On the Statistical Validity of Standards Used In Profile Monitoring of Health Care

WILLIAM E. MCAULIFFE, PHD

Abstract: In current methods of profile monitoring, standards of acceptability (cut-offs) are set either by consulting panels of experts, or by selecting an arbitrary point (e.g., the 75th percentile) on the profile (statistical distribution). However, experts have only vague ideas of what outcome rates ought to be, while profile statistics stem from samples for which unknown percentages of cases have received acceptable care. Poorly chosen standards could cause profile monitoring to be ineffective, inefficient, or unnecessarily disruptive. A new method proposes to set standards by using statistics for which the percentage of adequate care has been predetermined by examining the process of care. Plans to circumvent the pitfalls involved are described, as are two approaches to estimating the degree of process adequacy from routinely produced outcome rates. (Am. J. Public Health 68:645–651, 1978)

Introduction

Implementing the last phase of the government's Professional Standards Review Organization (PSRO) program of quality assurance will feature the statistical method of "profile monitoring," even though the methodological basis of the procedure has not yet been worked out fully. "Profile monitoring" seeks indications of poor quality care by inspecting distributions of statistics on the provision of health services or on patient outcomes. For example, if a hospital's rate of unfavorable outcomes exceeds a predetermined cut-off, say the 90th percentile, low-quality care may be involved, and further investigation is indicated. Although this sort of statistical monitoring has already become widespread with the growing computerization of health data, at least one fundamental element of the general approach—the setting of standards (cut-off points indicating the minimum level of acceptable performance)—still lacks a valid rationale.* This article will focus on the problem of setting standards for hospital *outcome* rates as an example, for presently there is no objective means for identifying which point on the statistical profile of outcomes indicates inadequate care.

Having valid standards profoundly affects the success and acceptability of a regulatory process since standards specify the expected level of performance of practitioners being monitored. Too stringent quality standards will result in low compliance with regulations and could erode the essential cooperation of physicians. Doctors disenchanted with unreasonable statistical standards can be expected to protest to legislators or the courts, or they may refuse to treat Medicare or Medicaid patients. Patients would also suffer if regulation forced physicians into adopting excessively conservative standards when considering high-risk procedures. Moreover, setting unattainable standards will cause unnecessary disruptions and will be administratively inefficient because too many cases that do not meet the statistical standards will prove satisfactory upon closer inspection. Standards set too low, on the other hand, could make regulation ineffective. Thus, it is crucial to select just the right performance levels as standards.

Unfortunately, while existing methods—the expert and the empirical—are easy to apply and relatively inexpensive, they do not produce outcome standards sufficiently valid to bolster the PSRO program's legitimacy in the eyes of practitioners. "Expert standards" are set by panels of specialists,[2] but the process is unsatisfactory because experts typically do not have the requisite facts at their disposal. For example, they often have only a rough idea of what percentage of poor outcomes to expect when care is faultless.[3] Conventional wisdom suggests, moreover, that experts tend to set unrealistically high standards. By contrast, "empirical standards," fixed by inspecting distributions of statistics derived

Address reprint requests to Dr. William E. McAuliffe, Department of Behavioral Sciences, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115. This paper, submitted to the Journal May 16, 1977, was revised and accepted for publication January 17, 1978.

*What is meant here by "statistical standards," some authors call "norms." Also, in this article, a "criterion" is a measure, and a "normative distribution" is a statistical distribution or profile specially constructed to provide a reference for interpretation of other scores; e.g., PAS standards are derived from normative distributions.[1]

from current practice and then selecting administratively reasonable cut-offs, are probably set too low in many cases. They would thereby serve only to justify the *status quo*. Moreover, although the use of empirical data gives the appearance of objectivity to standard-setting, the current method still involves considerable arbitrariness.[4, 5]

As Densen[5] has argued, "It is . . . an inherent responsibility of the standard setting body to . . . reduce the degree of arbitrariness in its decisions."; he added that health services research should guide the way. In this article I propose a method of setting standards for hospital outcome rates that may be more valid and informative than existing techniques. Although the new approach is illustrated by measuring the quality of care with outcome statistics, the rationale of validation applies as well to other health-regulatory problems for which statistical profiling has been used (e.g., setting length-of-hospital-stay targets and prospective reimbursement rates for hospitals or nursing homes). The new method's logic will be grasped more easily after a review of the current rationale for setting empirical standards for mortality rates.

The Rationale for Setting Profile Standards

It is necessary to set standards for hospital outcome rates because a patient's outcome is not a perfectly valid indicator of the quality of care he/she received. If death resulted only when care was inadequate, then the acceptable rate would simply be zero mortality, and there would be no need to consult experts or normative distributions. But since patient outcomes can reflect many causes (such as severity of disease, aging, multiple diagnoses) other than quality of care, the proper outcome standard for a hospital would be the proportion of suboptimal outcomes that is "normal" *for its patient population* even when care in every case was satisfactory.** Thus, a non-zero standard seeks to discount the effects of other causes so that outcome statistics can serve as a valid indicator of adequate medical care.

At present, however, no sound basis exists for identifying the "normal" rate. If all hospitals represented in a hypothetical distribution of outcome rates had provided *uniformly adequate care* to a *random sample of the patient population*, while any other systematic causes of outcome rates were held constant, the "normal" rate could then be determined easily. It would be the mean rate, and the variation in rates around it would stem only from random sampling and measurement errors. But real outcome-rate distributions currently available to regulators reflect at once variations in the inadequacy of care, systematic differences in patient mix, and other factors which affect patient outcome rates. And so, although the purpose of examining empirical distributions is to find the "normal" rate, in practice no rate stands out as the correct one.[4]

The problems resulting from differences in patient populations have received attention in a number of studies.[6–8] Because each hospital has its own special mix of patients, no one "normal" rate would actually apply to all of them. Each hospital would have its own "normal" rate, unless statistical adjustments could successfully equate all hospitals. To remove extraneous differences statistically, however, one must be able to identify them all, measure them perfectly, and correctly model their relationship with the dependent variable. Results from even the most sophisticated recent studies[7] suggest that the technology is still not well developed. Therefore, differences in patient mix remain a major obstacle to profile monitoring.

But even if there were no differences in patient mix across hospitals, regulators would still lack a valid statistical rationale for identifying the "normal" rate. The mean or median may be chosen arbitrarily,[4, 5] but an average obviously reflects merely what is, not what ought to be. Being higher or lower than the mean says nothing about how one stacks up absolutely. There is currently usually no information on the position or shape of the underlying distribution of quality. If most hospitals were providing poor care, as at least one study has found,[9] then the mean outcome rate would be unacceptably low, not the desired "normal" rate. The same problem holds for any arbitrarily chosen percentile, be it the 50th, the 75th, or the 95th. Similarly, marking off a range, such as falls between plus and minus two standard deviations from the mean, might seem a reasonable way to identify "the normal range of variation." But there is, in fact, no logical basis for employing standard deviations or for assuming that an outlier in an unfavorable direction necessarily reflects unacceptable performance.[5, 10] If one does not know how adequate care was in the hospitals, one cannot know which of their outcome rates reflects a desirable level of quality, nor is there any basis for believing that "skimming the worst five per cent" would be a satisfactory regulatory strategy.

Thus, two obstacles hinder detecting an acceptable rate of outcomes from existing profiles: 1) the effects of care and other factors are confounded; and 2) they are confounded in different proportions from one hospital to the next.

The essential arbitrariness of these empirical standards makes evaluating outcome statistics at present no more than a screening device. Failing to meet a statistical outcome standard does not convey definite information about the quality of care; it indicates only that poor care is more probable. Moreover, the success of that function rests on the assumption of a positive correlation of some magnitude between quality and good outcomes, when in fact even negative quality-outcome correlations are not too farfetched. A number of studies have reported negative process-outcome and structure-outcome correlations.[13, 14] And so, the standard at best merely alerts one to look more closely—ordinarily by reviewing the process of care documented in the medical record (called "process auditing"). More selective auditing would reduce the cost of regulation, but could also be more disruptive, since practitioners will view being singled out for scrutiny as accusatory. Thus, while potentially valuable as a screening device, the functioning of profile monitoring could be improved considerably by discovering a better way to set standards.

The crux of the difficulty with current empirical methods of setting standards is that "internal" features of a distri-

**Assuming that every patient is entitled to adequate care.

bution, such as its mean or median, are weak bases for designating a standard. Because a standard should be an operational definition (measure) of what is considered acceptable, something about how a standard is selected should lend that interpretation to it. It would be better to have some "external" basis, something outside the distribution itself, which argues for interpreting a particular point as the correct one. In most cases a valid statistical theory or additional independent research is needed; for instance, the external basis for validating a medical process standard might be the demonstration of optimal cost effectiveness. In the next section, a comparable basis for operationally defining acceptable outcome rates will be suggested.

A New Method of Setting Diagnosis-Specific Outcome Standards

The problems with existing methods of standard setting may be avoided by constructing normative distributions of outcome statistics, such as death rates, *from samples in which process evaluation has deemed the quality of care adequate in all cases*. Instead of using hospital death rates derived from *all* cases of a given diagnosis, one would base the normative rates only on samples of each hospital's cases for which care was found to be faultless. Since high-quality care would then be a constant, the variance of the sample rates would result entirely from other factors. In other words, the remaining variation would stem from patient differences, differences in the auditor's assessments of outcomes, and so on—all systematic and random measurement errors from the perspective of quality measurement. By looking at the distribution, one could then see how large the rates of unfavorable outcomes could be even when all care has been adequate. With the effects of care and other factors thus effectively disentangled, *the normative distribution becomes the benchmark against which to judge outcome rates of hospitals for which quality of care is unknown*. For example, if one selects as the standard rate the 99th percentile of the normative distribution, there would be only one chance in 100 that any actual hospital rate exceeding the standard resulted entirely from factors other than inadequate physician performance. One could then infer that hospitals exceeding the cut-off probably delivered some poor care, because the chances of not meeting the standard even though all patients received adequate care would be acceptably small.

Although the selection of the 99th percentile here might seem as arbitrary as the methods being criticized, it is not. The choice is not just a matter of administrative convenience or of one point's being as suitable as any other, but has a statistical interpretation in terms of the *known* rate of false positives one is willing to accept. The inference process is analogous to the process used in testing for statistical significance, where the exact probability of incorrectly rejecting the null hypothesis can be selected according to the substantive context. If the median were chosen arbitrarily as is currently done with other methods, the likelihood of falsely labeling an acceptable outcome rate as excessive would be 50 per cent. Since such errors are disruptive, using a higher percentile as the standard would make far more sense.

Outcome rates for cases receiving adequate care could be obtained prospectively or retrospectively. In a prospective design, *n* patients with a given diagnosis at each of the sampled hospitals would be selected randomly at admission and would be given special attention (concurrent review) to insure that care was optimal. In a retrospective design, previously treated cases would be sampled and audited. Only those for which care was satisfactory would be used. Although desirable from a number of perspectives, the retrospective method requires that quality of care be uncorrelated with other determinants of outcomes. If patients with poorer prognoses tended to receive better care, the normative rates might be too high and would result in overly lenient standards.

A Hypothetical Example

This section illustrates the proposed idea using a distribution of outcome rates in Figure 1a which is borrowed from a study of 34 hospitals by Moses and Mosteller[6] (assume the data refer to one diagnosis) and the data in Table 1 which are adapted (by moving decimal points) from a study by Brook.[3]‡ Brook's data for three diagnoses are outcome rates from one hospital broken down by process quality.

The key to understanding one problem with existing methods of setting outcome standards is to recognize that the Moses-Mosteller distribution is based on rates comparable to those in the "Total Sample" column in Table 1, based, that is, on cases unselected with respect to quality of care. We can see that "Total Sample" rates are profoundly affected by the underlying, but normally unknown, mix of adequate and inadequate care if we contrast them with the outcome rates of columns 1 ("Adequate Process") and 2 ("Inadequate Process"). For example, the rate of unsatisfactory outcomes for urinary tract infection is only .77 when care is adequate, but is 3.55, nearly five (4.6) times as great, when care is inadequate. The total sample's outcome rate for urinary tract infection (3.21) more nearly approximates the rate for inadequate care simply because 88 per cent of the sample's cases received inadequate care. However, the only way to know definitely that the total sample's rate reflects a substantial proportion of poor care rather than just the hospital's patient mix is either to conduct a process audit or to know that the rate for the total sample would be only .77 if all care were adequate.

Process auditing is expensive, and so it would be important to devise a less costly yet satisfactory method of monitoring quality. I propose we first generate a normative distribution of diagnosis-specific outcome data from cases receiving adequate care (as shown by process assessment) at a random sample of hospitals‡‡; we then use the normative

‡To keep the argument as general as possible and to anticipate practical difficulties associated with rare outcomes such as death, I refer to the rates simply as "outcomes rates".

‡‡The sampling frame would probably be regional or national. If research shows that these distributions change little over time, then the normative statistics could be used for more than one time period.

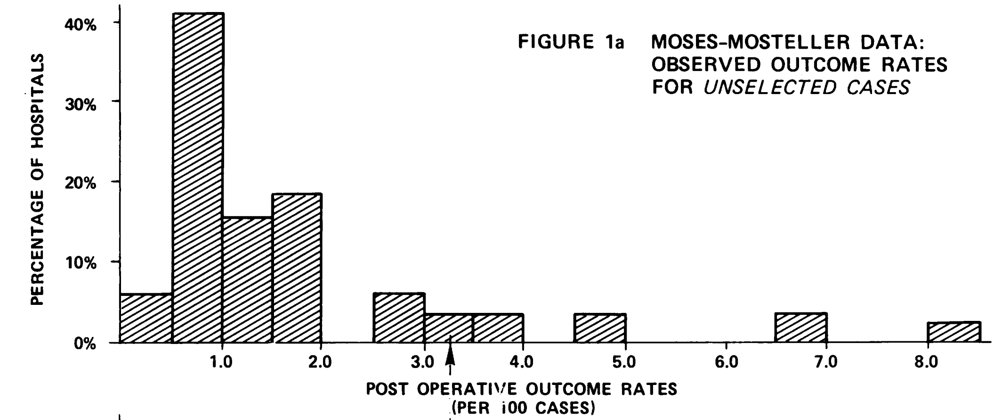

FIGURE 1a MOSES-MOSTELLER DATA: OBSERVED OUTCOME RATES FOR *UNSELECTED CASES*

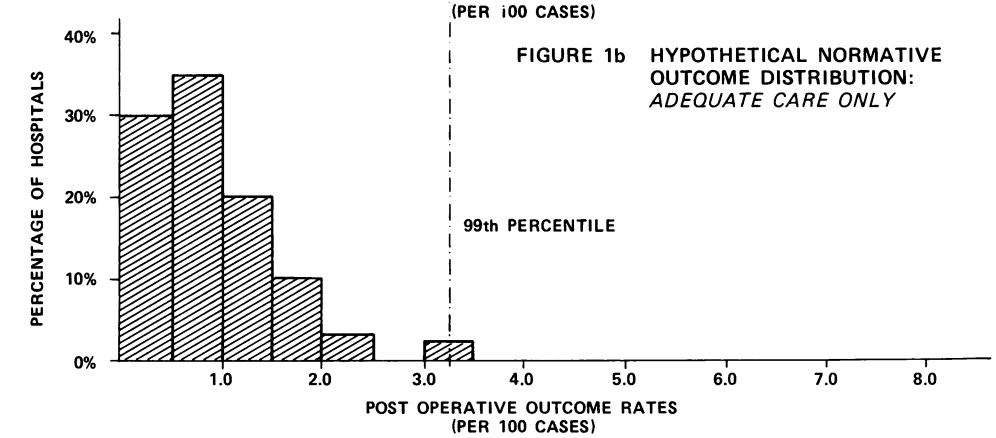

FIGURE 1b HYPOTHETICAL NORMATIVE OUTCOME DISTRIBUTION: *ADEQUATE CARE ONLY*

distribution to evaluate the corresponding rates of the entire population of hospitals. Since the normative distribution contains the variance of all factors affecting outcomes other than quality (including patient mix), using it as a benchmark is equivalent to controlling all non-quality determinants of outcomes at once. To illustrate the process, Figure 1b portrays a hypothetical example of such a normative distribution of rates for only cases receiving *adequate care*. The broken vertical line in the right-hand tail marks the 99th percentile of this normative distribution, a rate of 3.3 poor outcomes per 100 patients. Since some hospitals in the Moses-Mosteller distribution in Figure 1a had rates exceeding 3.3 per 100, one could infer with considerable confidence that those hospitals did not deliver acceptable care in all cases.

Statistical reliability would of course have to be considered. Although the details of a significance test are still to be worked out, the test would take into account both the sampling error of estimating the 99th percentile and the error of estimating the individual hospital's actual outcome rate. The statistical reliability of the method would therefore depend on the size of the samples (the same for all hospitals) used to generate the normative statistics and the number of cases represented in the rates being monitored. Consequently, the practical applicability of the method could be limited to larger hospitals and common diagnoses. Profile monitoring is sensitive to such considerations regardless of how standards are set.

Other writers, such as Lembcke,[11] have suggested basing outcome standards on the performance of teaching hospitals, where presumably the best care is practiced. Standards might also be based on the results of published clinical trials. The proposed method should be superior, however, since not all cases in even the best hospitals necessarily receive acceptable care, and the patient mix in teaching hospitals or clinical studies might be unique. One Health Maintenance Organization (HMO) recently found that their outcome rates were consistently better than standards based on the results of published clinical trials. As a result, their outcome "assessments offered little leverage over the participating physicians in terms of requiring a vigorous response to deficiencies in process criteria performance."[12]

The Effect of Invalidity in Process Evaluations

Since its validation mechanism (the external criterion) is process evaluation, success of the proposed approach depends ultimately on the validity of process auditing.* Although studies have purported to show that process audits of individual cases generally lack validity, in reviewing these studies elsewhere,[15] I have found them to be poorly designed and their conclusions unwarranted. The best existing evi-

*Whereas outcome measures may be used to validate process and structural measures of quality, the reverse is also true.[13] For example, Goss and Reed[14] attacked the validity of an outcome index devised by Roemer, *et al.*,[8] because the index failed to correlate with accepted structural indicators of quality.

dence suggests that process auditing has some validity. Moreover, correlations between percentage of adequate care by process evaluation and outcome rates *across hospitals*, which is the relevant level of aggregation here, are likely to be considerably stronger than correlations based on process and outcome *across individuals*, because aggregate statistics tend to be more reliable.

In any case, measurement validity is always a matter of degree, and having only moderate process validity is less serious than one might think. Even if "true quality" did not correlate at all with the process evaluations, the normative distribution would turn out to be identical to the "unselected" Moses-Mosteller distribution, except that the statistics would be less reliable. In effect, we would return to our current position of lacking an external validator. Moreover, if what was actually done to patients had no bearing on their measured outcomes, and if there was no correlation, then outcome rates would also not be valid measures of quality and should not be used to monitor quality. Lesser amounts of process invalidity would shift the normative distribution to the right, which moves the 99th percentile cut-off to the right and results in standards somewhat less stringent than they should be.** For example, the 99th percentile in Figure 1b would then fall higher in the range of the Moses-Mosteller distribution in Figure 1a. From one standpoint, however, the error is conservative, for the interpretation that some poor care is implicated by exceeding the standard would even more certainly be correct.

The best remedy for poor process data is to audit better by using better criteria, more judges, improved data sources, or more complete coverage of nonphysician-quality inputs, etc., (see McAuliffe[13, 15, 16] on the validity of process and outcome assessments and how they might be improved by special efforts). Since generating the normative distribution would be a special effort promising long-run efficiencies by eliminating the need to examine as many hospitals with unpreventably high rates, greater than usual resources could be spent obtaining valid process audits.[16]

The location of the normative distribution also depends on the stringency of the standards for process adequacy. Although I have chosen to focus on outcome standards, other writers[5, 10] have pointed out the arbitrary nature of current process standards, and they have recommended experimentation and cost-effectiveness analysis as external, validating bases for process standards. Clearly, valid process standards would be necessary for ultimately perfecting the proposed method of setting outcome standards; making do with less adequate process standards only diminishes the full contribution of the method.

Statistical Adjustments

The proposed method's sensitivity could be increased by statistically adjusting for patient mix *both* the normative rates and the true hospital rates. The method itself should be superior to statistical controls alone for removing the effects of patient mix because the method would not require that one know, measure, and model perfectly every cause of outcome variation. The effects of all non-quality causes, known and unknown, are contained in the normative distribution and are discounted when the normative distribution is used as suggested. Since our current understanding of what specifically needs to be controlled is limited and since our existing measures of control variables are admittedly crude, this alternative approach to control is a major advantage of the method. However, without explicit statistical adjustments, hospitals rendering relatively poor care might avoid detection if they had relatively healthy patients. That is so because the method sacrifices sensitivity to poor care to maximize the certainty that an "unfavorable" outcome rate is not due to patient mix or other non-quality factors. In general, the method's sensitivity to poor care is a function of the relative size of the outcome variance due to quality, as compared to the variance due to other factors. Therefore, fewer false positives will occur if the variance of major extraneous factors can be removed statistically.‡ In that sense, the proposed method can be seen as picking up where statistical adjustments leave off.

Statistical adjustments could also improve the method's sensitivity to structural inadequacies (e.g., poor equipment) that may not be measured by process audits of physician performance. For example, a hospital whose equipment was not up to standard might have an especially high mortality rate even for cases where the performance by the physician and nursing staff were rated as adequate. The normative rates could be adjusted to what they should have been if all hospitals had *adequate* equipment. (Of course, the actual rates would not be adjusted.) The result would be a more refined normative distribution, one which reflects a broader definition of quality.

Estimating Process Quality from Outcome Rates: Two Approaches

The basic rationale of the proposed method might be extended to generate a family of distributions which would specify the minimum amount of inadequate care reflected by a given outcome rate. That is, one would know not only that the rate reflected some inadequate care, but how much. The parameter of the distributions would be "quality mix", the percentage of cases in the normative samples that had an inadequate process. Instead of deriving a normative distribution only from samples in which all cases received adequate care, it would be possible to estimate what the distribution of rates would be if they had been calculated from samples where, say, 80 per cent of each sample had received inadequate care. The necessary empirical data would be a pair of sample outcome rates from each of *i* hospitals, one rate

*It is desirable to apply the same process criteria and standards in all the sampled hospitals, otherwise the validity of the process evaluations is reduced further, and the proposed method becomes less effective.

‡This use of statistical adjustments for achieving greater sensitivity is analogous to their use in "blocking" in randomized experiments.

TABLE 1—Rates per 100 of Unsatisfactory Outcomes by Process Quality of Care

		Process Quality of Care		
Diagnosis		Adequate Process	Inadequate Process	Total
Urinary Tract Infection		(n = 13) (12%)	(n = 93) (88%)	(n = 106) (100%)
	Rate of Unsatisfactory Outcomes. . . .	.77	3.55	3.21
Hypertension		(n = 31) (27%)	(n = 82) (73%)	(n = 113) (100%)
	Rate of Unsatisfactory Outcomes. . . .	2.58	5.12	4.42
Ulcer		(n = 28) (38%)	(n = 46) (62%)	(n = 74) (100%)
	Rate of Unsatisfactory Outcomes. . . .	3.57	7.61	6.08

Source: Brook.[3] For all rates, decimal points were moved one place to the left to make them compatible with the Moses-Mosteller rates.

TABLE 2—Hypothetical Outcome Standards for Different Percentages of Inadequate Care

	Rates of Suboptimal Outcomes (per 100 patients)		
Percent of Cases Receiving Inadequate Care	Urinary Tract Infection	Hypertension	Ulcer
0	*0.77*	*2.58*	*3.57*
10	1.05	2.83	3.97
20	1.33	3.09	4.38
30	1.61	3.34	4.78
40	1.88	3.60	5.19
50	2.16	3.85	5.59
60	2.44	4.10	5.99
70	2.72	4.36	6.40
80	2.99	4.61	6.80
85	3.13		
86	3.16		
87	3.18		
88	3.21		
89	3.24		
90	3.27	4.87	7.21
100	*3.55*	*5.12*	*7.61*

Note: The percentage of a hospital's cases receiving poor care is determined by locating the outcome rate next lower than the observed rate. For example, if the observed outcome rate for urinary tract infection were 3.15, the nearest lower standard in the Table is 3.13, which coincides with 85 per cent of the cases having received poor care.

(R_{si}) for satisfactory care (same as before) and one (R_{ui}) for unsatisfactory care.‡‡ Plugging these data into the formula,

$$R_{ji} = p_jR_{si} + (1 - p_j)R_{ui}, \qquad (1)$$

one could generate the desired family of 101 distributions of normative rates (R_{ji}) by systematically changing the quality-mix parameter (p_j), a proportion which specifies the percentage (0–100) of satisfactory care.

To show how these families of distributions might be used, I have calculated the values in Table 2 using the Brook data (underlined) from Table 1. Since his data came from only one hospital, it is necessary for illustration to assume that for all three diagnoses his outcome rates fell at the 99th percentile of hypothetical distributions for satisfactory and unsatisfactory care. Therefore, the values in Table 2 are estimates of the 99th percentiles of their respective quality mix distributions. With such a table, if a hospital's overall outcome rate exceeds the 99th percentile of the 80 per cent mix (2.99 for urinary tract infection), one could then infer that not only was there *some* poor care, but that *at least 80 per cent of the hospital's cases* had received poor care. Moreover, since it is possible to generate a distribution for all possible quality mixes, from 0 to 100 per cent adequate, one could estimate for any actual outcome rate the minimum amount of poor care it reflects. For example, one could infer from Table 2 that an observed outcome rate of 3.21 indicated that at least 88 per cent of the hospital's urinary tract infection cases had received less than satisfactory care.

Use of this formula assumes that patient mix differences at any one point along the range of quality mix would be estimated reasonably well by differences between the sampled hospitals, which probably span the entire range of quality. If this assumption were grossly in error, better estimates might be obtained by stratifying hospitals by quality and estimating normative distributions within the strata limits from only hospitals which actually fall within that range of quality mix.

Process quality could also be estimated from observed outcome statistics by regression analysis. The needed regression equation would be estimated using process and outcome data from a sample of hospitals. The percentage of patients receiving inadequate process would be the *dependent* variable, and the rate of unfavorable outcomes (after case-mix adjustments) would be the independent variable.* Once the regression equation were estimated, it could be used for predicting the results of process audits for other hospitals from their more easily collected outcome data.

Both methods seem worth pursuing. Data would be easier to obtain for the regression approach; but if the outcome rates lacked validity because of inadequate controls, process

‡‡To simplify the presentation, I have assumed that process quality was a dichotomy. If there are multiple possibilities for inadequacy and some inadequacies have more serious implications than others, then outcome variation due to differences in process quality is probably incompletely controlled using the present method. One possibility for further removing unwanted quality variation would be to develop a more refined measure of process adequacy. For example, process quality might be scored as a percentage of weighted elements of care. Then, for each hospital, patient outcomes would be regressed on their process scores. The regression equations could then be used instead of equation (1) to generate the estimated outcome rates, R_j, for each hospital. The outcome rate distributions would therefore reflect a mean percentage of adequate *care* rather than the percentage of adequate *cases*, but could be used in exactly the same manner.

*If process and outcomes were measured by more finely calibrated scales (as discussed above) rather than simply "adequate/inadequate", the regression variables could be the mean process score and the mean outcome score.

quality would be poorly predicted by the regression estimates. Also, the estimates from the first approach and the regression approach are not exactly comparable. Figures from the first approach estimate the *minimum* amount of poor process that would be found; whereas figures from the regression approach estimate the *most likely* amount of poor care that would be found. More conservative figures could be obtained from the regression approach by using confidence limits around the estimates. How the methods compare in practice remains to be studied, but it seems likely that both types of estimates could be useful to regulators.

Time-Series Monitoring of Individual Institutions

The principles described thus far for setting standards across hospitals would apply as well to a single hospital monitoring its own performance from year to year. The only change from current methods would involve developing the baseline outcome rate. Instead of arbitrarily selecting the outcome rate of one year as the baseline and then comparing rates in subsequent years with it and each other, one would obtain a baseline rate from a random sample of cases receiving adequate care and compare it with rates for all cases from subsequent years. With the addition of statistical controls and confidence intervals to control for systematic changes in patient mix and sampling errors, one would be in a position to determine not only whether performance had improved in subsequent years but also how close to target it had come. This knowledge would be especially useful for planning and evaluation.

Conclusion

Current statistical standard-setting methods fall down because they rely on "internal" features of distributions, such as means or ranges, to indicate acceptability, even though the implicit rationale for that interpretation is weak. More valid standards would result if standards were based on an "external" criterion that supplied a logical basis for locating the point of acceptable performance. Other authors[4, 9] having previously recommended that the external criterion for length-of-stay targets and process standards be the experimentally determined effect on patient outcomes, but up to now no comparable external criterion has been available for setting outcome standards.

I have proposed that valid outcome standards might be set using normative distributions based on process-validated statistics. The method requires recognizing the possibility of improving the validity of outcome statistics by prior process assessment, which would disentangle and then discount nonquality causes of poor outcomes.

Two possible techniques for estimating the degree of process adequacy (what an expensive process audit would produce) from routinely produced outcome rates have been described. Although generating the estimating tables and regression equations would entail some additional cost at first, the methods could result in more efficient and effective regulation and could thereby be cost-effective in the end. Moreover, if successful, the methods would offer regulators the advantage of understandable, scientifically defensible standards that would transform outcome monitoring from a mere screening technique into a positive method for assessing the quality of care.

Obviously, much work remains to determine whether the specific methods proposed here are workable and worth the effort—which of course is more or less true for all outcome methods of assessing quality of care.[16] But presenting these ideas now is valuable because they show clearly what is wrong with the current methods of setting standards for profile monitoring, and point the way toward new conceptual approaches. Since profile monitoring has become an increasingly popular tool for regulators and is scheduled for national implementation by PSRO, it is important that these issues be brought out.

ACKNOWLEDGMENTS

I received helpful comments on earlier drafts from Noha Applebaum, Ralph Berry, Howard Bloom, Penny Feldman, Robert Gordon, Robert Haggerty, Frederick Mosteller, Heather Palmer, Susan Radius, and Marc Roberts. Work on this article was supported by Contract No. N01-AH-44105 with the Bureau of Health Manpower, Health Resources Administration, DHEW.

REFERENCES

1. Commission on Professional and Hospital Activities. Hospital Mortality: PAS Hospitals, United States, 1972–73. Commission on Professional and Hospital Activities, Ann Arbor, Michigan, 1975.
2. Schonfeld, H. K. Standards for the audit and planning of medical care: A method for preparing audit standards for mixtures of patients. Medical Care 8(4):287–297, 1970.
3. Brook, R. H. Quality of Care Assessment: A Comparison of Five Methods of Peer Review. Department of Health, Education, and Welfare. Publication No. HRA-74-3100, 1973.
4. Sullivan, D. J. Algorithms for determining length of stay standards. Presented at the American Public Health Association Annual Meeting, Miami Beach, FL, October 20, 1976.
5. Densen, P. M. Standard setting, monitoring and health services research. Health Care Research, edited by D. E. Larsen and E. J. Love. University of Calgary Press, Calgary, 1974.
6. Moses, L. E., Mosteller, F. Institutional differences in postoperative death rates. JAMA 203(7):150–2, 1968.
7. Stanford Center for Health Care Research. Comparisons of hospitals with regard to outcomes of surgery. Health Services Research 11:111–127, 1976.
8. Roemer, M. I., Moustafer, A. T., Hopkins, C. E. A proposed hospital quality index: hospital death rates adjusted for case severity. Health Services Research 3:96–118, 1968.
9. Bird, D. Substandard care is found in majority of 105 hospitals. New York Times Briefing Papers for Public Affairs. No date.
10. Cochrane, A. L. The history of the measurement of ill health. International Journal of Epidemiology 1:89–92, 1972.
11. Lembcke, P. A. A scientific method of medical auditing. Hospitals 33:65, 68, 70–71, 1959.
12. Southern California Region Kaiser-Permanente Medical Care Program. Quality of Care Study, Final Report. Kaiser-Permanente Medical Care Program, Los Angeles, California, 1976.
13. McAuliffe, W. E. Validation of process and outcome measures of quality of health care. Department of Behavioral Sciences, Harvard School of Public Health, 1977.
14. Goss, M. E. W., Reed, J. I. Evaluating the quality of hospital care through severity-adjusted death rates: some pitfalls. Medical Care 12:202–13, 1974.
15. McAuliffe, W. E. Studies of process-outcome correlations in medical audits: a critique. Medical Care. In Press.
16. McAuliffe, W. E. On the validity of process and outcome measures of quality of medical care. Department of Behavioral Sciences, Harvard School of Public Health, 1977.