Cost and Quality Under Managed Care: Irreconcilable Differences?

Eugene Litvak, PhD; and Michael C. Long, MD

Abstract
Managed care companies contend there is still waste in the healthcare system that should be eliminated. Healthcare providers argue that further cuts will reduce quality. Which side is right? In order to answer this question it is necessary to determine the threshold implicit in the corollary question: How far can we go in reducing healthcare expense without diminishing quality? A new variability based methodology is proposed that has the potential to determine the threshold at which cost reduction will negatively impact quality. Illustrations of its specific application are provided.

(Am J Manag Care 2000;6:305-312)

The healthcare delivery system in the United States is in transition. The “good old days” of fee for service for physicians and cost-plus reimbursement for hospitals are being replaced by various forms of managed care. All healthcare providers are now under intense financial pressure, most recently as a result of the federal Balanced Budget Act, which has changed many hospital operating budgets from being comfortably positive to hemorrhaging red ink. According to the American Hospital Association (personal communication, 1999), the percentage of hospitals with negative Medicare margins has increased from 46% in 1997 to 57% in 1998 and is expected to reach 70% by 2002. Operating margins at hospitals in Massachusetts have moved from +1.6% in 1996 to -2.6% in the second quarter of 1999.1 We cannot go back, but do we really know how to go forward? Our greatest challenge in health care today is to maximally reduce cost without reducing quality.

Limitations of Cost-Effectiveness Analysis
The current gold standard for relating cost and quality is cost-effectiveness analysis. This type of analysis provides data on the incremental clinical cost per additional life-year, or quality-adjusted life-year, gained by the drug, intervention, medical technology, or policy under consideration.2,3 Cost-effectiveness analyses are necessary, but not sufficient, to answer the question, How much quality can you buy for a dollar? The following example illustrates the problem: a person with a limited budget needs to buy a washer and a dryer. His concerns are cost and quality as measured by capacity, size, performance, reliability, and other factors. Imagine that he investigates all available washers, reads all the consumer literature on washers, and then buys the best washer and the first dryer he sees. You would likely question his approach. This is the situation in healthcare, however, when cost-effectiveness analyses focus exclusively on clinical cost and benefit without taking into account the effect of management decisions on cost. One can argue that management and clinical care are apples and oranges and should not be integrated, but in reality there is only one healthcare “pocket” to pay for

From the Harvard Center for Risk Analysis, Harvard School of Public Health, Boston, MA (EL) and Massachusetts General Hospital, Boston, MA (MCL).
Address correspondence to: Eugene Litvak, PhD, Harvard Center for Risk Analysis, Department of Health Policy and Management, Harvard School of Public Health, 718 Huntington Avenue, Boston, MA 02115. E-mail: litvak@hsph.harvard.edu.
argue that there is still a lot of extra “fat” in budgets, accompanied by a reduced budget leads to diminishing quality of care. Managed care companies have not included a rigorous analysis of management, the poor cousin of cost-effectiveness analysis. The first reason is that until recently, the healthcare industry has never been business oriented. It has been accustomed to spending with limited budgetary oversight, seeking to preserve or enhance perceived quality. Even the introduction of global payments to hospitals through diagnosis-related groups did not decrease the rate of growth in healthcare cost. The introduction of managed care in the early 1990s did have some early success in decelerating the rate of growth in healthcare cost primarily through more effective purchasing strategies. Unfortunately, core issues relating cost and quality have not been adequately addressed, and the cost of healthcare has again begun increasing in the past few years. In this environment, operations research methodologies, widely used in many other industries (including banking, insurance, manufacturing, transportation, military, and telecommunications) to relate operational cost to service quality and to decrease costs, have been virtually ignored.

The second reason is more technical. Optimal management decision making is a new area for the healthcare industry, and the consultants on whom the industry relies have little direct experience in the field. It is also technically difficult to measure the cost and quality consequences of most healthcare management decisions. As a result, optimal management decision support systems are rare. This difficulty in integrating the effect of management decisions on the cost versus quality equation is both a problem and an opportunity. It currently prevents healthcare institutions from being globally cost effective but at the same time provides the increasingly important possibility of satisfying consumers’ expectations to simultaneously decrease cost and improve quality. It also makes for a complex answer to the provocative question, How much money can be squeezed out of healthcare providers?

What Is Enough?

This question is the main focus of attention in many current healthcare debates. Healthcare providers argue that global payments or capitation accompanied by a reduced budget leads to diminished quality of care. Managed care companies argue that there is still a lot of extra “fat” in budgets, which can be reduced by efficient management. Which side is right? The answer is both—and neither. In order to answer this question, one has to first determine the threshold implicit in the corollary question: How far can we go in reducing healthcare expense without diminishing quality? When this question is posed, further progress is usually impeded by our inability to decide what constitutes quality healthcare. The intrinsic, multidimensional nature of quality has made it difficult to develop a consensus on its definition or measurement. Fortunately, a comprehensive definition of quality is not necessary to determine the threshold at which cost and quality are positively related in a specific system. It is necessary, however, to either establish the required quality of care for the system under consideration or to at least accept the current historical level of quality as adequate. Once this threshold relating cost and quality is determined, the provider must then determine the optimal (ie, least expensive) way of managing hospital departments and physician offices within the quality-of-care constraints in order to reach the threshold. The debate among healthcare providers and managed care companies cannot be satisfactorily concluded as long as we cannot reach this threshold. Its importance has been widely underestimated by both sides. For healthcare providers, it means a continuing inability to manage efficiently. For managed care, it means setting cost standards under which the current ability to manage leads to diminished quality of care. Managed care organizations cannot just decrease payments to providers without teaching them how to provide quality care with reduced resources. To address these issues, we have developed and applied variability-based operations research methods that can be used to determine the threshold at which quality and cost are positively related and to provide specific management interventions to reach this threshold in any healthcare delivery system.

Before describing our variability based methodology, however, let us review some status quo management attempts to reach the cost and quality threshold.

Current Methods of Cost Reduction

Suppose you are a healthcare provider or manager in charge of cost reduction. What would you do if you do not completely understand your system and do not have the ability to reengineer it? Your attempts at cost control might include the following:

Negotiate lower prices for materials (“buying cheap gloves”). This simple, reasonable step can lead to substantial savings without affecting quality
of care. Extensively used to date, this step is unlikely to negatively affect the financial interests of healthcare providers or the managed care organization.

Fire vulnerable staff whose performance does not have an immediate noticeable effect on quality of care (“firing the cleaning staff”). This step is usually a crisis reaction when you desperately need to decrease your budget. This step may or may not reduce the quality of care, depending on your ability to “feel” the consequences. Remember that you do not have a tool to determine who, if anyone, really needs to be fired.

Cut the budget by intuition (“managing by feeling”). In many instances, this step can produce larger errors than across-the-board budget cuts. Some people believe that their experience gives them such a feeling, but optimal management decisions are often counterintuitive. Consider the following hypothetical situation: 2 physicians, Drs. A and B, share 2 office examination rooms, and some patients must wait an unacceptably long time to be examined. The physicians decide intuitively that adding an extra examination room would decrease the waiting time. The decision is costly but worth it to reduce the waiting time. Will it work? Not necessarily. Suppose the reason some patients are waiting is that Dr. A books the examination rooms 5 minutes before Dr. B. Dr. A’s patients arrive and occupy the rooms while Dr. B and his patients wait. If the same scenario of appointment booking and patient arrival is carried over to 3 rooms, then Dr. B and 3 patients, rather than 2 patients, will wait. The net effect is that increasing the number of examination rooms increased the number of patients waiting and increased the cost. Unfortunately, this scenario is not purely hypothetical. The interdependence between the 2 physicians is critical and becomes much more complicated in real life when 3 or 4 physicians or more are sharing common office facilities.

Hire management consultants (“the blind leading the blind”). Consultants can provide important advice to managers and are widely used in the healthcare industry. Managers have many reasons for using consultants, some of which include to take responsibility for wrong decisions, to compensate for their inability to formulate the problem, and to benchmark. Reasons such as these have led to the hackneyed definition of a consultant as one who “borrows your watch and then tells you the time.”

A primary reason to use a consultant should be to find the threshold at which further cost control will compromise quality. To date, consultants have been unable to accomplish this goal. When they do have the tools to determine this threshold and the methodologies to reach it, they will be able play a truly important role in helping healthcare institutions achieve maximal efficiency.

Promote clinical pathways (“following the yellow brick road”). The standardized approach to delivery of care inherent in clinical pathways does show merit in reducing waste and improving the quality of care delivered to some homogeneous groups of patients. Clinical pathways are not the goal, however, but merely a vehicle and cannot be applied to all patients. When patients are inappropriately placed on pathways to satisfy administrative goals, quality of care is at risk. In addition, the standardized approach of clinical pathways risks dragging down the performance of the most gifted caregivers and may stifle the clinical innovation so important to medical progress.

Having discussed the limitations of current approaches to cost control, let us now consider a new methodology based on variability analysis—one that first diagnoses the healthcare provider’s “cost disease” and then provides practice-specific treatments.

Variability Is the Universal Key

Let us consider a healthcare system without variability. Suppose all patients are homogeneous in disease process. That is, they all have the same disease, the same degree of illness, and the same response to therapy. Suppose they all appear for care at a uniform rate. Furthermore, suppose all medical practitioners and healthcare systems have the same ability to deliver quality care. In this best of all situations, it would be possible to achieve 100% efficiency in healthcare delivery. There would be no waste. Cost would be minimal and quality maximal within the boundaries of knowledge and technology. It would be easy to satisfy the goal of managed care—to provide the right care, to the right patient, at the right time.

In the real world, healthcare systems are expected to deliver quality care for patients with many different types of disease. Patients with the same disease exhibit significant differences in their degree of illness, choice of treatment alternatives, and response (clinical variability). They also usually appear for care in random fashion with different means and standard deviations of arrival rates (flow variability). In addition, medical practitioners and healthcare delivery systems are not uniform in their ability to provide the best treatment (professional variability). The constant challenge to the healthcare system is to efficiently convert a naturally variable incoming group of sick patients into a homogeneous
outgoing stream of well patients. The presence of these “natural” clinical, flow, and professional variabilities increases complexity and adds cost to the healthcare system. The goal then is to optimally manage natural variabilities. However, dysfunctional management often leads to the creation of a fourth type of variability—“artificial”—that unnecessarily increases the very cost and inefficiency we are trying to control. Let us give one of many examples of artificial variability in healthcare delivery systems.

One common problem at hospitals is extreme variation in daily bed occupancy. On days when occupancy is too high, quality of care decreases because it is too costly to staff for peak loads. On days when occupancy is too low, there is waste. No staffing system can be flexible enough to optimally manage these daily fluctuations. It is reasonable to assume that these variations in occupancy are related to a combination of the natural clinical variability of the patients’ response to therapy and the natural flow variability of their admission through physician offices or the emergency room. Surprisingly, this assumption is only partially correct. An additional source of admission and occupancy variability in many hospitals is through the operating rooms. Typically, 80% or more of this variability from the operating rooms is due to variations in the elective scheduled daily caseload. The variability is not related to unexpected changes in the operating room day from unscheduled emergencies, cancellations, or additions. It is artificial variability introduced into the system by the advance elective surgical scheduling process. Not only are there significant variations in the elective caseload among each day of the week but as much as a 50% difference in caseload on the same day of the week. Compared with natural variability, artificial variability is nonrandom. Yet it also is unpredictable, driven by numerous competing demands on the surgeons’ time that are usually unknown and therefore unaccounted for by the healthcare system. The net effect of this operating room artificial variability on occupancy variability may be greater than the effect of the variability of admissions through physician offices or the emergency room. In more dramatic terms, the predictability of the number of admissions to the hospital on any day from elective scheduled surgery may be worse than the purely random appearance of patients for emergent admission through the emergency room.

That variability is an obstacle to efficient delivery of healthcare has been previously appreciated.\(^5\) However, the innovative methodology proposed here to analyze the types and amounts of variability present in healthcare delivery systems and then to eliminate or optimally manage them gives us the potential to overcome it. It also gives us the answer to the previously posed contentious question, How much money can be squeezed out of healthcare providers?

The answer is that we can and should eliminate all system expense resulting from artificial variability in healthcare delivery. Using operations research methods, we must then optimally manage the remaining natural variabilities. Some of these methodologies may be borrowed from other industries where they have already been shown to be effective, but many will be unique because of the specifics of the healthcare system (see survey in Pierskalla and Brailer\(^6\)). We will then have both determined and achieved the threshold at which further attempts at cost reduction will invariably reduce quality. The irreducible floor of natural variability that must be accommodated is unique to each healthcare delivery system, and should be one factor used to determine reimbursement by payers. At this threshold, cost-effectiveness analysis, including the cost of management decision making, will be necessary to guide healthcare expenditure. Until this threshold is achieved, however, it will still be possible to decrease cost and simultaneously maintain or even improve quality.

An extensive discussion of variability methodology and its application across the many diverse models of healthcare delivery is beyond the scope of this paper. We can, however, illustrate the basic principles and their application in a brief analysis of 1 hospital system already cited for inefficiency—the operating room.

**An Example of Applying Variability Analysis**

The first principles of variability analysis, whether applied to a specific department or hospitalwide are to identify, classify, and measure variabilities. Hospitals experience substantial variability in the type of disease and the severity of illness of patients admitted for surgery. In most hospitals, all types of surgery are performed, including general, orthopedic, and cardiac, on patients ranging from those in good health to those near death. Some cases are elective and primarily performed during regular weekday hours, and some are urgent or emergent and performed any time of day. Hospitals also vary in caregiver ability, manifested primarily by the addition of resident house staff, medical students, and nurses-in-training to the surgical team.
These variabilities can be conveniently organized for further analysis into the 3 broad categories described previously (ie, clinical, flow, and professional variabilities). For the operating rooms, different diseases and severity of illness would be classified as clinical variability, elective versus urgent/emergent arrivals would be flow variability, and differing levels of expertise among providers would be professional variability.

In our variability analysis, the next step is to measure each variability as a deviation from an ideal, stable pattern. Measurement would be different for each type of variability and unique to the system being measured. For example, variability in disease severity would be measured as the deviation of health status from perfectly healthy, while variability in the flow of elective surgical procedures would be measured as a deviation from the mean daily caseload. When measuring the variability of a system, the total variability is not necessarily the sum of its parts, because they are usually mutually dependent. In the operating rooms, for example, if 2 services compete for operating time in the same rooms, then the total variability in caseload for those rooms is not the sum of the individual variabilities in flow for each service.

The next step is to eliminate all artificial components of each of the variabilities identified. Artificial variabilities are generally flow or professional variabilities caused by a dysfunctional process within the healthcare delivery system. These variabilities are usually easily identified during a thorough system operations analysis. For example, if a system receives steady input but the output is variable, there is artificial flow variability in the system. Another important sign is that any variability that is not predictable and at the same time is nonrandom should be measured as a deviation from the mean daily caseload. When measuring the variability of a system, the total variability is not necessarily the sum of its parts, because they are usually mutually dependent. In the operating rooms, for example, if 2 services compete for operating time in the same rooms, then the total variability in caseload for those rooms is not the sum of the individual variabilities in flow for each service.

The first step in managing natural variability is to identify homogeneous subgroups that may provide a basis for dividing the system into similar parts. These subgroups are then functionally separated, and optimal management scenarios are developed for each.

For the operating room, we have a number of potential ways to subdivide its patient population into fairly homogeneous groups. We have at least 2 flow groupings, elective and emergent/urgent. We could also group patients by disease type, such as cardiac, orthopedic, or neurologic. We could group them by disease severity, dividing groups into complex surgeries in sick patients and minor surgeries in healthy patients. Another division could be teaching versus nonteaching cases. Each time we identify a homogeneous subgroup and develop a strategy to optimally manage it, we have an opportunity to increase efficiency. For example, most hospitals currently separate complex cases, such as open-heart surgery, from simple ones, such as ambulato-
ry inguinal herniorrhaphy. They intuitively recognize that it is easier to achieve optimal professional performance if they have specialized operating room staff in dedicated rooms to care for these different types of patients. Theoretically, we could continue this separation process to the extreme endpoint of providing a dedicated team and operating room for each patient, standing by, immediately ready to provide optimal care. We stop substantially short of this because of the imposition of severe flow variability. The low arrival rate of any specific patient for surgery by his team in his operating room would mean that most of the team’s clinical time would be wasted. The overall effect of this severe flow variability would be to significantly increase cost and more than offset any increase in efficiency or quality from optimal staff performance.

Average flow and variability in flow to each of the homogeneous subgroups under consideration will be different for each hospital and will determine whether that potential subdivision of the operating room is likely to increase overall efficiency. For example, separation of ambulatory surgical operating rooms into a dedicated building would be a rational design only for a hospital that has a high, steady demand for this type of surgery. The conflict between flow and separation will result in an optimal and unique design for managing the operating rooms in individual hospitals.

A typical redesign of a large city hospital would likely include a main operating room area with a number of rooms in which all types of major surgeries are performed and a functionally separate ambulatory area for outpatient surgery. If the hospital had an active emergency room, providing separate rooms in the main operating room area to care for this randomly appearing patient population would be imperative. If patients appearing on an urgent or emergent basis are forced to compete with elective scheduled patients for the same rooms, it will be impossible for the main operating room area to simultaneously satisfy the performance objectives of high utilization and low waiting time for these different patient groups. This same problem exists in most emergency departments as well. After separating the main operating room into scheduled and unscheduled areas, subdividing by type of surgery would probably increase efficiency through decreased turnover time or shorter case length. The number and types of these specialized units would be determined by analysis of the tradeoff between increasing performance by reducing professional variability with specialization and decreasing performance from lower utilization caused by low average flow or high variability in flow to the specialty units.

For a small community hospital with primarily outpatient surgery and no emergency room, a design mixing inpatients and outpatients undergoing all types of surgery in a highly flexible set of operating rooms may provide the optimal tradeoff between patient flow and professional specialization.

Once the separation of the healthcare delivery system into homogeneous groups has been accomplished, the final step is to optimally manage them. One impediment to doing so has been developing methodology to allocate resources to those groups experiencing natural flow variability. In the main operating room, for example, we need to determine how many rooms to devote to an unscheduled area for emergent and urgent patients. Well-described operations research methodologies using queuing theory designed for random flow rates can be used to determine optimal resource allocation to these groups. This methodology has rarely been applied in healthcare practice but is widely used in nearly every other industry to optimally allocate resources to random demand.

Applying variability methodology at the hospital unit or departmental level is necessary but not sufficient for maximal cost reduction in a healthcare delivery system. We also need to conduct a systemwide analysis.

Systemwide Analysis. Suppose you consistently observe a traffic jam outside your window. You propose to avert future jams by widening the road. If the true dynamics of the traffic flow are not apparent to you, however, and the problem is a constriction at a distant exit, you will only worsen the jam at your area. To achieve maximal cost effectiveness in healthcare, we must understand the complete dynamics of patient interaction with all components of the delivery system and their mutual interdependencies. Much of the artificial variability in healthcare that is costly and should be eliminated is caused by poorly understood interdependencies between different hospital departments and healthcare professionals, who are simultaneously contributing to the delivery of healthcare. Simulation tools for modeling such interdependencies can be developed for the healthcare industry using network structures already extensively applied in many other areas.
Discussion
Inefficient management of the operating rooms in many hospitals is just one example of the thousands of healthcare business practices that lead to artificially inflated cost and sometimes to a lower quality of care. In recent years, appreciation of this waste in healthcare has led to the concept of managed care. One could expect that managed care would eliminate mismanagement scenarios and provide healthcare organizations with the necessary management tools and knowledge, thereby justifying the term “managed.” However, this does not appear to be the case. Managed care companies have no more management expertise than do healthcare practitioners or, if they do, are unwilling to share it.

The historical, pre-managed care reasons for the current healthcare situation are clear—unlimited expense accounts for providers and ignorance of business practices available to increase efficiency. Under managed care, providers are no longer reimbursed on a fee-for-service basis. Lack of knowledge of management methodologies, however, still prevails in managed care organizations. This combination of cost control and ignorance of optimal management decision making inexorably decreases quality of care as the industry struggles with inflation and the ever-increasing cost of technology.

How to solve this problem? Research to develop new management technologies and application of those already used in other industries to specific healthcare delivery problems needs to be supported. No one seems interested in doing so, however. Managed care companies have a clear role and only 2 choices to manage healthcare cost. They can do it through medical decision making or management decision making. In either case, managed care organizations become involved in the providers’ business. Medical decision making has been chosen so far because it seems technically (although not politically) easier to do than management decision making. This choice has been made even though it is economically much less beneficial than optimal management decision making and in some cases is in conflict with quality of care (management decisions are not). This conflict is an intrinsic part of this choice. Even when proper medical decisions have been made, mismanagement artificially drives costs up, and any attempt to decrease the cost without proper management will decrease quality of care. This fact has not yet been widely recognized. Sooner or later (sooner is cheaper both politically and economically), managed care organizations will realize it is in their best interest to help healthcare providers keep operational cost low through optimal management. We are unaware of any current effort by managed care companies to support research to develop methodologies to achieve this goal. Thus far, they prefer to blame healthcare providers for their inability to manage and show no interest in helping them manage instead of helping them treat their patients.

Similarly, most providers show little or no interest (or currently have no resources) in investing in healthcare management research. They cite the numerous examples of decreased quality from cost control in an attempt to return to the “good old times” of practically unlimited resources.

When the fragmentation of healthcare and conflict among competitors obstruct research and put quality of care at risk, we should look to the government to intervene. Because it is in the public domain, government-sponsored research is available to benefit all parties. Does it do so? To the best of our knowledge, the answer is no. Research support has been focused on benchmarking current healthcare delivery systems rather than developing new management strategies. Why? Because a comprehensive definition of quality in healthcare delivery is perceived to be a prerequisite to further efforts at cost control. The intense national focus on achieving consensus on this nearly insolvable problem has diverted attention and support from optimal management decision making. Must we continue to expend resources until we can define quality or should we work on decreasing management cost in parallel?

Even if interested, universities cannot do this research alone because it needs to be very practical and healthcare specific. It must, therefore, be a sponsored, collaborative effort between managed care companies and healthcare providers.

Conclusion
The results of variability analysis indicate that the goals of decreasing costs while maintaining or even improving quality of care are not irreconcilable. The application of variability-based methodology is in its infancy but has the potential to distinguish effective cost-control interventions in healthcare from those that would only waste money or even damage the healthcare system. More importantly, variability analyses can be used to determine the threshold at which further attempts at cost reduction will compromise quality of care. Specific application of variability analysis to current healthcare systems suggests that there are continuing opportunities to
reduce waste and that we have not yet reached the threshold at which there must be a decrease in quality to achieve further cost control. Eliminating artificial variabilities in healthcare systems has the most potential for reducing waste. It requires a commitment to change, however, and acknowledgment by providers that eliminating these variabilities will reduce their flexibility and convenience. Current efforts to implement these new management methods are impeded by the lack of an established track record of their use in healthcare, a lack of understanding by current managers, and the need for consensus among multiple hospital governing bodies with divergent objectives.

Significant variability in healthcare delivery is inevitable because of the changing nature of disease, the availability of new therapies, and the wide variety of patients’ psychological and physical responses. Unless we develop healthcare delivery models that can respond to this variability, we will never be able to maximize operating efficiency and quality. Because managed care as a healthcare delivery model is here to stay, we have no choice but to develop and use these new tools to reassure our patients that we know how to efficiently manage their care in a way that delivers the highest possible quality at the lowest possible cost.

Acknowledgments

We are indebted to Donald M. Berwick, MD from the Institute for Healthcare Improvement and Milton C. Weinstein, PhD, from the Harvard School of Public Health for their thoughtful comments and suggestions. We are especially grateful to the editor and reviewers for their recommendations.

REFERENCES