8

Understanding Supply-Sensitive Care

Our work in Vermont and Maine focused mainly on surgical procedures, but lurking in the background was another form of care that showed a very different pattern of variation. We found that surgical procedures displayed unique signatures in each location in Maine and Vermont. The rates of tonsillectomy and hysterectomy might be high and that of back surgery low in one place, and vice versa in another region, and a third region might show a low rate for all three—and this surgical signature was remarkably stable over time. Admission rates for nonsurgical care, however, appeared to be another matter entirely. It looked as if the rates in a community followed a consistent pattern: a region with high admission rates for one medical (nonsurgical) condition tended to have high rates for other medical conditions. We also had early evidence that the supply of medical resources, such as hospital beds and physicians, was related to the rates of hospitalization for medical conditions and to the use of imaging tests and electrocardiography. But our hypothesis was difficult to test in the early 1970s, because the myriad overlapping diagnostic codes hampered our ability to know with any precision which patients were admitted for medical conditions.

This limitation disappeared in the early 1980s, when the Health Care Financing Administration implemented the diagnosis-related group, or DRG, payment system, which reimbursed hospitals a set amount for each individual diagnosis, regardless of how long the patient stayed in the hospital.

The DRG system of coding the cause of hospitalization offered us a new tool for studying practice patterns. Using this system, we were able to group the literally thousands of diagnoses physicians use to classify their patients into clusters of related conditions. Moreover, because every patient who was admitted to the hospital was assigned a DRG, we were now able to study the entire population of hospitalized patients according to clinically meaningful causes for being hospitalized and according to whether they were medical or surgical patients. Because of the assistance of the Maine Health Information Center, we obtained access to hospitalization data covering a three-year period, from 1980 through 1982.

Our DRG research revealed that admission rates for virtually every medical condition varied to a remarkable degree.[1] We compared admission rates among thirty hospital service areas in Maine and used certain common surgical procedures as benchmarks for evaluating variation. Not one medical condition exhibited the low variation pattern seen for hospitalization for a fractured hip, the condition for which the admission rate closely follows the incidence of the condition itself. Indeed, the dial on our variation gauge was telling us that supply factors were likely playing a role in determining utilization rates for all medical conditions, some more than others. Only three medical conditions—heart attacks, strokes, and bleeding from the stomach or intestine—were moderately variable: they showed less variation than the admission rates for hysterectomy. The admission rates for over 90% of medical conditions were classified as "high variation medical conditions": they exhibited greater variation among Maine hospital service areas than hysterectomy, and about 40% were more variable than back surgery.

We realized that understanding the pattern of variation in admission rates was critical to health care policy—that "to be successful, cost-containment programs based on fixed, per admission hospital prices will need to assure effective control of hospitalization rates." It was also important for clinical reasons. By focusing on specific medical conditions, we hoped to be able to connect our epidemiology of variation, which we were measuring at the level of populations, to the clinical experience of physicians, and to interest them in working to reduce unwarranted variation. But while our results gained the attention of physicians in Maine, they did not seem to make much of a stir elsewhere. Skepticism was particularly evident among physicians in the nation's teaching hospitals, who found it all too easy to dismiss the findings from this largely rural state as having no relevance to modern scientific medicine. My counterattack was to take the study of practice variation to the citadels of America's academic medical centers.

The Boston–New Haven Studies

Boston and New Haven occupy a special place in my portfolio of small area analyses for illuminating the supply-sensitive care phenomenon. These two communities are served by some of the nation's finest teaching hospitals, and most patients who are hospitalized there go to the principal academic medical centers of Yale University, Harvard University, Boston University, and Tufts University. Moreover, Boston and New Haven are remarkably similar in the demographic characteristics of their populations that predict the need for health care. Yet how different is the amount of acute hospital resources allocated to those populations! Over the years and until very recently, the number of acute care hospital beds per 1,000 used by residents for Boston has exceeded that of New Haven by about 55%. The number of hospital employees per 1,000 serving Bostonians generally ran about 90% higher, and hospital expenditures per capita in Boston were about twice those of New Haven.

My curiosity about the clinical purposes for which these "extra" acute care resources in Boston were used was first aroused by a small area study we conducted using data for 1978. This study, which I published in 1984 in *Health Affairs*, showed that residents of Boston used 4.4 beds per 1,000 residents, while New Haven residents used 2.7—a difference of 1.7 beds per 1,000.[2] At that time, we could not distinguish between surgical versus medical admissions, because the data we had did not include a diagnosis. By the mid-1980s, however, we obtained hospital discharge information similar to the Maine data, allowing us to use the DRG classification system to study the situation in some detail for hospitalizations that happened in 1982. We found that the physicians in Boston used 739 more hospital beds per 1,000 in 1982 in treating their patient populations than predicted by the New Haven benchmark.[3] As predicted by the Maine DRG study, most of those beds (71%) were used to care for adult patients with medical conditions. Seven diagnoses, all of them for chronic conditions, accounted for about 30% of the excess bed use for medical conditions: low back pain (not treated surgically) accounted for the largest portion, followed by gastroenteritis, congestive heart failure, pneumonia, diabetes, cancer of the lung (not treated surgically), bronchitis, and asthma. Five percent of the beds were used for pediatric patients with medical conditions; 12% for minor surgery (the kind of surgery that today is mostly done in the outpatient setting); and 12% were for major surgery.

For those patients hospitalized for medical conditions and minor surgery, the difference in bed use between the two communities was explained largely by a higher rate of admission to hospitals for Boston patients, not by longer

lengths of stay in the hospital. By contrast, rates of admission were the same in both communities for major surgery, so the difference in bed use in that case was explained entirely by the Boston hospitals' longer lengths of stay. Once again, beds per 1,000 exerted a powerful influence on medical admission rates, but it had little effect on rates of admission for surgery, with the exception of minor surgery, which was more often performed in the inpatient setting in Boston than in New Haven.

Evidence for a Subliminal Effect of Capacity

By the time we began studying New Haven and Boston, we already suspected that more beds led to higher hospitalization rates for medical conditions. The question was, were physicians aware of it? Before the results of the study were published, I sought interviews with physicians who practiced in Boston and New Haven. I wanted to learn whether the physicians who were actually making the decisions to hospitalize were aware that their practice patterns were different in the two communities and that the availability of beds seemed to be influencing their decisions. I was particularly interested in learning whether, in supply-constrained New Haven, physicians sensed that beds were scarce—whether they ever felt a need to hospitalize patients but could not, because all the beds were full. In short, w*ere they consciously rationing hospital care because of a lack of hospital beds?*

What I learned from these interviews helped me gain insight into the largely *unconscious* nature of demand induction for supply-sensitive treatments. At first, I did not show the physicians our results, but simply asked them if they were aware that there were differences in the rates of hospitalization between the two communities. They were not. Indeed, a number of New Haven physicians I talked with who had previously practiced in Boston said that they did not think local practice styles were different, or that they had changed when they moved to New Haven. The clinicians of New Haven denied that they were rationing care, and once I informed them about the relative differences between Boston and New Haven, they seemed to take pride in their more conservative practice style.

The study, which was published in *The Lancet* in May 1987, bore the rhetorical title: "Are hospital services rationed in New Haven or over-utilized in Boston?" The study showed conclusively that even among communities served by famous academic medical centers, there were large differences in population-based hospitalization rates. Moreover, for the care that we were calling supply-sensitive, physicians with strong academic credentials were

quite unaware that they practiced differently or that they might actually change their practice styles, depending on the number of hospital beds available.

A Look at Outcomes

Toward the end of the 1980s, our research group acquired access to Medicare data for New England, allowing us to search more closely for evidence that differences in the supply of resources might be leading either to rationing or overuse of health care. We revisited Boston and New Haven to compare hospital use and mortality, and to see if the difference in utilization was associated with a difference in overall population mortality rate.[4] First, we confirmed that the chance of being hospitalized still varied substantially between the two locations. It did. In 1982, 21% of the Medicare population living in Boston was hospitalized at least once compared to 16% for New Haven, and 33% of the hospitalized patients in Boston were readmitted one or more times within the study year compared to 25% for New Haven. We then looked at overall population mortality—all deaths that occurred in the hospital plus all deaths that occurred elsewhere—and found that Medicare death rates for Boston and New Haven were virtually identical.

Might New Haven patients have lived longer had their physicians admitted more of them to the hospital? We could not know from this study, but at least this much of the outcomes puzzle was becoming clear: the lower rate of hospital use in New Haven was not associated with a *higher* overall mortality rate.

The study also provided further insight into how hospital capacity may influence utilization rates. A common hypothesis ran something like this: clinicians hospitalize patients based on *sickness.* The sickest get hospitalized first, then the next sickest, and so on until beds are exhausted. Regions with fewer beds per capita run out first, so in these regions the "case-mix" of hospitalized patients will include a greater proportion of the severely ill than the mix in regions with more beds. We tested this theory by comparing the population-based hospital statistics for Boston and New Haven. We found that on an annual basis, a greater proportion of Medicare patients were admitted once or more to hospitals in Boston than those in New Haven which had fewer beds, suggesting that capacity influences the decision to admit, leading to more hospitalizations for those who were less severely ill in Boston. The lower case-fatality rates in Boston hospitals were also consistent with this interpretation.[5] On the other hand, the bed effect also seemed to influence the hospitalization rate for those who were the most severely

ill: on a population basis, Boston patients were much more likely to die in the hospital than someplace else, such as at home or in hospice care. For Bostonians, 40% of all deaths occurred in the inpatient setting, compared to 32% for New Havenites. It was as if Boston hospitals were a giant vacuum, hoovering patients of varying levels of sickness into beds, but not necessarily making a difference in their outcomes compared with New Haven.

A New Way to Study Practice Variations

We conducted yet another test of our theory that in Boston the clinical threshold for admitting patients was lower for a broad spectrum of medical conditions when compared to New Haven. This study, published in 1994 in the *New England Journal of Medicine*, used a new method for measuring practice variations based on a cohort design.[6] It focused on patients who all experienced a specific clinical condition, and followed them over time. (See Box 8.1 for a description of the advantages of cohort studies.)

The first part of the study was conducted on residents of Boston who were hospitalized for one of the handful of clinical conditions that are more or less uniformly diagnosed, and for which, once the diagnosis is made, virtually all physicians recommend hospitalization. To become part of the study, a resident of Boston or New Haven had to have been hospitalized for one of these "index events," a hip fracture; a surgical procedure for cancer of the colon, lung, or breast; an acute myocardial infarction; a stroke; or gastrointestinal bleeding. For these conditions, the hospitalization rates were about the same for residents of Boston and New Haven (because the rates of the conditions were about the same for the two cities). The goal of the study, however, was not to compare the rate for the initial hospitalization among Bostonians and New Havenites—we already knew that they were pretty much the same. Rather, we were interested in comparing the pattern for *subsequent* hospitalizations, to test the hypothesis that Bostonians with identified chronic illnesses were being hospitalized much more frequently than similarly ill patients in New Haven. To do this, we first identified all patients hospitalized for an index event over a two-year period, and then linked the initial record for each patient to all subsequent hospitalizations that occurred for that patient during a period of time that extended up to three years. We then analyzed the records for each of the six cohorts (groups of patients with hip fractures, cancer, etc.) to calculate the admission rate for each six-month period of follow-up.

The results confirmed our hypothesis. Overall, the risk for *subsequent* hospitalization following the index event was 1.6 times higher for patients

Box 8.1. *The Advantages of Cohort Studies*

While very useful for studying patterns of variation, cross-sectional geographic studies are less useful for studying outcomes of care, particularly questions concerning the impact of treatment on specific types of patients—say the survival of heart attack patients who receive (or do not receive) a particular drug. For such questions, epidemiologists typically use cohort studies, which "enroll" patients who experience a given event and observe what happens to them subsequently, depending, say, on the medical community where they live. An important advantage of the cohort approach is that it can include everyone with the disease, and not just those accepted into the conventional randomized trial. (Often randomized trials exclude patients with complications, or older patients.) Furthermore, the use of cohort studies allows for far more patients—often in the thousands—which increases the statistical precision of the results. Cohort studies do lack pure randomization, but we have found that populations of heart attack patients, for example, tend to be similar regardless of where they live. Furthermore, the Medicare data allows for adjustment for comorbidities (other conditions the patient may have had during the index hospitalization), as well as demographic factors, such as age, sex, and race, that may affect the individual's level of illness and outcome. This allows us to compare the outcomes of similar patients (apples to apples!) who live in different regions and experience different intensity of care.

living in Boston compared to New Haven—an almost exact replication of our study published in *The Lancet*, which used the classic small area analysis design showing that population-based rates of hospitalization were different between the two cities. Moreover, as predicted by our previous small area variation studies, the large majority of the readmissions were for medical, not surgical conditions. A patient who had been first admitted for a heart attack, for example, might be readmitted for congestive heart failure. The effect of bed capacity on clinical decision making seemed about equal for all cohorts. In other words, *the effect did not depend on the initial diagnosis*; for the cancer cohorts, the risk of subsequent admission for Bostonians was 1.6 times greater than for New Havenites; when the initial condition was a hip fracture, it was 1.6 times greater; and for acute myocardial infarction and for stroke, it was also 1.6 times greater.

Indeed, the threshold effect of beds worked to influence the risk of hospitalization for all patient subgroups. Women in the Boston cohorts (regardless of initial diagnosis) were 62% more likely to be hospitalized than their counterparts in New Haven. For men, the rates were 67% higher; for white patients, 66% higher; for nonwhite (mostly black) patients, rates were 43% higher; for older patients (75 years of age or older), 69% higher; and for younger Medicare patients (aged 65–74), rates were 54% higher. As predicted by previous studies, the threshold effect influenced primarily medical and minor surgery cases, rather than major surgery. Virtually every acute and chronic illness diagnostic group was affected.

Evaluating Hospital-Specific Performance

The second part of our study broke further ground in advancing the methods for evaluating patterns of care. The cohort method was adapted to provide hospital-specific estimates, allowing, for the first time, an investigation into the rates of admission according to the hospital most often used by the patient, rather than the region as a whole. (See Box 8.2.) We uncovered considerable differences in the risks of hospitalization for individual teaching hospitals within Boston. Compared to the most conservative teaching hospital, the Yale-New Haven Hospital, the rates of admission were substantially higher for all Boston teaching hospitals. Some were below the increased risk factor of 1.6 measured for the area as a whole, while others were well above it.

Armed now with hospital-specific data, I once again sought the opportunity to see if clinicians in Boston teaching hospitals, whose decision making was responsible for determining which patients were hospitalized, were aware that their practice styles varied according to where they practiced. Keeping the identify of each Boston teaching hospital hidden, I first showed them data comparing the admission rates of the six major teaching hospitals in Boston to the Yale-New Haven Hospital. For these institutions, admission rates were between 50 and 98% higher than Yale-New Haven. Here are the ratios compared to Yale-New Haven:

Hospital A	1.98
Hospital B	1.86
Hospital C	1.62
Hospital D	1.61
Hospital E	1.57
Hospital F	1.50
Yale-New Haven	1.00

Box 8.2. *Measuring Hospital-Specific Performance*

If each Boston hospital were like a prepaid, staff model HMO, such as Kaiser Permanente, we would know from the enrollment files the exact size of the populations they serve. The cohort method provided us with a way of estimating the population at risk for the vast majority of U. S. providers, who were not (and still are not) organized in this way. We assigned patients to the hospital where the index event occurred: the hip fracture, cancer surgery, etc. We then analyzed the data to determine which hospitals were used for subsequent admissions. There was a high degree of loyalty among the Medicare patients, as most subsequent hospitalizations occurred at the same hospital as the initial one. (Among the 11 hospitals in the study, between 62% and 90% of readmissions were to the index hospital.) Thus, we could calculate the rate for subsequent hospitalizations—using the number of patients with hip fractures, cancers, and so on as denominators—with assurance that the clinical decisions that led to hospitalization were primarily made by clinicians associated with specific teaching hospitals. It thus became possible to compare the rates for specific hospitals in Boston and New Haven.

I then asked them to guess, based on their personal experience, where their own institution was in the spectrum of variation, and to name the other Boston hospitals. None were aware of their own, much less any other institution's, relative frequency of hospitalizing patients. Many guessed that Hospital A, with admission rates 1.98 times greater than Yale-New Haven, was the Massachusetts General Hospital. As it turned out, Massachusetts General was Hospital F, the Boston hospital that was closest to New Haven in its rate of admission, though it still exceeded the Yale-New Haven Hospital's practice pattern by 50%.

The most interesting case concerned the admission rates for Hospitals A and C. One is Boston City Hospital, the hospital serving the indigent of Boston; the other is Boston University Medical Center. At the time of this study, these two "hospitals" were in fact a single building separated into two separate hospital wings, each with its own complement of beds relative to the size of the population it served. The physicians attending at the Boston City Hospital also served the Boston University Medical Center and vice versa. The data showed

that the rate of admission for patients loyal to Hospital A was significantly higher (statistically and clinically) than for Hospital C. I asked them to guess which hospital was which. Although some were onto my game by then, most guessed that Hospital A, with the highest rate of admission per 1,000 had to be Boston City Hospital, primarily because it served the poorest—and therefore the sickest—segments of the Boston population. They were wrong. The admission rate at Boston University Medical Center (Hospital A) was the highest of all Boston teaching hospitals, almost twice that of Yale-New Haven, and 22% greater than the admission rate among patients loyal to Boston City Hospital, suggesting that the complement of beds available for the insured population was greater than the complement of beds for the indigent. This natural experiment provided important insight into the subliminal, yet powerful effect that bed supply exerts on physician decisions. *The physicians were simply unaware of the changes in their own practice styles that occurred when they crossed the firewall dividing the two wings of the hospital complex.* (The assumption that poverty—and illness—is the most important determinant of variation in admission rates persists, and it was raised again in 2009 during the debate over health care reform, as discussed in subsequent chapters.)

What about the outcomes of care? An important advantage of the cohort methodology is its ability to measure survival following an initial admission event such as a heart attack or hip fracture. Using this method, we could directly address important questions about health care rationing that could not be answered by small area correlation studies. Were New Haven physicians keeping patients out of the hospital that would have lived longer had they been admitted? To answer this question, we followed our heart attack, stroke, hip fracture, cancer, and intestinal bleeding patients for up to three years. While Bostonians with these conditions received about 60% more hospitalizations, they did not live any longer. The overall mortality for the cohorts during the entire period of follow-up was essentially the same in the two cities. The implications of this finding were both clear and arresting: for these two cities—and their constituent academic medical centers—the extra care delivered to patients in Boston did not appear to improve life expectancy. The variation in supply-sensitive care appeared to be a case of overuse in Boston, not rationing in New Haven.

The Invisible Hand of Capacity

The idea that the supply of resources "causes" an increase in utilization of services is not a new one. Indeed, in the health care policy world, it is often held as the truism known as "Roemer's Law," named for Milton Roemer, who

concluded in the 1960s that a hospital bed, once built and available, will be used no matter how many beds there are.[7]

With the completion of the first round of Dartmouth Atlas studies in the 1990s, we were able to conduct the first national study of the association between available hospital beds and hospitalization rates. Among the 306 Dartmouth Atlas regions, as predicted by our earlier studies, hospitalization rates for hip fracture showed virtually no relationship with hospital bed capacity (R^2 = .06). By contrast, having more hospital beds was directly associated with higher hospitalization rates for patients with acute and chronic medical conditions. Indeed, the association between beds and admission rate was quite strong. More than half—54%—of the variation was associated with bed capacity (Figure 8.1).

The pattern makes medical sense. When, as in the case of hip fracture, the incidence of disease is the most important determinant of variation in hospitalization, the supply of resources is not closely associated with the utilization of care. The market is "cleared" of need, as every case of hip fracture has a priority claim on hospital beds, no matter what the per capita supply of beds. For most medical conditions, however, the clinical decision to hospitalize

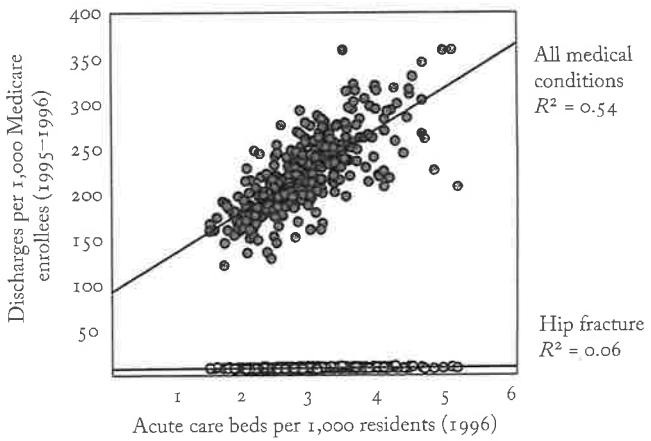


Figure 8.1. The association between hospital beds (1996) and discharges for medical conditions and for hip fracture (1995 through 1996) among hospital referral regions. (Source: Wennberg, J. E., and E. S. Fisher, eds. 2006. *The Care of Patients with Severe Chronic Illness: A Report on the Medicare Program by the Dartmouth Atlas Project. The Dartmouth Atlas of Health Care 2006.* Hanover, NH: The Center for the Evaluative Clinical Sciences [online].)

a patient is not so clear-cut and the "supply" of cases that current medical practice labels as appropriate for admission nearly always exceeds capacity. In other words, there are nearly always more sick people than there are beds. For most acute and chronic illnesses, the diagnosis is not in itself sufficient grounds for hospitalization. The clinician is forced to make decisions on the hospitalization of individual cases that have a place on a spectrum of severity—to distinguish between shades of grey, not the binary black-and-white hip fracture decision. Physicians make these decisions within the context of available beds. The key idea here is that when a physician faces uncertainty concerning medical prognosis, the dominant cultural bias is to err on what is perceived to be the side of safety—to prescribe hospitalization when a bed is available. Moreover, under fee-for-service Medicare, economic incentives are squarely in sync with the "more is better" assumption, even when the physician does not directly benefit financially from the decision to hospitalize.

In the absence of explicit theory and useful rules of thumb, decision making is often guided by a general assumption that when in doubt, more health care is better. Both doctors and patients assume that the acute hospital setting, with all of its resources and concentrated medical skills, is a better place to deal with sick patients with guarded or uncertain prognoses than are other settings, like the patient's home or even the nursing home, where care is seemingly less organized and there are fewer physicians and nurses available. Under such an assumption, the availability of beds becomes critical. Among teaching hospitals in Boston and New Haven, the occupancy rates were all quite high, but beds were always available for the "low variation" conditions like hip fracture, or cancer patients needing surgery, cases that everyone agrees require hospitalization. But these conditions comprise only a small proportion of patients using beds—even in regions with constrained beds per 1,000 people. Thus, at any given point in time, the patient population of the hospital with medical diagnoses is composed mostly of patients with acute and chronic illnesses that are susceptible to the threshold effect of capacity. And when there are more beds per capita, there are more opportunities to place the patient in the "safer" inpatient environment.

The reader will recall that our studies in Maine found that each hospital service area had a surgical signature, its own peculiar pattern of surgical rates for different conditions—high rates for some, low rates for others. Moreover, the overall rate of surgery (the total discharge rate) is not closely correlated with the rate for any given surgical procedure. By contrast, the rate of hospitalization for a specific high variation medical condition tends to be closely associated with total discharge rates; and within a given region, hospitalization rates tend to be more or less uniform across all high variation medical

conditions. The medical signatures for Boston and New Haven, as reported in the 1998 Atlas, are illustrated in Figure 8.2.

We found a similar pattern when we looked at the frequency of physician visits. Patients had more physician visits per capita in regions where the per capita supply of physicians was higher, particularly for physicians that spend most of their practice time on older, chronically ill patients, such as general internists and cardiologists (Figure 8.3). This association between supply and utilization makes sense in the outpatient setting, given what is known about the way patients are scheduled for follow-up visits. Most physician visits are revisits, scheduled by the physician (or, more likely, their office personnel), who typically fill most available hours with established patients. Most patients with chronic illnesses are assumed to need monitoring, and the only real question the physician faces in rescheduling is the relative need among the individual patients for whom he routinely provides care. (The sicker ones, of course are seen more often.) But if physicians have fewer patients in their patient population, the frequency of revisits will be higher for all patients with chronic illness—the sickest and less sick as well.

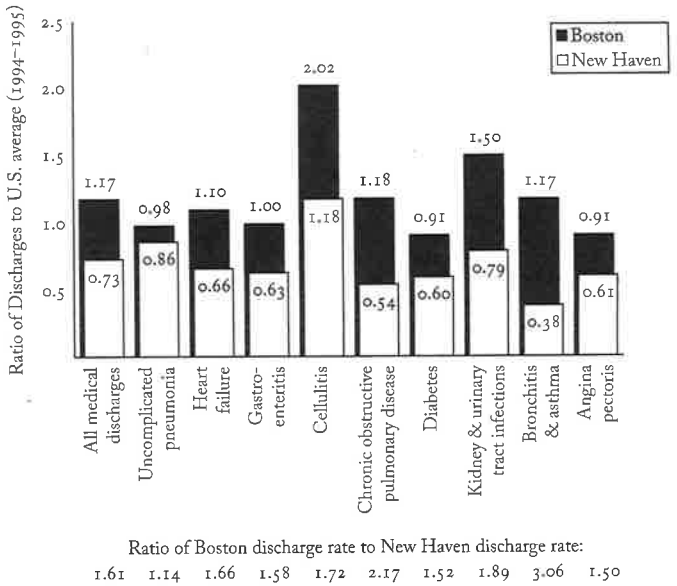


Figure 8.2. The medical signatures of the Boston and New Haven hospital service areas (1994 through 1995). (Source: Wennberg, J. E., and M. M. Cooper, eds. 1998. *The Dartmouth Atlas of Health Care 1998*, Chicago, IL: American Hospital Publishing.)

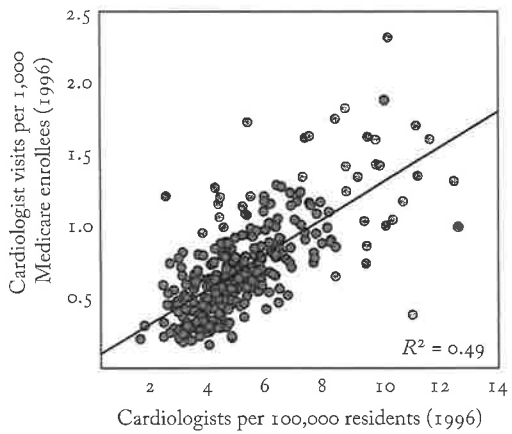


Figure 8.3. The association between cardiologists and visits to cardiologists among hospital referral regions (1996). (Source: Wennberg, J. E., and E. S. Fisher, eds. 2006. *The Care of Patients with Severe Chronic Illness: A Report on the Medicare Program by the Dartmouth Atlas Project. The Dartmouth Atlas of Health Care 2006.* Hanover, NH: The Center for the Evaluative Clinical Sciences [online].)

It's tempting for some to believe that physicians are acting as self-serving, even cynical inducers of demand, hospitalizing patients and scheduling revisits so that they can make more money. But this cannot be the explanation, as there are no normative scientific standards for rescheduling or hospitalization to be transgressed. Astonishing as it may seem to many patients and even some health care policy analysts, medical science provides no guidance on what the best practice interval between visits should be or when to hospitalize. There is remarkably little medical theory and almost no medical evidence concerning the optimum frequency of interventions for supply-sensitive services. This was evident both through my personal interviews with academic clinicians in Boston and New Haven, and also in the lack of formal discourse in medical texts concerning best practices regarding the appropriate frequency of the use of supply-sensitive services. In the standard medical texts that inform the practice of both primary and medical specialty care, and in the practice guidelines that constrain clinical decision making, one searches in vain for even the briefest discussion concerning the criteria for admitting chronically ill patients to the hospital and to intensive care, or the optimal interval between revisits for patients with established disease.

The lack of guidelines, or evidence, or any form of normative scientific constraint on physician decision making for supply-sensitive care has a profound impact on the health care economy. The number of physician office hours available for monitoring and managing the care of the population living

in a region is closely dependent on the supply of clinically active physicians per 100,000 residents. Take a hypothetical case. In region A, which has twice as many cardiologists as region B, twice as many hours will be available for a cardiologist to schedule. On average, region A's population will experience twice as many visits per person compared with region B, and the mean interval between visits will be about half that of Region B. Neither the patients nor the clinicians in regions A and B will be aware of the differences in practice style. The patients will assume that their medical need determines the schedule for revisits. Physicians will allocate their time to patients on the basis of relative illness, with the sicker patients experiencing more frequent visits. Most physicians in both communities will be working long hours, believing that the care they provide is necessary care, and totally unaware that capacity differs—or that capacity influences their clinical decision making. Only the epidemiologist, peering at health care from 30,000 feet, can see the patterns of practice and make the connection between capacity and utilization.

What Accounts for Variation in Capacity?

Understanding supply-sensitive care requires an understanding of why capacity itself varies so much from region to region (and from hospital to hospital). In my experience, satisfactory answers to the question, "Why do some hospitals in some regions grow more rapidly in relation to the size of the local population than do others?" do not emerge from the 30,000-foot perspective or statistical correlations. Epidemiology has in its book of methods what traditionalists call "shoe-leather" research—that is, getting out on the streets and looking for explanations that might solve a mystery. The most famous example remains John Snow's careful charting of the outbreak of cases in the London cholera epidemic of 1854, when he pinpointed contaminated drinking water supplied through the Broad Street pump by the Vauxhall Water Company as the source of contagion. Following in the footsteps of Snow, I have had the opportunity to undertake two shoe-leather investigations of the dynamics of hospital construction, both of which illuminated how the capacity of local health care markets became established.

Consider first the example of Boston and New Haven, where different regulatory regimes influenced the growth of hospital capacity. Consistently over the years, the capacity of the acute care hospital sector in Massachusetts exceeded that of Connecticut. For example, the number of acute care hospital beds per 1,000 allocated to the health of Bostonians exceeded that of New Havenites by about 55%; the numbers of hospital employees per 1,000

serving Bostonians generally ran about 90% more, and hospital expenditures per capita were about twice those of New Haven. These differences can be traced to the period shortly after World War II, when the hospital industry enjoyed a period of growth, stimulated in part by the Hill-Burton Act.

Passed in 1946, the act required states to develop "state health plans" on the need for beds, in order to receive federal subsidies for hospital construction. In many states, including Massachusetts, Hill-Burton grants were tied to a planning methodology designed to ensure that the occupancy of hospitals did not exceed a given level. Thus, the more pressure that was placed on available beds, the more "need" there was determined to be, independent of the actual numbers of hospital beds per 1,000 in the community or region.

As I learned from a 1987 interview with John Thompson, who had recently retired from his professorship in hospital administration at Yale, the evolution of the Hill-Burton planning process in Connecticut was quite different from Massachusetts. In Connecticut, the decision process was dominated to a large extent by the CEOs of the existing hospitals. Their basic strategy was to keep new competitors out of their local markets, using the state's Certificate of Need, or CON, legislation to thwart attempts to establish new hospitals. Thompson, who had been part of the process, believed that this was the primary reason why over the years Connecticut has been at the low end of the national spectrum in hospital beds per 1,000. He cited two specific examples of how the process responded to keep capacity low in the New Haven area. One was the reaction to a petition by several dissident physicians who wished to leave the teaching hospital to start a suburban hospital in a neighboring community. The other was a proposal to build a Jewish hospital. Both were turned down during the CON process (as were similar applications in other parts of the state).

The CON process in Massachusetts, by contrast, was much more open to the influence of various interests that wanted to expand the hospital industry. Thompson cited competition between Boston teaching hospitals as a major reason for the expansion of capacity in that region: each hospital required its full complement of services and obtained the needed approvals from the CON administrators (and capital from banks, bondholders, and federal subsidies) without difficulty. Growth of the hospital sector in the greater Boston area was also susceptible to the pressure for a place to practice medicine from physicians who did not win, or did not want, appointments at a Boston teaching hospital, but who stayed in the area and were welcomed on the staffs of community hospitals. This pressure was particularly strong in the Boston area because of the many academic training programs that produced new medical residents (who characteristically seek to practice medicine in the region where they train).

In other communities, hospital capacity is built up for different reasons. Take Augusta and Waterville, two neighboring communities in central Maine, where competitive dynamics and religious preference created the pressure to build more beds. The following facts emerged from our studies in Maine in the 1970s. At that time, Augusta and Waterville had about 50,000 persons each, but very different supplies of acute care hospital beds: about 3.5 beds per 1,000 for residents of Augusta and about 5.5 beds per 1,000 for Waterville. In Waterville, there were three hospitals: one an osteopathic hospital, the second an allopathic Catholic hospital, and the third nonsectarian and allopathic. (The Catholic hospital and nonsectarian allopathic hospital have since merged.) In Augusta, history produced but one nonsectarian hospital that, from the beginning, welcomed allopathic and osteopathic physicians as well as all religions. Having three hospitals netted 60% more beds per capita for Waterville—and higher per capita spending and utilization.

Why did the three hospitals in Waterville not come to some market equilibrium, with each taking care of its share of the population, and none building more beds than necessary? I have already made the case that the physician is ineffective as society's agent for constraining the overuse of supply-sensitive care, largely because he or she is almost entirely unaware of the effect of supply on his or her discretionary decisions, and because clinical science imposes no significant constraint on physician decision making in ways that might also place limits on their use of resources. One can be quite sure that in 1970, the administrators and boards of trustees of the three hospitals in Waterville, or anyone else in a position to influence decisions on capacity, were not at all concerned about the possibility of excess beds per capita in their community; it would never have crossed their minds, for any number of reasons. There was little recognition that supply could drive utilization, and a widespread assumption that more medical services led to better outcomes. In addition, several "system-level" factors were at work to reduce awareness of the consequences of any decision to increase capacity. First, key information was lacking: because population-based data on resource capacity was unavailable, administrators and boards of trustees of hospitals were unaware of hospital capacity relative to the size of the resident population in their own region, much less the number of beds their own hospital used in caring for its loyal population. Second, the capital for expanding the acute care sector was readily available, no matter how many hospital beds per capita there already were. During the 1970s and 80s, the federal Hill-Burton Act subsidized the construction of hospitals, but its planning methods were flawed. Again, the problem can be traced in part to lack of population data: the signal that planners relied upon for measuring scarcity of beds was the occupancy rate—the

percentage of available beds that on average are filled. But the occupancy rate is an unreliable measure of the needs of the population, because it is largely uncorrelated with either prevalence of illness or the existing bed supply. In Vermont, for example, as we documented in our 1973 paper in *Science*, the use of this measure to determine need resulted in paradoxical decisions on the part of the state health planning agency, calling for additional bed construction in regions that already had a high per capita number of beds.

Finally, there were no direct economic consequences to employers and individuals living in Waterville in terms of the price they paid for health insurance. Those who buy insurance are insulated from the true cost of care in their local communities, so they do not put pressure on hospitals to constrain utilization or the growth in capacity that can drive it. In the late 1970s, Blue Cross was the dominant provider of health insurance in Maine and the price Blue Cross charged for a policy was the same throughout the state, no matter what the actual level of per capita utilization, and thus spending, was in a given region. Furthermore, hospitals are viewed as desirable to the community, both in generating local jobs and in attracting new residents.

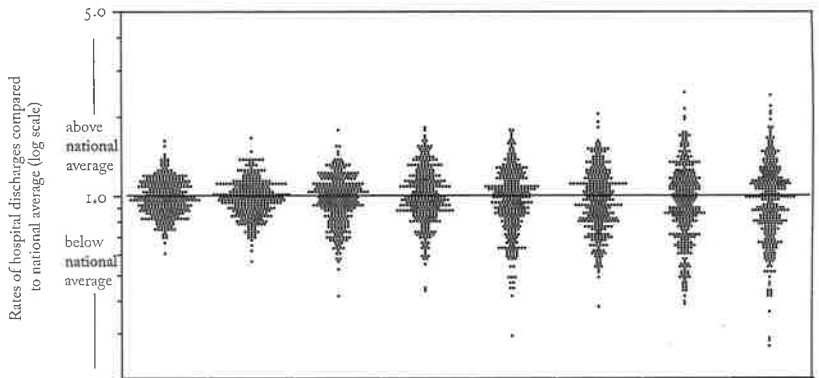
In our Maine research, we documented striking differences in per capita reimbursements by Blue Cross in Maine and then compared how much the residents in different communities had paid out for insurance versus how much care they received. In 1979, Blue Cross paid the providers in Waterville $221 per subscriber on average, 1.46 times greater than the $151 it paid per subscriber in Augusta, meaning that the 22,800 Waterville subscribers received nearly $1 million worth of care more than they (or their employers) paid to Blue Cross. Residents of Augusta, by contrast, received $750,000 less care than they paid for.[8] One has to wonder, if the price of health insurance had been adjusted to reflect local market per capita costs, would the citizens of Waterville have come to a different conclusion concerning the need for three hospitals, and taken steps to reduce their excess capacity? These may seem like small numbers today, but given the dramatic increase in the cost of health care, both in terms of utilization and price per unit of service, the magnitude of dollar transfers from low to high cost communities now reaches into the billions of dollars. (In Chapter 12, I provide an estimate of the amount of transfer payments under traditional Medicare.)

The Patterns of Practice Today

The pattern of practice for supply-sensitive care today is very much the same as it was when I first began my studies in New England some 30 years ago. In

preparing this chapter, I repeated as closely as I could the 1980–1982 Maine study of variation in medical conditions discussed earlier in this chapter, using Medicare data from 2005. I looked at the pattern of variation in discharge rates among the 306 Atlas hospital referral regions for 59 medical conditions identified through the DRG coding system. With one exception, the story across the nation in 2005 is essentially the same as it was in Maine in the 1980s. Back then, 90% of medical discharges in Maine were high variation—as variable or more so than hysterectomy; in 2005, most Medicare patients in the country—88% of Medicare discharges for medical conditions—were hospitalized with high variation medical conditions—more variable than knee replacement.[9]

Figure 8.4 illustrates the pattern of variation for eight medical conditions, selected because they are the most common in terms of frequency of hospitalization: each accounts for about 250,000 or more of patients hospitalized for medical DRGs in 2005 among Medicare recipients. Together, the eight conditions account for 40.9% of all medical conditions. Discharge rates for stroke and bleeding from the gastrointestinal tract exhibit moderate variation among the 306 regions, with a coefficient of variation that lies between hip fracture hospitalizations and knee replacement. The discharge rate for



	Stroke	Gastro-intestinal hemorr-hage	Cardiac arrhythmia	Simple pneu-monia	Gastro-enteritis	Conges-tive heart failure	Septi-cemia	COPD
Ratio highest to lowest region	2.6	2.9	4.2	4.2	6.2	5.5	6.5	9.2
Coefficient of variation (x100)	16.0	16.4	20.7	24.2	27.1	27.6	34.1	34.3

Figure 8.4. The pattern of variation of hospitalization for eight common medical conditions among hospital referral regions in 2005. (Source: Dartmouth Atlas Project database.)

cardiac arrhythmia is on the boundary between high and medium variation, with a coefficient of variation similar to knee replacement. Discharge rates for patients with pneumonia, gastroenteritis, and congestive heart failure are more variable than knee replacement; discharge rates for chronic pulmonary obstructive disease and septicemia are more variable than back surgery.

The exception was the change I noted in the pattern of variation for patients with acute myocardial infarction. In the Maine study, heart attack discharge rates followed the moderate variation pattern. In the 2005 Medicare study, however, heart attacks classified as medical conditions were highly variable, in fact more variable than the rates for knee replacements. The increase is explained in part by the DRG coding convention. The Maine study was conducted before the advent of percutaneous coronary intervention, or PCI—a procedure involving using a catheter to expand a coronary artery, such as stents. By 2005, heart attack victims were often treated with PCI and thus, under the DRG convention, they became classified as "surgical patients." But this is not the only reason why variation increased. Diagnostic practice also changed. In the 1980s in Maine, the diagnosis of a heart attack was made primarily on the basis of a blood test and changes in the electrocardiogram caused by damage to the heart muscle. By 2005, the availability of methods to improve blood flow and prevent damage to the heart muscle, and more sensitive blood tests, had led to earlier interventions, often in patients for whom the diagnosis of acute myocardial infarction is less certain. Depending on how hard they look, more patients will be diagnosed with a heart attack in some hospitals than in others.[10]

Recent years have brought about some interesting changes in discharge rates for medical conditions in New Haven. Across the United States, discharge rates for medical conditions rose from 224 per 1,000 in 1995 to 244 per 1,000 in 2005, a 9.0% increase. During the same period of time, discharge rates for residents of Boston increased 6.5%—a roughly similar increase. New Haven rates, however, rose dramatically. In 1995, the discharge rate was 166 discharges per 1,000; by 2005, the rates had risen 41.4% to 234 per 1,000. The high rate of growth in utilization among New Haven hospitals went a long way to closing the Boston-New Haven gap: in 1995, discharge rates in Boston were 59% higher than New Haven; by 2005, they were only 20% higher.

At the time of this writing, we are still investigating the question of why, after years of stability, the New Haven profile changed so dramatically. Between then and now, New Haven built more beds, increasing its capacity by about 5.6%, even though the Medicare population did not grow. The New Haven increase in discharge rates was associated with a 29% decline in length of stay. (The drop in lengths of stay in essence released beds that were

then used for new admissions.) The changes in traditional Medicare were also associated with a striking rise, and then a fall, in Medicare HMO enrollment in the intervening years, rising from essentially zero in 1995, peaking at 30% of the Medicare population in 1999–2000, and falling back to 9% by 2003–2005. Unfortunately we do not have records for hospitalizations for the HMO population, nor for the patient population under 65, which are likely essential for fully understanding the sudden shift in practice patterns.

* * *

By the end of the 1980s, our research projects were well on the way to building the factual basis for understanding practice variations for supply-sensitive care. Beginning with the Vermont survey, we saw that while illness obviously influenced patient behavior in seeking medical care—and sicker patients on average got more care than the less sick—illness did not explain the variation in the amount of care patients received in different regions of the state. In Maine, we saw that hospitalization rates for conditions such as hip fractures, which clinicians all agree need to be hospitalized, showed little variation. On the other hand, hospitalizations for conditions such as pneumonia, chest pain, and congestive heart failure varied substantially, much more than seemed plausible on the basis of differences in lung or heart disease.

We continued these studies in Boston and New Haven, where we followed patients when they were hospitalized for heart attacks, hip fractures, and a few other conditions for which the initial hospitalization was considered mandatory. Although it was unlikely that Bostonians with these conditions were sicker than New Havenites, they nonetheless experienced 60% more hospitalizations over a three-year period of follow-up after the index hospitalization, mostly for such medical conditions as pneumonia, chest pain, and congestive heart failure, for which there is no guidance for physicians about when to hospitalize.

We also accumulated evidence that patients living in regions with fewer resources and lower utilization of hospitals were not experiencing worse outcomes. In Vermont, we found no correlation between hospitalization or medical spending and mortality; in Boston and New Haven, mortality rates were similar, even though hospitalization rates were much lower in New Haven. And when we followed victims of heart attacks, stroke, hip fracture, gastrointestinal bleeding, and colon cancer for up to three years, we found no differences in survival between Boston and New Haven patients, despite dramatic differences in their hospitalization rates for high variation medical conditions.

More recently, thanks to the Dartmouth Atlas Project, the scope of our research has expanded. Our findings, summarized in the next two chapters, confirm that the prevalence of illness plays only a minor role in driving practice variation across the United States; that patient preferences do not explain care intensity; and that patient survival, patient satisfaction, and quality of care tend to be worse in regions where care is more intense.

9

Chronic Illness and Practice Variation

The idea that the supply of medical resources can influence utilization is not new—Milton Roemer said it in the 1960s—yet it has proved to be one of the most contentious aspects of our research. Physicians are often deeply threatened by the notion that the supply of everything from hospital beds to slots in their appointment books can influence their day-to-day decisions about their patients, decisions they prefer to believe are grounded in rational medical judgment and sound science. Hospital administrators and boards of trustees do not want to acknowledge that their expensive expansion plans may not always be in the best interests of patients, or society. Nor have economists always been receptive to the specter of systematic market failure resulting from a mismatch between the supply of medical resources and the medical needs and wants of patient populations.

The principal argument made against our characterization of the role of supply factors in influencing utilization has been that regions and hospitals that deliver more services do so because they have sicker patient populations, or they have more demanding patients than regions and hospitals that deliver fewer services per capita. It is certainly possible that residents of a region like, say, Los Angeles want more care than residents of San Francisco. But can patient demand explain the extraordinary variation in utilization that we see between regions? And it is true that sicker patients access the health care system more frequently than less sick patients. This has been evident since the