NATIONAL QUALITY OF CARE FORUM

BRIDGING THE GAP
BETWEEN

THEORY+PRACTICE

EXPLORING CONTINUOUS QUALITY IMPROVEMENT

Sponsored by
The Hospital Research and
Educational Trust and The
Parke-Davis Division of the
Warner-Lambert Company

NATIONAL QUALITY OF CARE FORUM

BRIDGING THE GAP BETWEEN THEORY PRACTICE

EXPLORING CONTINUOUS QUALITY IMPROVEMENT

AHA Catalog no. 169510
ISBN 0-87258-587-5

Printed in the U.S.A.
10M-7/92-IN24108

Editorial Services: Carole J. Bolster
Production: Susan Smith
Design: David Prihoda

Contents

Good Enough? Standards and Measurement in Continuous Quality Improvement

Brent C. James, M.D., M.Stat

Several years ago a 14 month-old girl entered a major U.S. children's hospital. The child suffered from several congenital anomalies. Although she was admitted to the pediatric neurosurgery service for placement of a shunt to treat her hydrocephalus, she also received daily doses of digoxin elixir for a heart abnormality.

A neurosurgery fellow, already board-eligible in general surgery, completed the child's history, physical, and presurgical orders the day before surgery. The fellow had joined the pediatric neurosurgery service less than two weeks earlier to complete the single three-month pediatric rotation in his fellowship. It was his first experience with the children's hospital. He was unfamiliar with the hospital's written policy concerning digoxin orders for pediatric patients. The policy required that, for any digoxin order, the physician calculate the digoxin dose on the order sheet, based on the patient's weight, then sign it. The patient's nurse was required to independently complete and sign a similar calculation, reaching the same dose, on the same order sheet. If both calculations and signatures were not complete, the clerk was to refuse to process the order, and the pharmacy was to refuse to fill it.

Unfortunately, the neurosurgery fellow did not calculate the dose. Instead he transposed it from the bottle of digoxin elixir that the patient's parents brought to the hospital. That bottle listed 0.0345 milligrams of digoxin per day, divided into two doses. But in writing the order the fellow transposed the decimal. He wrote for 0.345 milligrams of digoxin per day—a 10-times overdose. The nurse never completed a parallel calculation. The clerk did process and submit the order. The pharmacist, upon filling the order. recognized that it was impractical to administer the amount of elixir required to deliver that large a dose. In place of the elixir, he sent adult-sized digoxin tablets. Upon receiving the tablets, the patient's nurses expended the extra effort necessary to prepare the drug in a form that was palatable to the patient: They ground the tablets into a powder then mixed them with sugar water in order to administer the twice-daily doses.

The night before surgery the patient's cardiac condition destabilized. She never went to surgery. Three days later, when the patient was on the verge of death from a sudden, mysterious cardiac ailment, someone discovered the order error. The attending physician immediately corrected the digoxin order and the patient quickly recovered.

As is often the case with process breakdowns that result in quality failures. this error increased the cost of the patient's care—she received three extra days of hospitalization, testing, and treatment that would have been avoided if the original error had not occurred (ref. 1). But the real problem was, during the four years immediately preceding this episode, three children died at that hospital from digoxin overdoses. In fact, rare but

recurrent digoxin order errors were the reason that the hospital's clinical professionals developed a written protocol. It was one among more than 800 clinical procedures and policies in force at the hospital, which together occupied two four-inch-thick binders.

In 1976, McDonald demonstrated an important fact of medical practice that any clinician recognizes: Humans are inherently fallible information processors (ref. 2). Any practitioner, regardless of training or commitment. will occasionally make mistakes. In fact, digit or decimal transpositions when reading or recording numbers are among the most common types of errors. But how did the hospital's administration, working through their quality assurance department and the medical staff organization, respond to this process failure? What systems did they build to address the hospital's capacity to train all new house staff physicians in the hospital's standard policies (an insurmountable task, given the frequency with which house staff change and the length of their rotations), or to deal with the fact that only about 40 percent of all digoxin orders, whether written by house staff or attending faculty, followed the written policy? They placed a letter of reprimand in the neurosurgery fellow's file.

Focus on the Tail

The foregoing example demonstrates American medicine's customary approach to quality assurance. That approach attempts to identify individual providers (clinicians or hospitals) that show unacceptable behavior relative to their peers, then take action to eliminate the offending behavior. Figure 1 illustrates the quality assurance (QA) theory that underlies corrective actions aimed at individuals. Traditional QA measures some important quality factor then uses a threshold to separate acceptable quality (which does not require immediate action) from unacceptable quality (which requires action). The Joint Commission on Accreditation of Healthcare Organizations' (JCAHO) ten-step model for evaluating quality lists as its fifth step, "Establish thresholds ... for the indicators that trigger evaluation of the care" (ref. 3). For example, national norms regarding postoperative deep wound infections function as thresholds: Under a traditional QA approach, if a hospital's rate is at or below 2 percent, then its performance is excellent—although improvement action is encouraged, it is not required. But if the hospital's rate creeps above 4 percent, then action must be taken. As a second example, representatives of the JCAHO originally used a threshold to illustrate the use of quality indicator measures in their Agenda for Change: If a trauma patient enters an emergency room, and during the course of the workup requires a CAT scan. then the scan should be obtained within four hours of admission to

the emergency department. Cases that exceed four hours are flagged for further review (ref. 4).

Traditional QA's use of thresholds creates an inspection system that tries to identify low-quality events, practitioners, or providers. Berwick used the term "bad apple" to describe such systems (ref. 5). Others have called it "sort and shoot": The aim is to separate the good from the bad, then act to remove or change ("shoot") the bad. A more apt appellation might be "focus on the tail": Such systems use thresholds to establish a statistical tail, then concentrate their improvement activities in that tail. But threshold systems have a number of theoretical defects that can be traced directly to practical failings. Three of traditional QA's most important deficiencies (in order of increasing importance) are (1) normal human reaction to disincentive systems with arbitrary thresholds (the "cycle of fear"). (2, inappropriate conclusions drawn from screening examinations, and (3) the underlying philosophy that drives threshold-based quality—being "good enough" instead of "the best possible"—which often results in threshold, becoming artificial quality floors or ceilings.

The Cycle of Fear

How does the medical community choose threshold levels in a traditional QA inspection system? For example, why is the accepted national threshold for postoperative deep wound infections 2 percent? Why is it not less than 1 percent, as some hospitals consistently achieve, or perhaps 2.5 percent, a rate more acceptable to those hospitals that struggle with the 2 percent level'? Why is it *exactly* two percent?

In fact, nearly all threshold levels are inherently arbitrary. With very rare exception, there is no scientific justification for selecting one specific threshold level over some other level in the same general range. Thresholds are usually chosen based on statistical measures from published reports (for example, the average rate for a small group of respected hospitals or for a large group of typical hospitals).

Once an arbitrary threshold is in place, most traditional QA programs, apply some sort of disincentive to providers who fall into the tail. Such disincentives can be extensive (reprimands, privilege restriction,. or licensing actions) or limited (requirements for further education. being identified to one's colleagues as a quality outlier, or failure to receive real or perceived rewards that accrue to those who do not fall in the tail. But Scherkenbach argues that, when faced with disincentives applied through arbitrary thresholds, human beings react in a predictable way. He describes their typical reaction as the "cycle of fear" (ref. 6). It contains three

sequential phases, all of which consume effort and resources while failing to produce better outcomes.

Phase 1: Denial (also known as "kill the messenger" or "shift the blame"). Most clinicians' first reaction, upon discovering that they have exceeded some threshold and been flagged as an outlier, is denial. They challenge the measurement system and the aims, techniques, and accuracy of the study that finds their performance to be inferior. That response is so common that it has become cliche. All hospital quality assurance managers have heard: "The sample size wasn't large enough." "My patients are sicker (the study failed to properly adjust for patient factors)." "The study's focus was inappropriate. It overlooked critical elements that outweighed the elements measured." "That bad outcome may have occurred, but it was the nurse's (or administration's, or some other department's) fault." "It was a fluke. That kind of bad result happens to everybody sometime (it was a random event)." Consider a recent newspaper headline: "VA hospital…gets low marks in investigation. Survey: Center rated among the 14 worst in 1989. Spokesman calls the study flawed" (ref. 7). A follow-up article, printed a few days later, ran under the headline: "VA hospital outraged over media image. Vets: Congressional allegations of substandard care called unfair, false." The article goes on to state that the "report ranks. . . (the) VA center as among the 10 VA hospitals most likely to have serious care problems." Representatives of the VA Medical Center then point out that their hospital may do a better job of reporting (resulting in higher apparent failure rates compared to other centers that fail to even detect bad outcomes). They criticize the use of raw rates for interim quality indicators, rather than ultimate outcomes following final peer review. They even question the motives of the congressional examiners, suggesting a possible "hidden agenda" in the year before an election, and lament the fact that individuals testifying in congressional hearings cannot be sued for "false allegations." Finally, the VA's spokesman notes, "The problems the VA hospital does have are the result of poor funding" (ref. 8).

Under a system that combines disincentives with arbitrary thresholds, the primary aim is to avoid being flagged as an outlier. Unfortunately, the most direct response is denial, not actions to improve care. It is a rare human indeed who will routinely react with positive words or actions to such criticisms. Clinicians who have been singled out as quality outliers, or those who fear that they someday may be, often erect barriers to future measurement and resist participating in quality assurance efforts. Many deny that quality assurance has anything to do with actual care quality. They privately opine that the true aim of the quality assurance department is to show sufficient sanctions against providers as to demonstrate a vigorous "bad apple" program to regulators--"blood in the water"—regardless of the actual level of care achieved.

Phase 2: Filter the Data (also known as "game the system"). Many clinical assessments include an element of judgment. It is not surprising that threshold/disincentive systems can bias clinical judgments, even for conscientious, honest practitioners. In the extreme, a series of thresholds create a rule system that practitioners legally can, and arguably should, manipulate in order to maximize benefits and minimize risks, for themselves and for their patients.

For example, Iezzoni and others have documented a consistent shift in ICD-9 coding since 1983 (refs. 9-11). Iezzoni traces those changes to the Medicare prospective pricing system. What hospital can afford an administrator or medical records director who cannot upgrade ICD-9 codes to produce the highest-paying, defensible DRG classification? Similarly, most physicians routinely receive mailings that invite them to bring their office staffs to special conferences in order to learn how to upcode CPT-4 reporting to those clinically-defensible codes that will maximize reimbursement.

Phase 3: Micromanagement. In some instances practitioners may fall beyond a quality assurance threshold but not know how they came to that position—the quality assurance measure demonstrated that an outcome was unacceptable but did not trace the outcome to a process step failure. In such a circumstance, the practitioners may try any number of changes in their processes of care in an unfocused effort to "do something." “Micromanagement” describes detailed, unfocused interventions taken within a process to closely control some outcome, even though a clear relationship between the desired outcome and the intervention does not exist. As Berwick correctly notes, "How can the root causes of differences in results be discovered? This is the most significant question. Unless there is a method for discovering the reasons for differences, the knowledge of results is useful only for judgment, not for improvement. Information on results, alone, is not enough. Knowing how well something works is different from knowing how it works (ref. 12).

For example, several investigators have suggested that defensive medicine—extra tests that physicians order with a hope to prevent lawsuits from arising, or to protect the physician in court if they do—may not be effective (refs. 13, 14). But physicians order the extra tests regardless. They are not familiar with other, more effective, actions they could pursue in order to avoid falling into the "litigation" tail.

Inappropriate Conclusions from Screening Tests

Few tests are perfect. Nearly every test occasionally shows a positive result when, in fact, the underlying factor that the test was designed to detect did not exist (a false positive), or shows a negative result when the underlying factor does exist (a false negative). For example, some individuals who are really on the "acceptable" side of the-threshold in figure I may, through the influence of random errors inherent in the testing environment, be carried to the "unacceptable" side of the threshold. Similarly some individuals, whose true performance falls on the "unacceptable" side may be carried to the "acceptable" side of the threshold by random chance. One would, therefore, expect that quality assurance professionals would be very interested in the false positive and false negative rates of the tests they employ. This is especially true when a test is used to screen a large population for some relatively rare factor. In such circumstances, even a very good test may generate a high false positive rate—its positive predictive value depends on the underlying prevalence of the factor in the population, as well as the test's specificity (ref. 15).

For example, in 1989 the Hospital Research and Education Trust's (HRET) Quality Measurement and Management Project (QMMP) task force launched a study to compare risk-adjusted mortality following acute myocardial infarct (AMI) for a large group of hospitals. About 2,500 QMMP member hospitals gave permission for the study group, headed by Dr. Mark Blumberg, to use their MEDPAR UB-82 data for that purpose. Almost 700 hospitals were eliminated from the main study group because their AMI data didn't pass computer screens for completeness and internal consistency. Among the remaining 1,800 hospitals, Blumberg constructed a risk model and then generated adjusted mortality rates for each hospital. In comparing the resulting rates, the QMMP task force chose (arbitrarily) a 2 percent X2 threshold to identify outlier hospitals. Thirty-five hospitals fell beyond that threshold and were classified as high-mortality outliers.

But Blumberg knew something about the statistical models he had employed and the underlying variation in the hospital data. He estimated that about 20 hospitals could fall into his 2 percent X^2 tail by random chance alone. That is, he estimated that the test's false positive rate could exceed 60 percent (ref. 16). If the QMMP "shot" the hospitals who had fallen into the tail—say, by publishing their risk-adjusted mortality rates in the newspaper-then for every guilty hospital that was flagged more than one innocent hospital would have been punished.

Rolph, Kravitz, and McGuigan recently analyzed the use of physician malpractice claims histories to target individual practitioners for education

or sanctions, with an aim to improve care delivery. They concluded that such techniques have only modest predictive power and are likely to be ineffective in eliminating bad care. (ref. 17). Similarly, Park and others used independent quality measures to estimate that the false positive rate for HCFA's initial annual mortality report ranged from 56 percent to 82 percent (ref. 18). Green, Passman, and Wintfeld estimated that the false positive rate approached 50 percent when cross-checking the 1989 HCFA mortality reports (ref. 19). It is not surprising that many hospital leaders find the HCFA mortality reports to be of very little use in improving care delivery (ref. 20).

Screening tests are *concentrators.* They are designed to extract a small subgroup of test subjects within which the concentration of true positive cases is much higher than in the total population, while not overlooking too many true positive cases that is, with a low false negative rate (ref. 15). They are not designed to make black-and-white determinations, as is implied by the manner in which threshold values are commonly used to make "sort and shoot" decisions. A positive screening result is a question. not an answer. A screening test can form a subgroup with a relatively high concentration of true positives, but that group always requires further investigation before conclusions can be accurately drawn. Those who use threshold quality screening methods to reach "good" or "bad" conclusions fail to understand the nature of the tests that they employ. But in most traditional quality assurance settings, falling into the tail results in immediate disincentives—the contumely associated with being labelled an outlier, criticism, closer scrutiny (with inherent bias arising from the intensity of the search for questionable practices), demands for education. or even sanctions (ref. 21).

'Good Enough' or 'The Best Possible'?

Regardless of its other failings, the most troubling aspect of threshold-based quality assurance is its underlying philosophy. By its very structure it assumes that, if the egregiously bad are detected and eliminated, then what remains is somehow excellent. But that is just not true. What's left is average. It is hard to imagine a better system to consistently ensure mediocrity.

Consider Figure 1. Some (theoretical) hospitals function at a quality level marked by point B. If it is possible to function at point B, why is it acceptable to function at point A (adjacent to the threshold that separates "unacceptable" from "acceptable" care)? Each time a hospital fails to function at point B it wastes resources (through the mechanisms of quality waste and low efficiency) and harms patients (through excess morbidity

and mortality) (ref. 1). While wasting resources may only be poor management, allowing patients to come to preventable harm is unconscionable. Continued reliance on thresholds means that we are asking the wrong question. We ask whether we are "good enough" instead of asking whether we are "the best we can be." Because every time we fail to be the best we can be, we waste resources and we fail to deliver the level of care that our patients deserve.

Finally, "good enough" thresholds can become ends in themselves; they become artificial floors or ceilings that limit potential quality improvements. For example, many hospitals seek to participate in comparative studies with other hospitals (or regulators insist that hospitals participate in such studies) in order to identify areas in which they fall below some group threshold. Often the group average is used for that purpose. Such studies usually examine only outcomes, and track process-of-care factors only on rare occasion. The assumption, of course, is that the hospital will be able to identify areas of concentration—areas within which they fall below the group threshold, indicating a need for additional effort. But the (usually unspoken) corollary of such thinking is that, in areas where the hospital does not fall below the threshold, no effort is needed. Thresholds, with their concentration on the tail, thus have the effect of damping the added effort required to improve already above-average performance. They become ceilings (or floors) that limit excellence.

Thresholds also reinforce average performance by increasing the degree of scrutiny applied to rates or outcomes that fall far from the threshold on either side. For example, if a hospital demonstrates a postoperative deep wound infection rate near the 2 percent national standard, little additional attention is usually focused on the infection detection systems used at the hospital. But if a hospital posts a much better rate than the national threshold—for example, under 1 percent—then the hospital's infection detection systems are likely to be subjected to much closer review and criticism. Gilovich has documented other areas in which deviation from accepted norms leads to closer scrutiny and inappropriate attribution of cause, with the effect of incorrectly reinforcing the original norm (ref. 2 1). In a similar vein, Caplan and others demonstrated that physician peer reviewers were much more likely to judge that a patient's record showed inappropriate care if that record also showed a permanent adverse outcome, as opposed to their findings regarding the same patient record if it documented only a temporary adverse outcome (ref. 22).

It should be said in defense of the JCAHO that its publications explicitly state that its threshold-based standards represent minimum acceptable performance. They strongly urge hospitals to move beyond them. But for most hospitals, the minimum mandated effort is "good enough." To further

cite Berwick: "The dilemma is painful. On the one hand, improvement depends on learning from information about performance. Yet, on the other hand, that same information can easily be used to make and enforce judgments that Provoke fear and prevent learning. In many arenas, American culture seems addicted to forms of surveillance, measurement, comparison, judgment, and reaction that frighten people away from learning. In the health care system, we pay for that addiction in endless. wasteful debates about the fairness, accuracy, and meaning of measurements, in costly liability litigation, and in the erection of barriers among functions that ought to cooperate with each other. Instead of asking how we could be better, we spend our time and money proving we are good enough" (ref. 12).

Focus on the Whole Group

The alternative to "good enough" quality thinking is continuous quality improvement (Ql). QI is an organized system to continually improve processes, outcomes, and service, regardless of prior excellence, in order to "be the best we can be." Clinical practice improvement combines Ql's management philosophy with a scientifically-defensible measurement methodology (ref. 23). It uses those tools to manage clinical processes, systematically improve medical outcomes, and reduce medical resource use. To that end it propounds the following two axioms:

- Eliminate inappropriate variation (usually in process steps).
- Document continuous improvement (usually in outcomes).

The utility of these axioms becomes apparent when they are contrasted to traditional QA's threshold-based approach. The top half of figure 2 shows a theoretical "focus on the tail" system that worked perfectly. As shown in the figure, a traditional QA system selected and measured some quality indicator. It established a threshold; for the sake of illustration, assume that it marks a 5 percent tail. The QA system then concentrated all of its improvement resources on remedial actions aimed at those providers who fell in the tail (for example, it sponsored education programs or undertook sanctions). Assume that the remedial program was completely successful—the traditional QA approach entirely eliminated the tail. In the case of perfect success, how much did quality improve? With a 5 percent tail, the mean performance of the entire system will improve by approximately 5 percent, as shown by the dashed line that marks the center of the new distribution. Figure 2 also shows a "bad apple" provider, assuming that such providers exist, lying immediately adjacent to the threshold. The "bad apple" provider has taken advantage of random chance or a vigorous criticism of the threshold-based measurement system

to avoid sanctions. It represents those circumstances, experienced by any quality manager in a real-world setting, when a clinician attempts to hide in the statistical tail. Clinical QI redirects the focus of the quality improvement effort. Rather than concentrating its resources on the few who provide poor care, it directs the same resources at the many who provide good care—the 95 percent who entered medicine, and who come to work each day, primarily to provide excellent service to patients. Rather than looking for and attacking bad care, quality improvement attempts to identify and establish excellent care. To achieve that end, it eliminates inappropriate variation and documents continuous improvement.

As QI eliminates variation, the distribution narrows. The curve becomes tall and (compared to the original curve) slender. In order to document continuous improvement, Q1 uses the same amount of resources as used by the "bad apple" approach, but applies them to the entire curve, not just the lower tail, with an aim to institutionalize best practices, as opposed to eliminating bad practices. The combined effect of "eliminate inappropriate variation" and "document continuous improvement" is shown in the bottom half of Figure 2. Compare the quality improvement represented in the top half of the figure, under traditional QA, to that shown in the bottom half for QI.

While Figure 2 shows a theoretical case, Figure 3 shows the same effect in a real setting: a quality improvement project for total hip arthroplasty. Notice how the variation declines (the two dashed lines, representing variation in length of stay, come closer together) as time progresses. Notice too the shift in average performance. Figure 4 demonstrates the impact that the quality management project had on the procedure's cost—a not unexpected effect based on an understanding of cost's dependence on quality (ref. 1).

Figure 5 shows a second real example of variation narrowing as a group of physicians use credible clinical data to find best care (refs. 24, 25). It displays protocol compliance rates among 14 intensivists who generated an extensive flowchart (about 40 pages in length, with an average of 3 decision nodes per page) for the management of ventilator support for patients with severe adult respiratory distress syndrome (ARDS). The structure of the protocol-improvement process required that, if a clinician failed to follow the protocol at some point, the corresponding protocol element was automatically placed on the agenda for the next clinical team meeting. The group always started from the assumption that the protocol was incorrect or incomplete with regard to the clinical point under discussion. Their reasoning was that, if the protocol were correct, then all clinicians in the group would follow it. The clinician who had disagreed with the protocol step had an opportunity to present his or her reasoning to the group, in the context of a real patient, so that the group could modify

the protocol step or reach consensus that the protocol, as written, did represent best practice.

The team observed four preliminary results as variation decreased and they centered on "best care":

1. Physician time to manage these complex patients fell. Routine patient management activities were moved into the system, where physicians could think about them systematically, freeing physicians' time that could then be used to concentrate on critical patient management issues or to care for other patients.

2. For patients who survived, those treated with the stabilized protocol left the intensive care unit faster than those treated before the protocol was introduced.

3. Patient survival increased from historical rates of less than 10 percent to more than 40 percent.

4. The cost of caring for these extremely ill patients, when compared with the cost of the next most effective therapy ($ECCO_2$ R), for those who lived, fell from more than $160,000 per patient to about $120,000 per patient.

The theoretical QI example in the bottom half of figure 2 also shows the potential of clinical practice improvement to address "bad apple" providers, if such truly exist. As clinical QI narrows and shifts the quality distribution, it moves the tail. Focusing on the whole group may deal more effectively with "bad apples" than focusing on the tail. It slides the tail off them, leaving them nowhere to hide. Unless they change, "bad apple" providers become glaring outliers.

Anecdotally, two principles apply when a QI approach is used to slide the tail off physicians who might otherwise have been classified as "bad apples": (1) Physicians learn from physicians, and (2) the data speak for themselves. Said another way, as part of their medical training physicians are taught to recognize and respond to credible data (the data speak for themselves). They are also socialized, through the residency training process, to be very uncomfortable if their practice patterns are different from those of their professional peers without some defensible explanation (physicians learn from physicians). For example, within Intermountain Health Care (IHC), even though the clinical QI management team laid contingency plans to deal with potential "bad apple" physicians found through quality improvement projects, those plans have never been activated. When confronted with credible clinical data, potential "bad apple" physicians changed their practice patterns. Most continued to

practice in the lower tail of the quality distribution, but of the new curve. In fact, they showed more improvement than most of their colleagues. They had further to move. Finally, because the QI process produced a new, narrower curve with less variation, their practices were closer to those of their peers than they were before. Among more than 2,000 independent physicians who practice within IHC, the QI management team did not identify any "bad apples."

Where Does It Lead?

The final step in any quality improvement process is to communicate the resulting superior performance to consumers. Within health care it is almost an ethical issue: Who would seriously consider sending patients to some other, less capable, provider, as measured by outcomes? Providers who can consistently provide better medical outcomes and services deserve the patients, just as those patients deserve the excellent care that such providers can deliver. And, by taking advantage of quality's direct ties to cost, those same hospitals and physicians that are able to document improved outcomes will be able to charge less for the care they deliver. In an increasingly competitive environment, they will get the patients because they deserve the patients. They understand that being "good enough" is never good enough, Regardless of how well they do today, they will always strive to be better tomorrow—they will constantly redefine "the best they can be" in service to their patients.

References

1. James, B. C. *Quality Management for Health Care Delivery* (monograph). Chicago, IL: Hospital Research and Educational Trust (American Hospital Association), 1989.

2. McDonald, C. J. Protocol-based computer reminders, the quality of care and the nonperfectability of man. *New England Journal of Medicine.* 295(24):13525, Dec. 9, 1976.

3. Joint Commission on Accreditation of Healthcare Organizations. A*ccreditation Manual for Hospitals,* 199 1. Chicago: JCAHO, 199 1, p. 218.

4. Prevost, J. Interpretation and uses of clinical indicator data. *Agenda for Change Update: National Invitational Forum on Clinical Indicator*

Development, March 31, 1989. Chicago: Joint Commission on Accreditation of Healthcare Organizations, 1989. Oral conference presentation (Section 4).

5. Berwick, D. M. Continuous improvement as an ideal in health care. *New England Journal of Medicine.* 320(1):53-56, Jan. 5, 1989.

6. Scherkenbach, W. *W. The Deming Route to Quality and Productivity Road Maps and Roadblocks.* Washington, DC: CEE Press Books, George Washington University, 1991, (11th printing), p. 7 1.

7. Davidson, L., and Thompson, J. VA hospital in S.L. gets low marks in investigation. *Deseret News. Nov.* 20, 199 1.

8. Thompson, J. VA hospital outraged over media image. *Deseret News* Nov. 2728, 1991.

9. Iezzoni, L. I., Burnside, S., and others. Coding of acute myocardial infarction: clinical and policy implications. *Annals of Internal Medicine* 109:745-51, Nov. 1, 1988.

10. Hsia, D. C., Krushat, W. M., and others. Accuracy of diagnostic coding for Medicare patients under the prospective-payment system. *New England Journal of Medicine.* 318(6):352-55, Feb. 11, 1988.

11. Simborg, D. W. DRG creep: a new hospital-acquired disease (sounding board). *New England Journal of Medicine.* 304(26):1602-4, 1984.

12. Berwick, D. M. The double edge of knowledge (editorial). *JAMA.* 266(6):841-42, Aug. 14, 199 1.

13. Harris, J. E. Defensive medicine: it costs, but does it work? (editorial). *JAMA.* 257(20):2801-2, May 22/29,1987.

14. Kapp, M. B. Letter to the editor. *JAMA.* 258(9):1176, Sept. 4, 1987.

15. Weinstein, M. C., Fineberg, H. V., and others. *Clinical Decision Analysis.* Philadelphia: W. B. Saunders Company; 1980, pp. 79-92. 112-13.

16. Blumberg, M. S., and Binns, G. S. *Risk-Adjusted 30-Day Mortality of Fresh Acute Myocardial Infarctions: The Technical Report* (monograph). Chicago: Hospital Research and Educational Trust (American Hospital Association), 1989, pp. 42-45.

17. Rolph, J. E., Kravitz, R. L., and McGuigan, K. Malpractice claims data as a quality improvement tool. 11. Is targeting effective? *JAMA.* 266(15):2093-97, Oct 16, 1991.

18. Park, R. E., Brook, R. H., and others. Explaining variations in hospital death rates: randomness, severity of illness quality of care. *JAMA.* 264(4):484-90, July 25, 1990.

19. Green, J., Passman, L. J., and Wintfeld, N. Analyzing hospital mortality: the consequences of diversity in patient mix. *JAMA.* 265(14):1849-53, Apr. 10, 1991.

20. Berwick, D. M., and Wald, D. L. Hospital leaders' opinions of the HCFA mortality data. JAMA. 263(2):247-49, Jan. 12,1990.

2 1. Gilovich, *T. How We Know What Isn't So: The Fallibility of Human Reason in Everyday Life.* New York City: The Free Press (A Division of Macmillan, Inc.), 1991, pp. 49-87.

22. Caplan, R. A., Posner, K. L., and Cheney, F. W. Effect of outcome on physician judgments of appropriateness of care. *JAMA.* 265(15):1957-60, Apr. 17, 1991.

23. James, B. C., Horn, S. D., and Stephenson, R. A. Management by fact: the relationship of quality improvement to outcomes management, practice guidelines, and randomized clinical trials. (In press).

24. Morris, A. H., Wallace, C. J., and others. A computerized protocol-controlled randomized clinical trial of new therapy including PCIRV and extracorporeal C02 removal for ARDS. 1. Clinical trial results. (In press).

25. Henderson, S. E., Crapo, R. 0., and others. Computerized clinical protocols in an intensive care unit: How well are they followed? *Proceedings of the Fourteenth Annual Symposium on Computer Applications in Medical Care (SCAMC).* Los Alamitos, CA: IEEE Computer Society Press; 1990, pp. 28488.

Figure 1. A theoretical example of traditional quality assurance methods that use thresholds. (See page 20.)

The abscissa (X-axis) represents some quality measurement—for example, the wound infection rate for a group of providers. A lower rate (to the left) is better. The curve shows the number of providers that achieve each particular rate. For the sake of demonstration, it is assumed to be perfectly risk-adjusted, and to follow a bell-shaped curve. But the same arguments apply for any relatively continuous curve.

Traditional quality assurance usually establishes a performance threshold. Cases or providers that fall beyond the threshold are subject to intensive investigation or other, more active, disincentives, while those that do not exceed the threshold are usually ignored.

Point A represents a provider whose performance falls quite close to the threshold but still in the region considered "acceptable." It may actually represent unacceptable performance that was carried across the threshold by random measurement error. Point B represents a provider whose performance is significantly better than most other providers shown on the graph.

Figure 2. A theoretical example of the comparative effect of successful quality assurance and quality improvement programs. (See page 21.)

As with Figure 1, each curve represents a theoretical quality outcome—for example, an infection rate. For the sake of illustration the outcome is assumed to be perfectly risk-adjusted and to follow a bell-shaped curve. although the same arguments hold for any relatively smooth curve. The solid center line represents the original ("before") average of each distribution. The vertical dashed lines show the average of the new distribution that results after traditional quality assurance or quality improvement have successfully been applied.

The round dot, just to left of the threshold line in the upper two graphs, ostensibly represents a "bad apple" provider. The dot is at exactly the same position in all four graphs, comparing the potential effect of traditional quality assurance techniques versus quality improvement techniques with regard to "bad apples" who attempt to hide in a quality distribution's statistical tail.

Figure 3. Stabilization of one process-of-care factor (length of stay) for nonfracture total hip arthroplasty at one hospital. (See page 22.)

Eligibility criteria were used to form a cohort of comparable patients in terms of presenting disease, complications, and short-term medical outcomes. Each point represents the average length of stay during one month. The solid center lines show the average length of stay for 1988 (10.0 days). 1989 (7.50 days), and the first three months of 1990 (7.26 days). The upper and lower dashed lines are set at plus and minus two standard deviations in per-case length of stay, by year. An initial study comparing practice patterns for comparable patients was shared with the hospital's orthopedic surgeons in a subspecialty medical staff meeting in early February 1988. In that study, physicians were identified only by blinded codes. Each physician was given a sealed letter that contained his or her unique code, so that each could recognize his or her own practice

measurements within the group. Following the initial meeting, follow-up comparative practice pattern data were mailed to each physician every three months, with all physician identities blinded except that of the physician for whom the report was prepared.

In 1989 case-to-case variation in length of stay fell to about half the level displayed in 1988. While average length of stay declined between 1988 and 1989, during 1988 some surgeons already demonstrated performance similar to the 1989 rate. The change in length of stay between 1988 and 1989 reflected the entire group of surgeons coming to a common rate, already demonstrated by some members of the group, rather than a new level of performance for all members of the group.

By 1990, statistically significant differences in length of stay among the surgeons could no longer be found, although that measurement may have been affected by the relatively small sample size represented by only three months of cases.

The complete quality monitor tracked many more process and outcome factors than are shown here. Included among them were medical outcomes, as reflected in complication rates, number of physical therapy sessions, distance walked at discharge, pain relief, and functional status. Major short-term patient outcomes remained stable over the time period shown in the study.

Figure 4. Per-case hospital resource consumption for nonfracture total hip arthroplasty at one hospital. (See page 23.)

Length of stay was only one among several factors that the hospital's clinicians examined, stabilized, and improved during the course of the project. Others included the type and timing of physical therapy, the use of transitional care units, and standardization of prostheses and operating room techniques.

Figure 5. Percentage of protocol-based recommendations followed for adult respiratory distress syndrome (ARDS) patients. (See page 24.)

Starting with the first patient (patient number 29, admitted to LDS Hospital's pulmonary ICU on August 14, 1988) for whom the protocol was applied, then tracking the next 30 consecutive patients (through patient number 59, admitted on January 1, 1990). A typical treatment episode involved about 208 protocol-based treatment recommendations. About four months elapsed between patient number 29 and patient number 37. From patient number 37 on, most protocol noncompliances occurred (1) when a patient was removed from the ICU (usually for either surgery or imaging), (2) as further improvements to the protocol were tested, or (3) as a consequence of the fact that few protocols are perfect-nearly all guidelines

show some level of random noncompliance as clinicians address patient factors not anticipated by the protocol or factors that are so rare as to not justify inclusion in the protocol.

(Figure modified from Henderson and others (ref. 25).

Figure 1. A theoretical example of traditional quality assurance methods that use thresholds.

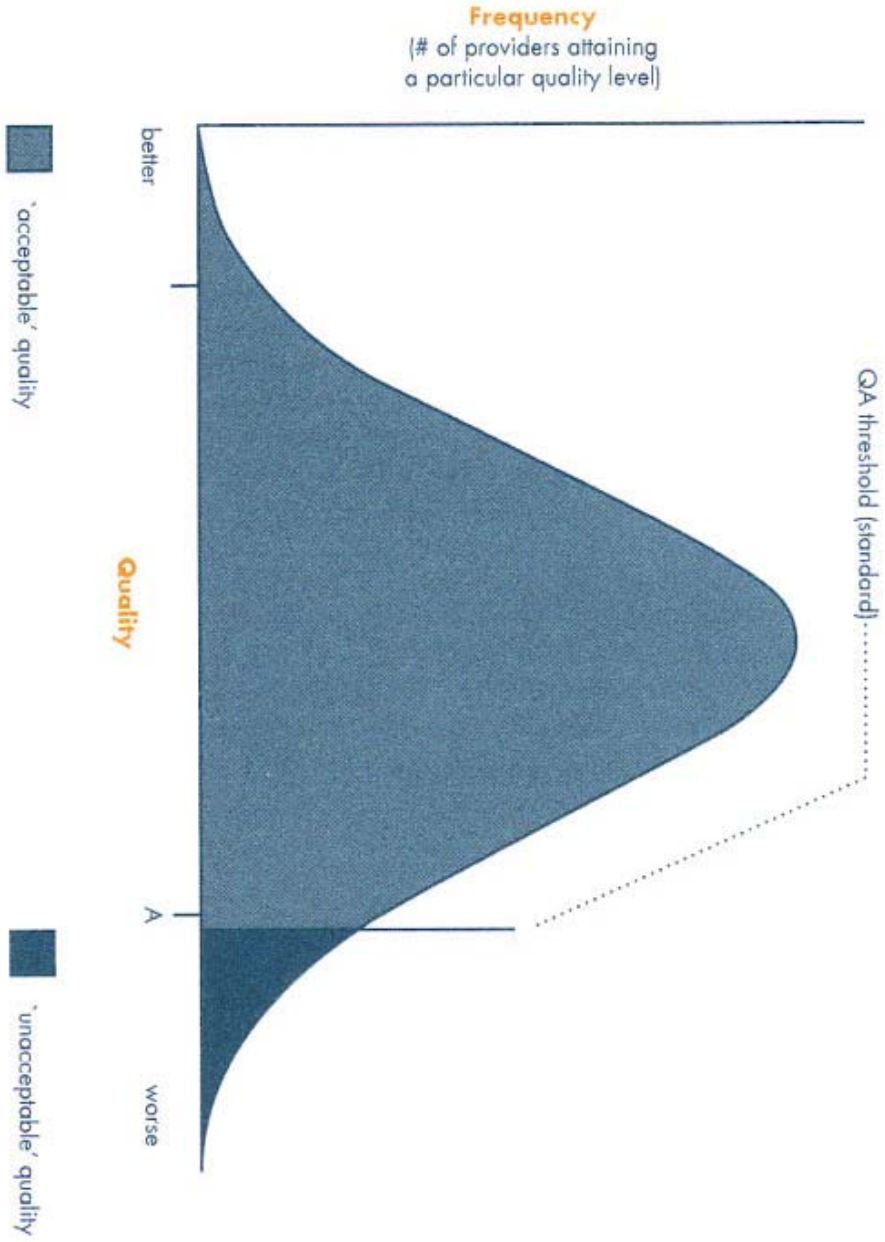

Figure 2. A theoretical example of the comparative effect of successful quality assurance and quality improvement programs.

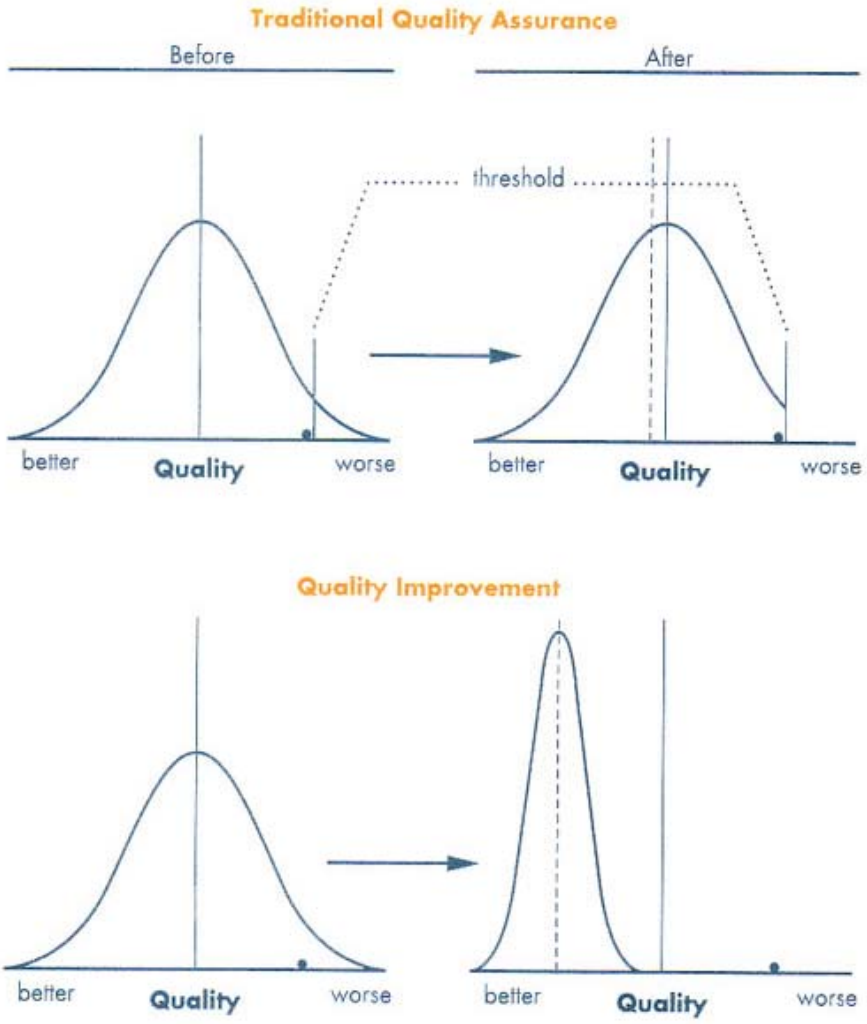

Figure 3. Stabilization of one process of care factor (Length of stay) for non-fracture total hip arthroplasty at one hospital.

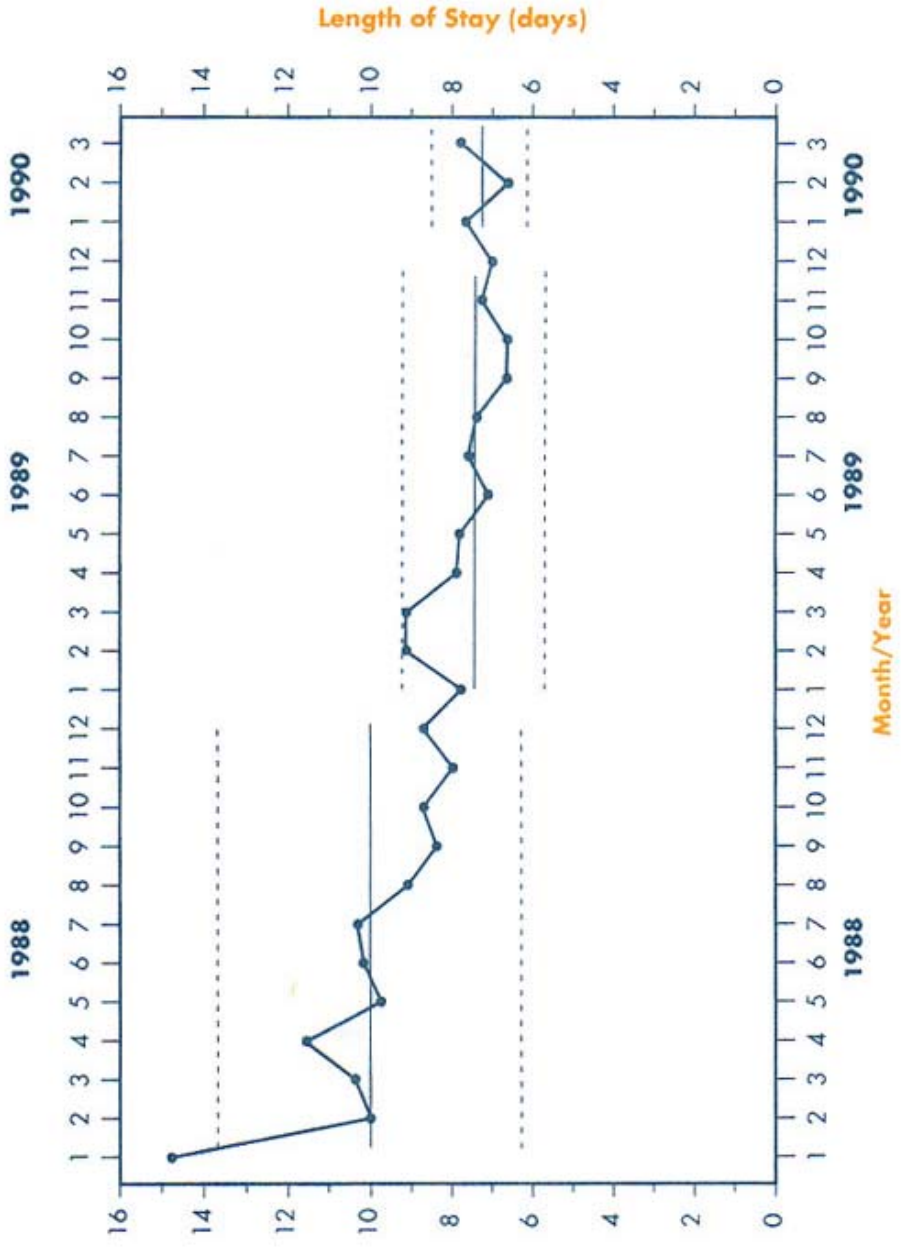

Figure 4. Per case hospital resource consumption for non-fracture total hip arthroplasty at one hospital.

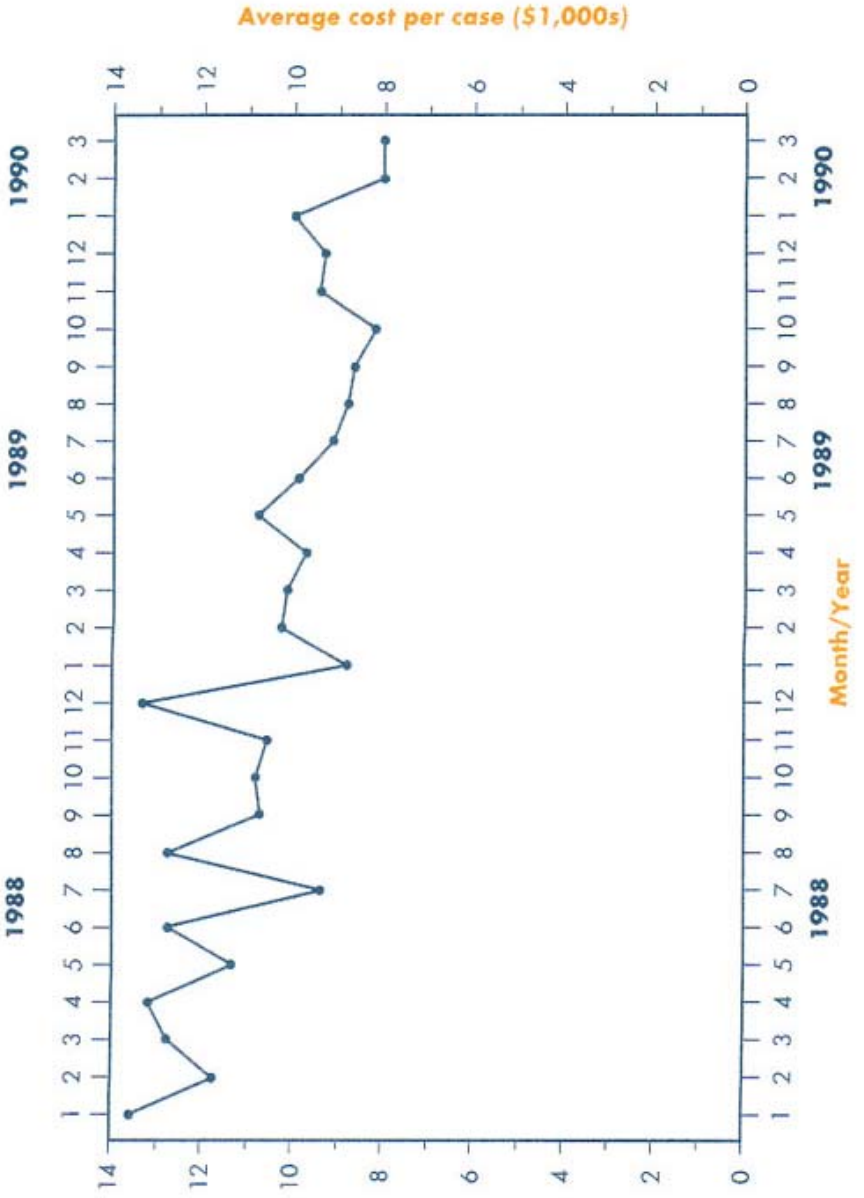

Figure 5. Percentage of protocol-based recommendations followed for Adult Respiratory Distress Syndrome (ARDS) patients.