Britain, Department of Health and Social Security, 1969).

Aside from the legal ramifications surrounding these forms of

duce poor peoples' demand for medically "necessary" services by as much or more than it will reduce the utilization of less appropriate services by the rich.[1] Under "normal" conditions (that is, when demand is inversely related to price), a rise in price must lead to a decline in overutilization to whatever extent it exists, but if this is accomplished at the expense of extreme vertical inequity then one must question the efficacy of the approach.

The price mechanism, however, is a highly flexible tool. Because each medical procedure may be individually priced, it is possible to use selective increases in areas where overutilization is a known or potential problem. While selective pricing does not eliminate the distributional problems associated with a general shift in the price level, at least it offers one way of reducing the economic burden on low- and middle-income groups without necessarily reducing the effectiveness of the price constraint facing the well-to-do.[2] The importance of this fact has not been lost on the health insurance industry, where selective pricing techniques (which take the form of benefit limitations, restrictions, exclusions, deductibles, and coinsurance) represent the most common form of utilization control. It would be naive, of course, to assume that profit-making firms are guided by the principle of minimizing the social costs of inappropriate usage (although this may represent an appropriate goal for nonprofit and governmental insurers). In fact, such techniques allow the carriers to offer considerably lower premiums than would otherwise be the case, and to increase the marketability of their policies in the process. However, differences in motivation need not compromise the value of the end result.

Deductibles

Of all the types of selective pricing, deductibles are applied to the widest variety of medical services and are found under both private

[1] For this to occur, one need only assume (1) that an individual's aggregate demand for all forms of medical care reflects an increasing proportion of "essential" services as price rises, and (2) that the demand for medical care by the poor is more price-elastic than the demand for care by the rich.

[2] Whether a price constraint will affect the rich depends upon the price-elasticity of their demand for the service in question. Since it is likely that their demand for most services will be inelastic at low price levels, only a relatively "expensive" constraint will force the rich to forgo consumption. On the other hand, if their demand for nonessential or "unnecessary" care is elastic, then a price constraint may prove effective.

a spurious relationship since the former group was composed of lower-risk members who may have preferred deductible coverage and the lower premiums that go with it (Andersen and Riedel, 1967).

Coinsurance

Insurance policies which use deductibles also typically include elements of coinsurance. An example is the $18 paid by hospitalized Medicare patients for each day in excess of 60 (and up to 90) days spent in a hospital during the year. As opposed to a deductible which allows the price that patients pay for covered medical services to drop to zero after the deductible is met, coinsurance guarantees that a certain positive price is associated with service use. This in turn may influence the consumption or utilization of care at all levels of expenditure.

Various types of coinsurance are in use today, and they vary widely in the burden placed upon the patient. In some, the insured pays a fixed price per unit of service; in others, he pays a fixed percentage of all insured expenses; and in a third type, he pays all expenses over and above a fixed price per unit of service. Michael Crew (1969) has shown that under conditions of monopoly, coinsurance is necessary for a socially optimum output of health services. But even under relatively competitive conditions, the effect of coinsurance depends upon the manner of application. In an unpublished paper, Karen Davis (undated) has analyzed each of the three basic methods and concludes that only the third would result in an optimum output. The fact that under this scheme the patient is concerned with the upper limits of his total bill provides him with an incentive to reduce utilization (particularly high-cost treatment) to whatever extent he is able.

There have been a number of empirical investigations into the effects of coinsurance. A study of Blue Cross of Western Pennsylvania (Hardwick et al., 1971) concluded that a five-dollar-per-day copayment had no significant impact on hospital use as measured by average length of stay, average benefits per admission, average benefits per day, and admissions and patient days per 1,000 members. A Connecticut Blue Cross study (Heaney, 1969) found that subscribers with full coverage enter a hospital more frequently and stay longer than if they are forced to pay part of the expenses them-

plans, places no direct restriction on the use of medical care but requires that all services be channeled through a designated primary physician chosen by the patient. This process not only systematizes delivery, but also prevents duplication by other physicians who may not even know they are treating the same patient. A variant of this system, known as a "lock-in," is found in the Kentucky Medicaid program. Under a pilot project begun in January, 1971, 36 Medicaid recipients in 10 counties were restricted to the use of a single primary physician and one pharmacy (of the patient's choice) during each 30-day period. But, unlike the prepaid plans, the Kentucky project also set limits on utilization at four physician visits and four prescriptions per month (additional services were permitted, but only after prior authorization by the physician). Before the lock-in, patients averaged 19.4 prescriptions and five physician visits per month. During the first six months of the program, utilization dropped to 6.3 prescriptions and 4.0 visits, and, in the next six-month period (when 26 additional patients were added to the program), it dropped further to 2.03 prescriptions and 1.0 visits.[8] As a result, the dollar cost per Medicaid beneficiary per month fell from $91.56 to $46.51 and finally to $10.43 at the end of the first year. This latter figure was substantially below the statewide average Medicaid expenditure of $17.00 per enrollee for comparable services during 1971 (Kentucky State Department of Health, 1972).

Although the Kentucky program represents a limited experiment, it may be expanded to cover the whole state, in which case a more reliable estimate of its effectiveness can be made. The alleged abuses under Medicaid and Medicare provide a rationale for such limits, but there are problems with the approach. First, there is a question of equity in imposing limits on but a single group of patients. Second, limitations must be flexible enough to assure that services will always be available to those with a clearly demonstrable need. Finally, limits must not encourage use of a more expensive service than necessary, simply because it is fully covered by the policy while a less expensive service is not covered or is severely

[8] Although it is probable that the reduction in Medicaid claims for drugs and physician visits represented an actual reduction in use, it is possible that certain individuals in the group purchased additional services from their own funds. Such out-of-plan use has been noted under the Medicare program, but in this case the poverty of the individuals involved certainly precluded extensive non-Medicaid medical service purchases.

Control Through Authorization Requirements

Aside from the issue of form, financial disincentives will fail to curb the inappropriate use of medical services whenever individuals follow the dictates of physicians who themselves are unconcerned with overutilization or misutilization. While this argument should not be overdrawn, it is generally true that when a patient decides to visit a physician, he makes an implicit decision to do as the physician recommends. Under these circumstances, selective pricing techniques are relevant only for initial patient-physician encounters and cannot be relied upon to ensure the appropriateness of prescribed treatments.

There are, of course, a number of ways in which the appropriateness of treatment may be controlled. At the extreme is the threat of malpractice litigation should negligence be involved in the treatment procedure. A second and less severe method entails peer review to pinpoint questionable procedures. Both of these mechanisms rely upon negative incentives (penalties), but it is also possible to exercise control through positive inducements (for example, the "profits" that a prepaid group practice obtains by lowering the level of unnecessary surgery). A final method, and by far the most direct (as well as the most controversial) is control through authorization requirements.

Authorization requirements are of two basic types. The first and most common is certification, whereby payment for a medical procedure is made only if the attending physician testifies in writing that the procedure is medically necessary. Limited primarily to hospital and nursing-home admissions, certification requirements are often extended to encompass the entire inpatient stay. In such cases, the physician must recertify the continued necessity of care at periodic intervals throughout the treatment process. In any event, the general approach is characterized by the fact that the physician who provides the treatment is also responsible for attesting to its necessity. This arrangement clearly violates an intrinsic principle of regulation, and apart from acting as a potential deterrent to fraudulent practices, certification requirements have not been noted for their success in limiting overutilization (Fitzpatrick, 1966; Commerce Clearing House, 1972a: 9897–9912).

The second basic type of authorization requirement operates

in 1970, Medi-Cal required prior authorization from a state-employed physician or medical consultant for all nonemergency hospital admissions. In the beginning, only admissions were subject to prior authorization, but at present the expected length of stay is also set prior to admission and any extensions are subject to recertification. During its first few months, the program did reduce the incidence of hospitalizations. After this initial period the number of admissions rose, but at a far lower rate than the increase in the number of persons eligible for Medi-Cal benefits. Despite a 23 percent increase in eligibility, the Medi-Cal program paid for only 3.5 million patient days in 1970, as opposed to 3.6 million in 1969. Overall, the admission rate dropped from 17.4 per 100 eligible persons in 1969 to 16.1 in 1970 (Brian, 1971; California Health Data Corporation, 1971). Although criticized for its occasional harshness and its political motivation (Gordon, 1972), the California program has shown that prior authorization can achieve more effective control over hospital utilization. Such evidence should provide a hopeful sign to administrators of public medical programs elsewhere, even though the approach may have limited applicability to private insurers.

Insurance carriers do have alternatives in this regard. In Michigan, for example, Blue Cross has experimented with a modified preadmittance screening procedure for hospital admissions. The procedure (though never implemented on an ongoing basis) worked through the hospital, which called upon a special clerical division of the Michigan Hospital Service to ensure that the patient to be admitted had the required coverage. Such a review process was unique, but it also represented an extremely limited form of prior authorization because the use of services was not determined on the basis of medical necessity, nor were the most important alternatives to inpatient care (notably clinic and outpatient benefits) covered under most Blue Cross subscriber contracts at the time of the experiment.

A modified approach to prior authorization which does not rely upon any overt administrative structure is found in United Mine Workers health plan. The United Mine Workers (UMW), which provides comprehensive medical care to about one-half million coal miners and their dependents, has a dual incentive to insure high quality care and to reduce unnecessary utilization because it is both provider and beneficiary. In achieving these ends, the union

stays.[5] Therefore, any method of recertification which excludes the majority of hospital cases from review on the basis of length of stay compromises the effectiveness of this utilization control.

The experience of most Blue Cross plans indicates that recertification programs have little or no net effect on the average length of stay (Fitzpatrick, 1966). One study of the New Jersey Blue Cross Approval by Individual Diagnosis (AID) program, however, is of special interest. Recertification limits under AID were established by diagnosis, and the resultant lengths of stay were analyzed for 308 types of treatment representing 94 percent of the 313,000 Blue Cross–paid claims in all New Jersey hospitals during 1963 (Bailey and Riedel, 1968). The study showed that although length of stay decreased during the first two quarters of the experiment, the effects of the AID program largely disappeared after six months as physicians reverted to old and established patterns of care. This short-term trend (which has been observed elsewhere) may be credited to the "Hawthorne effect," meaning that the results were not produced by the test factor (recertification) but by the fact that the subjects knew they were being observed and altered their behavior accordingly. An alternative and equally plausible explanation is that at the beginning of the program physicians believed that recertification decisions would be actively reviewed. When it became clear that utilization review committees were either unable or unwilling to apply sanctions, behavior returned to "normal."

The issue of Committee review of physician recertification has recently been cited as a major administrative problem facing the Medicare program. In a 1971 report, the U.S. General Accounting Office (GAO) concluded that although committee review of Medicare patients tended to be more impartial than the attending physician's determination of need for continued care, many questionable cases slipped by because "neither utilization review committees nor the administrative staffs at hospitals and extended care facilities (ECF) had taken timely action" (Commerce Clearing House, 1972a: 9897). The GAO study recommended a tightening of re-

[5] It has been argued that recertification is appropriate only in the case of long stays and that other controls should be used to reduce overutilization in shorter stays. However, since the admitting physician usually determines the length of stay in any event, logic would suggest that the same control mechanism be used in both cases.

1. Internal study of medical records, i.e., analyzing all records or a sample of discharges or admissions.
2. Ongoing studies of selected diagnoses or therapeutic categories. Studies of the more common diseases or treatment procedures found in the facility are often used as a UR technique.
3. Use of the services of organizations with computer facilities to compile pattern statistics, design profiles, and provide comparative data. Professional Activity Study–Medical Audit Program (PAS-MAP) is an example.
4. Cooperation with a fiscal intermediary. Blue Cross and Blue Shield are able to accumulate data through claims processing, which can help UR committees.
5. Any combination of the above. The selection of the method will depend upon data available and the type of study being made. Staff time and the availability of clerical help are other considerations.

Whichever approach is employed, the normal procedure involves evaluation of both in-process and post-treatment cases. The typical procedure for in-process review is the medical audit in which a clerk in the admitting office flags patients' medical records for recertification review. Nurses or ward clerks then fill in the charts to be reviewed and, either directly or through the medical records librarian, call to the doctors' attention those which require recertification.

Post-treatment review typically begins when supervised personnel in the medical records department record and review the charts of all discharged patients. The medical records librarian brings to the UR committee any questionable charts, a sample of patient records, or any cases which are significantly different from predefined norms such as those provided by PAS-MAP. The UR committee then proceeds to evaluate those cases, problems, and discharged patients' records brought to its attention.

The traditional approach to utilization review is for the UR committees to examine individual cases, sometimes sampled randomly, without using any uniform methods of evaluation and often without having any predetermined criteria at hand. Under such conditions, UR represents a hit-or-miss affair. A far superior method is the patterns-of-care approach which involves a review of selected categories of disease or operations. Under the patterns system, the UR committee first examines hospital practices in the aggregate, then looks at the clinical department, the diagnosis or oper-

There is some empirical evidence to this effect. Bonner, Decker, and Kasten (1972), in their study of 46 medical facilities, found that UR was much more successful in those institutions with high occupancy, low rate of turnover, and high initial length of stay. It proved less successful where facilities were being underutilized. Roemer and Shain (1959a; 1959b) examined the same phenomenon and concluded that 70 percent of the differences in hospital utilization rates by state and county could be explained by bed supply, and that, beyond a certain point, bed use becomes less and less reflective of illness levels.

In addition to the question of institutional motivation, the effectiveness of UR depends both on the quality of medical records and the ability of review committees to agree upon what constitutes appropriate usage. It is generally conceded that deficiencies in medical record keeping are pervasive enough to seriously hamper UR activities (Tufo and Spiedel, 1971; Peterson et al., 1956). But even when adequate records are kept, the concept of appropriateness of care can change from context to context and is usually defined, as Donabedian (1966: 167) points out as "almost anything anyone wishes it to be." At the outbreak of World War II, for example, when England was anticipating heavy war casualties, British hospitals were told to accelerate discharges, retaining only those patients for whom "institutional treatment is essential" (Titmuss, 1950: 153). As a result of this directive, between 40 and 50 percent of all patients were discharged immediately.

An empirical study done in Michigan (Riesser, 1969) points up another aspect of the same problem: Two pairs of doctors were asked to review 100 cases of hyperbilirubinemia, a blood disease related to Rh blood factor. Though care was deemed inappropriate in 15 cases, not one case was singled out by more than one physician.

Overcoming this inability to agree upon common criteria of evaluation is critical to the success of utilization review activities. Basically, the problem hinges upon two difficulties common to many if not most review efforts. First, the concept of "appropriateness" is ultimately based upon the individual physician's own value system, so that a great deal of variation exists among reviewers. This problem is exacerbated by the fact that the rules of evaluation are often imprecise. Reviewers may be simply told (as were those who conducted a special study for the Teamsters) to "use as a

McClain, 1972) have also shown that an examination of intra-reviewer unreliability can be useful in locating the source of most errors.

But even if the problem of defining common evaluative criteria can be overcome, there are other practical shortcomings evident in the operation of review committees. These shortcomings are clearly seen in an analysis of utilization review in 35 short-term Connecticut general hospitals three years after UR requirements were mandated under the Medicare program (Berman, 1969). The study found a high degree of heterogeneity among UR techniques in the hospitals and suggested that a majority were intent upon justifying patient admissions and extended lengths of stay. Other problems noted were a lack of communication between members of the utilization review committees, a tendency to concentrate on one segment of the patient population (such as Medicare patients) to the exclusion of other segments, inefficient use of physicians' time due to an inadequate use of paraprofessional and other medical personnel, and an ineffective use of computerized statistics due to a general lack of understanding of these statistics. The study concluded that few of the hospitals were receiving benefits commensurate with the cost of physician energy expended and that most failed to use UR committee findings in a way designed to change hospital policy or to improve the efficiency and quality of patient care.

Two other studies analyzing the effectiveness of utilization review in Pennsylvania hospitals offer a somewhat more optimistic view. The first (Marcom, 1965: 11), based upon an analysis of hospitals in Pittsburgh over a 10-year period ending in 1963, reached the following conclusion:

> If one accepts the premise that a high occupancy rate and lower average length of stay constitute more effective utilization of hospital facilities, there can be little doubt that intense utilization committee activity can contribute to more effective utilization of hospitals. There is no evidence whatever to indicate that utilization committees have succeeded in curbing the quantity of utilization, as measured in admissions per thousand persons and patient days per thousand persons.

The second study, conducted in 1969, analyzed the operation of 23 hospitals in central Pennsylvania and concluded that most UR committees have been successful in reducing the overall length of pa-

American Medical Association's *Peer Review Manual,* for example, argues that "quality control is the prime objective of Peer Review and cannot be allowed to become secondary to cost control" (American Medical Association, 1972, ch. 3: 16). In contrast to this rather narrow interpretation, the AMA Council of Medical Service has stated that peer review represents "medicine's efforts to assure high quality of health services at reasonable cost, slowing the rate of escalation in health care charges, stimulating health insurance organizations to make broader protection available to more people, and regaining professional control in patient-physician fiscal and economic relationships" (Delaware Medical Journal, 1970: 248).

Some form of peer review is practiced by virtually all health-related professional societies including dentistry, pharmacy, and physical therapy. The professional society undertakes review either when the professional's place of business does not perform such a function (i.e., for work performed outside the hospital) or as a "court of appeals" for decisions reached by an institutional UR committee. Peer review is typically initiated with the referral of a complaint from a patient, physician, insurance carrier, or government agency to a local committee which acts as a fact-finding board. In most cases this board has no disciplinary power but can make recommendations for action. If no decision is made or if one of the parties objects to the decision, the case may be referred or appealed to a regional or even a state peer review committee.

In recent years, most physician organizations have come to accept and support the concept of peer review. To an undetermined but probably significant extent, this acceptance has been motivated by a desire to preclude government regulation or supervision. The AMA Committee on Health Insurance Legislation, for example, concluded that "if medicine does not provide for a mandatory review procedure . . . this responsibility will pass to government by default" (AMA, 1972, app. A: 12–13). In Mississippi (AMA, 1972, app. B: 2), the state medical association was even more direct:

> The private practice of medicine is under siege. Sadly, it is a focal point of public wrath, and consequently, politically an inviting target. Many advocate the wresting away of control from the physician. Peer Review is a positive program designed to establish and maintain control of medical practice in the hands of physicians.

Commission which recommended that peer review be performed at the local level with professional societies acting as sponsors and supervisors (Commerce Clearing House, 1972b). The law as enacted, however, allows only physicians to form or hold memberships in PSROs even though the services of nonphysicians are subject to review. This creates the potential for a serious conflict of interest both within the health sector and between the medical profession and the public. While it is true that accountability for PSRO activities will be strengthened by a provision in the law for the development of sophisticated systems to detect inappropriate utilization, it is not at all clear that this benefit will outweigh the dangers inherent in providing to any group the statutory authority to regulate itself. Peer review is by most accounts a rather conservative UR device in any event, and there seems little reason to predict that the activities of PSROs will necessitate a change in this judgment. The provisions in the law for self-regulation, part-time and rotating memberships, and the exclusion of nonphysicians all suggest that PSROs may well give the appearance without the substance of control.

Claims Review

A final type of ex post facto control over utilization is claims review. All third-party payers (including Blue Cross–Blue Shield, commercial carriers, and the intermediaries for Medicare and Medicaid) employ some kind of claims review which typically involves the use of both prepayment edits and postpayment audits. Factors considered when claims are reviewed include completeness of information, internal consistency, lack of obvious error or misrepresentation, the extent of recipient coverage, the reasonableness of charges, and such utilization characteristics as appropriateness of admission, length of stay, and efficiency of scheduling procedures in the facility. In addition, some insurers use manual or computerized prepayment screens to scan providers and subscribers for unusual patterns of care, such as the physician who gives the same type of x-ray or pathology tests to a large number of patients, or the patient who receives an unusual number of physician visits while in a nursing home.

Although claims review clearly involves many functions besides utilization review, it is difficult to isolate such functions as the detection of fraud, improper billing, and the identification of unnecessary utilization. For example, a claim may be rejected by the insur-

In Michigan, for example, Blue Shield review of Medicaid expenditures in 1971 led to the recovery of almost $23 million or 7.4 percent of total payments (Michigan Blue Shield, 1972). Review of utilization, however, represented a relatively small component of this recovery process, amounting to only 3.3 percent of all monies recovered or 0.2 percent of all monies paid out. Moreover, even this figure is high, since utilization review as defined by Blue Shield includes recoveries from both unnecessary treatment and fraudulent claims. During the same year, Michigan Blue Cross surveillance of Medicaid expenditures led to the recovery of $1.3 million (of a total payout to hospitals of $112 million) from claims rejected on the basis of program exclusions and limitations (Michigan Blue Cross, 1973). According to Blue Cross, "Many questionable hospital cases are referred to hospital utilization review committees for their review and recommendations as to the level or type of care provided" (Michigan Blue Cross, 1973: 3). But again, something less than the full savings of $1.3 million can be attributed to claims for "unnecessary" care.

Part of the problem arises from the fact that few third-party payers have integrated systems designed both to provide full-scale surveillance of patient utilization and provider performance, and to provide the policing necessary when irregularities are discovered. A step in this direction was taken several years ago when New York City developed its "Watchdog" program over Medicaid expenditures. The "Watchdog" system, based on a rather complex combination of standard setting, committee and professional review, complaint procedures, spot checks, and other methods of surveillance has been credited with saving the city millions of dollars annually (U.S. Senate Committee on Finance, 1970: 249-252). In recent years, interest in supervision and surveillance over public expenditures has moved toward the development of sophisticated computer programs of claims review designed to far surpass the capabilities of such predecessor systems as the PAS-MAP approach to data generation and information retrieval. The 1972 amendments to the Social Security Act call for the creation of regional and national data-processing centers which will provide this type of computer analysis to federal and state agencies. At the state level, several Medicaid agencies have shown an independent interest in systems development. Illinois, for example, has just recently developed a program known as the Hospital Admission and Surveillance Program

tic tests. Defensive medicine, according to Hershey (1972: 72), consists of two undesirable patterns of behavior:

> First, when a test or procedure is performed because the physician fears that if he does not perform it, and the patient has a bad result, some medical expert might testify that it was unnecessary, and . . . second when a test or procedure is not performed because the physician believes that the risk of legal difficulty from a complication arising from the procedure is substantial, although the physician's view is that the patient would be better off if it were performed.

The extent to which defensive medicine is practiced in America is debated by knowledgeable observers, but almost all agree that such a phenomenon does exist. A study of the attitudes and opinions of 500 physicians in the late 1950s is indicative (Brenner, 1960). The physicians sampled were asked whether fear of malpractice suits led to any change in their own practices. The results were dramatic: 54 percent said they kept more detailed office records; 47 percent said they ordered more x-rays; 43 percent used more consultations; 37 percent ordered more diagnostic tests; 40 percent gave less telephone advice; and 36 percent permitted fewer prescription refills. Other effects cited by less than 30 percent of those interviewed related to caution with new procedures (28 percent), more frequent hospitalization (20 percent), and screening of new patients for legal reasons (25 percent).

A more recent survey of physician attitudes conducted by the staff of the *Duke Law Journal* (1971) suggests that the threat of malpractice has a relatively small effect on *positive* defense practices (i.e., ordering excessive laboratory tests), but has a potentially greater impact in terms of *negative* defensive practices (such as the failure to apply new technological advances in diagnosis and treatment). However, in the case of both these and other studies (U. S. Department of Health, Education, and Welfare, 1973; Rice, 1971; U. S. Senate Subcommittee on Executive Reorganization, 1969) it is important to realize that a physician's stated behavior may not correspond with his actual practices. As Hershey notes, "[m]ost physicians point out specialities other than their own as examples of those most influenced by liability considerations. These same physicians also seem to imply that others within their own specialty are generally practicing with more concern about liability than them-

sions and financing mechanisms contained in each. What is of interest is how they propose to limit unnecessary utilization.

Taken as a group, these national health insurance proposals contain surprisingly little in the way of innovative approaches to utilization control. With the single exception of the Health Security plan, they all contain the deductibles, coinsurance, benefit limitations, and other features so common to private insurance polices (and, in fact, the Medicredit and National Healthcare bills are explicitly designed to subsidize private health insurance). Every one of the proposals requires utilization review procedures similar if not identical to those mandated under current Medicare and Medicaid regulations. Two of the bills (the AMA and Nixon proposals) specify payment mechanisms (such as full cost reimbursement for health facilities and payment of customary and prevailing fees for physicians) which offer no new incentives for providers to reduce unnecessary utilization. There are, however, some new twists. The Ullman bill, for example, would create a network of "health care corporations" or HCCs to coordinate community health resources. Depending upon local circumstances, these HCCs might be in a position to apply some of the more positive forms of supply limitation considered previously. Several of the bills (including the Burleson-McIntire proposal and the Kennedy-Griffiths plan) make provision for experimentation with incentive reimbursement schemes. When this is combined with support for alternative delivery mechanisms (a primary element in the Kennedy-Griffiths plan), the result might eventually serve to reverse the incentive to overprovide certain types of medical services. Of all the bills, the Health Security proposal offers the most innovative package of utilization controls, including prospective budgeting for hospitals, preferential treatment for physicians operating in prepaid groups, the establishment of a quasi-independent Commission on the Quality of Health Care to establish standards, and a Federal Health Security Board with responsibility for determining spending priorities.

Variable Subsidy Insurance

The utilization controls envisaged under the Health Security proposal are also likely to be extremely complex from an administrative standpoint. In part, this complexity reflects the price one must pay when consumers are not forced to regulate their own demand

Prepaid Health Care Centers

While most proposals for reducing overutilization are designed to operate within the context of the present fee-for-service delivery system, there are alternative models which suggest that a change in provider reimbursement, together with a change in the delivery mechanism, is the most promising approach to utilization control. The model most commonly considered in this regard is the prepaid health care center—or in more popular terms, the health maintenance organization (HMO). An HMO may be defined as a group of medical providers which offers a comprehensive package of health services at a centralized location to a defined population for a fixed monthly fee per enrolled individual or family. It is argued that an HMO can reduce overutilization and unnecessary treatments through an alteration in provider incentives: (1) because providers receive a fixed income regardless of the amount of services offered, there is no incentive to provide treatments not dictated by medical need; (2) because the organization is responsible for comprehensive care, there is likely to be more preventive care, more early diagnosis, and therefore greater savings in curative treatments than is the case in a fee-for-service system; and (3) because of the principle of shared risk, there is an incentive to prescribe less expensive methods of treatment when the choice arises.

Each of these assumptions regarding provider incentives has been questioned (Roth, 1972; Greenberg and Rodburg, 1971), and, particularly in the case of preventive treatments, there does appear to be reasonable doubt whether an HMO can be expected to produce better results than more traditional modes of delivery. But the incentive for preventive practices notwithstanding, there is ample evidence from a number of matched-population studies that the HMO can be quite effective in reducing hospital admissions and inpatient stays (Gaus et al., 1972; Donabedian, 1969; Saward, 1969; Shapiro, 1964). The savings obtained in this one area alone are enough to warrant giving serious consideration to the HMO model as a utilization control. Furthermore, the fact that the savings generated are the result of the delivery system itself rather than of some external rule or regulatory body increases its attractiveness from an administrative point of view. But the HMO has the disadvantage of being limited in practice to urban and suburban areas with relatively stable populations, and even in these places the start-up costs associated with new HMOs are substantial enough to

ductive capacity necessary to meet the medical needs of a given population. Because this represents a very difficult task even without consideration of technological advances in medical science, it is unlikely that present efforts to use supply limitations for utilization control (such as certificate-of-need requirements for hospital construction) will succeed except in instances where overcapacity is an obvious problem.

Second, financial disincentives to overuse may have an appropriate place in a system of utilization controls, but they must not be considered as surrogates for mechanisms designed to contain provider-induced overutilization. Also, it should be realized that the form of financial disincentive imposed upon the consumer is an important determinant in whether these control devices can be relied upon to serve their proper function. The present widespread use of flat-rate deductible and coinsurance provisions in public medical programs and private insurance contracts probably has little effect on the overutilization of the rich but may serve to curtail the demand for necessary care on behalf of the poor. While there is no administratively feasible way to eliminate such differential impacts, the process could be made both more efficient and more equitable by scaling the disincentives on the basis of income class through a system of variable subsidy insurance.

Third, efforts to control unnecessary utilization through administrative devices such as review boards or UR committees have demonstrated a rather checkered pattern of success and failure which can be expected to continue unless corrective steps are taken. Where the administrative approach has proven ineffective, failure may be attributed to one or more of four factors: the lack of adequate surveillance systems, the inability of UR committee members to agree upon a course of action, role confusion and conflict of interest when review members are drawn from the group to be regulated, and a general lack of enforcement powers. With the exception of the first factor (which requires substantial sophistication in computer technology) these problems are more political than technical in nature, and thus may be amenable to improvement through organizational change.

One organizational approach specifically applicable to public medical programs involves the creation of teams of professional career review agents assigned to each hospital service area in a state. These agents could be government-employed physicians or else em-

nificantly motivated to control utilization levels because the fee-for-service payment mechanism together with the spread of health insurance have created a situation where "someone else" assumes the burden of unnecessary medical treatment. Arguments in favor of changing the basic incentive structure are thus intuitively appealing to those concerned with the proper allocation of medical resources.

Perhaps the best, and certainly the most discussed alternative in this regard is the concept of the prepaid health care center. Such organizations have demonstrated that the utilization of hospital services may be significantly reduced through a holistic, coordinated approach to treatment; and, even more important, they have shown that a delivery system can operate successfully under the constraints imposed by a predetermined and fixed income base. A logical extension of this last point is regional and federal budgeting for health expenditures under national health insurance. Given the ever-increasing proportion of gross national product which goes for health services, this may prove to be the only long-range solution to the question of utilization control.

Bruce Stuart, PH.D.
Director, Health Research Division
Michigan Department of Social Services
Commerce Center Building
300 S. Capitol Avenue
Lansing, Michigan 48926

Ronald Stockton, PH.D.
Research Analyst, Health Research Division
Michigan Department of Social Services
Commerce Center Building
300 S. Capitol Avenue
Lansing, Michigan 48926

The views expressed in this paper are those of the authors and do not necessarily reflect the position of the Michigan Department of Social Services.

References

American Medical Association
1970 Distribution of Physicians, Hospitals and Hospital Beds in the U.S. Chicago: American Medical Association.
1972 Peer Review Manual. Chicago: American Medical Association.

Burnes, E. M.
1971 "A critical review of national health insurance proposals." HSMHA Health Reports 86 (February): 111–120.

Burrow, J. G.
1966 AMA: Voice of American Medicine. Baltimore: The Johns Hopkins Press.

California Health Data Corporation
1971 Examination of Average Stay Patterns: Hospitals in Sacramento County, January 1969–December 1970. San Francisco: California Health Data Corporation.

Commerce Clearing House
1972a Medicare and Medicaid Guide. New York: Commerce Clearing House.
1972b 1973 Social Security and Medicare Explained. Chicago: Commerce Clearing House.

Crew, M.
1969 "Coinsurance and the welfare economics of medical care." American Economics Review 59 (December): 906–908.

Davis, K.
no date "Efficiency and distributional effects of alternative coinsurance schemes: the monopoly case." Washington, D.C.: The Brookings Institution.

Delaware Medical Journal
1970 "Peer review, utilization, and fees." Delaware Medical Journal 42 (September): 248–249.

deRouville, W. H.
1971 "Peer review in bilary tract surgery." New York State Journal of Medicine 71 (June): 1544–1548.

Dickerson, O. D.
1968 Health Insurance. Homewood, Illinois: Richard D. Irwin.

Donabedian, A.
1966 "Evaluating the quality of medical care." Milbank Memorial Fund Quarterly 44 (July, Part 1): 166–203.
1969 "An evaluation of prepaid group practice." Inquiry 6 (September): 3–27.

Duke Law Journal
1971 "The medical malpractice threat: a study of defensive medicine." Duke Law Journal 1971 (December): 939–993.

Gordon, J.
1972 How Standard Setting, Peer Review, Prior Authorization, and Administrative Regulations Reduce the Quality of Care in a Government Sponsored Medical Program. Paper presented at annual convention of the American Public Health Association, Atlantic City, November 13.

Great Britain, Department of Health and Social Security
1969 Annual Report for the Year 1968, CMND No. 4100. London: Her Majesty's Stationery Office.

Greenburg, I. G., and M. L. Rodburg
1971 "The role of prepaid group practice in relieving the medical care crisis." Harvard Law Review 84 (February): 887–1001.

Grimes, Thomas F.
1970 The Effectiveness of Utilization Review Since the Advent of Medicare. Unpublished Master's thesis, George Washington University.

Hall, C. B.
1966 "Deductibles in health insurance: an evaluation." Journal of Risk and Insurance 33 (June): 253–263.

Hansen, W. L.
1970 "An appraisal of physician manpower projections." Inquiry 7 (March): 102–113.

Hardwick, C. P., S. B. Myers, S. Barnoon, and L. Shuman
1971 The Effect of a Co-payment Agreement on Hospital Utilization. Pittsburgh: Blue Cross of Western Pennsylvania, Series no. 9.

Harrison, J. L., and K. D. Nash
1972 "A new approach to the physician shortage." Inquiry 9 (March): 34–39.

Heaney, C. T.
1969 An Analysis of the Changing Use Patterns of Blue Cross Subscribers After a Change to a Comprehensive Benefit Package. New Haven, Connecticut: Yale University Press.

Hershey, N.
1972 "The defensive practice of medicine—myth or reality." The Milbank Memorial Fund Quarterly 50 (January, Part 1): 69–98.

Holtman, A. G.
1965 "Another look at the shortage of physicians." Industrial and Labor Relations Review 18 (April): 423–424.

Marcom, E. M.
1965 Medical Staff Utilization Committees: Development and Effectiveness. Unpublished Master's thesis, Duke University.

McClain, J. O.
1972 "Physician confidence and reliability in utilization review." Medical Care 10 (November): 463–469.

Michigan Blue Cross
1973 A Report on Medicaid, 1971. Detroit: Michigan Blue Cross.

Michigan Blue Shield
1972 Medicaid and Michigan Blue Shield. Detroit: Michigan Blue Shield.

Morehead, M. A.
1967 "The medical audit as an operational tool." American Journal of Public Health 57 (September): 1643–1656.

Owens, A.
1970 "General surgeons: too many in the wrong places." Medical Economics 47 (July): 128–133.

Pauly, M. V.
1971 Medical Care at Public Expense: A Study in Applied Welfare Economics. New York: Praeger.

Peters, E. N.
1972 "Practical versus impractical peer review" Medical Care 10 (November): 516–521.

Peterson, O. L., L. P. Andrews, R. S. Spain, and B. G. Greenberg
1956 An Analytical Study of North Carolina General Practice. Chicago: Association of American Medical Colleges.

Phelps, C. P., and T. P. Newhouse
1972 "The effect of coinsurance: a multivariate analysis." Social Security Bulletin 35 (June): 20–28.

Rayak, E.
1965 "The shortage of physician services." Industrial and Labor Relations Review 18 (July): 584–588.
1967 Professional Power and American Medicine. Cleveland: World Publishing Company.

Somers, H. M., and A. R. Somers
1967 Medicare and the Hospitals. Washington, D.C.: The Brookings Institution.

Spontz, L. G.
1972 "The economics of financing medical care: a review of the literature." Inquiry 9 (December): 3–21.

State of Maryland
1964 Report of the Commission to Study Hospital Costs. Baltimore, Maryland: State of Maryland.

Stuart, B.
1972 "National health insurance and the poor." American Journal of Public Health 62 (September): 1252–1259.

Stuart, B., and L. Bair
1971 Health Care and Income: The Distributional Impacts of Medicaid and Medicare Nationally and in the State of Michigan, 2nd edition. Lansing, Michigan: Michigan Department of Social Services.

Titmuss, R. M.
1950 Problems of Social Policy. London: Longmans, Green.

Tufo, H. M., and J. J. Speidel
1971 "Problems with medical records." Medical Care 9 (November): 509–517.

Turner, W.
1969 "How can the individual hospital assess its use and costs?" Rhode Island Medical Journal 49 (June): 348–352.

U. S. Department of Health, Education, and Welfare
1968 Report of the Secretary's Advisory Committee on Hospital Effectiveness, Washington, D.C.: Government Printing Office.
1969 Utilization Review: A Selected Bibliography 1933–1969. Washington, D.C.: Government Printing Office.
1973 Report of the Secretary's Commission on Medical Malpractice. Washington, D.C.: Government Printing Office.

U. S. Senate Committee on Finance
1970 Medicare and Medicaid: Problems, Issues, and Alternatives. Washington, D.C.: Government Printing Office.

U. S. Senate Subcommittee on Executive Reorganization
1969 Medical Malpractice: The Patient Versus the Physician. Washington, D.C.: Government Printing Office.