Milbank Memorial Fund Quarterly/*Health and Society, Vol. 57, No. 1, 1979*

Measuring the Quality of Medical Care: Process Versus Outcome

WILLIAM E. MCAULIFFE

Department of Behavioral Sciences,
Harvard University School of Public Health

IN AUGUST 1974, the Institute of Medicine (IOM) of the National Academy of Sciences issued a policy statement, "Advancing the Quality of Health Care," which was prepared by a committee of experts on the topic. The Committee's first conclusion was (1974:2):

> *Quality should be measured by results.* There is great need in the gauging of quality to move beyond structure and process and toward the measurement of outcomes of care. The committee believes strongly that the goal of quality assurance can only be achieved by relating assessments of quality to the measurements of results.

But one Committee member, Mildred Morehead, dissented in an appendix (ibid., 53):

> There is overemphasis on outcome measurements and undue restriction on process evaluation . . . [which] will, in my opinion, have a deleterious effect on efforts to improve present medical practice on a nation-wide basis.

Morehead's dissent represents just one instance of a growing controversy concerning whether quality of care is measured better by focusing on the process of care (what is done) or the outcome (patient's health, disability, etc.). In one of the most authoritative

0026-3745-79-5701-0118-35/$01.00/0

discussions of quality assessment, Donabedian (1966: 168–169) granted that "Outcomes . . . remain the ultimate validators of the effectiveness and quality of medical care." But then he argued that process criteria "may, however, be more relevant to the question at hand: whether medicine is properly practiced . . . 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." A discussant disputed Donabedian's view, declaring that: "The overall social circumstances in which medical care is provided today requires concentration on . . . outcomes rather than process" (ibid., 205). Other arguments favoring an end-result approach and calls for its adoption were made from the beginning of the modern quality assessment movement by Codman (Lembcke, 1967: 112), and more recently by Shapiro (1967), Williamson (1970), Brook (1973), Jacobs, Christoffel, and Dixon (1976), and many others. Nevertheless, other observers (Ginzberg, 1975; Rosenberg, 1977) have joined with Morehead and Donabedian in favor of a process approach, and Brook has recently qualified his advocacy of an outcome approach (Brook, Davies-Avery, Greenfield et al., 1977). Thus, the desirability of process versus outcome assessment remains a major unsettled issue among experts on quality assurance.

Resolution of this controversy should be important for the success of quality regulation by the new Professional Standards Review Organizations (PSRO). To effect improvements in the quality of care, PSROs must be able to identify poor care, and have confidence in their assessments when imposing sanctions. The PSROs would be reluctant to act otherwise, as they would quickly lose the essential cooperation of physicians if quality assessments could not withstand challenge. Without grassroots support, effective regulation would be impossible. Valid measurement of quality is also a prerequisite for demonstrating the impact of regulation. So, establishing the validity of quality assessments is no mere academic issue.

This article examines relevant empirical evidence and the logic of major arguments relating to process versus outcome measurement. The arguments include assertions concerning practical data problems, impacts on medicine and the public interest, and measurement validity. Analysis reveals that outcome measures are not clearly superior: they are less direct than process measures, they have major practical problems, and their validity has rarely been

tested empirically. Although process measures have been studied more often than outcome measures, the extent of the validity and effectiveness of process assessments is also virtually unknown because the research methods used up to now have been inadequate. Thus, there is little reason for favoring outcome assessments over process.

Outcome Measures of Quality Care

The main argument for outcome measurement is simply that, since the goal of care is health, one should concentrate on measuring the achievement of health (Brook, 1974: 29; Palmer, 1976: 33; Thompson and Osborne, 1974: 808). Thus, McClure (1973: 334) and Brooke, Davies-Avery, Greenfield et al. (1977) have explained that an outcome approach would be more *direct* and skirt squabbles over whose process is most effective by letting the results speak for themselves. Inspecting outcomes would also insure attention to the cost effectiveness of care, which is a central concern of decision makers. Many authors have asserted that an outcome approach would therefore be superior (Schroeder and Donaldson, 1976; Osborne and Thompson, 1975: 627).

Also, few questions have even been raised concerning the measurement validity of assessing objective end results such as death, disease, or disability; these measures have been accepted on face value (Donabedian, 1969: 34). Even proponents of process assessment have often conceded the validity of outcome measures (Schroeder and Donaldson, 1976: 50), and have granted that process and structural elements are ultimately "validated" by their "correlation with outcomes" (Donabedian, 1966: 169; De Geyndt, 1970: 36). Another argument for outcome measures (Donabedian, 1969) is that the intangibles of care (e.g., a physician's judgment), which are seemingly difficult to measure directly with process techniques based on the medical record, are revealed in the patient's outcome.

Conceptual Arguments for Outcome Measurement

Close examination reveals major flaws in the logic of the arguments for the superiority of outcome measures. While the ultimate *goal* of most medical care and quality-of-care *regulation* is improved health, it does not follow that the quality of medical care in any particular

case can be defined by whether health was attained. The best attempts can fail, sometimes even in a majority of cases, whereas at other times patients routinely recover in spite of substandard treatment. No regulatory body can insist that patient outcomes be positive, nor do positive outcomes insure that care was appropriate or skillful.

The goal of quality *assessment* is not to produce health, at least not directly; it is to determine whether acceptable *care* was rendered. Presumably, if proper care is given, the best achievable outcome under the circumstances will result. The direct approach to assessment would be to observe care (the process) first hand (Donabedian 1978: 856–857). A *less* direct method of assessment would be to observe whether the patient had a good outcome *as the result of the process.* Unfortunately, it is often unclear whether the outcome was primarily a result of the process.

One reviewer has objected that to judge the quality of care by direct observation of process assumes that one knows which process results in the best outcomes, which he asserts is seldom the case. The necessary experimental evidence of efficacy is absent for most medical procedures. Consequently, he claims that one must examine the outcomes to determine whether proper care was rendered.

Assessing uncontrolled outcomes, however, is no solution for the absence of experimental evidence regarding process or structure. Randomized, controlled experimentation is desired in place of medical opinion for determining the efficacy of medical procedures precisely because the effects of factors affecting outcomes other than medical care (such as disease severity) must be eliminated before one can safely infer that outcome variation reflects the effects of care. The same extraneous factors are operative (and uncontrolled) in medical audit studies. If, in clinical research, the connection between process (or structure) and outcome is too ambiguous to infer causality unless experimental controls are employed, then what epistemological basis can there be in an uncontrolled audit study for inferring that an undesirable outcome resulted from inadequate care rather than from other factors? Thus, examining outcomes directly does not offer a way around the constraints imposed by the limits of medical knowledge.

In theory, the outcome variance associated with irrelevant factors could be eliminated statistically, but developing and testing satisfactory statistical models would probably not be much easier

than conducting randomized trials. Below, I shall discuss the many methods that have been proposed for refining outcome measures. Here, it is enough to point out that constructing a statistical model of outcomes that successfully identifies the variance in outcomes associated with the effects of care is, practically speaking, equivalent to making nonexperimental causal inferences between the process and outcome of care.

If quality of care does not always correlate with patient outcomes, then, one might ask, why bother assuring "quality?" The answer requires recognition that, even when a process of care is efficacious, patient outcomes may still have no correlation or even a negative correlation with process. This seeming paradox is explained by the distinction between a correlation in an experiment (indicating causality) and a correlation in a descriptive audit. Existence of a causal connection between process and outcome implies a significant correlation (although the correlation need not be *strong*) in a properly designed randomized, controlled trial; in an uncontrolled audit, that correlation can be completely obscured by other factors. Thus, assuring performance of efficacious care is desirable even if outcome measures lacking needed controls do not correlate with it in a medical audit.

In sum, the conceptual arguments for selecting outcome measurement over process prove to be rather weak when examined closely. While the goal of medical care is health, the achievement of health by any particular patient in uncontrolled conditions does not define or even necessarily indicate that the care received was acceptable. End results do not speak for themselves. Outcome assessment is not an adequate regulatory solution when medical knowledge is inadequate. On the basis of logic alone, care itself (the process) should be the prime object of quality-of-care measurement, but other considerations besides logic must be weighed before determining which type of measure would be best in any given situation.

Practical Obstacles to Outcome Assessment

Proponents of outcome measurement admit that its full adoption hinges on finding solutions to a number of practical problems (Institute of Medicine, 1974). Large samples are needed when evaluating rare outcomes; follow-up surveys for gathering data on post-hospitalization outcomes can be expensive; there may be a long

time lag between treatment and important final outcomes; setting standards for outcome measures may be difficult (McAuliffe, 1978a). Schroeder and Donaldson (1976) have described the difficulties one Health Maintenance Organization (HMO) encountered locating patients and judging outcomes when implementing an outcome-oriented quality assessment. What has not been immediately obvious is that these "practical" constraints cause *invalidity* in outcome assessments. Below, I shall explain why.

The Validity of Outcome Measures

Writers on quality assessment generally agree that outcome measures have validity, a form known as "face validity," but the basis for this conclusion can be questioned. Face validity simply appeals to one's intuitive judgment: "Does the measure seem valid?" Although it coincides with commonsense notions of validity, measurement experts take a dim view of face validation; they consider it as untrustworthy compared to empirically-based strategies. "Obviously" valid measures often fail to stand up under empirical testing (Cronbach, 1971: 453; Selltiz, Jahoda, Deutsch et al., 1963: 151), and *outcome measures have rarely been subjected to empirical validation.*

Yet what could be wrong with the observations of death or even disease or disability? Surely, they objectively measure what they purport to, and their "empirical validity" could be shown if one made the effort. Perhaps. But one must avoid a common misunderstanding here. The validity of an indicator may vary depending on which concept it seeks to measure. Even when death or disease measure *health status* validly, they may measure *quality of care* much less well. For example, the outcomes of care depend in part upon patient compliance, and research shows that *non*compliance occurs in a substantial percentage of cases (Wilson, 1973; Marston, 1970).[1] While many have recognized that end results have determinants in addition to quality of care, they have not recognized that *the variance associated with these other factors is systematic measurement error, a type of invalidity.* Thus, the validity of measures of quality of care based on

[1]The effects of patient compliance on outcome is a quality factor that is especially difficult to assess, since compliance depends, in part, on satisfactory performance by the medical team and in part on factors beyond its control.

health status or patient satisfaction is less than entirely obvious since they reflect extraneous factors to some extent.[2]

If outcome measures have less than perfect validity, exactly how strong is the likely connection between quality of care and outcome measures? As quality of care is a hypothetical construct, there is no way to determine the answer directly. But there are reasons to believe that nonquality determinants of outcomes could be substantial (see McAuliffe, 1978a, for numerous examples of questionable outcome measures), and so the validity of outcome measures cannot be taken for granted. Contrary to current practice, *outcome measures must be empirically validated just as process measures must,* for outcome measures of quality are not obviously valid.

Because experts in quality of care assessment have almost always taken "validation" to mean "correlation with outcomes," many readers may still have difficulty understanding how outcome measures of quality of care could be invalid, or how outcome measures might be validated empirically. This difficulty is just one of the many reasons for believing that the definition, "correlation with outcomes," is too narrow for validation of measures of quality, be they structural, process, or outcome (see McAuliffe, 1978b, for a detailed discussion). There is a broader theory of measurement validity, developed primarily by psychologists, which offers an analytical framework appropriate for assessing the validity of outcome measures.

According to psychometric measurement theory, validity is defined as the amount of correspondence between a concept (such as quality of care) and a measure (such as an outcome index). The measure is valid insofar as it is *pure* (excludes extraneous factors), *complete* (covers all relevant aspects), and *representative* (has the proper balance or mix of relevant aspects). Validity is expressed quantitatively as the proportion of a measure's variance that is associated with the concept of interest (Cronbach, 1971; Nunnally, 1978; Kerlinger, 1965). The remaining variance represents either systematic or random measurement error.

---

[2]The most forceful arguments for outcome assessments of quality have come from economists who are typically concerned with evaluating the performance of industrial organizations that theoretically should have a high degree of control over the quality of their output. But one cannot uncritically generalize these arguments to measuring hospital performance, since hospitals have much less control over outcomes.

Extraneous Outcome Variance. Although few investigators have explicitly evaluated the validity of outcome measures of quality of care, there have been some important exceptions. Roemer, Moustafa, and Hopkins (1968) noted that crude hospital mortality rates have been viewed with much skepticism as measures of quality because patient characteristics—their diagnosis, severity of illness, and general health status—may vary greatly from one hospital to the next, and patient-mix differences may be more important than quality-of-care differences in determining mortality rates. The skepticism appeared well founded since Roemer et al. found that crude death rates were higher in teaching hospitals than in nonteaching hospitals, higher in accredited hospitals than in nonaccredited hospitals, and higher in more technologically sophisticated hospitals than in less technologically sophisticated hospitals. Goss and Reed (1974) have reported similar results. Roemer et al. asserted that hardly anyone would suggest that these results mean that the quality of care was superior in the nonteaching, nonaccredited, or less technologically sophisticated hospitals. A more probable explanation is that the crude death rate had low validity as an indicator of hospital quality.

Other studies have verified that substantial proportions of the variance in mortality are associated with factors other than hospital quality. In a study of surgical mortality rates as an outgrowth of the National Halothane Study (Bunker, Forrest, Mosteller et al., 1969), Moses and Mosteller (1968) showed that much of the variance in rates was associated with nonquality factors. Standardization for patient differences, type of operation, and patient physical status explained 24.3%, 68.6% and 40.8% respectively of the variance in mortality rates among the study's 34 hospitals (calculated from Table 5, Bunker et al., 1969: 196). A composite of type of operation and physical status explained 74.4% of the variance. Moreover, the authors could not determine the precise proportion of variance attributable to quality (that is, they could not show that any of the outcome variance was valid) because they had no independent measures of quality of care (e.g., process measures). So, although these findings do not prove conclusively that uncontrolled death rates are largely invalid as measures of quality of care, they are consistent with doubts raised about crude mortality rates as such measures.

Following up the Moses-Mosteller inquiry, the Stanford Center for Health Care Research (Flood, Scott, Ewy et al., 1977; Scott,

Forrest, and Brown 1976) undertook another study of hospital differences in postoperative patient mortality and morbidity. Recognizing the need to remove the effects of extraneous variables, the researchers statistically adjusted the outcomes for differences due to stage of disease (severity), patient's age, sex, physical status, cardiovascular status, and whether the surgery was elective or emergency. The percentage of variance accounted for by those controls varied by diagnosis from a low of 2% to a high of 44%. Flood et al. (1977) then introduced measures of quality inputs, including hospital characteristics (size, teaching status, and expenditures) and surgeon characteristics (specialization, certification, number of residencies, etc.). At best, all surgeon and hospital characteristics combined accounted for no more than a total of 1% of the outcome variance, even though the variance due to patient characteristics had already been removed. These results, should they be confirmed by subsequent research, raise serious questions concerning the validity of existing outcome measures of quality of care and the current viability of outcome approaches to quality assessment.

Finally, Martini, Allan, Davison et al. (1977) have analyzed the percentages of British regional variations in rates of mortality, complications, and morbidity that are explained by socio-demographic factors (e.g., age, socioeconomic status) and the structure of regional medical care systems (e.g., expenditures, percentage of care occurring in teaching hospitals). The authors concluded that "indexes constructed from the traditional outcome measures are more sensitive to sociodemographic circumstances . . . than to the amount of medical care provided and/or available" (ibid., 306). Although the study's focus was not quality of care in hospitals, its sample was small (15), and the quality of inputs and process was measured only crudely, the consistency of its results with those of the other studies already reviewed nevertheless helps build a case against the uncritical acceptance of outcome measures of quality of care.

Data Quality. Outcome measures can also be impure, and therefore invalid, as a result of random rather than systematic measurement error. In psychometric theory, random measurement error is defined as unreliability, which in turn sets a ceiling for validity (Nunnally, 1978): to the extent that a measure is unreliable, it is invalid.

Although outcome assessments based on mortality are frequently preferred on the grounds that they are more objective (objectivity is one component of reliability), the other aspects of health status (e.g., symptoms, functional level) are less objective. Brook (1973) found that physicians often disagreed when judging patients' outcomes from follow-up interview data. The average correlation among the 10 judges was 0.61 (Brook, 1973: 38, my calculation).

Random measurement errors can contaminate outcome measurements in other ways as well. Patients' physical condition or subjective reports of symptoms may fluctuate from one day to the next, pathology laboratory reports may be in error (Donabedian, 1969: 28–29), physiological measures, such as urine cultures or blood pressure readings, are sometimes in error (Maskell and Pead, 1976; Labarthe, Hawkins, and Remington, 1973), outcome data in medical records may be incomplete (Fessel and Van Brunt, 1972, Table 3), and errors may be made in the process of abstracting data from the charts.

Linn, Linn, Greenwald et al. (1974) correlated assessments of outcome (13 categories of "impairment") based on record review with comparable assessments made by the patients' attending physicians at discharge. The 13 correlations ranged from 0.19 to 0.66, with a median of 0.46. The disagreements in assessment were not entirely due to poor record-keeping, however, since the attending physicians' assessments were shown to be somewhat unreliable, and the medical-record-based assessments predicted death at follow-up slightly more accurately.

Incomplete Outcome Measures. Another potential source of invalidity in outcome measures is the incompleteness of assessments based on only some of the relevant effects of medical care. This incompleteness is in part the methodological upshot of the "practical" obstacles to gathering data on long-term and other difficult-to-observe effects of care. Since much medical care is directed toward outcomes occurring after discharge, large components of quality could remain unassessed if one were to rely solely on "intermediate outcomes" from inpatient medical records.

Outcome measures are also often not sensitive to many diagnostic aspects of care (McAuliffe, 1978b). Since most medical audit studies sample cases by diagnosis, they often exclude from consideration patients incorrectly diagnosed as a result of inadequate

process (see Greenfield, Nadler, Morgan et al., 1977, for a study of such a sample). If such patients are discharged and have a poor outcome as a result of not receiving needed care, their cases could easily be overlooked because the patients ended up in different hospitals. Furthermore, many diagnostic procedures are designed for detecting special (but often rare) management problems (such as, allergic drug reactions). Failure to perform these essential procedures for *all* patients will affect *only* the outcomes of patients having the problem. Often there will be no such patients in samples as small as the usual 50 cases examined in medical care evaluation studies, and if so, an outcome assessment would fail to reflect important diagnostic inadequacies in care (see McAuliffe, 1978b, for examples). In general, if a medical process includes medically warranted procedures whose effects are unrepresented in the study's outcome data, the outcome data are incomplete and therefore somewhat invalid as a measure of quality of care.

Proponents of outcome measurement might nevertheless counter that outcome measures are still more complete than process or structural measures, because most relevant components of care—including unrecorded aspects of surgical or medical care, as well as the performance of other segments of the medical care system—affect patients' outcomes. But the concept "outcome" is itself a broad and complex construct, and if few extant outcome measures cover its domain satisfactorily, then outcome measures may not be more complete than process measures. For example, death rates often may not detect differences in care as higher levels of performance and skill are achieved, or where a disease is rarely life-threatening. Consequently, data on mortality should be supplemented by data on morbidity, functional status, subjective distress, and so on, if the outcome assessment is to approach completeness.

How differences in completeness affect outcome measures is illustrated by the Stanford study described earlier (Scott, Forrest, and Brown, 1976), which employed five measures of outcome reflecting different combinations of data on mortality, severe morbidity, moderate morbidity, and postoperative complications (see Table 1). Measure 1 (death) and Measure 5 (death or incomplete return to function) showed no significant differences between hospitals, whereas the other three outcome measures (death or severe morbidity; death or moderate or severe morbidity; death or monitors or

catheters) resulted in significant interhospital differences. Thus, the conclusions one draws regarding quality of care would depend upon the outcome measure one chose to examine.

A deeper understanding of why the results varied from one measure to another can be gained by examining the correlations presented in Table 1. For the correlations above the main diagonal, the hospital outcome rates were standardized (statistically adjusted) for the patient's age, sex, physical status, cardiovascular status, stage of disease and type of operation; for the correlations below the diagonal, the rates were subjected to an additional statistical (Bayesian) adjustment for differential reliability (due to different numbers of cases at the hospitals). The latter measures were used in the study's analyses.

Measure 4 (death or severe or moderate morbidity) and Measure 5 (death or incomplete return to function) are apparently most complete, but they do not correlate significantly (n = 17) with any of the other measures or with each other. In fact, there are a number of negative correlations, and all of the correlations are small. The less complete measures, which focus only on death (Measure 1) or severe morbidity (Measures 2 and 3), have more respectable correlations among themselves, but even those correlations may be smaller than many might have assumed. Thus, differences in the completeness of outcome indexes, even when the

TABLE 1
Correlations Among Standardized and Bayes-Adjusted Outcome Measures Used in the Stanford Study

Outcome Measures	1	2	3	4	5
1. Death	—	0.85*	0.46*	-0.19	0.21
2. Death or severe morbidity	0.36	—	0.40	0.07	0.09
3. Death or catheters or monitors	0.23	0.41*	—	0.05	-0.08
4. Death or severe or moderate morbidity	-0.22	0.24	0.05	—	-0.23
5. Death or incomplete return to function	—	—	—	—	—

*Correlations are significant at $p<0.05$.
Source: Stanford Center for Health Care Research (1974). See Tables 17 and 18, pp. 169-170.
Note: Correlations among the standardized outcome measures are above the diagonal. Correlations below the diagonal are among the standardized outcomes after an additional adjustment for differential reliability using a Bayesian technique described in the study. Bayes-adjusted correlations with Measure 5 were not reported in the study.

indexes draw upon overlapping parts of the same data base, can have profound effects on the picture the indexes convey regarding quality of care.

Three other quality-of-care studies have also reported correlations among alternative outcome measures. Romm, Hulka, and Mayo (1976) alternately used activity levels and subjective symptoms as the outcome (dependent variables) in parallel regression analyses of the process of care for congestive heart failure. Although the two outcome measures correlated reasonably well,[3] they were sufficiently different so that the results from the two regressions led to somewhat different conclusions regarding the relationship between process and outcome. Other outcomes included in the study were patient satisfaction, compliance, and knowledge; only patient satisfaction correlated significantly with activity levels (0.25) and symptoms (0.26). In a study of the organizational determinants of quality of care in 42 hospitals, Shortell, Becker and Neuhauser (1977) collected data on six measures of hospital quality: 1) the medical-surgical death rate; 2) the postoperative complication rate; 3) Medicare patients' death rate; 4) the match between pathologists' reports and preoperative diagnoses for each hospital's last 50 appendectomy and 50 cholecystectomy patients, 5) the percentage of single-unit blood transfusions, and 6) outside expert ratings of the hospitals. There was "little intercorrelation among these measures" (Shortell et al., 1977, footnote 8), with the strongest correlation (0.30; nonsignificant) being between Measures 2 and 5. The authors nevertheless concluded that the medical surgical death rate and the postoperative complication rate "best reflected hospital-wide activities," and employed the two measures as alternative dependent variables in regression analyses. That some of the results of the regressions conflicted is not surprising, since the two dependent variables correlated only 0.09 with each other (Shortell, 1978). Finally, Brook (1973) studied the correlations among a number of different process and outcome measures. Most of the process-

[3]This zero-order correlation was 0.70, but it overstates the measures' "convergent" validity (Nunnally, 1978) because the correlation reflects pre-existing patient differences as well as the effects of care. A second-order partial correlation controlling simultaneously for both initial symptoms and initial activity levels would be more appropriate. Dr. Romm has furnished me with the necessary correlations, and I calculated the second-order partial correlation to be 0.58.

outcome correlations were low, but the correlations among the outcome measures (death, subjective symptoms, activity levels, and physiological evidence) were even lower on average (McAuliffe, 1978b). For example, assessments of quality of outcome based separately on subjective symptoms and on activity levels correlated 0.19 with each other, and correlated 0.04 and 0.01, respectively, with physiological evidence of disease (McAuliffe, 1978b).

The results of these four studies thus furnish additional weight against the uncritical acceptance of the validity of outcome measures of quality of care. It is likely that up to now researchers have employed obviously incomplete outcome measures on the assumption that the different dimensions of outcome (death, disease, disability, etc.) were highly correlated, and therefore the omitted data would be mostly redundant. The results of the studies just reviewed show that such an assumption is probably unwarranted in many cases.

In passing, it is important to point out some of the implications of these findings for research on quality assessment. Clearly, different measures of outcome are not interchangeable, and consequently researchers should develop a rationale for selecting an indicator (single or composite index) appropriate to their pursuits. Results based on a study of one outcome measure need not generalize to studies using other measures, and therefore findings should be described in terms appropriate to the specific type of measure. If multiple measures are employed in composite indices, one would be wise to examine profiles of individual measures as well. Finally, if different outcome measures do not correlate highly with each other, it is impossible for any structural or process measure to correlate highly with all of the outcome measures. For example, if a process measure that covered the quality domain rather completely were correlated separately with each one of a set of incomplete outcome measures, it might correlate only weakly with any one of them (for an example, see McAuliffe, 1978b).

Techniques for Increasing the Validity of Outcome Measures

Since most of the problems with outcome measures have long been recognized even if not labelled as invalidity, over the years many techniques have been proposed for improving outcome measures.

The techniques include statistical adjustments (e.g., age-adjusted mortality rates, multiple regression), examining *patterns* of care because they are more reliable than individual cases (Jacobs and Jacobs, 1974: 46), judgmentally discounting unpreventable poor outcomes (ibid., 40), using statistically-derived standards (cut-offs) for acceptable outcome rates (McAuliffe, 1978a), and focusing on "tracers" (Kessner, Kalk, and Singer, 1973) or "sentinel" outcomes (Rutstein, Berenberg, Chalmers et al. 1976) that are known to be relatively "pure" measures of quality. Each of these techniques seeks in its own way to maximize the valid proportion of the outcome variance.

However promising the techniques may be, none has yet been shown to be both practical and effective. For example, application of advanced methods of statistical adjustments such as those employed in the Stanford study (Scott et al., 1976) requires considerable expertise that is not widely available, may demand elaborate data collection efforts, and has not yet been proven to be effective. Use of these techniques does not guarantee that the resulting measure will possess acceptable validity; the final outcome assessments must still be validated. Because this point is so important, but routinely missed, I shall describe a specific instance.

In the study of hospital death rates mentioned above, Roemer et al. (1968) hypothesized that validity might be increased if the rates were adjusted for case severity. Because ideal adjustments would require collecting extensive data on diagnosis and disease severity, Roemer et al. chose instead to adjust the death rates for occupancy-corrected length of stay, which the authors considered a practical "approximate measure" of case severity. However, make-do or proxy measures usually reduce the effectiveness of statistical controls, and therefore the index's validity was still in doubt. Roemer et al. compared their index to measures of hospital technological adequacy, Joint Commission on Accreditation of Hospitals (JCAH) accreditation, and voluntary versus proprietary status. For most comparisons, but not all, the statistical adjustment successfully reversed the previous relationships between these structural measures and the uncorrected death rate, and thus appeared to increase validity. But Goss and Reed (1974) were unable to replicate Roemer et al.'s findings on a sample of 97 hospitals. Goss and Reed argued that length-of-stay adjusted death rates, like crude death rates, have doubtful validity and need further refinement.

The Reciprocal Validation of Structure, Process, and Outcomes

How could outcomes be such poor measures of quality if they are the "ultimate validators" of process and structural criteria? First of all, as explained earlier, one cannot assume that an outcome measure would necessarily be as valid in an uncontrolled audit study as it would be in a randomized, controlled experiment designed to assess the efficacy of structure or process.

It is also incorrect to assume that a measure employed to validate (in the measurement sense) another indicator is necessarily superior in validity. In fact, whenever a new, more refined measure is developed, its initial validation usually includes comparison with existing and accepted, but ultimately less valid, measures of the concept.

Finally, while outcome measures are used to validate structural and process criteria, the reverse is also true, as was shown in studies by Roemer et al. (1968) and by Kisch and Reeder (1969).[4] Using a measure as a validator assumes validity but does not convey it: structure, process, and outcomes can validate each other only because theoretically each can be assumed to possess some validity. If, let us say, process and outcome measures do agree in a properly designed study, then our faith in the validity of both is strengthened. Should they fail to agree, other information (e.g., their respective correlations with structural indices) is needed to interpret the failure. Again, which measure was formally designated as the "validator" means nothing by itself.

It should now be clear that claims of validity for outcome measurement on the grounds of greater objectivity and completeness were not solidly based. Whatever "obvious" validity outcome measures seem to possess fades when examined closely, for there are many ways outcome measures could have low validity.

---

[4]For an application of this principle when setting standards for monitoring outcome profiles, see McAuliffe (1978b). Moreover, most outcome methods of quality assessment judge a poor outcome as indicative of inadequate care *only* after a subsequent process audit has established that the outcome followed some irregularity in care. Consequently, the ultimate test of quality appears to be assessment by both process and outcome.

Are Outcome Measures Clearly Superior After All?

The advocacy of outcome measures was based almost entirely on a theoretical, rather than empirical, analysis of the measures, and the main propositions in the argument have now been examined in detail. In response to the contention that outcomes are the proper object for quality assessment because the goal of medical care is health, I have argued that the quality of care is more directly gauged by focusing on medical care, the process; outcomes are less direct manifestations of quality. Outcome measures do not possess the face validity claimed for them, because it is apparent that outcomes usually reflect more than just the effects of care, often do not include many relevant effects of care, and are based on data which are poor in quality. Moreover, validity cannot be assumed for outcome measures just because they serve to validate structural and process criteria, since the reverse is also true.

At present, there is also little empirical evidence that demonstrates the high validity of outcome measures that proponents have assumed. Because the validity of outcome measures had typically been taken for granted, few studies sought to provide the necessary empirical confirmation. Examination of limited, existing data has shown that doubts about the validity of outcome measures may be well founded. Factors unrelated to the medical care system accounted for substantial proportions of outcome variance, far more, in fact, than did measures of medical inputs. In addition, alternate measures of outcome often failed to correlate with each other. Various statistical techniques, such as multiple regression, have been proposed as possible solutions to these problems, but none has yet been clearly demonstrated to be effective or practical. Clearly, further research is needed before the validity of operational outcome measures can be decided. Thus, outcome measures are not demonstrably superior to other types of measures.

Process Measurement

If outcomes are not clearly best, how do they compare to process measures, which have long been under attack? In this section I examine the relevant arguments and evidence, and finding previous interpretations either incorrect or overstated, I conclude that process measures are at least as promising as outcome measures.

Criticisms of Process Measurement

After decades of using structural criteria and a brief period of flirting with the idea of a process-oriented audit, the JCAH recently adopted an outcome-oriented system for measuring quality of care. According to Jacobs and Jacobs (1974: 32), who designed JCAH's outcome audit, they passed over process-auditing because of the following reasons:

> . . . [It] is *cumbersome;* a list of the processes of care for all but the simplest diagnoses can include many dozens of items, and each of these must be checked off for each chart reviewed. . . . The relationships of many health care processes to desired health care results is *questionable* or, at best, *unverified by empirical evidence.* . . . The uncritical use of process measures runs the danger of penalizing practitioners who obtain satisfactory patient outcomes by routes other than those prescribed by process criteria, thus *stifling innovations* in treatment. In response, practitioners may order tests and procedures to satisfy criteria lists, rather than on the basis of their best clinical judgement, thereby *increasing the use of ancillary services.* (My italics)

Process-auditing from medical records or abstracts has also been criticized because the data are often incorrect or incomplete (Zuckerman, Starfield, Hochreiter et al., 1966), and therefore the data fail to reflect what actually happened to the patient. Kelman (1976) also questioned whether recording "Done/Not Done" for various procedures does not overlook the intangible "true quality" or skill facets of medical care. Just because a procedure was performed does not mean it was done well. Brook, Appel, Avery et al. (1976: 17) pointed out that process-auditing based on medical records would also routinely miss psychosocial aspects of care (patient satisfaction). Finally, current audits focus on physician or nurse performance only and ignore the various other aspects of patient care.

To summarize, process measurement has been criticized as: 1) *impractical* because criteria are difficult to develop and cumbersome to apply; 2) *undesirable* in its impact on innovations and medical costs; and 3) *invalid* since it covers limited aspects of care, many processes have not been proven effective, and data sources contain errors.

Analysis of the Criticisms

Practical Problems in Process-Auditing. Although applying many process criteria may seem to be more trouble than applying a few outcome criteria, the advantage to outcome measures would hold only as long as essential outcome data were readily obtainable from medical records. But outcome data in medical records are often quite limited, and if follow-up surveys are needed to fill the gap, then an outcome approach could be as much trouble as process-auditing. In fact, even proponents of outcome measurement (Starfield, 1974) recommend process assessment as more convenient when outcomes are long-term (e.g., immunizations; for other examples see Rosenberg, 1977: 1936). Added to the cost of collecting outcome data are the difficulties of performing the sophisticated statistical analyses outcome measures seem to require.

Outcome systems typically escape the burdens of collecting follow-up survey data by sacrificing the completeness of their assessments, and process assessments permit similar trade-offs. The cumbersome aspects can be reduced by focusing on only key process criteria, ideally the criteria that most clearly differentiate between adequate and inadequate care (see Richardson, 1972, for such an example).

Developing process criteria may currently consume large amounts of audit committees' energies, but it is likely that the committees concentrate on process criteria more because of the abilities and interests of their members than because of inherent differences in the types of measure. Audit committee members are medical personnel who are more interested in and better trained for evaluating the predominantly medical issues raised when developing process criteria than for evaluating the statistical issues more commonly raised by the selection of outcome criteria. If the committees included more statistically-oriented measurement experts, and if the validity of outcome measures received the amount of attention it deserved, then developing outcome measures could easily require as much time and effort as is currently spent on process criteria.

It is also noteworthy that outcome-oriented methods, such as the JCAH's, require process assessment to determine why a patient's outcome was unsatisfactory and how care should be improved, and so these methods require developing and applying both process and outcome criteria. In principle, at least, if the outcome criteria are

complete, *developing* process criteria for verifying unfavorable outcomes should be just as difficult as it would be for a normal process audit.

Impacts on Innovation and Medical Efficiency. Process-auditing might hinder true innovation and lead to "defensive medicine" (e.g., ordering unnecessary laboratory tests), but the extent would hinge on how rigidly audit committees adhere to prescribed criteria and standards, and apply sanctions. Current trends are clearly toward flexible criteria, numerous reviews by peers before finding fault, and many opportunities for appeal. Yet some infringements on the freedom of clinicians is inevitable in any system of regulating care, regardless of the method of assessment. Even in the JCAH's outcome system, physicians must be prepared to justify decisions that deviate from standard practice whenever outcomes are poor. It is therefore hard to see why a system based on process assessment would be more stifling than one that combined outcome and process.[5]

Validity: Data Quality. Ultimately, problems of data quality may profoundly affect how quality of care assessments are conducted. Researchers performing retrospective process audits have found that *both* process and outcome data are incompletely recorded (e.g., Lindsay, Hermans, Nobrega et al., 1976; Fessel and Van Brunt, 1972; Zuckerman, Starfield, Hochreiter et al., 1966), and the missing data no doubt reduce the validity of the measures, especially if missing data sometimes reflect negative findings and other times reflect noncompliance with criteria. However, Roos, Henteleff, and Roos (1977: 3) argue that validation studies have demonstrated that medical records are more accurate than responses to questionnaire surveys (one important source of outcome data), and since physicians now know that records are subject to review, their records should become more complete. Also, medical records could be improved by standardizing recording formats. There is a danger, however, that a heightened awareness of the role of medical records in regulation could result in instances of falsification; different data sources for both process and outcome may eventually be needed in special cases.

---

[5]Of course, the threat of malpractice suits has probably already caused far more defensive medicine than any system of peer review could.

At present, it is difficult to say precisely to what extent low data quality affects the validity of process-auditing because there are no completely adequate studies. For example, Zuckerman et al. (1966) have documented the incompleteness of medical record data by comparing the content of medical record data with audio tape recordings of the same patient-physician encounter. But to estimate the effect of the missing information on the validity of process assessments, one would have to go one step further than Zuckerman et al. by process-auditing the medical records and the tapes separately, and then correlating the two independent assessments.

Validity: The Limits of Medical Record Data. The validity of process assessment has been questioned because it concentrates on physicians' technical performance, which is just one component of care, and measures even that component crudely, since it counts merely what is done (e.g., a surgical procedure) but not how well.

These appear to be rather serious shortcomings, but without further study it is difficult to estimate the comparative disadvantage or its importance to regulators. We must remember that outcomes are also far from perfect as measures of high-level medical skill. Furthermore, federal quality regulation tends to be concerned with determining whether the *minimum* rather than highest standards of care have been met, and therefore the upper ranges of medical skill are probably beyond regulators' range of interest. And so, performance or nonperformance of essential procedures may represent the lion's share of what regulators want to know. In any case, the "intangibles of care" actually can be measured indirectly by process assessment, just as they are by outcome measurements.

Indirect measurement is achieved as long as the characteristics being measured correlate or overlap with those unmeasured, and there is evidence that the elements of good care (including taking adequate histories, ordering necessary tests, as well as the "intangibles") do correlate with one another (Peterson, 1956: 19; Lyons and Payne, 1974; Rosenfeld, 1957: 862). So, if a physician orders the correct diagnostic tests, and if ordering correlates with skillfulness and thoroughness in taking a history, then the odds are that he or she will have taken a good history; failing to observe the history-taking session may therefore cause little harm. Usually, pair-wise correlations between process criteria are modest, but the multiple correlation between a single criterion and the usual large number of

other criteria in a process composite may be quite high—so high that including that one criterion (such as a measure of skill in history-taking) in the composite might add virtually no new information (see Richardson, 1972, for a demonstration). Thus, process measures are theoretically just as capable of measuring the unmeasurable as are outcome measures. Whether in practice either type of measure validly reflects these aspects of care is unknown.

If current process data prove to be too skimpy for some aspects of care, modifications in assessment methods may be necessary. For example, process data in the medical record might be too insensitive to the aspects of care that produce high rates of postoperative infection. In those instances, either the process data source could be improved (by direct observation, improved recording of relevant data, for example), or outcome measures added. Cost and relative validity would dictate the choice.

Validity: Correlations with Outcomes. The chief charge against process assessment is that it lacks validity because many procedures are ineffective, as shown by studies that have failed to find strong correlations between process and outcomes. I have recently reviewed nine published studies of process-outcome correlations, but found little to support the claim that process-auditing is generally invalid (McAuliffe, 1978b). The review is summarized in Table 2. Three studies reported nonsignificant correlations, two had mixed results, and four reported significantly positive correlations. But drawing conclusions from these results is difficult because the studies had serious methodological flaws (discussed in detail in McAuliffe, 1978b) which either made obtaining positive correlations difficult or otherwise left unclear the meaning of a nonsignificant correlation.

The most obvious shortcoming of the studies was their outcome measurement. Since a correlation between two measures depends on the strengths and weaknesses of both, one must first rule out the possibility that the outcome measure is invalid before one can safely infer that the process measure is at fault when a correlation is low. But, I found numerous apparent shortcomings in the specific outcome measures employed in the studies, and therefore the nonsignificant correlations could have been entirely attributable to those weaknesses.

The studies were also improperly designed for the purpose of validation. To obtain a correlation, for example, one must have

TABLE 2
Summary of Studies of Process-Outcome Correlations in Medical Units

Authors	Sample	Process Measures	Outcome Measures	Statistical Technique	Reported Results	Evaluation
Brook, 1973	107 urinary infection, 114 hypertensives, 75 ulcer patients	Implicit; 2 explicit composites	Specific outcomes: death, activity, symptoms, physiologic evidence; implicit judgments; composite index	Correlation	All process measures correlated significantly with implicit outcomes; outcome index not related to explicit process; mixed results for implicit process data and specific outcomes	Process validity underestimated; extraneous causes of outcomes uncontrolled; components of outcome index had questionable validity; patient compliance not taken into account in outcome measures
Fessel and Van Brunt, 1972	I: 50 appendix operations each from 3 hospitals	I: "Empirical," diagnostic	I: Rate of confirmed appendicitis	I: No statistical measures used	I: Qualitative judgment of no correlation	Results of all 3 studies marred by numerous faults in methods
	II: 50 discharged alive after myocardial infarction	II: % of 44 empirical, diagnostic	II: 5 post-hospital outcomes	II: Significance test, between favorable and unfavorable outcome categories	II: No difference in process scores for any favorable vs. unfavorable outcome	

TABLE 2 *(Continued)*

Authors	Sample	Process Measures	Outcome Measures	Statistical Technique	Reported Results	Evaluation
Fessel and Van Brunt, 1972	III: 50 alive after myocardial infarction versus 50 dead	II: 26 empirical	III: Death	III: Significance test, between alive and dead groups for each process criterion	III: No differences for relevant criteria	
Greenfield et al., 1977	137 patients with chest pain who were discharged from emergency room	"Criteria map" on decision to discharge	Death or rehospitalization at 21 day follow-up	Fisher's Exact Test	Significant correlation ($r=0.31$, my calculation) between process and outcome	Study's topic narrow, but one ignored by other audits
Kane et al., 1977	I: 251 acutely ill patients (7 diagnoses)	I: Explicit (5 to 35 criteria, depending on diagnosis)	Composite index of death, activity level, subjective symptoms, and physiologic data; good outcome,	I: Compared process scores for good versus bad outcomes	I: Good outcomes in 5 diagnoses had higher average process scores; overall significance was not tested	Controlled for general health status only
	II: 162 acutely ill patients	II: implicit assessment	defined as final health status at least equal to status before illness; patient satisfaction with care and outcome	II: Chi-square	II: Significantly better process scores associated with good outcomes. Patients were slightly more satisfied with good outcomes but not with good care	

TABLE 2 *(Continued)*

Authors	Sample	Process Measures	Outcome Measures	Statistical Technique	Reported Results	Evaluation
Lindsay et al., 1976	42 cystitis	10 explicit	4 measures of "recurrence"	Chi-square for each process criterion by outcome	No relationship between process and recurrence except "notation of prior infection"	Diagnostic criteria often relevant to only excluded patients; no variation in therapy criterion; "outcomes" could be new illness episodes; acute outcome not studied; inappropriate one-criterion-at-a-time analysis
Nobrega et al., 1977	138 hypertensive	% of 89 explicit criteria (79 diagnostic)	Blood pressure at follow-up	Correlation; multiple regression	Neither correlations nor regressions significant	Failed to control for initial status in correlation; specification error in regression analysis; outcomes irrelevant to diagnostic criteria
Romm et al., 1976	122 congestive heart failure	"MD awareness," "communication," "drug error," "management," "satisfaction," "utilization"	Activity levels; subjective symptoms	Correlation; step-wise multiple regression	Correlations mixed: "satisfaction" and "MD awareness" significant regressors for both outcomes	Failed to control for initial status in correlations; regression analysis incorrectly controlled other outcome measures and did not include key process measures in one analysis

TABLE 2 *(Continued)*

Authors	Sample	Process Measures	Outcome Measures	Statistical Technique	Reported Results	Evaluation
Rubenstein et al., 1977	Patients presenting with symptoms of urinary infection (approximately) 126)	Weighted composite process	Index including symptoms, satisfaction, understanding, and compliance	Correlation	Significant correlation of 0.38	Outcome index may be too broad; sample selection superior; subjective outcome defined to be relevant to diagnostic process criteria
Starfield and Scheff, 1972	46 anemic children	Iron therapy and follow-up	Adequate hemoglobins	Chi-square	Significant correlation (0.33) between combined processes and outcome	Specificity of process-outcome relationship is key feature; No controls for initial severity

Source: Adapted from McAuliffe, W.E. 1978. Studies of Process-Outcome Correlations in Medical Care Evaluations: A Critique. *Medical Care* 16(11): 907-930.

variation in both variables, but in these audit studies there was often little or no variation in either process, or outcomes, or both. Consequently, if the designs of some of the studies were improved, they might have completely different results, as my analysis showed that the best designed studies were those reporting positive correlations.

Impact and Efficiency of Health Care. A criticism related to the validity question is that "regulation based on process data is likely to . . . increase the cost of medical care . . . but is unlikely to improve the component of health under control of the medical system" (Brook, 1973: 57). Brook reasoned that many procedures included in criteria sets are ineffective, a point supposedly confirmed in his own study by the "weak" and often nonsignificant correlations between composite process scores and outcomes. Requiring the procedures would therefore increase costs without improving health.

Although it is surely true that the cost effectiveness of a medical care regulated by process methods would be maximized by focusing measurement on only effective medical procedures, Brook's conclusion can be questioned. Effectiveness should not be inferred from correlations, but from regression coefficients (or their analog, percentage differences). To show how different the resulting conclusions can be, I have presented in Table 3 selected data from Brook's own studies.[6] All the correlations are significant but "weak," according to Brook (1973: 57). Nevertheless, when the process was judged adequate, the proportion of satisfactory outcomes increased *substantially*. The smallest percentage difference was 25.4%. As discussed earlier, inadequate controls prevent one from concluding that improving the process of care would definitely result in such large apparent improvements in patients' health. Still, even these data do not support the claim that regulation based on process-auditing would have little impact on health.

Evaluating the Case Against Process Assessment

The arguments against process measurement are not persuasive. Process-auditing currently involves more data elements than does

[6] I selected what I consider the most valid of his data; his other outcome measures show less impact. See McAuliffe (1978b) for the basis of my evaluation of his measures.

outcome assessment, and thereby process audits seem to be more trouble than outcome audits. But eventually, equivalent amounts of effort may be needed for outcome measurement, if follow-up surveys and sophisticated statistical analysis are required to bolster outcome validity. Some observers have charged that process-auditing will

TABLE 3
Impact of Therapeutic Process on Outcome at Follow-Up

Diagnosis	Outcome Measures	Process Measures		Percentage Difference
		Adequate	Inadequate	
	Implicit Outcome Judgment	Implicit Process Judgment		
Three diagnoses combined*		(n=69)	(n=227)	
	Satisfactory	89.9%	55.1%	34.8
	Unsatisfactory	10.1	44.9	
	$\chi^2=27.5$, $p<0.001$; $r=0.30$			
	Specific Outcomes			
Urinary tract infection*		(n=13)	(n=93)	
	Negative culture	92.3%	64.5%	27.8
	Positive culture	7.7	35.5	
	$\chi^2=4.05$, $p<0.05$; $r=0.20$			
Hypertension*		(n=31)	(n=82)	
	Controlled blood pressure	74.2%	48.8%	25.4
	Uncontrolled blood pressure	25.8	51.2	
	$\chi^2=5.89$, $p<0.05$; $r=0.23$			
		Selected Explicit Process		
Ulcer*		(n=28)	(n=46)	
	Asymptomatic	64.3%	23.9%	40.4
	Symptomatic	35.7	76.1	
	$\chi^2=10.27$, $p<0.01$; $r=0.40$			
	Health Status	Follow-up Process		
Unselected diagnoses†		(n=263)	(n=108)	
	Improved	59.7%	33.4%	26.3
	No change or slightly improved	24.7	37.0	
	Worsened	15.6	29.7	
	$\chi^2=24.4$, $p<0.01$; $r=0.23$			

*Selected data from Brook, R.H. 1973, *Quality of Care Assessment: A Comparison of Five Methods of Peer Review.* DHEW HRA-74-311. See Tables 11, 21, 22, and Figure 5-7.
†Selected data from Brook, Appel, Avery et al. 1971. Effectiveness of Inpatient Follow-up Care. *The New England Journal of Medicine* 285 (27): 1509-1514. See Table 4.

stifle innovation and result in defensive medicine, but those effects are not unique to process-auditing. The validity of process-auditing has been challenged largely on the basis of studies of process-outcome correlations, but the studies' results were not universally negative, and the most negative studies were so poorly designed that drawing firm conclusions from them is virtually impossible. The studies have also led to doubts regarding the cost-effectiveness of process regulation, but re-analysis of relevant data revealed that the results were somewhat more promising than previously thought. Final appraisal of the validity and impact of process regulation must await better research, but the best existing studies leave room for optimism. In any event, although many potential pitfalls of process auditing can be identified, there is little definitive evidence at this point that warrants rejecting the process approach in favor of outcome measurement.

Conclusions

At present, there is little solid basis for the widespread view that outcome measures are superior to process measures for assessing the quality of medical care. Although many leading authorities have been arguing vigorously that outcome measurement is ultimately preferable—and they have apparently convinced most other observers—the logic of their arguments and the supporting empirical evidence had heretofore never been examined closely.

Analysis shows that there are parallel sets of problems encountered whether one measures quality by process or outcome. Practically speaking, both types of measures require a base of knowledge concerning the medical relevance of criteria. At present, we have relatively little scientific evidence on the efficacy of medical procedures or on the relevance of outcome variance to the effects of care. Outcomes assessment is clearly not a solution to the problems created by the lack of evidence on efficacy, nor does regulating by outcomes insure that medical care will become cost effective.

Practical problems of data collection and data quality affect both process and outcome measures. Both measures draw heavily on medical records data, and therefore suffer similarly from the incompleteness and inaccuracies in medical records. Also, medical records abstractors disagree when coding both process and outcome.

The relative costs of the two approaches cannot be weighed without taking into account validity and the efforts needed to insure validity. Up to now, only process validity seems to have received adequate attention.

At present, it is unclear which type of measure is likely to be more valid. Although the validity of outcome measures has rarely been investigated, quantitative evidence from a number of relevant studies tended to confirm the suspicion that many outcome measures may be largely invalid as indexes of quality.

In contrast to outcome measures, process measures have been attacked repeatedly on the grounds of validity. However, the attacks were primarily based on studies of process-outcome correlations that were methodologically unsound. Process indexes should not be faulted if they do not correlate highly with outcome measures which have doubtful validity themselves.

Obviously, we currently do not know enough to make a clear choice between process and outcome measures as the best method of assessing quality of care. Up to now, discussions of the relative merits of the measures have been almost entirely lacking in empirical evidence. Conceptual discussions of possible pitfalls of various measures are a useful first step, but more refined assessments are now needed since neither process nor outcome is obviously superior. Quality of care refers conceptually to optimal performance by the medical care system to produce the best possible outcome under the circumstances. Because it is so difficult to determine in any particular case precisely what constitutes optimal performance or the best possible outcome, quality of care will be difficult to measure no matter what approach or blend of approaches is employed. If we are to learn how quality can be measured most validly and practically, further research using improved validation methods is essential.

References

Brook, R. H. 1973. *Quality of Care Assessment: A Comparison of Five Methods of Peer Review.* DHEW Publication No. HRA-74-3100. Washington, D.C.; U.S. Government Printing Office.

———. 1974. A Skeptic Looks at Peer Review. *Prism* 2 (10): 29–32.

———, Appel, F. A., Avery, C. et al. 1971. Effectiveness of Inpatient Follow-up Care. *The New England Journal of Medicine* 285 (27): 1509–1514.

———, Davies-Avery, A., Greenfield, S. et al. 1976. *Quality of Medical Care Assessment Using Outcome Measures: An Overview of the Method.* Santa Monica, Calif.: Rand Corporation.

———, Davies-Avery, A., Greenfield, S. et al. 1977. Assessing the Quality of Medical Care Using Outcome Measures: An Overview of the Method. *Medical Care* 15 (9): Supplement.

Bunker, J. P., Forrest, W. H., Mosteller, F. et al. 1969. *The National Halothane Study.* Washington, D.C.: U.S. Government Printing Office.

Cronbach, L. J. 1971. Test Validation. In Thorndike, R. L., ed., *Educational Measurement.* 2nd edition. Washington, D.C.: American Council on Education.

De Geyndt, W. 1970. Five Approaches for Assessing the Quality of Care. *Hospital Administration* 15 (Winter): 21–42.

Donabedian, A. 1966. Evaluating the Quality of Medical Care. *Milbank Memorial Fund Quarterly* 44 (3): 166–206.

———. 1969. *A Guide to Medical Care Administration. Volume 2: Medical Care Appraisal—Quality and Utilization.* New York: American Public Health Association.

———. 1978. The Quality of Medical Care. *Science* 200 (4344): 856–864.

Fessel, W. J., and Van Brunt, E. E. 1972. Assessing Quality of Care from the Medical Record. *The New England Journal of Medicine* 286 (3): 134–138.

Flood, A. B., Scott, W. R., Ewy, W. et al. 1977. Effectiveness in Professional Organizations: The Impact of Surgeons and Surgical Staff Organizations on the Quality of Care in Hospitals. Stanford, Calif.: Stanford Center for Health Care Research.

Ginzberg, E. 1975. Notes on Evaluating the Quality of Medical Care. *The New England Journal of Medicine* 292 (7): 366–368.

Goss, E. W., and Reed, J. I. 1974. Evaluating the Quality of Hospital Care Through Severity-Adjusted Death Rates: Some Pitfalls. *Medical Care* 12 (3) 202–213.

Greenfield, S., Nadler, M. A., Morgan, M. T. et al. 1977. The Clinical Investigation and Management of Chest Pain in an Emergency Department: Quality Assessment by Criteria Mapping. *Medical Care* 15 (11): 898–905.

Institute of Medicine. 1974. *Advancing the Quality of Health Care: A Policy Statement by a Committee of the Institute of Medicine.* Washington, D.C.: National Academy of Sciences.

Jacobs, C. M., and Jacobs, N. D. 1974. *The PEP Primer: The JCAH Performance Evaluation Procedure for Auditing and Improving Physician Care.* Chicago, Ill.: Quality Review Center, Joint Commission on Accreditation of Hospitals.

———, Christoffel, T. H., and Dixon, N. 1976. *Measuring the Quality of Patient Care: the Rationale for Outcome Audit.* Cambridge, Mass.: Ballinger Publishing Company.

Kane, R. L., Woolley, F. R., Gardner, J. H. et al. 1976. Measuring Outcomes of Care in an Ambulatory Primary Case Population: A Pilot Study. *Journal of Community Health* 1 (4): 233–240.

Kelman, S. 1976. *Improving Doctor Performance: A Study of the Use of Information and Organizational Change.* New York: Center for Policy Alternatives.

Kerlinger, N. 1965. *Foundations of Behavioral Research.* New York: Holt, Rinehart and Winston.

Kessner, D. M., Kalk, C. E., and Singer, J. 1973. Assessing Health Quality—the Case for Tracers. *The New England Journal of Medicine* 288 (4): 189–194.

Kisch, A. I., and Reeder, L. G. 1969. Client Evaluation of Physician Performance. *Journal of Health and Social Behavior* 10 (1): 51–58.

Labarthe, D. R., Hawkins, C. M., and Remington, R. D. 1973. Evaluation of Performance of Selected Devices for Measuring Blood Pressure. *American Journal of Cardiology* 32 (Sept. 20): 546–553.

Lembcke, P. A. 1967. Evolution of the Medical Audit. *Journal of the American Medical Association* 199 (8): 111–8.

Lindsay, M. I., Hermans, P. E., Nobrega, F. T. et al. 1976. Quality of Care Assessment. I. Outpatient Management of Acute Bacterial Cystitis as a Model. *Mayo Clinic Proceedings* 51 (May): 307–312.

Linn, B. S., Linn, M. W., Greenwald, S. R. et al. 1974. Validity of Impairment Ratings Made from Medical Records and from Personal Knowledge. *Medical Care* 12 (4): 363–386.

Lyons, T. F., and Payne, B. C. 1974. The Relationships of Physicians' Medical Recording Performances to their Medical Care Performance. *Medical Care* 12 (5): 463–469.

Marston, M. 1970. Compliance with Medical Regimens: A Review of the Literature. *Nursing Research* 19 (4): 312–323.

Martini, C. J. M., Allan, G. J. B., Davison, J. et al. 1977. Health Indexes Sensitive to Medical Care Variation. *International Journal of Health Services* 7 (2): 293–309.

Maskell, R. M., and Pead, L. J. 1976. Urinary Infection in Children in General Practice: A Laboratory View. *Journal of Hygiene* 77: 291–298.

McAuliffe, W. E. 1977. *Validation of Process and Outcome Measures of Quality of Care.* Boston, Mass. Harvard School of Public Health.

_____. 1978a. On the Statistical Validity of Standards Used in the Profile Monitoring of Health Care. *American Journal of Public Health* 68 (7): 645–651.

_____. 1978b. Studies of Process-Outcome Correlations in Medical Care Evaluations: A Critique. *Medical Care* 16 (11): 907–930.

McClure, W. 1973. Four Points on Quality Assurance. In Regional Medical Programs Service, ed., *Quality Assurance of Medical Care.* Washington, D.C.: Health Services and Mental Health Administration, DHEW.

Mitchell, J. H., Hardacre, J. M., Wenzel, F. S. et al. 1975. Cholecystectomy Peer Review: Measurement of Four Variables. *Medical Care* 13 (5): 409–416.

Moses, L. E., and Mosteller, F. 1968. Institutional Differences in Postoperative Death Rates: Commentary on Some of the Findings of the National Halothane Study. *Journal of the American Medical Association* 203 (Feb. 12): 150–152.

Nobrega, F. T., Morrow, G. W., Smoldt, R. K. et al. 1977. Quality Assessment in Hypertension: Analysis of Process and Outcome Methods. *The New England Journal of Medicine* 296 (3): 145–148.

Nunnally, J. C. 1978. *Psychometric Theory.* Revised edition. New York: McGraw-Hill.

Osborne, C. E., and Thompson, H. C. 1975. Criteria for Evaluation of Ambulatory Child Health Care by Chart Audit: Development and Testing of a Methodology. *Pediatrics* 56 (Suppl.): 625–692.

Palmer, R. H. 1976. Quality Assessment. In Green, R., ed., *Assuring Quality in Medical Care.* pp. 11–136. Cambridge, Mass.: Ballinger Publishing Company.

Peterson, O. L. 1956. An Analytical Study of North Carolina General Practice: 1953–54. *Journal of Medical Education* 31 (pt. 2, December): 1–165.

Richardson, F. M. 1972. Methodological Development of a System of Medical Audit. *Medical Care* 10 (6): 451–462.

Roemer, M. I., Moustafa, A. T., and Hopkins, C. E. 1968. A Proposed Hospital Quality Index: Hospital Death Rates Adjusted for Case Severity. *Health Services Research* 3 (1): 96–118.

Romm, F. J., Hulka, B. S., and Mayo, F. 1976. Correlates of Outcomes in Patients with Congestive Heart Failure. *Medical Care* 14 (9): 765–776.

Roos, N. P., Henteleff, P. D., and Roos, L. L. 1977. A New Audit Procedure Applied to an Old Question: Is the Frequency of T & A Justified? *Medical Care* 15 (1): 1–18.

Rosenberg, E. W. 1977. Medical Audit JCAH-Style: A Negative View. *Journal of the American Medical Association* 237 (18): 1935–1937.

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

Rubenstein, L., Mates, S., and Sidel, V. W. 1977. Quality-of-Care Assessment by Process and Outcome Scoring: Use of Weighted Algorithmic Assessment Criteria for Evaluation of Emergency Room Care of Women with Symptoms of Urinary Tract Infection. *Annals of Internal Medicine* 86 (5): 617–625.

Rutstein, D. D., Berenberg, W., Chalmers, T. C. et al. 1976. Measuring the Quality of Medical Care: A Clinical Method. *The New England Journal of Medicine* 294 (11): 582–588.

Schroeder, S. A., and Donaldson, M. 1976. The Feasibility of an Outcome Approach to Quality Assurance—A Report from One HMO. *Medical Care* 14 (1): 49–56.

Scott, W. R., Forrest, W. H., and Brown, B. W. 1976. Hospital Structure and Postoperative Mortality and Morbidity. In Shortell, S. M., and Brown, M., eds., *Organizational Research in Hospitals.* pp. 72–89. Chicago, Ill.: *Inquiry Book,* Blue Cross Association.

Selltiz, C., Jahoda, M., Deutsch M. et al. 1963. *Research Methods in Social Relations.* Revised edition. New York: Holt, Rinehart, and Winston.

Shapiro, S. 1967. End Result Measurements of Quality of Medical Care. *Milbank Memorial Fund Quarterly* 45 (2): 7–30.

Shortell, S. E., Becker, S. W., and Neuhauser, D. 1976. The Effects of Management Practices on Hospital Efficiency and Quality of Care. In Shortell, S. M., and Brown, M., eds., *Organizational Research in Hospitals.* Chicago, Ill.: *Inquiry Book,* Blue Cross Association.

———. 1978. Personal communication.

Stanford Center for Health Care Research. 1974. *Study of Institutional Differences in Postoperative Mortality.* Springfield, Va.: National Technical Information Service.

Starfield, B. 1974. Measurement of outcome: A Proposed Scheme. *Milbank Memorial Fund Quarterly* 52 (Winter): 39–50.

———, and Scheff, D. 1972. Effectiveness of Pediatric Care: The Relationship Between Process and Outcome. *Pediatrics* 49 (4): 547–552.

Thompson, H. C., and Osborne, C. E. 1974. Development of Criteria for Quality Assurance of Ambulatory Child Health Care. *Medical Care* 12 (10): 807–827.

Williamson, J. W. 1970. Outcomes of Health Care: Key to Health Improvement. In Hopkins, C. E., ed., *Outcomes Conference I-II: Methodology of Identifying, Measuring, and Evaluating Outcomes of Health Service Programs, Systems and Subsystems.* pp. 75–101. Washington, D.C.: Health Services and Mental Health Administration, Department of Health, Education, and Welfare.

Wilson, J. T. 1973. Compliance with Instructions in the Evaluation of Therapeutic Efficacy: A Common but Frequently Unrecognized Major Variable. *Clinical Pediatrics* 12 (6): 333–340.

Zuckerman, A. E., Starfield, B., Hochreiter, C. et al. 1966. Validating the Content of Pediatric Outpatient Medical Records by Means of Tape-Recording Doctor-Patient Encounters. *Pediatrics* 56 (3): 407–411.

This research was supported by the Executive Programs in Health Policy and Management, Contract N01-AH-44105 with the Bureau of Health Manpower, Health Resources Administration, U.S. Department of Health, Education, and Welfare.

Acknowledgments: I gratefully acknowledge the helpful comments of Noha Applebaum, Barbara Hulka, Nathan Keyfitz, Harry Marks, Heather Palmer, Marc Roberts, and the participants in the Executive Program in Policy, Planning, and Regulation.

Address correspondence to: William E. McAuliffe, Ph.D., Department of Behavioral Sciences, Harvard University School of Public Health, 677 Huntington Avenue, Boston, Massachusetts 02115.