

Hospitals get serious about **operations**

Paul D. Mango and Louis A. Shapiro

The income statements of hospitals have been ailing. The cure?
Serious attention to operating efficiency.

Walk into most hospitals in the United States today, and you observe a true logistical anachronism. Patients arrive in the admissions area at 5:00 in the morning, only to wait two hours before they are checked in and two more before someone preps them for surgery. At lunchtime, the traffic jam spills over into the operating rooms, where patients routinely arrive late because of the admissions delays. The surgeons, anticipating this, come later than scheduled for operations in order to avoid wasting time. By midafternoon the bottleneck has shifted to the recovery area and the intensive care unit (ICU), forcing groggy patients to wait back in the operating rooms. Frederick Winslow Taylor was worrying about inefficiencies of this kind in factories a century ago.



To understand why such problems persist in health care, you must go back at least to the mid-1980s. In those days, insurers still paid whatever fee hospitals demanded, and federal and state governments still subsidized the expansion of capacity by adding a “capital pass-through” term to their Medicaid and Medicare payments. Although much of US industry was applying modern logistics techniques, hospitals—like many service providers—felt no competitive pressure to do so.

Hospital care becomes a commodity

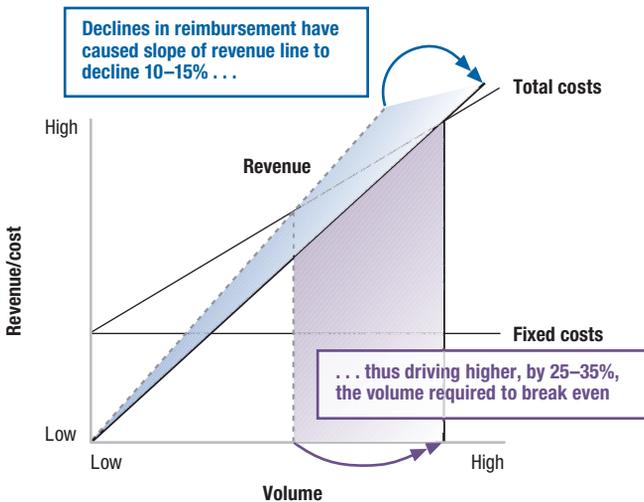
In the late 1980s, managed-care organizations began negotiating lower fees and sharper incentives. Medicaid and Medicare followed this lead—a transformation that culminated in the Balanced Budget Act of 1997. In reaction to this more austere environment, the \$400-billion-a-year hospital industry made almost every large-scale change it could think of, from mergers and acquisitions to slash-and-burn cost cutting. A few hospitals even launched their own insurance plans. But none of these measures worked very well. Mergers in particular neither improved the productivity of hospitals nor

helped them achieve enough local-market bargaining power to offset the influence of either the mammoth health maintenance organizations (HMOs) or the essentially bargain-proof federal and state governments of the United States. Meanwhile, at least in many of the hospitals we have seen, reimbursement rates per unit of activity dropped markedly in the latter half of the 1990s (Exhibit 1).

EXHIBIT 1

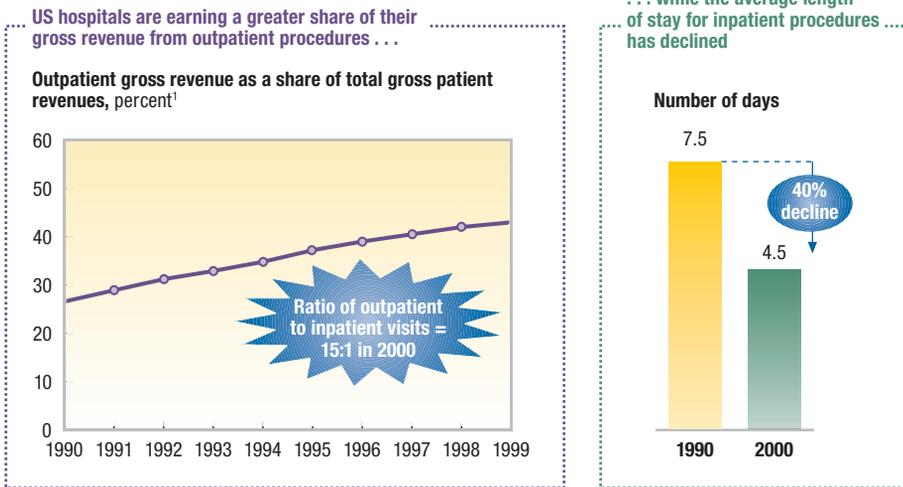
A difficult case

Change in economic structure of US hospitals, 1995–2000



Competitive prices were accompanied by price structures that rewarded fast turnaround times. No longer would HMOs and the government pay hospitals on a “per-patient, per-night” basis; instead, they began paying largely by the illness. That and other related changes led to a sharp decline in the proportion of patients who spent the night in a hospital bed and to shorter stays for those who did (Exhibit 2). This shift in patient activity to the front

EXHIBIT 2

The need for speed

¹For 50th percentile of US hospitals; includes government hospitals.
Source: Solucient; American Hospital Association; McKinsey analysis

end of the process placed huge demands on those parts of hospitals that deliver acute care: the emergency rooms, the operating rooms, and the ICUs. Such demands led in turn to the bottlenecks and long delays those departments endure today.

Once upon a time, the chief executive officers of many hospitals would have responded by expanding capacity: new operating rooms, for instance. But these days, many hospitals are already losing money on ongoing operations, so CEOs naturally shy away from capital-intensive expansion for fear of losing still more. The few who want to expand find that debt markets are far less welcoming to hospitals than they were in bygone years. Just in the past 18 months, the ubiquity and seriousness of these problems have become clear to people in the industry. The average US hospital runs its operations in the red, suffers from overcrowding in critical areas, and can't expand without ratcheting up financial risk.

What will help hospitals escape this quandary? The answer is the one cure that their CEOs have been too distracted to attempt: detailed, day-to-day attention to operations and logistics. Like it or not, hospitals are being reimbursed in pretty much the same way that commodity sellers are and, like them, will rise or fall largely on the strength of operational performance. Stocks and flows, queuing theory, just-in-time processes—all of the notions associated with the factory floor—are exactly what modern hospitals most sorely need. Manufacturing industries have used these ideas for decades.

More recently, service industries such as retail banking, fast food, and telecommunications have followed suit. Now it is time for hospitals to start sweating the details.

The rewards can be great. In less than a year, a hospital can increase the number of patients it serves by as much as 20 percent while eliminating lengthy wait times—gains that can be achieved just by implementing the basic end-to-end process improvement tools that Ford and GM used in the early 1980s. In most cases, this much growth in volume suffices to restore a hospital's long-term financial viability and requires no new expenditures on buildings, equipment, or employees. Everything results from deploying current assets more effectively.

A simple model goes a long way

As the experience of many industries shows, end-to-end processes of all kinds—the flow of patients through a hospital or of visitors through an amusement park—are no better than their weakest link. If too few nurses work in the ICU, for example, patients will be stuck in the operating rooms waiting to get into it, and this problem will in turn delay the preparation of the operating area for the next patient, thus throwing off the surgical schedule, and so on. The cardinal task of an operations manager is thus to identify potential bottlenecks—places where the system is subjected to greater demand than it can handle—and to alleviate them by balancing the components of the process, smoothing demand as far as possible and shifting capacity during peak periods.

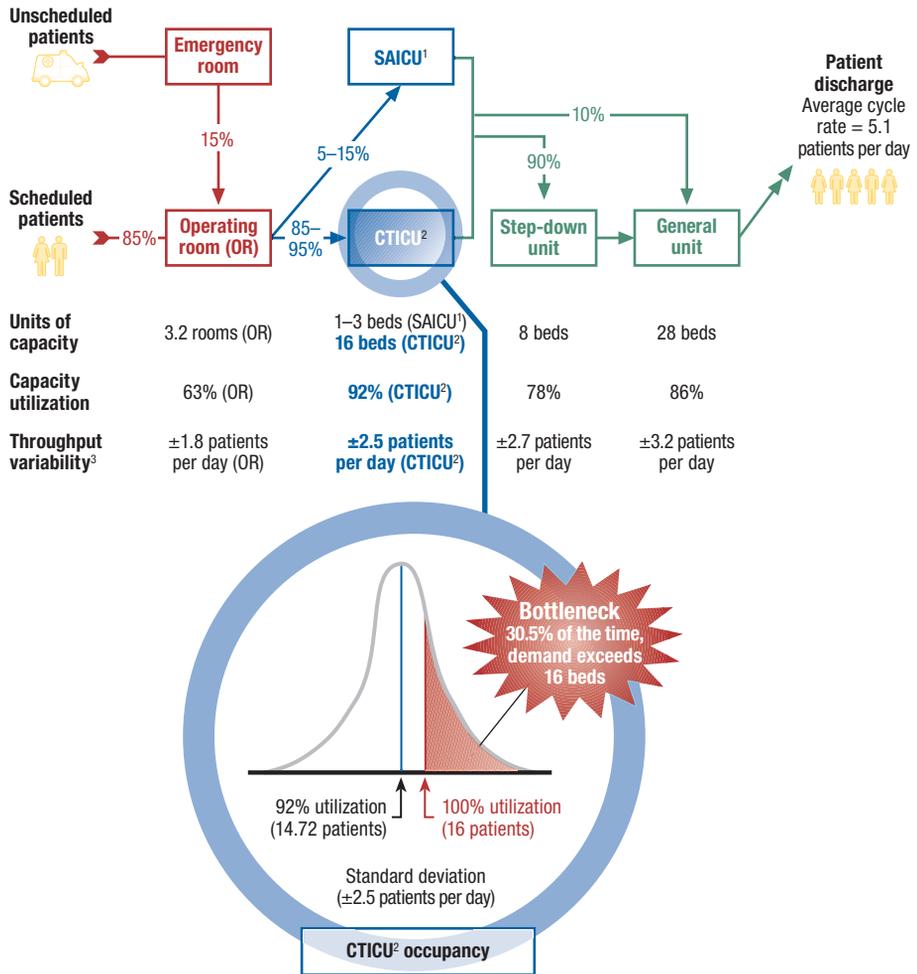


Improving the process efficiency of a hospital begins with identifying the main stages in a typical patient's visit. (Of course, the stages may differ for each major category of patient.) The list need not be very detailed. A surgical inpatient, for instance, might pass through four or five stages: scheduling and registration, preoperative care, the operation itself, recovery in the ICU, and perhaps several nights in a general medical unit.

For each stage, the manager must analyze the demand for services, the system's ability to supply them, and the variations associated with both, measured in the number of patients per unit of time. Demand, which can be reduced to the arrival rates of patients, might involve some day-to-day (and hour-to-hour) uncertainty or variability. Capacity means the number of slots open for service at a given stage of the process. To find the capacity of that stage, you must know how much time is needed to finish it, and this too may vary from one case to the next.

EXHIBIT 3

Anatomy of a bottleneck



¹Surgical anesthesia intensive care unit (as it applies to cardiothoracic patients only).

²Cardiothoracic intensive care unit.

³Measured as standard deviation.

Source: Disguised client example; McKinsey analysis

In general, both demand and capacity are functions of just a few gross parameters, such as the number, type, and arrival rates of cases, as well as the number of beds, nurses, and doctors available in various departments. Starting with these basic ideas, you can predict the maximum patient flow throughout the day and identify bottlenecks under various assumptions. Whenever demand outstrips the available capacity, a bottleneck occurs (Exhibit 3). The challenge is to anticipate these events.

Running such a model repeatedly exposes one of the basic features of end-to-end systems: the exponential relationship between capacity utilization and waiting times (or cancellations) at any stage of the process. Consider a

highway at rush hour. In theory, the amount of traffic could be maximized if, say, there were only a one-inch gap between cars and each car traveled at 65 miles an hour. But with disruptions such as on-ramps, off-ramps, and finite

reaction times, traffic grinds to a halt—just as it does in many hospitals today. When capacity is strained, tiny variations ripple through the whole process (Exhibit 4).

These problems are almost certainly not caused by a few negligent employees or by the failure of any individual department to make itself efficient. Indeed, one of the main causes of bottlenecks in hospitals is the insistence of these semi-autonomous departments on optimizing their own throughput—the number of patients or procedures per hour—without considering how

their actions affect the performance of upstream or downstream departments. In many cases, it would be better to sacrifice local speed for global predictability. When a doctor, anticipating upstream delays, shows up 15 minutes late for surgery, the doctor's personal productivity may well rise. But the patient may be only 10 minutes late, so by hindering upstream improvements, the tardy surgeon or anesthesiologist contributes to the vexing delays.

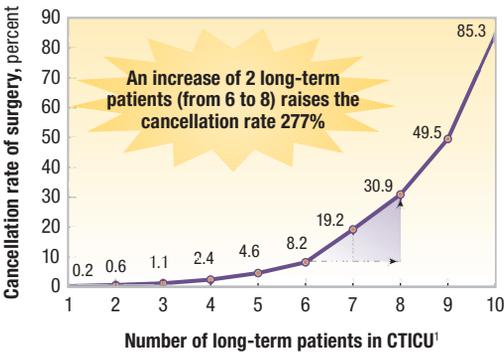
Ten years ago, a certain amount of logistical communication among departments was undertaken informally by senior charge nurses, who each oversaw several hospital departments. These nurses served as traffic cops, discovering incipient bottlenecks and redeploying resources to smooth out the flow of patients. But charge nurses didn't care directly for patients and thus were often among the first casualties of the cost-cutting efforts of the mid-1990s. Without these nurses, a system that was already poorly coordinated became even more so.

We like things nice and smooth

As the example of the highway illustrates, the archenemy of process control is variability. Walk into a modern, hyperefficient automobile factory, and you will observe this principle in action. Everything is calm, orderly, highly choreographed. Automotive companies have spent years—and billions of dollars—wringing variability out of their systems and balancing their lines.

EXHIBIT 4

Straining the system



¹Cardiothoracic intensive care patients with length of stay >15 days. Source: Disguised client example; McKinsey analysis

The less the variability at a given stage of a process, the closer the utilization of capacity can come to 100 percent. Obviously, a hospital isn't an automobile factory, and people—especially sick ones—are less predictable than car parts. Nevertheless, hospitals, which usually have far fewer discrete stages to worry about than do major manufacturers, can often reduce their variability a good deal.

A simple model of the supply and demand at each stage of the process is often enough to identify the sources of variability most likely to cause bottlenecks. On the demand side, the important thing is variability in the number of patients arriving for service at a given time. A hospital can control these variations by such means as changing the schedule of its operating rooms. On the supply side, variability often boils down to poor deployment of people and resources—for example, inconsistencies in the times when doctors conduct patient rounds or poor estimates of how long it takes the housekeeping staff to turn over a patient room.

Once it becomes clear where the problems lie, it is usually possible to eliminate them solely through better planning. Even a little bit of it goes a long way, since hospitals can often eliminate much of their idle capacity by reducing the variability of just three or four parameters in the system.



Know what is really happening

One particularly low-cost way of reducing variability is to look for patterns in what seem like random fluctuations. When we examined the hourly arrival rate of patients to an emergency room over a 90-day period, we observed significant variability. But when we organized the data by time of day, much of it disappeared: at a given hour, the rate was largely predictable (Exhibit 5, on the next page). The hospital could manage this variability by adjusting staffing levels over the course of the day to match the expected demand.

Similarly, when we measured the time needed for triple-bypass operations at a hospital, we found that they took from 283 to 368 minutes. But when we grouped these data according to the particular surgeon performing the operation, the variance fell significantly. Some surgeons take longer than others, and the hospital, recognizing this reality, began to schedule more time for those who needed it, dramatically reducing the unused capacity of the operating rooms.

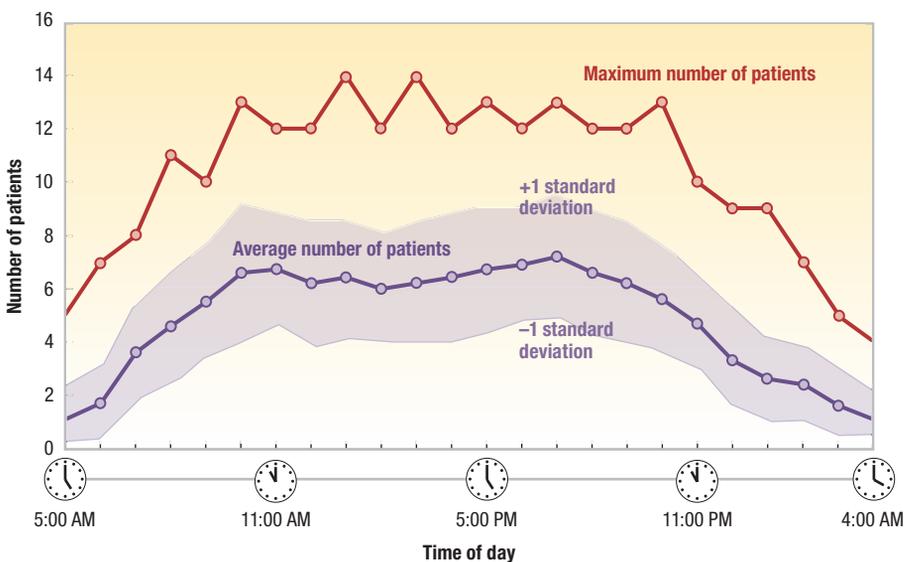
Other, less intuitive, forms of systematic variability might come to light only through a multistep model of the process. One hospital saw large spikes in demand for cardiovascular surgery on Fridays, for example, with all of the attendant bottlenecks and costly overtime staffing. The problem, it turned out, really originated not in the operating rooms but back in the cardiac-catheterization lab—the part of the hospital that diagnoses heart problems that often lead to surgery. The lab processed an unusually large number of cases on Wednesdays and Thursdays, probably because people who have mild chest pains over the weekend first see their primary-care doctors on Monday or Tuesday and arrive for testing two or three days later. Since bypasses tend to occur about a day after the diagnosis of arterial blockage, this pattern created the Friday operating-room bottleneck. The hospital eliminated it by moving catheterizations that could safely be rescheduled to early in the following week, thus spreading out demand for the operating rooms more evenly.

Set limits, but carefully

In addition to reducing variance through these process-oriented techniques, a hospital can also apply systematic, policy-oriented methods. It might sometimes want to establish narrower windows of time for physicians' rounds or discharges of patients, for example, since limits of this sort help contain the more high-variability processes, making it less likely that they will spill over

EXHIBIT 5

Order from chaos: Predicting the patient flow of an emergency room



Source: Disguised client example; McKinsey analysis

into other parts of the system. A firm upper limit on the length of time allotted to cleaning operating rooms and to preparing for the next patient can also reduce variations in their turnover times—variations that may not be large but are certainly disruptive. One hospital created “swat teams” to clean and sterilize any operating room in danger of sitting empty for more than 30 minutes.

Of course, such measures must be implemented with care. Constraining doctors’ rounds to ridiculously brief periods may eliminate some variability but will also annoy doctors and lead to lower admissions rates and, possibly, to inferior care. Nonetheless, a great deal of variability can be eliminated without producing such outcomes.

Less variability means shorter waiting times, and that alone is a great boon to patients and doctors alike. In fact, another good reason to improve logistics is the manifest desire of patients to save time: in a recent survey of 75 people, respondents said that they would drive farther, pay more (in the form of higher co-payments), and even switch doctors if it meant getting faster service.

If medical care is **a commodity**, quality of care will rarely distinguish a particular hospital, at least within a class of competing institutions

In any case, if medical care is largely a commodity, quality of care will only rarely distinguish a particular hospital, at least within a class of competing institutions in a given region. Any major teaching hospital in New York City, for instance, is capable of performing a triple bypass competently. It is benefits such as short waiting times and fast turnarounds that can distinguish one hospital from another.

Inverting the pyramid

Competing on the basis of logistics may require a radical change in the attitudes of hospital executives. In the past, hospitals followed a strict hierarchy. Doctors were treated with kid gloves. Managers occupied the next level down, followed by nurses. At the bottom of the pyramid were semiskilled workers such as those who check patients in and out, clean their rooms, get them positioned in X-ray machines, and transport them around the hospital. The work of these employees was thought to have little strategic importance for the organization.

Redesigning the process stands that traditional pyramid on its head. For identifying and eliminating bottlenecks, the most important people are those closest to the patient and information flow: the frontline workers. In some

cases, they may be physicians, but more often they are less-specialized employees at the bottom of the hospital pecking order, who can often provide keen insights into the detailed process-improvement opportunities that are the lifeblood of logistics.

As any manufacturer can tell you, logistics problems are never fully solved: you find an approximate accommodation and then continue to make improvements as conditions change and better information comes to light. To make the detailed adjustments that are needed to avoid bottlenecks, frontline employees must therefore not only record a hospital's demand and capacity information but also use that information to troubleshoot operations. In addition, it is important to give the frontline the authority to solve problems before they even occur. No longer can the hospital treat these employees as mere ciphers.

In redesigning the processes of hospitals, their managers will have to create new positions that are likely to become pivotal in improving hospital operations. One of them is akin to the job of a production manager in the industrial sector: a person who has a good overview of the entire process and can quickly redeploy resources to solve day-to-day logistics problems that defy more systematic solutions. This person plays the "traffic cop" role that once fell, informally, to the senior charge nurse. Another key employee is the process analyst, who in a manufacturing setting might be described as an industrial engineer: an operations expert who oversees the evolving computer model of the hospital's end-to-end processes and continually recommends changes that might improve the system.

Survival of the fittest

Often, these changes involve information technology. For instance, a new system called "bed tracking" keeps tabs on which beds are vacated and when, thus making the entire process much more transparent. Whoever removes a patient for discharge starts by dialing a number on the room telephone. That number automatically pages a housekeeper, whose activities are logged on the computer, and so on. In this way, the computer displays each room's current state: "empty" or "occupied" and "clean," "needs cleaning," or even "taking too long to clean." Emergency room staffers can follow the process on-screen and adjust their own work accordingly. If a room is about to open up, they know that they can prepare the patient for admission. Under the current system, by contrast, the emergency room staff calls admissions to ask for an open bed, admissions calls a particular floor, the floor supervisor pages a nurse, and the nurse walks down the hall to see whether the bedroom is clean.

These decades-old processes are a luxury that hospitals can't afford at a time when, like banks and insurance companies, they provide a largely commodified service, in the sense that marginal increases in the excellence of medical techniques rarely deliver much competitive advantage. What can deliver it—and will ultimately divide the hospitals that survive from those forced to exit the market—is the efficient use of assets. And that means drastically improved operations.

The prospect may sound daunting, especially to hospital executives whose heads are still spinning from a decade of mergers and cost reductions. But if these executives spend a few hours in their emergency rooms, they may become more sanguine, for they will find almost unlimited room for improvement. *M*
Q

Paul Mango is a principal and **Lou Shapiro** is a consultant in McKinsey's Pittsburgh office. Copyright © 2001 McKinsey & Company. All rights reserved.