Evaluating the Quality of Medical Care

AVEDIS DONABEDIAN

THIS PAPER IS AN ATTEMPT TO DESCRIBE AND evaluate current methods for assessing the quality of medical care and to suggest some directions for further study. It is concerned with methods rather than findings, and with an evaluation of methodology in general, rather than a detailed critique of methods in specific studies.

This is not an exhaustive review of the pertinent literature. Certain key studies, of course, have been included. Other papers have been selected only as illustrative examples. Those omitted are not, for that reason, less worthy of note.

This paper deals almost exclusively with the evaluation of the medical care process at the level of physician-patient interaction. It excludes, therefore, processes primarily related to the effective delivery of medical care at the community level. Moreover, this paper is not concerned with the administrative aspects of quality control. Many of the studies reviewed here have arisen out of the urgent need to evaluate and control the quality of care in organized programs of medical care. Nevertheless, these studies will be discussed only in terms of their contribution to methods of assessment and not in terms of their broader social goals. The author has remained, by and large, in the familiar territory of care provided by physicians and has avoided incursions into other types of

The Milbank Quarterly, Vol. 83, No. 4, 2005 (pp. 691–729)
© 2005 Milbank Memorial Fund. Published by Blackwell Publishing.

Reprinted from The Milbank Memorial Fund Quarterly, Vol. 44, No. 3, Pt. 2, 1966 (pp. 166–203). Style and usage are unchanged.

health care. Also, consideration of the difficult problem of economic efficiency as a measurable dimension of quality has been excluded.

Three general discussions of the evaluation of quality have been very helpful in preparing this review. The first is a classic paper by Sheps which includes an excellent discussion of methods.[1] A more recent paper by Peterson provides a valuable appraisal of the field.[2] The paper by Lerner and Riedel discusses one recent study of quality and raises several questions of general importance.[3]

Definition of Quality

The assessment of quality must rest on a conceptual and operationalized definition of what the "quality of medical care" means. Many problems are present at this fundamental level, for the quality of care is a remarkably difficult notion to define. Perhaps the best-known definition is that offered by Lee and Jones[4] in the form of eight "articles of faith," some stated as attributes or properties of the process of care and others as goals or objectives of that process. These "articles" convey vividly the impression that the criteria of quality are nothing more than value judgments that are applied to several aspects, properties, ingredients or dimensions of a process called medical care. As such, the definition of quality may be almost anything anyone wishes it to be, although it is, ordinarily, a reflection of values and goals current in the medical care system and in the larger society of which it is a part.

Few empirical studies delve into what the relevant dimensions and values are at any given time in a given setting. Klein et al.,[5] found that 24 "administrative officials," among them, gave 80 criteria for evaluating "patient care." They conclude that patient care, like morale, cannot be considered as a unitary concept and ". . . it seems likely that there will never be a single comprehensive criterion by which to measure the quality of patient care."

Which of a multitude of possible dimensions and criteria are selected to define quality will, of course, have profound influence on the approaches and methods one employs in the assessment of medical care.

Approaches to Assessment: What to Assess

The outcome of medical care, in terms of recovery, restoration of function and of survival, has been frequently used as an indicator of the

quality of medical care. Examples are studies of perinatal mortality,[6,7] surgical fatality rates[8] and social restoration of patients discharged from psychiatric hospitals.[9]

Many advantages are gained by using outcome as the criterion of quality in medical care. The validity of outcome as a dimension of quality is seldom questioned. Nor does any doubt exist as to the stability and validity of the values of recovery, restoration and survival in most situations and in most cultures, though perhaps not in all. Moreover, outcomes tend to be fairly concrete and, as such, seemingly amenable to more precise measurement.

However, a number of considerations limit the use of outcomes as measures of the quality of care. The first of these is whether the outcome of care is, in fact, the relevant measure. This is because outcomes reflect both the power of medical science to achieve certain results under any given set of conditions, and the degree to which "scientific medicine," as currently conceived, has been applied in the instances under study. But the object may be precisely to separate these two effects. Sometimes a particular outcome may be irrelevant, as when survival is chosen as a criterion of success in a situation which is not fatal but is likely to produce suboptimal health or crippling conditions.[10]

Even in situations where outcomes are relevant, and the relevant outcome has been chosen as a criterion, limitations must be reckoned with. Many factors other than medical care may influence outcome, and precautions must be taken to hold all significant factors other than medical care constant if valid conclusions are to be drawn. In some cases long periods of time, perhaps decades, must elapse before relevant outcomes are manifest. In such cases the results are not available when they are needed for appraisal and the problems of maintaining comparability are greatly magnified. Also, medical technology is not fully effective and the measure of success that can be expected in a particular situation is often not precisely known. For this reason comparative studies of outcome, under controlled situations, must be used.

Although some outcomes are generally unmistakable and easy to measure (death, for example) other outcomes, not so clearly defined, can be difficult to measure. These include patient attitudes and satisfactions, social restoration and physicial disability and rehabilitation.[11] Even the face validity that outcomes generally have as criteria of success or failure, is not absolute. One may debate, for example, whether the prolongation of life under certain circumstances is evidence of good medical

care. McDermott et al., have shown that, although fixing a congenitally dislocated hip joint in a given position is considered good medicine for the white man, it can prove crippling for the Navajo Indian who spends much time seated on the floor or in the saddle.[12] Finally, although outcomes might indicate good or bad care in the aggregate, they do not give an insight into the nature and location of the deficiencies or strengths to which the outcome might be attributed.

All these limitations to the use of outcomes as criteria of medical care are presented not to demonstrate that outcomes are inappropriate indicators of quality but to emphasize that they must be used with discrimination. Outcomes, by and large, remain the ultimate validators of the effectiveness and quality of medical care.

Another approach to assessment is to examine the process of care itself rather than its outcomes. This is justified by the assumption that one is interested not in the power of medical technology to achieve results, but in whether what is now known to be "good" medical care has been applied. Judgments are based on considerations such as the appropriateness, completeness and redundancy of information obtained through clinical history, physical examination and diagnostic tests; justification of diagnosis and therapy; technical competence in the performance of diagnostic and therapeutic procedures, including surgery; evidence of preventive management in health and illness; coordination and continuity of care; acceptability of care to the recipient and so on. This approach requires that a great deal of attention be given to specifying the relevant dimensions, values and standards to be used in assessment. The estimates of quality that one obtains are less stable and less final than those that derive from the measurement of outcomes. They may, however, be more relevant to the question at hand: whether medicine is properly practiced.

This discussion of process and outcome may seem to imply a simple separation between means and ends. Perhaps more correctly, one may think of an unbroken chain of antecedent means followed by intermediate ends which are themselves the means to still further ends.[13] Health itself may be a means to a further objective. Several authors have pointed out that this formulation provides a useful approach to evaluation.[14,15] It may be designated as the measurement of procedural end points and included under the general heading of "process" because it rests on similar considerations with respect to values, standards and validation.

A third approach to assessment is to study not the process of care itself, but the settings in which it takes place and the instrumentalities of which

it is the product. This may be roughly designated as the assessment of structure, although it may include administrative and related processes that support and direct the provision of care. It is concerned with such things as the adequacy of facilities and equipment; the qualifications of medical staff and their organization; the administrative structure and operations of programs and institutions providing care; fiscal organization and the like.[16,17] The assumption is made that given the proper settings and instrumentalities, good medical care will follow. This approach offers the advantage of dealing, at least in part, with fairly concrete and accessible information. It has the major limitation that the relationship between structure and process or structure and outcome, is often not well established.

Sources and Methods of Obtaining Information

The approach adopted for the appraisal of quality determines, in large measure, the methods used for collecting the requisite information. Since these range the gamut of social science methods, no attempt will be made to describe them all. Four, however, deserve special attention.

Clinical records are the source documents for most studies of the medical care process. In using them one must be aware of their several limitations. Since the private office practice of most physicians is not readily accessible to the researcher, and the records of such practice are generally disappointingly sketchy, the use of records has been restricted to the assessment of care in hospitals, outpatient departments of hospitals and prepaid group practice. Both Peterson[18] and Clute[19] have reported the prevailing inadequacies of recording in general practice. In addition, Clute has pointed out that, in general practice, ". . . the lack of adequate records is not incompatible with practice of a good, or even an excellent quality . . . ." On the other hand, a recent study of the office practice of a sample of members of the New York Society of Internal Medicine[20] suggests that abstracts of office records can be used to obtain reproducible judgments concerning the quality of care. But to generalize from this finding is difficult. It concerns a particular group of physicians more likely to keep good records than the average. Moreover, for one reason or another, the original sample drawn for this study suffered a 61 per cent attrition rate.

Assuming the record to be available and reasonably adequate, two further issues to be settled are the veracity and the completeness of the record. Lembcke[10] has questioned whether key statements in the record

can be accepted at face value. He has questioned not only the statements of the physician about the patient and his management, but also the validity of the reports of diagnostic services. The first is verified by seeking in the record, including the nurses' notes, what appears to be the most valid evidence of the true state of affairs. The second is verified by having competent judges re-examine the evidence (films, tracings, slides) upon which diagnostic reports are made. Observer error tends to be a problem under the best of circumstances.[21] But nothing can remove the incredulity from the finding by Lembcke, in one hospital, that the true incidence of uterine hyperplasia was between five and eight per cent rather than 60 to 65 per cent of uterine curettages, as reported by the hospital pathologist. In any case, the implications of verification as part of the assessment of quality must be carefully considered. Errors in diagnostic reports no doubt reflect particularly on the quality of diagnostic service and on the care provided by the hospital, in general. But the physician may be judged to perform well irrespective of whether the data he works with are or are not valid. This is so when the object of interest is the logic that governs the physician's activities rather than the absolute validity of these activities.

Much discussion has centered on the question of the completeness of clinical records and whether, in assessing the quality of care based on what appears in the record, one is rating the record or the care provided. What confuses the issue is that recording is itself a separate and legitimate dimension of the quality of practice, as well as the medium of information for the evaluation of most other dimensions. These two aspects can be separated when an alternative source of information about the process of care is available, such as the direct observation of practice.[18,19] In most instances, however, they are confounded. Rosenfeld[22] handled the problem of separating recording from care by examining the reasons for downrating the quality of care in each patient record examined. He demonstrated that the quality of care was rated down partly because of what could have been poor recording ("presumptive" evidence) and partly for reasons that could not have been a matter of recording ("substantial" evidence). He also found that hospitals tended to rank high or low on both types of errors, showing that these errors were correlated. Since routine recording is more likely to be complete in the wards, comparison of ward and private services in each hospital by type of reason for downrating might have provided further information on this important question. Other investigators have tried to allow for incompleteness

in the record by supplementing it with interviews with the attending physician and making appropriate amendments.[23–25] Unfortunately, only one of these studies (length of stay in Michigan hospitals) contains a report of what difference this additional step made. In this study "the additional medical information elicited by means of personal interviews with attending physicians was of sufficient importance in 12.6 per cent of the total number of cases studied to warrant a reclassification of the evaluation of the necessity for admission and/or the appropriateness of length of stay."[3,25] When information obtained by interview is used to amend or supplement the patient record, the assumption may have to be made that this additional information has equal or superior validity. Morehead, who has had extensive experience with this method, said, "Many of the surveyors engaged in the present study employed the technique of physician interview in earlier studies without fruitful results . . . . The surveyor was . . . left in the uncomfortable position of having to choose between taking at face value statements that medical care was indeed optimal, or concluding that statements presented were untrue."[26] Even in an earlier study, where supplementation by interview is reported to have been used,[24] verbal information was discarded unless it was further corroborated by the course of action or by concrete evidence.[27]

Another question of method is whether the entire record or abstracted digests of it should be used as a basis for evaluation. The question arises because summaries and abstracts can presumably be prepared by less skilled persons allowing the hard-to-get expert to concentrate on the actual task of evaluation. Abstracting, however, seemingly involves the exercise of judgment as to relevance and importance. For that reason, it has been used as a first step in the evaluation of quality only in those studies that use very specific and detailed standards.[10] Even then, little information is available about how reliable the process of abstracting is, or how valid when compared with a more expert reading of the chart. The study of New York internists, already referred to, demonstrated a high level of agreement between physicians and highly trained non-physicians abstracting the same office record.[20]

While the controversy about the record as a source of information continues, some have attempted to reduce dependence on the physician's recording habits by choosing for evaluation diagnostic categories which are likely to be supported by recorded evidence additional to the physician's own entries.[28] This explains, in part, the frequent use of surgical operations as material for studies of quality.

In general practice, patient records are too inadequate to serve as a basis for evaluation. The alternative is *direct observation* of the physician's activities by a well qualified colleague.[18,19] The major limitation of this method would seem to be the changes likely to occur in the usual practice of the physician who knows he is being observed. This has been countered by assurances that the physician is often unaware of the true purpose of the study, becomes rapidly accustomed to the presence of the observer, and is unable to change confirmed habits of practice. Even if changes do occur, they would tend to result in an overestimate of quality rather than the reverse. These assurances notwithstanding, measuring the effect of observation on practice remains an unsolved problem.

Those who have used the method of direct observation have been aware that the problem of completeness is not obviated. The practicing physician often knows a great deal about the patient from previous contacts with him—hence the need to select for observation "new" cases and situations that require a thorough examination irrespective of the patient's previous experience. Moreover, not all of the managing physician's activities are explicit. Some dimensions of care, not subject to direct observation, must be excluded from the scheme of assessment. Selective perception by the observer may be an additional problem. The observer is not likely to be first a neutral recorder of events and then a judge of these same events. His knowledge and criteria are likely to influence what he perceives, and thus to introduce a certain distortion into perception.

An indirect method of obtaining information is to study *behaviors* and *opinions* from which inferences may be drawn concerning quality. A *sociometric* approach has been reported by Maloney et al., which assumes that physicians, in seeking care for themselves and their families, exhibit critical and valid judgments concerning the capacity of their colleagues to provide care of high quality.[29] Such choices were shown to identify classes of physicians presumed to be more highly qualified than others. But both sensitivity and specificity, using as a criterion more rigorous estimates of the quality of care, lack validation. Georgopoulos and Mann[30] used what might be called an *autoreputational*[31] approach in assessing the quality of care in selected community hospitals. This grew out of previous studies showing that people are pretty shrewd judges of the "effectiveness" of the organizations in which they work.[32] The hospitals were rated and ranked using opinions concerning the quality of medical care, and other

characteristics, held by different categories of managerial, professional and technical persons working in, or connected with, each hospital, as well as by knowledgeable persons in the community. The responses were sufficiently consistent and discriminating to permit the hospitals to be ranked with an apparently satisfactory degree of reliability. This is in spite of the generally self-congratulatory nature of the responses that classified the quality of medical care in the hospitals as "very good," "excellent," or "outstanding" in 89 per cent of cases, and "poor" in almost none. The authors provide much evidence that the several opinions, severally held, were intercorrelated to a high degree. But little evidence supports the validity of the judgments by using truly external criteria of the quality of care.

Sampling and Selection

The first issue in sampling is to specify precisely the universe to be sampled, which, in turn, depends on the nature of the generalizations that one wishes to make. Studies of quality are ordinarily concerned with one of three objects: (1) the actual care provided by a specified category of providers of care; (2) the actual care received by a specified group of people and (3) the capacity of a specified group of providers to provide care. In the first two instances representative samples of potential providers or recipients are required, as well as representative samples of care provided or received. In the third instance a representative sample of providers is needed, but not necessarily a representative sample of care. A more important aspect is to select, uniformly of course, significant dimensions of care. Perhaps performance should be studied in certain clinical situations that are particularly stressful and therefore more revealing of latent capacities or weaknesses in performance. Hypothetical test situations may even be set up to assess the capacity to perform in selected dimensions of care.[33–35] The distinctions made above, and especially those between the assessment of actual care provided and of the capacity to provide care, are useful in evaluating the sampling procedures used in the major studies of quality. By these criteria, some studies belong in one category or another, but some seem to combine features of several in such a way that generalization becomes difficult. For example, in the first study of the quality of care received by Teamster families, the findings are meant to apply only to the management of specific categories of hospitalized

illness in a specified population group.[28] In the second study of this series, somewhat greater generalizability is achieved by obtaining a representative sample (exclusive of seasonal variation) of all hospitalized illness in the same population group.[26] Neither study is meant to provide information about all the care provided by a representative sample of physicians.

The degree of homogeneity in the universe to be sampled is, of course, a matter of great importance in any scheme of sampling or selection. The question that must be asked is to what extent the care provided by a physician maintains a consistent level. Do specific diagnostic categories, levels of difficulty or dimensions of care exist in which a physician performs better than in others? Can one find, in fact, an "overall capacity for goodness in medical care,"[18] or is one dealing with a bundle of fairly disparate strands of performance? One might, similarly, ask whether the care provided by all subdivisions of an institution are at about the same level in absolute terms or in relation to performance in comparable institutions. Makover, for example, makes an explicit assumption of homogeneity when he writes, "No attempt was made to relate the number of records to be studied to the size of enrollment of the medical groups. The medical care provided to one or another individual is valid evidence of quality and there should be little or no chance variation which is affected by adjusting the size of the sample."[23] Rosenfeld began his study with the hypothesis "that there is a correspondence in standards of care in the several specialties and for various categories of illness in an institution."[22]

The empirical evidence concerning homogeneity is not extensive. Both the Peterson and Clute studies of general practice[18,19] showed a high degree of correlation between performance of physicians in different components or dimensions of care (history, physical examination, treatment, etc.). Rosenfeld demonstrated that the differences in quality ratings among several diagnoses selected within each area of practice (medicine, surgery and obstetrics-gynecology) were not large. Although the differences among hospitals by area of practice appeared by inspection to be larger, they were not large enough to alter the rankings of the three hospitals studied.

The two studies of care received by Teamster families[26,28] arrived at almost identical proportions of optimal and less than optimal care for the entire populations studied. This must have been coincidental, since the percent of optimal care, in the second study, varied greatly by diagnostic

category from 31 per cent for medicine to 100 per cent for ophthalmology (nine cases only). If such variability exists, the "diagnostic mix" of the sample of care must be a matter of considerable importance in assessment. In the two Teamster studies, differences in "diagnostic mix" were thought to have resulted in lower ratings for medicine and higher ratings for obstetrics-gynecology in the second study than in the first. That the same factor may produce effects in two opposite directions is an indication of the complex interactions that the researcher must consider. "The most probable explanation for the ratings in medicine being lower in the present (second) study is the nature of the cases reviewed." The factor responsible is less ability to handle illness "which did not fall into a well recognized pattern." For obstetrics and gynecology the finding of the second study ". . . differed in one major respect from the earlier study where serious questions were raised about the management of far more patients. The earlier study consisted primarily of major abdominal surgery, whereas this randomly selected group contained few such cases and had more patients with minor conditions."[26] In studies such as these, where the care received by total or partial populations is under study, the variations noted stem partly from differences in diagnostic content and partly from institutionalized patterns of practice associated with diagnostic content. For example, all nine cases of eye disease received optimal care because "this is a highly specialized area, where physicians not trained in this field rarely venture to perform procedures."[26]

Sampling and selection influence, and are influenced by, a number of considerations in addition to generalization and homogeneity. The specific dimensions of care that interest one (preventive management or surgical technique, to mention two rather different examples) may dictate the selection of medical care situations for evaluation. The situations chosen are also related to the nature of the criteria and standards used and of the rating and scoring system adopted. Attempts to sample problem situations, rather than traditional diagnoses or operations, can be very difficult, because of the manner in which clinical records are filed and indexed. This is unfortunate, because a review of operations or established diagnoses gives an insight into the bases upon which the diagnosis was made or the operation performed. It leaves unexplored a complementary segment of practice, namely the situations in which a similar diagnosis or treatment may have been indicated but not made or performed.

Measurement Standards

Measurement depends on the development of standards. In the assessment of quality standards derive from two sources.

Empirical standards are derived from actual practice and are generally used to compare medical care in one setting with that in another, or with statistical averages and ranges obtained from a larger number of similar settings. The Professional Activities Study is based, in part, on this approach.[36]

Empirical standards rest on demonstrably attainable levels of care and, for that reason, enjoy a certain degree of credibility and acceptability. Moreover, without clear normative standards, empirical observations in selected settings must be made to serve the purpose. An interesting example is provided by Furstenberg et al., who used patterns of prescribing in medical care clinics and outpatient hospitals as the standard to judge private practice.[37]

In using empirical standards one needs some assurance that the clinical material in the settings being compared is similar. The Professional Activities Study makes some allowance for this by reporting patterns of care for hospitals grouped by size. The major shortcoming, however, is that care may appear to be adequate in comparison to that in other situations and yet fall short of what is attainable through the full application of current medical knowledge.

Normative standards derive, in principle, from the sources that legitimately set the standards of knowledge and practice in the dominant medical care system. In practice, they are set by standard textbooks or publications,[10] panels of physicians,[25] highly qualified practitioners who serve as judges[26] or a research staff in consultation with qualified practitioners.[22] Normative standards can be put very high and represent the "best" medical care that can be provided, or they can be set at a more modest level signifying "acceptable" or "adequate" care. In any event, their distinctive characteristic is that they stem from a body of legitimate knowledge and values rather than from specific examples of actual practice. As such, they depend for their validity on the extent of agreement concerning facts and values within the profession or, at least, among its leadership. Where equally legitimate sources differ in their views, judgments concerning quality become correspondingly ambiguous.

The relevance of certain normative standards, developed by one group, to the field of practice of another group, has been questioned.

For example, Peterson and Barsamian report that although spermatic fluid examination of the husband should precede surgery for the Stein-Leventhal syndrome, not one instance of such examination was noted, and that this requirement was dropped from the criteria for assessment.[38] Dissatisfaction has also been voiced concerning the application to general practice of standards and criteria elaborated by specialists who practice in academic settings. The major studies of general practice have made allowances for this. Little is known, however, about the strategies of "good" general practice and the extent to which they are similar to, or different from, the strategies of specialized practice in academic settings.

Some researchers have used both types of standards, normative and empirical, in the assessment of care. Rosenfeld used normative standards but included in his design a comparison between university affiliated and community hospitals. "Use of the teaching hospital as a control provides the element of flexibility needed to adjust to the constantly changing scientific basis of the practice of medicine. No written standards, no matter how carefully drawn, would be adequate in five years."[22] Lembcke used experience in the best hospitals to derive a corrective factor that softens the excessive rigidity of his normative standards. This factor, expressed in terms of an acceptable percent of compliance with the standard, was designed to take account of contingencies not foreseen in the standards themselves. It does, however, have the effect of being more realistically permissive as well. This is because the correction factor is likely to be made up partly of acceptable departures from the norm and partly of deviations that might be unacceptable.

Standards can also be differentiated by the extent of their specificity and directiveness. At one extreme the assessing physician may be very simply instructed as follows: "You will use as a yardstick in relation to the quality of care rendered, whether you would have treated this particular patient in this particular fashion during this specific hospital admission."[26] At the other extreme, a virtually watertight "logic system" may be constructed that specifies all the decision rules that are acceptable to justify diagnosis and treatment.[38,39] Most cases fall somewhere in between.

Highly precise and directive standards are associated with the selection of specific diagnostic categories for assessment. When a representative sample of all the care provided is to be assessed, little more than general guides can be given to the assessor. Lembcke, who has stressed the need for specific criteria, has had to develop a correspondingly detailed

diagnostic classification of pelvic surgery, for example.[10] In addition to diagnostic specificity, highly directive standards are associated with the preselection of specific dimensions of care for evaluation. Certain diagnoses, such as surgical operations, lend themselves more readily to this approach. This is evident in Lembcke's attempt to extend his system of audits to nonsurgical diagnoses.[40] The clear, almost rule-of-thumb judgments of adequacy become blurred. The data abstracted under each diagnostic rubric are more like descriptions of patterns of management, with insufficient normative criteria for decisive evaluation. The alternative adopted is comparison with a criterion institution.

Obviously, the more general and nondirective the standards are, the more one must depend on the interpretations and norms of the person entrusted with the actual assessment of care. With greater specificity, the research team is able, collectively, to exercise much greater control over what dimensions of care require emphasis and what the acceptable standards are. A great deal appears in common between the standards used in structured and unstructured situations as shown by the degree of agreement between "intuitive" ratings and directed ratings in the Rosenfeld study,[22] and between the "qualitative" and "quantitative" ratings in the study by Peterson et al.[18] Indeed, these last two were so similar that they could be used interchangeably.

When standards are not very specific and the assessor must exercise his own judgment in arriving at an evaluation, very expert and careful judges must be used. Lembcke claims that a much more precise and directive system such as his does not require expert judges. "It is said that with a cookbook, anyone who can read can cook. The same is true, and to about the same extent, of the medical audit using objective criteria; anyone who knows enough medical terminology to understand the definitions and criteria can prepare the case abstracts and tables for the medical audit. However, the final acceptance, interpretation and application of the findings must be the responsibility of a physician or group of physicians."[41] The "logic system" developed by Peterson and Barsamian appears well suited for rating by computer, once the basic facts have been assembled, presumably by a record abstractor.[38,39]

The dimensions of care and the values that one uses to judge them are, of course, embodied in the criteria and standards used to assess care.[42] These standards can, therefore, be differentiated by their selectivity and inclusiveness in the choice of dimensions to be assessed. The dimensions

selected and the value judgments attached to them constitute the operationalized definition of quality in each study.

The preselection of dimensions makes possible, as already pointed out, the development of precise procedures, standards and criteria. Lembcke[10] has put much stress on the need for selecting a few specific dimensions of care within specified diagnostic categories rather than attempting general evaluations of unspecified dimensions which, he feels, lack precision. He uses dimensions such as the following: confirmation of clinical diagnosis, justification of treatment (including surgery) and completeness of the surgical procedure. Within each dimension, and for each diagnostic category, one or more previously defined activities are often used to characterize performance for that dimension as a whole. Examples are the compatibility of the diagnosis of pancreatitis with serum amylase levels or of liver cirrhosis with biopsy findings, the performance of sensitivity tests prior to antibiotic therapy in acute bronchitis, and the control of blood sugar levels in diabetes.

In addition to the extent to which preselection of dimensions takes place, assessments of quality differ with respect to the number of dimensions used and the exhaustiveness with which performance in each dimension is explored. For example, Peterson et al.,[18] and Rosenfeld[22] use a large number of dimensions. Peterson and Barsamian,[38,39] on the other hand, concentrate on two basic dimensions, justification of diagnosis and of therapy, but require complete proof of justification. A much more simplified approach is illustrated by Huntley et al.,[43] who evaluate outpatient care using two criteria only: the percent of work-ups not including certain routine procedures, and the percent of abnormalities found that were not followed up.

Judgments of quality are incomplete when only a few dimensions are used and decisions about each dimension are made on the basis of partial evidence. Some dimensions, such as preventive care or the psychological and social management of health and illness, are often excluded from the definition of quality and the standards and criteria that make it operational. Examples are the intentional exclusion of psychiatric care from the Peterson study[18] and the planned exclusion of the patient-physician relationship and the attitudes of physicians in studies of the quality of care in the Health Insurance Plan of Greater New York.[27] Rosenfeld[22] made a special point of including the performance of specified screening measures among the criteria of superior care; but care was labeled good in the absence of these measures. In the absence of specific instructions

to the judges, the study by Morehead et al.,[26] includes histories of cases, considered to have received optimal care, in which failure of preventive management could have resulted in serious consequences to the patient.

Another characteristic of measurement is the level at which the standard is set. Standards can be so strict that none can comply with them, or so permissive that all are rated "good." For example, in the study of general practice reported by Clute,[19] blood pressure examinations, measurement of body temperature, otoscopy and performance of immunizations did not serve to categorize physicians because all physicians performed them well.

Measurement Scales

The ability to discriminate different levels of performance depends on the scale of measurement used. Many studies of quality use a small number of divisions to classify care, seen as a whole, into categories such as "excellent," "good," "fair" or "poor." A person's relative position in a set can then be further specified by computing the percent of cases in each scale category. Other studies assign scores to performance of specified components of care and cumulate these to obtain a numerical index usually ranging from 0–100. These practices raise questions relative to scales of measurement and legitimate operations on these scales. Some of these are described below.

Those who adhere to the first practice point out that any greater degree of precision is not possible with present methods. Some have even reduced the categories to only two: optimal and less than optimal. Clute[19] uses three, of which the middle one is acknowledged to be doubtful or indeterminate. Also, medical care has an all-or-none aspect that the usual numerical scores do not reflect. Care can be good in many of its parts and be disastrously inadequate in the aggregate due to a vital error in one component. This is, of course, less often a problem if it is demonstrated that performance on different components of care is highly intercorrelated.

Those who have used numerical scores have pointed out much loss of information in the use of overall judgments,[38] and that numerical scores, cumulated from specified subscores, give a picture not only of the whole but also of the evaluation of individual parts. Rosenfeld[22] has handled this problem by using a system of assigning qualitative scores to component parts of care and an overall qualitative score based on

arbitrary rules of combination that allow for the all-or-none attribute of the quality of medical care. As already pointed out, a high degree of agreement was found between intuitive and structured ratings in the Rosenfeld study[22] and between qualitative and quantitative ratings in the study by Peterson et al.[18]

A major problem, yet unsolved, in the construction of numerical scores, is the manner in which the different components are to be weighted in the process of arriving at the total. At present this is an arbitrary matter. Peterson et al.,[18] for example, arrive at the following scale: clinical history 30, physical examination 34, use of laboratory aids 26, therapy 9, preventive medicine 6, clinical records 2, total 107. Daily and Morehead[24] assign different weights as follows: records 30, diagnostic work-up 40, treatment and follow-up 30, total 100. Peterson et al., say: "Greatest importance is attached to the process of arriving at a diagnosis since, without a diagnosis, therapy cannot be rational. Furthermore, therapy is in the process of constant change, while the form of history and physical examination has changed very little over the years."[18] Daily and Morehead offer no justification for their weightings, but equally persuasive arguments could probably be made on their behalf. The problem of seeking external confirmation remains.[44]

The problem of weights is related to the more general problem of value of items of information or of procedures in the medical care process. Rimoldi et al.,[34] used the frequency with which specified items of information were used in the solution of a test problem as a measure of the value of that item. Williamson had experts classify specified procedures, in a specified diagnostic test setting, on a scale ranging from "very helpful" to "very harmful." Individual performance in the test was then rated using quantitative indices of "efficiency," "proficiency" and overall "competence," depending on the frequency and nature of the procedures used.[35]

A problem in the interpretation of numerical scores is the meaning of the numerical interval between points on the scale. Numerical scores derived for the assessment of quality are not likely to have the property of equal intervals. They should not be used as if they had.

Reliability

The reliability of assessments is a major consideration in studies of quality, where so much depends on judgment even when the directive types

of standards are used. Several studies have given some attention to agreement between judges. The impression gained is that this is considered to be at an acceptable level. Peterson et al.,[18] on the basis of 14 observer revisits, judged agreement to be sufficiently high to permit all the observations to be pooled together after adjustment for observer bias in one of the six major divisions of care. In the study by Daily and Morehead, "several cross-checks were made between the two interviewing internists by having them interview the same physicians. The differences in the scores of the family physicians based on these separate ratings did not exceed 7 per cent."[24] Rosenfeld[22] paid considerable attention to testing reliability, and devised mathematical indices of "agreement" and "dispersion" to measure it. These indicate a fair amount of agreement, but a precise evaluation is difficult since no other investigator is known to have used these same measures. Morehead et al.,[26] in the second study of medical care received by Teamster families, report initial agreement between two judges in assigning care to one of two classes in 78 per cent of cases. This was raised to 92 per cent following reevaluation of disagreements by the two judges.

By contrast to between-judge reliability, very little has been reported about the reliability of repeated judgments of quality made by the same person. To test within-observer variation, Peterson et al.,[18] asked each of two observers to revisit four of his own previously visited physicians. The level of agreement was lower within observers than between observers, partly because revisits lasted a shorter period of time and related, therefore, to a smaller sample of practice.

The major mechanism for achieving higher levels of reliability is the detailed specification of criteria, standards and procedures used for the assessment of care. Striving for reproducibility was, in fact, a major impetus in the development of the more rigorous rating systems by Lembcke, and by Peterson and Barsarmian. Unfortunately, no comparative studies of reliability exist using highly directive versus nondirective methods of assessment. Rosenfeld's raw data might permit a comparison of reliability of "intuitive" judgments and the reliability of structured judgments by the same two assessors. Unreported data by Morehead et al.,[26] could be analyzed in the same way as those of Rosenfeld[22] to give useful information about the relationship between degree of reliability and method of assessment. The partial data that have been published suggest that the post-review reliability achieved by Morehead et al., using the most non-directive of approaches, is quite comparable

with that achieved by Rosenfeld who used a much more directive technique.

Morehead et al., raised the important question of whether the reliability obtained through the detailed specification of standards and criteria may not be gained at the cost of reduced validity. "Frequently, such criteria force into a rigid framework similar actions or factors which may not be appropriate in a given situation due to the infinite variations in the reaction of the human body to illness . . . The study group rejects the assumption that such criteria are necessary to evaluate the quality of medical care. It is their unanimous opinion that it is as important for the surveyors to have flexibility in the judgment of an individual case as it is for a competent physician when confronting a clinical problem in a given patient."[26]

The reasons for disagreement between judges throw some light on the problems of evaluation and the prospects of achieving greater reliability. Rosenfeld found that "almost half the differences were attributable to situations not covered adequately by standards, or in which the standards were ambiguous. In another quarter differences developed around questions of fact, because one consultant missed a significant item of information in the record. It would therefore appear that with revised standards, and improved methods of orienting consultants, a substantially higher degree of agreement could be achieved."[22] Less than a quarter of the disagreements contain differences of opinion with regard to the requirements of management. This is a function of ambiguity in the medical care system and sets an upper limit of reproducibility. Morehead et al., report that in about half the cases of initial disagreement "there was agreement on the most serious aspect of the patient's care, but one surveyor later agreed that he had not taken into account corollary aspects of patient care."[26] Other reasons for disagreement were difficulty in adhering to the rating categories or failure to note all the facts. Of the small number of unresolved disagreements (eight per cent of all admissions and 36 per cent of initial disagreements) more than half were due to honest differences of opinion regarding the clinical handling of the problem. The remainder arose out of differences in interpreting inadequate records, or the technical problems of where to assess unsatisfactory care in a series of admissions.[27]

A final aspect of reliability is the occasional breakdown in the performance of an assessor, as so dramatically demonstrated in the Rosenfeld study.[22] The question of what the investigator does when a well defined

segment of his results are so completely aberrant will be raised here without any attempt to provide an answer.

Bias

When several observers or judges describe and evaluate the process of medical care, one of them may consistently employ more rigid standards than another, or interpret predetermined standards more strictly. Peterson et al.,[18] discovered that one of their observers generally awarded higher ratings than the other in the assessment of performance of physical examination, but not in the other areas of care. Rosenfeld[22] showed that, of two assessors, one regularly awarded lower ratings to the same cases assessed by both. An examination of individual cases of disagreement in the study by Morehead et al.,[26] reveals that, in the medical category, the same assessor rated the care at a lower level in 11 out of 12 instances of disagreement. For surgical cases, one surveyor rated the care lower than the other in all eight instances of disagreement. The impression is gained from examining reasons for disagreement on medical cases that one of the judges had a special interest in cardiology and was more demanding of clarity and certainty in the management of cardiac cases.

The clear indication of these findings is that bias must be accepted as the rule rather than the exception, and that studies of quality must be designed with this in mind. In the Rosenfeld study,[22] for example, either of the two raters used for each area of practice would have ranked the several hospitals in the same order, even though one was consistently more generous than the other. The Clute study of general practice in Canada,[19] on the other hand, has been criticized for comparing the quality of care in two geographic areas even though different observers examined the care in the two areas in question.[45] The author was aware of this problem and devised methods for comparing the performance of the observers in the two geographic areas, but the basic weakness remains.

Predetermined order or regularity in the process of study may be associated with bias. Therefore, some carefully planned procedures may have to be introduced into the research design for randomization. The study by Peterson et al.,[18] appears to be one of the few to have paid attention to this factor. Another important source of bias is knowledge, by the assessor, of the identity of the physician who provided the care or of the hospital in which care was given. The question of removing identifying features from charts under review has been raised,[3] but little

is known about the feasibility of this procedure and its effects on the ratings assigned. Still another type of bias may result from parochial standards and criteria of practice that may develop in and around certain institutions or "schools" of medical practice. To the extent that this is true, or suspected to be true, appropriate precautions need to be taken in the recruitment and allocation of judges.

Validity

The effectiveness of care as has been stated, in achieving or producing health and satisfaction, as defined for its individual members by a particular society or subculture, is the ultimate validator of the quality of care. The validity of all other phenomena as indicators of quality depends, ultimately, on the relationship between these phenomena and the achievement of health and satisfaction. Nevertheless, conformity of practice to accepted standards has a kind of conditional or interim validity which may be more relevant to the purposes of assessment in specific instances.

The validation of the details of medical practice by their effect on health is the particular concern of the clinical sciences. In the clinical literature one seeks data on whether penicillin promotes recovery in certain types of pneumonia, anticoagulants in coronary thrombosis, or corticosteroids in rheumatic carditis; what certain tests indicate about the function of the liver; and whether simple or radical mastectomy is the more life-prolonging procedure in given types of breast cancer. From the general body of knowledge concerning such relationships arise the standards of practice, more or less fully validated, by which the medical care process is ordinarily judged.

Intermediate, or procedural, end points often represent larger bundles of care. Their relationship to outcome has attracted the attention of both the clinical investigator and the student of medical care organization. Some examples of the latter are studies of relationships between prenatal care and the health of mothers and infants[46,47] and the relationship between multiple screening examinations and subsequent health.[48] An interesting example of the study of the relationship between one procedural end point and another is the attempt to demonstrate a positive relationship between the performance of rectal and vaginal examinations by the physician, and the pathological confirmation of appendicitis in primary appendectomies, as reported by the Professional Activities Study.[49]

Many studies reviewed[18,19,23,26,28] attempt to study the relationship between structural properties and the assessment of the process of care. Several of these studies have shown, for example, a relationship between the training and qualifications of physicians and the quality of care they provide. The relationship is, however, a complex one, and is influenced by the type of training, its duration and the type of hospital within which it was obtained. The two studies of general practice[18,19] have shown additional positive relationships between quality and better office facilities for practice, the presence or availabilty of laboratory equipment, and the institution of an appointment system. No relationship was shown between quality and membership of professional associations, the income of the physician or the presence of x-ray equipment in the office. The two studies do not agree fully on the nature of the relationship between quality of practice and whether the physician obtained his training in a teaching hospital or not, the number of hours worked or the nature of the physician's hospital affiliation. Hospital accreditation, presumably a mark of quality conferred mainly for compliance with a wide range of organizational standards, does not appear, in and of itself, to be related to the quality of care, at least in New York City.[26]

Although structure and process are no doubt related, the few examples cited above indicate clearly the complexity and ambiguity of these relationships. This is the result partly of the many factors involved, and partly of the poorly understood interactions among these factors. For example, one could reasonably propose, based on several findings[26,38] that both hospital factors and physician factors influence the quality of care rendered in the hospital, but that differences between physicians are obliterated in the best and worst hospital and express themselves, in varying degrees, in hospitals of intermediate quality.

An approach particularly favored by students of medical care organization is to examine the relations between structure and outcome without reference to the complex processes that tie them together. Some examples of such studies have been cited already.[6–9] Others include studies of the effects of reorganizing the outpatient clinic on health status,[50] the effects of intensive hospital care on recovery,[51] the effects of home care on survival[52] and the effect of a rehabilitation program on the physical status of nursing home patients.[53,54] The lack of relationship to outcome in the latter two studies suggests that current opinions about how care should be set up are sometimes less than well established.

This brief review indicates the kinds of evidence pertaining to the validity of the various approaches to the evaluation of quality of care. Clearly, the relationships between process and outcome, and between structure and both process and outcome, are not fully understood. With regard to this, the requirements of validation are best expressed by the concept, already referred to, of a chain of events in which each event is an end to the one that comes before it and a necessary condition to the one that follows. This indicates that the means-end relationship between each adjacent pair requires validation in any chain of hypothetical or real events.[55] This is, of course, a laborious process. More commonly, as has been shown, the intervening links are ignored. The result is that causal inferences become attenuated in proportion to the distance separating the two events on the chain.

Unfortunately, very little information is available on actual assessments of quality using more than one method of evaluation concurrently. Makover has studied specifically the relationships between multifactorial assessments of structure and of process in the same medical groups. "It was found that the medical groups that achieved higher quality ratings by the method used in this study were those that, in general, adhered more closely to HIP's Minimum Medical Standards. However, the exceptions were sufficiently marked, both in number and degree, to induce one to question the reliability[56] of one or the other rating method when applied to any one medical group. It would seem that further comparison of these two methods of rating is clearly indicated."[23]

Indices of Medical Care

Since a multidimensional assessment of medical care is a costly and laborious undertaking, the search continues for discrete, readily measurable data that can provide information about the quality of medical care. The data used may be about aspects of structure, process or outcome. The chief requirement is that they be easily, sometimes routinely, measurable and be reasonably valid. Among the studies of quality using this approach are those of the Professional Activities Study,[36] Ciocco et al.,[57] and Furstenberg et al.[37]

Such indices have the advantage of convenience; but the inferences that are drawn from them may be of doubtful validity. Myers has pointed out

the many limitations of the traditional indices of the quality of hospital care, including rates of total and postoperative mortality, complications, postoperative infection, Caesarian section, consultation and removal of normal tissue at operation.[58] The accuracy and completeness of the basic information may be open to question. More important still, serious questions may be raised about what each index means since so many factors are involved in producing the phenomenon which it measures. Eislee has pointed out, on the other hand, that at least certain indices can be helpful, if used with care.[36]

The search for easy ways to measure a highly complex phenomenon such as medical care may be pursuing a will-o'-the-wisp. The use of simple indices in lieu of more complex measures may be justified by demonstrating high correlations among them.[1] But, in the absence of demonstrated causal links, this may be an unsure foundation upon which to build. On the other hand, each index can be a measure of a dimension or ingredient of care. Judiciously selected multiple indices may, therefore, constitute the equivalent of borings in a geological survey which yield sufficient information about the parts to permit reconstruction of the whole. The validity of inferences about the whole will depend, of course, on the extent of internal continuities in the individual or institutional practice of medicine.

Some Problems of Assessing Ambulatory Care

Some of the special difficulties in assessing the quality of ambulatory care have already been mentioned. These include the paucity of recorded information, and the prior knowledge, by the managing physician, of the patient's medical and social history. The first of these problems has led to the use of trained observers and the second to the observation of cases for which prior knowledge is not a factor in current management. The degree of relevance to general practice of standards and strategies of care developed by hospital centered and academically oriented physicians has also been questioned.

Another problem is the difficulty of defining the segment of care that may be properly the object of evaluation in ambulatory care. For hospital care, a single admission is usually the appropriate unit.[59] In office or clinic practice, a sequence of care may cover an indeterminate number of visits so that the identification of the appropriate unit is open to question. Usually the answer has been to choose an arbitrary time

period to define the relevant episode of care. Ciocco et al.,[57] defined this as the first visit plus 14 days of follow-up. Huntley et al.,[43] use a four-week period after the initial work-up.

Conclusions and Proposals

This review has attempted to give an impression of the various approaches and methods that have been used for evaluating the quality of medical care, and to point out certain issues and problems that these approaches and methods bring up for consideration.

The methods used may easily be said to have been of doubtful value and more frequently lacking in rigor and precision. But how precise do estimates of quality have to be? At least the better methods have been adequate for the administrative and social policy purposes that have brought them into being. The search for perfection should not blind one to the fact that present techniques of evaluating quality, crude as they are, have revealed a range of quality from outstanding to deplorable. Tools are now available for making broad judgments of this kind with considerable assurance. This degree of assurance is supported by findings, already referred to, that suggest acceptable levels of homogeneity in individual practice and of reproducibility of qualitative judgments based on a minimally structured approach to evaluation. This is not to say that a great deal does not remain to be accomplished in developing the greater precision necessary for certain other purposes.

One might begin a catalogue of needed refinements by considering the nature of the information which is the basis for judgments of quality. More must be known about the effect of the observer on the practice being observed, as well as about the process of observation itself—its reliability and validity. Comparisons need to be made between direct observation and recorded information both with and without supplementation by interview with the managing physician. Recording agreement or disagreement is not sufficient. More detailed study is needed of the nature of, and reasons for, discrepancy in various settings. Similarly, using abstracts of records needs to be tested against using the records themselves.

The process of evaluation itself requires much further study. A great deal of effort goes into the development of criteria and standards which are presumed to lend stability and uniformity to judgments of quality;

and yet this presumed effect has not been empirically demonstrated. How far explicit standardization must go before appreciable gains in reliability are realized is not known. One must also consider whether, with increasing standardization, so much loss of the ability to account for unforeseen elements in the clinical situation occurs that one obtains reliability at the cost of validity. Assessments of the same set of records using progressively more structured standards and criteria should yield valuable information on these points. The contention that less well trained assessors using exhaustive criteria can come up with reliable and valid judgments can also be tested in this way.

Attention has already been drawn, in the body of the review, to the little that is known about reliability and bias when two or more judges are compared, and about the reliability of repeated judgments of the same items of care by the same assessor. Similarly, very little is known about the effects on reliability and validity, of certain characteristics of judges including experience, areas of special interest and personality factors. Much may be learned concerning these and related matters by making explicit the process of judging and subjecting it to careful study. This should reveal the dimensions and values used by the various judges and show how differences are resolved when two or more judges discuss their points of view. Some doubt now exists about the validity of group reconciliations in which one point of view may dominate, not necessarily because it is more valid.[1] The effect of masking the identity of the hospital or the physician providing care can be studied in the same way. What is proposed here is not only to demonstrate differences or similarities in overall judgments, but to attempt, by making explicit the thought processes of the judges, to determine how the differences and similarities arise, and how differences are resolved.

In addition to defects in method, most studies of quality suffer from having adopted too narrow a definition of quality. In general, they concern themselves with the technical management of illness and pay little attention to prevention, rehabilitation, coordination and continuity of care, or handling the patient-physician relationship. Presumably, the reason for this is that the technical requirements of management are more widely recognized and better standardized. Therefore, more complete conceptual and empirical exploration of the definition of quality is needed.

What is meant by "conceptual exploration" may be illustrated by considering the dimension of efficiency which is often ignored in studies

of quality. Two types of efficiency might be distinguished: logical and economic. Logical efficiency concerns the use of information to arrive at decisions. Here the issue might be whether the information obtained by the physician is relevant or irrelevant to the clinical business to be transacted. If relevant, one might consider the degree of replication or duplication in information obtained and the extent to which it exceeds the requirements of decision making in a given situation. If parsimony is a value in medical care, the identification of redundancy becomes an element in the evaluation of care.

Economic efficiency deals with the relationships between inputs and outputs and asks whether a given output is produced at least cost. It is, of course, influenced by logical efficiency, since the accumulation of unnecessary or unused information is a costly procedure which yields no benefit. Typically it goes beyond the individual and is concerned with the social product of medical care effort. It considers the possibility that the "best" medical care for the individual may not be the "best" for the community. Peterson et al., cite an example that epitomizes the issue. "Two physicians had delegated supervision of routine prenatal visits to office nurses, and the doctor saw the patient only if she had specific complaints."[18] In one sense, this may have been less than the best care for each expectant mother. In another sense, it may have been brilliant strategy in terms of making available to the largest number of women the combined skills of a medical care team. Cordero, in a thought provoking paper, has documented the thesis that, when resources are limited, optimal medical care for the community may require less than "the best" care for its individual members.[60]

In addition to conceptual exploration of the meaning of quality, in terms of dimensions of care and the values attached to them, empirical studies are needed of what are the prevailing dimensions and values in relevant population groups.[5] Little is known, for example, about how physicians define quality, nor is the relationship known between the physician's practice and his own definition of quality. This is an area of research significant to medical education as well as quality. Empirical studies of the medical care process should also contribute greatly to the identification of dimensions and values to be incorporated into the definition of quality.

A review of the studies of quality shows a certain discouraging repetitiousness in basic concepts, approaches and methods. Further substantive progress, beyond refinements in methodology, is likely to come from a

program of research in the medical care process itself rather than from frontal attacks on the problem of quality. This is believed to be so because, before one can make judgments about quality, one needs to understand how patients and physicians interact and how physicians function in the process of providing care. Once the elements of process and their interrelationships are understood, one can attach value judgments to them in terms of their contributions to intermediate and ultimate goals. Assume, for example, that authoritarianism-permissiveness is one dimension of the patient-physician relationship. An empirical study may show that physicians are in fact differentiated by this attribute. One might then ask whether authoritarianism or permissiveness should be the criterion of quality. The answer could be derived from the general values of society that may endorse one or the other as the more desirable attribute in social interactions. This is one form of quality judgment, and is perfectly valid, provided its rationale and bases are explicit. The study of the medical care process itself may however offer an alternative, and more pragmatic, approach. Assume, for the time being, that compliance with the recommendations of the physician is a goal and value in the medical care system. The value of authoritarianism or permissiveness can be determined, in part, by its contribution to compliance. Compliance is itself subject to validation by the higher order criterion of health outcomes. The true state of affairs is likely to be more complex than the hypothetical example given. The criterion of quality may prove to be congruence with patient expectations, or a more complex adaptation to specific clinical and social situations, rather than authoritarianism or permissiveness as a predominant mode. Also, certain goals in the medical care process may not be compatible with other goals, and one may not speak of quality in global terms but of quality in specified dimensions and for specified purposes. Assessments of quality will not, therefore, result in a summary judgment but in a complex profile, as Sheps has suggested.[1]

A large portion of research in the medical care process will, of course, deal with the manner in which physicians gather clinically relevant information, and arrive at diagnostic and therapeutic decisions. This is not the place to present a conceptual framework for research in this portion of the medical care process. Certain specific studies may, however, be mentioned and some directions for further research indicated.

Research on information gathering includes studies of the perception and interpretation of physical signs.[61,62] Evans and Bybee have shown, for example, that in interpreting heart sounds errors of perception (of rhythm

and timing) occurred along with additional errors of interpretation of what was perceived. Faulty diagnosis, as judged by comparison with a criterion, was the result of these two errors.[62] This points to the need for including, in estimates of quality, information about the reliability and validity of the sensory data upon which management, in part, rests.

The work of Peterson and Barsamian[38,39] represents the nearest approach to a rigorous evaluation of diagnostic and therapeutic decision making. As such, it is possibly the most significant recent advance in the methods of quality assessment. But this method is based on record reviews and is almost exclusively preoccupied with the justification of diagnosis and therapy. As a result, many important dimensions of care are not included in the evaluation. Some of these are considerations of efficiency, and of styles and strategies in problem solving.

Styles and strategies in problem solving can be studied through actual observation of practice, as was done so effectively by Peterson et al., in their study of general practice.[18] A great deal that remains unobserved can be made explicit by asking the physician to say aloud what he is doing and why. This method of *réflexion parlée* has been used in studies of problem solving even though it may, in itself, alter behavior.[63] Another approach is to set up test situations, such as those used by Rimoldi et al.,[34] and by Williamson,[35] to observe the decision making process. Although such test situations have certain limitations arising out of their artificiality,[64] the greater simplicity and control that they provide can be very helpful.

At first sight, the student of medical care might expect to be helped by knowledge and skill developed in the general field of research in problem solving. Unfortunately, no well developed theoretical base is available which can be exploited readily in studies of medical care. Some of the empirical studies in problem solving might however, suggest methods and ideas applicable to medical care situations.[63–67] Some of the studies of "troubleshooting" in electronic equipment, in particular, show intriguing similarities to the process of medical diagnosis and treatment. These and similar studies have identified behavioral characteristics that might be used to categorize styles in clinical management. They include the amount of information collected, rate of seeking information, value of items of information sought as modified by their place in a sequence and by interaction with other items of information, several types of redundancy, stereotypy, search patterns in relation to the part known to be defective, tendencies to act prior to amassing sufficient information

or to seek information beyond the point of reasonable assurance about the solution, "error distance" and degrees of success in achieving a solution, and so on.

Decision making theory may also offer conceptual tools of research in the medical care process. Ledley and Lusted,[68,69] among others, have attempted to apply models based on conditional probabilities to the process of diagnosis and therapy. Peterson and Barsamian[38,39] decided against using probabilities in their logic systems for the very good reason that the necessary data (the independent probabilities of diseases and of symptoms, and the probabilities of specified symptoms in specified diseases) were not available. But Edwards et al.,[70] point out that one can still test efficiency in decision making by substituting subjective probabilities (those of the decision maker himself or of selected experts) for the statistical data one would prefer to have.

A basic question that has arisen frequently in this review is the degree to which performance in medical care is a homogeneous or heterogeneous phenomenon. This was seen, for example, to be relevant to sampling, the use of indices in place of multidimensional measurements, and the construction of scales that purport to judge total performance. When this question is raised with respect to individual physicians, the object of study is the integration of various kinds of knowledge and of skills in the personality and behavior of the physician. When it is raised with respect to institutions and social systems the factors are completely different. Here one is concerned with the formal and informal mechanisms for organizing, influencing and directing human effort in general, and the practice of medicine in particular. Research in all these areas is expected to contribute to greater sophistication in the measurement of quality.

Some of the conventions accepted in this review are, in themselves, obstacles to more meaningful study of quality. Physicians' services are not, in the real world, separated from the services of other health professionals, nor from the services of a variety of supportive personnel. The separation of hospital and ambulatory care is also largely artificial. The units of care which are the proper objects of study include the contributions of many persons during a sequence which may include care in a variety of settings. The manner in which these sequences are defined and identified has implications for sampling, methods of obtaining information, and standards and criteria of evaluation.

A final comment concerns the frame of mind with which studies of quality are approached. The social imperatives that give rise to

assessments of quality have already been referred to. Often associated with these are the zeal and values of the social reformer. Greater neutrality and detachment are needed in studies of quality. More often one needs to ask, "What goes on here?" rather than, "What is wrong; and how can it be made better?" This does not mean that the researcher disowns his own values or social objectives. It does mean, however, that the distinction between values, and elements of structure, process or outcome, is recognized and maintained; and that both are subjected to equally critical study. Partly to achieve this kind of orientation emphasis must be shifted from preoccupation with evaluating quality to concentration on understanding the medical care process itself.

References

1. Sheps, M.C. September, 1955. Approaches to the Quality of Hospital Care. *Public Health Reports* 70:877–886.
   This paper represents an unusually successful crystallization of thinking concerning the evaluation of quality. It contains brief but remarkably complete discussions of the purposes of evaluation, problems of definition, criteria and standards, various approaches to measurement, the reliability of qualitative judgments and indices of quality. The bibliography is excellent.
2. Peterson, L. December 5, 1963. Evaluation of the Quality of Medical Care. *New England Journal of Medicine* 269:1238–1245.
3. Lerner, M., and D.C. Riedel. January, 1964. The Teamster Study and the Quality of Medical Care. *Inquiry* 1:69–80.
   The major value of this paper is that it raises questions concerning methods of assessment including the sampling of populations and diagnostic categories, the use of records and the need for supplementation by interview, the value of detailed standards, the need for understanding the auditing process, the definition of terms and concepts (of "unnecessary admission," for example), and the problems of defining the relevant episode of care.
4. Lee, R.I., and L.W. Jones. 1933. *The Fundamentals of Good Medical Care*. Chicago: University of Chicago Press.
5. Klein, M.W., et al. Summer, 1961. Problems of Measuring Patient Care in the Out Patient Department. *Journal of Health and Human Behavior* 2:138–144.
6. Kohl, S.G. 1955. *Perinatal Mortality in New York City: Responsible Factors*. Cambridge: Harvard University Press.
   This study, sponsored by the New York Academy of Medicine, was an examination by an expert committee of records pertaining to a representative sample of perinatal deaths in New York City.

Preventable deaths were recognized and "responsibility factors" identified, including errors in medical judgment and technique. The incidence of both of these was further related to type of hospital service, type of professional service and type of hospital, indicating relationships between structure and outcome as modified by the characteristics of the population served.

7. Shapiro, S., et al. September, 1960. Further Observations on Prematurity and Perinatal Mortality in a General Population and in the Population of a Prepaid Group Practice Medical Care Plan. *American Journal of Public Health* 50:1304–1317.
8. Lipworth, L., J.A.H. Lee, and J.N. Morris. April–June, 1963. Case Fatality in Teaching and Nonteaching Hospitals, 1956–1959. *Medical Care* 1:71–76.
9. Rice, C.E., et al. May, 1961. Measuring Social Restoration Performance of Public Psychiatric Hospitals. *Public Health Reports* 76:437–446.
10. Lembcke, P.A. October 13, 1956. Medical Auditing by Scientific Methods. *Journal of the American Medical Association* 162:646–655. (Appendices A and B supplied by the author.)

    This is perhaps the single best paper that describes the underlying concepts as well as the methods of the highly structured approach developed by Lembcke to audit hospital records. Also included is an example of the remarkable effect that an "external audit" of this kind can have on surgical practice in a hospital.
11. Kelman, H.R., and A. Willner. April, 1962. Problems in Measurement and Evaluation of Rehabilitation. *Archives of Physical Medicine and Rehabilitation* 43:172–181.
12. McDermott, W., et al. January 22 and 29, 1960. Introducing Modern Medicine in a Navajo Community. *Science* 131:197–205 and 280–287.
13. Simon, H.A. 1961. *Administrative Behavior*, 62–66. New York: The Macmillan Company.
14. Hutchinson, G.B. May, 1960. Evaluation of Preventive Services. *Journal of Chronic Diseases* 11:497–508.
15. James, G. 1960. *Evaluation of Public Health*, 7–17. Report of the Second National Conference on Evaluation in Public Health. Ann Arbor: The University of Michigan, School of Public Health.
16. Weinerman, E.R. September, 1950. Appraisal of Medical Care Programs. *American Journal of Public Health* 40:1129–1134.
17. Goldmann, F., and E.A. Graham. 1954. *The Quality of Medical Care Provided at the Labor Health Institute, St. Louis, Missouri*. St. Louis: The Labor Health Institute.

    This is a good example of an approach to evaluation based on structural characteristics. In this instance, these included the layout and equipment of physical facilities, the competence and stability of medical staff, provisions made for continuity of service centering

around a family physician, the scheduling and duration of clinic visits, the content of the initial examination, the degree of emphasis on preventive medicine and the adequacy of the medical records.

18. Peterson, O.L., et al. December, 1956. An Analytical Study of North Carolina General Practice: 1953–1954. *The Journal of Medical Education* 31:1–165, Part 2.

    Already a classic, this study is distinguished by more than ordinary attention to methods and rather exhaustive exploration of the relationship between quality ratings and characteristics of physicians, including education training and methods of practice. The findings of this study, and others that have used the same method, raise basic questions about traditional general practice in this and other countries

19. Clute, K.F. 1963. *The General Practitioner: A Study of Medical Education and Practice in Ontario and Nova Scotia*. Toronto: University of Toronto Press, chapters 1, 2, 16, 17 and 18.

    Since this study uses the method developed by Peterson, et al., it offers an excellent opportunity to examine the generality of relationships between physician characteristics and quality ratings. In addition, the reader of this elegantly written volume gets a richly detailed view of general practice in the two areas studied.

20. Kroeger, H.H., et al. August 2, 1965. The Office Practice of Internists, I. The Feasibility of Evaluating Quality of Care. *The Journal of the American Medical Association* 193:371–376.

    This is the first of a series of papers based on a study of the practice of members of the New York Society of Internal Medicine. This paper reports findings concerning the completeness of office records, their suitability for judging quality and the degree of agreement between abstracts of records prepared by physicians and by highly trained non-physicians. Judgments concerning the quality of care provided are not given. Other papers in this series currently appearing in the *Journal of the American Medical Association* concern patient load (August 23), characteristics of patients (September 13), professional activities other than care of private patients (October 11), and background and form of practice (November 1).

21. Kilpatrick, G.S. January, 1963. Observer Error in Medicine. *Journal of Medical Education* 38:38–43. For a useful bibliography on observer error *see* Witts, L.J. (Editor), *Medical Surveys and Clinical Trials*, London, Oxford University Press, 1959, pp. 39–44.

22. Rosenfeld, L.S. July, 1957. Quality of Medical Care in Hospitals. *American Journal of Public Health* 47:856–865.

    This carefully designed comparative study of the quality of care in four hospitals addresses itself to the problems of methods in the assessment of quality. Here one finds important information about the use of normative and empirical standards, reliability and bias in judgments based on chart review, the correlation between defects in

recording and defects in practice and homogeneity in quality ratings within and between diagnostic categories.

23. Makover, H.B. July, 1951. The Quality of Medical Care: Methodological Survey of the Medical Groups Associated with the Health Insurance Plan of New York. *American Journal of Public Health* 41:824–832.

    This is possibly the first published report concerning an administratively instituted, but research oriented, program of studies of the quality of care in medical groups contracting with the Health Insurance Plan of Greater New York. Unfortunately much of this work remains unpublished. A particular feature of this paper is that it describes, and presents the findings of simultaneous evaluation of structure (policies, organization, administration, finances and professional activities) and process (evaluation of a sample of clinical records).

24. Daily, E.F., and M.A. Morehead, July, 1956. A Method of Evaluating and Improving the Quality of Medical Care. *American Journal of Public Health* 46:848–854.

25. Fitzpatrick, T.B., D.C. Riedel, and B.C. Payne. 1962. Character and Effectiveness of Hospital Use in *Hospital and Medical Economics*, edited by W.J. Mc Nerney, et al., 495–509. Chicago: Hospital Research and Educational Trust, American Hospital Association.

26. Morehead, M.A., et al. 1964. *A Study of the Quality of Hospital Care Secured by a Sample of Teamster Family Members in New York City*. New York: Columbia University, School of Public Health and Administrative Medicine.

    This study and its companion[28] perform a very important social and administrative function by documenting how frequently the care received by members of a union through traditional sources proves to be inadequate. These studies also make a major contribution to understanding the relationships between hospital and physician characteristics and the quality of care they provide. Considered are physician classifications by specialty status and admission privileges, as well as hospital classifications by ownership, medical school affiliation, approval for residency training and accreditation status. The interactional effects of some of these variables are also explored. In addition, the second of the two studies pays considerable attention to questions of method, including representative versus judgmental sampling of hospital admissions and the reliability of record evaluations by different judges.

27. Morehead, M.A., Personal communication.

28. Ehrlich, J., M.A. Morehead, and R.E. Trussell. 1962. *The Quantity, Quality and Costs of Medical and Hospital Care Secured by a Sample of Teamster Families in the New York Area*. New York: Columbia University, School of Public Health and Administrative Medicine.

29. Maloney, M.C., R.E. Trussell, and J. Elinson. November, 1960. Physicians Choose Medical Care: A Sociometric Approach to Quality Appraisal. *American Journal of Public Health* 50:1678–1686.

This study represents an ingenious approach to evaluation through the use of "peer judgments" in what is believed to be a particularly revealing situation: choice of care for the physician or members of his own family. Some of the characteristics of the physicians and surgeons selected included long-standing personal and professional relationships, recognized specialist status, and medical school affiliation. An incidental pearl of information is that although nine out of ten physicians said everyone should have a personal physician, four out of ten said they had someone whom they considered their personal physician, and only two out of ten had seen their personal physician in the past year!

30. Georgopoulos, B.S., and F.C. Mann. 1962. *The Community General Hospital*. New York: The Macmillan Company.

The study of quality reported in several chapters of this book is based on the thesis that if one wishes to find out about the quality of care provided, all one might need to do is to ask the persons directly or indirectly involved in the provision of such care. Although physicians may find this notion rather naive, the stability and internal consistency of the findings reported in this study indicate that this approach deserves further careful evaluation. A second study of a nationwide sample of general hospitals will attempt to confirm the validity of respondent opinions by comparing them to selected indices of professional activities in each hospital. The findings will be awaited with great interest.

31. One of the author's students, Mr. Arnold D. Kaluzny, helped the author to coin this word.

32. Georgopoulos, B.S., and A.S. Tannenbaum. October, 1957. A Study of Organizational Effectiveness. *American Sociological Review* 22:534–540.

33. Evans, L.R., and J.R. Bybee. February, 1965. Evaluation of Student Skills in Physical Diagnosis. *Journal of Medical Education* 40:199–204.

34. Rimoldi, H.J.A., J.V. Haley, and H. Fogliatto. 1962. *The Test of Diagnostic Skills*. Loyola Psychometric Laboratory Publication Number 25. Chicago: Loyola University Press.

This study is of interest because it uses a controlled test situation to study the performance of medical students and physicians. Even more intriguing is the attempt to approach the question of the value or utility of diagnostic actions in a systematic and rigorous manner. While this particular study does not appear to contribute greatly to understanding the quality of care, this general approach appears to be worth pursuing.

35. Williamson, J.W. February, 1965. Assessing Clinical Judgment. *Journal of Medical Education* 40:180–187.

This is another example of the assessment of clinical performance using an artificial test situation. The noteworthy aspect of the work is the attachment of certain values ("helpful" or "harmful") to a set of diagnostic and therapeutic actions and the development of measures of "efficiency," "proficiency" and "competence" based on which actions are selected by the subject in managing the test case. Differences of performance between individual physicians were detected using this method. An unexpected finding was the absence of systematic differences by age, training or type of practice in groups tested so far.

36. Eislee, C.W., V.N. Slee, and R.G. Hoffmann. January, 1956. Can the Practice of Internal Medicine Be Evaluated? *Annals of Internal Medicine* 44:144–161.

The authors discuss the use of indices from which inferences might be drawn concerning the quality of surgical and medical management. The indices described include tissue pathology reports in appendectomies, diabetes patients without blood sugar determinations and without chest x-rays, and pneumonia without chest x-rays. A striking finding reported in this paper, and others based on the same approach, is the tremendous variation by physician and hospital in the occurrence of such indices of "professional activity."

37. Furstenberg, F.F., et al. October, 1953. Prescribing as an Index to Quality of Medical Care: A Study of the Baltimore City Medical Care Program. *American Journal of Public Health* 43:1299–1309.

38. Peterson, O.L., and E.M. Barsamian. October 7–11, 1963. An Application of Logic to a Study of Quality of Surgical Care. Paper read at the Fifth IBM Medical Symposium, Endicott, New York.

This paper and its companion[39] present a fairly complete description of the "logic tree" approach to the evaluation of quality. Examples are given of the logic systems for the Stein-Leventhal Syndrome and uterine fibromyoma. No data are given on empirical findings using this method.

39. Peterson, O.L., and E.M. Barsamian. April, 1964. Diagnostic Performance. In *The Diagnostic Process*, edited by J.A. Jacquez, 347–362. Ann Arbor: The University of Michigan Press.

40. Lembcke, P.A., and O.G. Johnson. 1963. *A Medical Audit Report*. Los Angeles: University of California, School of Public Health (Mimeographed).

This is an extension of Lembcke's method of medical audit to medical diagnostic categories as well as a large number of surgical operations. Although this volume is a compendium of fairly raw data, careful study can provide insights and limitations of the method used by the author.

41. Lembcke, P.A. 1959. A Scientific Method for Medical Auditing. *Hospitals* 33:65–71, June 16 and 65–72, July 1.
42. The dimensionality of the set of variables incorporating these standards remains to be determined.
43. Huntley, R.R., et al., December, 1961. The Quality of Medical Care: Techniques and Investigation in the Outpatient Clinic. *Journal of Chronic Diseases* 14:630–642.

    This study provides an example of the application of a routine chart review procedure as a check on the quality of management in the outpatient department of a teaching hospital. Fairly often routine procedures were not carried out and abnormalities that were found were not followed up. A revised chart review procedure seemed to make a significant reduction in the percent of abnormalities not followed up.
44. Peterson, et al., loc. cit., attempted to get some confirmation of weightings through the procedure of factor analysis. The mathematically sophisticated are referred to their footnote on pp. 14–15.
45. Mainland, D., August 24, 1964. Calibration of the Human Instrument. Notes from a Laboratory of Medical Statistics, Number 81 (Mimeographed).
46. Joint Committee of the Royal College of Obstetricians and Gynecologists and the Population Investigation Committee. 1948. *Maternity in Great Britain*. London: Oxford University Press.
47. Yankauer, A., K.G. Goss, and S.M. Romeo. August, 1953. An Evaluation of Prenatal Care and Its Relationship to Social Class and Social Disorganization. *American Journal of Public Health* 43:1001–1010.
48. Wylie, C.M. July, 1961. Participation in a Multiple Screening Clinic with Five-Year Follow-Up. *Public Health Reports* 76:596–602.
49. Commission on Professional and Hospital Activities. October, 1957. *Medical Audit Study Report 5: Primary Appendectomies*. Ann Arbor: The Commission on Professional and Hospital Activities.
50. Simon, A.J. September, 1959. Social Structure of Clinics and Patient Improvement. *Administrative Science Quarterly* 4:197–206.
51. Lockward, H.J., G.A.F. Lundberg, and M.E. Odoroff. August, 1963. Effect of Intensive Care on Mortality Rate of Patients with Myocardial Infarcts. *Public Health Reports* 78:655–661.
52. Bakst, J.N., and E.F. Marra. April, 1955. Experiences with Home Care for Cardiac Patients. *American Journal of Public Health* 45:444–450.
53. Muller, J.N., J.S. Tobis, and H.R. Kelman. February, 1963. The Rehabilitation Potential of Nursing Home Residents. *American Journal of Public Health* 53:243–247.
54. These studies also include data on the relationships between structural features and procedural end points. Examples are the effect of

clinic structure on the number of outpatient visits,[50] and the effect of a home care program on hospital admissions.[52]

55. Getting, V.A., et al., October 5, 1964. Research in Evaluation in Public Health Practices. Paper presented at the 92nd Annual Meeting, American Public Health Association, New York.
56. Assuming the direct evaluation of process to be the criterion, the issue becomes one of the implications of reliability measures for validity.
57. Ciocco, A., H. Hunt, and I. Altman. January 27, 1950. Statistics on Clinical Services to New Patients in Medical Groups. *Public Health Reports* 65:99–115.

    This is an early application to group practice of the analysis of "professional activities" now generally associated with the evaluation of hospital care. The indices used included the recording of diagnosis and treatment, the performance of rectal and vaginal examinations, the performance of certain laboratory examinations and the use of sedatives, stimulants and other medications subject to abuse. As is true of hospitals, the groups varied a great deal with respect to these indicators.
58. Myers, R.S. July, 1954. Hospital Statistics Don't Tell the Truth. *Modern Hospital* 83:53–54.
59. Even for hospital care the appropriate unit may include care before and after admission, as well as several hospital admissions.[3]
60. Cordero, A.L. April–June, 1964. The Determination of Medical Care Needs in Relation to a Concept of Minimal Adequate Care: An Evaluation of the Curative Outpatient Services in a Rural Health Center. *Medical Care* 2:95–103.
61. Butterworth, J.S., and E.H. Reppert, September 3, 1960. Auscultatory Acumen in the General Medical Population. *Journal of the American Medical Association* 174:32–34.
62. Evans, L.R., and J.R. Bybee. February, 1965. Evaluation of Student Skills in Physical Diagnosis. *Journal of Medical Education* 40:199–204.
63. Fattu, N.C., February, 1964. Experimental Studies of Problem Solving. *Journal of Medical Education* 39:212–225.
64. John, E.R. 1957. Contributions to the Study of the Problem Solving Process, *Psychological Monographs* 71.
65. Duncan, C.P., November, 1959. Recent Research in Human Problem Solving. *Psychological Bulletin* 56:397–429.
66. Fattu, N.A., E. Mech, and E. Kapos. 1954. Some Statistical Relationships between Selected Response Dimensions and Problem-Solving Proficiency. *Psychological Monographs* 68.
67. Stolurow, L.M., et al., Winter, 1955. The Efficient Course of Action in "Trouble Shooting" as a Joint Function of Probability and Cost. *Educational and Psychological Measurement* 15:462–477.

68. Ledley, R.S., and L.B. Lusted. July 3, 1959. Reasoning Foundations of Medical Diagnosis. *Science* 130:9–21.
69. Lusted, L.B., and W.R. Stahl. 1964. Conceptual Models of Diagnosis. In *The Diagnostic Process*, edited by J.A. Jacquez, 157–174. Ann Arbor: The University of Michigan Press.
70. Edwards, W., H. Lindman, and L.D. Phillips. 1965. Emerging Technologies for Making Decisions. In *New Directions in Psychology, II*, edited by T.M. Newcomb, 261–325. New York: Holt, Rinchart & Winston, Inc.

Acknowledgments: Included among the reviewed authors who read the manuscript and made corrections or comments are Georgopoulos, Makover, Morehead, Peterson, Riedel, Rosenstock, Rosenfeld, Sheps and Weinerman. The author is especially indebted to Dr. Mildred A. Morehead and to professors Basil S. Georgopoulos, Herbert E. Klarman and Charles A. Metzner for taking time to make extensive comments. The official critics, Mr. Sam Shapiro and Dr. Jonas N. Muller, were helpful in sharpening some of the issues in the assessment of quality. Since the author was unable to use all the excellent advice he received, he alone is responsible for defects in this paper.

This review has been supported, in part, by Grant CH-00108 from the Division of Community Health Services, United States Public Health Service.