Linda V. Green

How Many Hospital Beds?

For many years, average bed occupancy level has been the primary measure that has guided hospital bed capacity decisions at both policy and managerial levels. Even now, the common wisdom that there is an excess of beds nationally has been based on a federal target of 85% occupancy that was developed about 25 years ago. This paper examines data from New York state and uses queueing analysis to estimate bed unavailability in intensive care units (ICUs) and obstetrics units. Using various patient delay standards, units that appear to have insufficient capacity are identified. The results indicate that as many as 40% of all obstetrics units and 90% of ICUs have insufficient capacity to provide an appropriate bed when needed. This contrasts sharply with what would be deduced using standard average occupancy targets. Furthermore, given the model's assumptions, these estimates are likely to be conservative. These findings illustrate that if service quality is deemed important, hospitals need to plan capacity based on standards that reflect the ability to place patients in appropriate beds in a timely fashion rather than on target occupancy levels. Doing so will require the collection and analysis of operational data—such as demands for and use of beds, and patient delays—which generally are not available.

In the face of diminishing government subsidies and regulations, increasing competition to obtain contracts with payers, and forecasted decreases in demand for acute care, hospitals are being forced to restructure. In recent years, hospitals increasingly have engaged in mergers, affiliations, downsizings, closings, and the creation of health care networks (Barro and Cutler 1997). One result has been an approximate 25% reduction in the number of hospital beds nationwide during the last 20 years (American Hospital Association 2000).

Much of the current activity is due to the widespread perception that there are currently "too many" hospital beds, and that given decreasing lengths of stay and fewer inpatient admissions the excess will continue to grow. Historically, the supply of hospital beds has been at the center of the debate about the status and future of health care delivery systems in many parts of the country (Billings, Kaplan, and Mijanovich 1996). Health policy analysts, government officials, and others regularly point to the "excess" number of hospital beds in the United States as one of the major reasons for persistently high health care costs (Pasley, Lagoe, and Marshall 1995).

Are there too many hospital beds in the Unit-

Linda V. Green, Ph.D., *is the Armand G. Erpf Professor of Business in the Graduate School of Business, Columbia University. Address correspondence to Prof. Green at Columbia University, Graduate School of Business, 423 Uris Hall, 3022 Broadway, New York, NY 10027-6902.*

Inquiry 39: 400–412 (Winter 2002/2003). © 2002 Excellus Health Plan, Inc.
0046-9580/02/3904–0400$1.25

ed States, in any given community, in any given hospital? More often than not, the assessment by politicians, policymakers and hospital administrators has been: "Yes." Yet, these conclusions do not rely on any service performance measures—such as the availability of an appropriate bed when needed—which generally are not even collected and reported. Furthermore, recent reports in the news media indicate a nationwide increase in the number of hospitals turning away ambulances due to a lack of inpatient beds, an increase in the frequency and duration of such diversions, and an increase in time spent by patients in emergency rooms and hallways waiting for a bed (Goldberg 2000; Shute and Marcus 2001; *New York Times* 2002). So from what criterion is a hospital bed surplus inferred?

Capacity Planning and the Regulation of Hospital Beds

Hospital capacity decisions traditionally have been made, both at the government and institutional levels, based on target occupancy levels—the average percentage of occupied beds. Historically, the most commonly used occupancy target has been 85%. Estimates of the number of "excess" beds in the United States, as well as in individual states and communities, usually have been based on this "optimal" occupancy figure (Brecher and Speizio 1995, p.55). (The current average occupancy rate for nonprofit hospitals is about 64% [American Hospital Association 2000].) The original goal of setting these occupancy targets was to control the supply of hospital beds in order to control costs.

Until recently, the number of hospital beds was regulated in most states by the certificate of need (CON) process, under which hospitals could not be built or expanded without state review and approval. (In the last few years, most of these states have either relaxed or totally eliminated CON requirements.) Though CON procedures may include detailed forecasting methodologies, most are based on the use of average occupancy level targets to ultimately determine the desired number of beds. For example, in New York state, the target occupancies for adult acute care beds have been 85% for urban counties and 80% for rural counties (New York State Department of Health 1993). These target occupancy levels originally were developed at the federal-government level in the 1970s as a response to accelerating health care costs and the perception that more hospital beds resulted in greater demand for hospital care. These occupancy targets were the result of analytical modeling for "typical" hospitals in various size categories and were based on estimates of "acceptable" delays (McClure 1976). Furthermore, occupancy targets have been, and continue to be, the primary measure for determining bed size at the individual hospital level, and even at the hospital unit level (Pendergast and Vogel 1988). Faced with increased pressure to be more cost efficient, some hospitals now are setting target levels that exceed 90%.

Problems with Reported Occupancy Levels

Hospital occupancy levels have been falling largely as a result of two trends: fewer admissions due to technological advances that have allowed for more procedures to be performed on an outpatient basis; and a decrease in average length of stay (ALOS) due largely to prospective payment and managed care, as well as advances in technology. Though current occupancy numbers are generally low, leading to the widespread perception of excess beds, they must be regarded with suspicion for several reasons.

First, hospital occupancy is defined as the ratio of occupied beds to the total number of beds. However, both the numerator and denominator of this ratio have associated measurement problems. First, what is a "bed"? Published occupancy levels usually are based on the total number of *certified* or licensed beds (i.e., beds officially approved by the state). However, internal data used by hospitals typically include both certified beds and beds "in service," where the latter is generally less than the former. For example, a report obtained from Beth Israel Deaconess Medical Center in Boston showed 495 certified beds and 445 beds in service.[1] This is because certified beds often are taken out of service (not staffed) when demand drops. Beds also may be taken out of service, either permanently or temporarily, for reasons of maintenance, construction, patient isolation, or staff shortages. For example, recent renovations in the obstetrics units at Maimonides Hospital in Brooklyn, New York, have resulted in an 11% reduction in their postpartum beds.[2] At other hospitals, inpatient beds have been converted for use as outpatient beds. Yet there is no incentive for hospitals to have these beds decertified. Therefore, the denominator used to calculate occupancy level is often larger than the actual number of beds in service.

Similarly, what is "occupied"? Reported occupancy levels generally are based on the average "midnight census." This refers to the time when hospitals count patients for billing purposes. However, the midnight census usually measures the lowest occupancy level of the day. One reason is the phenomenon known as the "23-hour patient"—a person who is admitted in the morning and discharged in the evening. Managed care companies have encouraged this practice as a way of allowing evaluation of a patient while avoiding unnecessary hospitalization. More generally, patients typically are discharged during the day shift when attending physicians are present. One hospital administrator estimated that when the official occupancy (i.e., the midnight census) rises to what he considers the precariously high level of 87%, the actual peak occupancy during the day is about 95%. At Maimonides Hospital, the average midnight census in the postpartum units is about 10% less than the daily average. Even larger discrepancies have been observed in other hospitals (LaPierre et al. 1999).

Finally, the use of hospital facilities is far from uniform across the week or across the year. Specifically, very few procedures are scheduled for weekends, so elective patients are usually not admitted on weekends when the average daily census is considerably lower. Summer and holiday periods are also slower (Baker et al. 2000) and other seasonal effects have been observed in specific hospitals and/or specific units. Reported occupancy levels are yearly averages, and hence do not reflect significantly higher levels that may exist for extensive periods of time. For all of these reasons, reported occupancy levels are not reliable measures of general bed utilization.

The aforementioned demonstrates that actual occupancy levels are probably higher than reported ones, implying that the current reported national average bed utilization of 64% is an underestimate. Yet, even if the actual number is higher, it is likely below the "desired" level of 85%, indicating there may still be a considerable number of "excess" beds. But is 85% a "good" target occupancy level?

Target Occupancy Levels and Hospital Size and Organization

Although the 85% target is the one most often cited in the literature and in the media, it has long been recognized that smaller hospitals may need to have lower target levels since they do not have the economies of scale of larger institutions. From queueing theory, we know that larger service systems can operate at higher utilization levels than smaller ones while attaining the same level of delays (Whitt 1992). However, there is another critical factor that needs to be considered in evaluating hospital occupancy levels: the number of different *types* of beds. Staffed beds are not all the same.

In most general care hospitals, beds are organized into nursing units. A nursing unit generally corresponds to a specific physical location with a dedicated nursing staff headed by a general nurse manager. For example, at Beth Israel Deaconess, the 445 staffed beds are organized into 17 nursing units, ranging from 20 to 40 beds each. Each nursing unit is used for one clinical service or more (i.e., medical, surgical, pediatric, obstetrics, cardiology, neurology). For convenience and a variety of legal, clinical, and cost reasons, patients are assigned to specific nursing units on the basis of their age and clinical diagnosis. In addition, some units have telemetry beds, which are needed for a significant fraction of patients.[3] Therefore, capacity and utilization must be evaluated for each distinct type of nursing unit in a hospital. In some teaching hospitals, beds are assigned on an even more fragmented basis because they may be controlled by specific physicians or research programs. Thus, for any given hospital, the greater the number of distinct types of beds, the lower will be the resulting utilization that corresponds to some desirable level of bed availability.

This leads to yet another question, which is arguably the most important: What is an "acceptable" delay for a bed? Surprisingly, delays in obtaining beds for patients have almost never been mentioned in the reports and literature on the excess number of hospital beds in the United States. Even at the individual hospital level, delays often are not recorded, nor are there standards for bed availabiity. Yet, within the last couple of years, stories from newspapers, magazines, and television on hospital emergency departments (EDs) report long and increasing delays and severe overcrowding (Shute and Marcus 2001; *New York Times* 2002).

It is important to note that "delay," or more generally bed unavailability, actually manifests itself in a number of different, often complex ways dependent on the specific hospital unit, type of patient, and hospital policy. Most basically, patients can be divided into scheduled and unscheduled admissions. For example, most surgical patients are scheduled, while most patients entering a neurological intensive care unit (NICU) are unscheduled. A day's delay may have little clinical consequence for a surgical patient who is in a post-anesthesia care unit or surgical intensive care unit and is waiting for a bed in the regular surgery unit. However, a delay of a half-hour or even less may have devastating medical consequences for a patient who arrives at the ED and is experiencing some loss of neurological functioning and needs prompt diagnosis and treatment from appropriate specialists.

The unavailability of a bed in one unit may impact the functioning of other parts of the hospital. The most common impact is on the ED. This is the one area of the hospital where bed delays are most likely to be recorded, since the level of ED congestion affects the likelihood that the hospital will have to go "on diversion," that is, send ambulances away to another hospital. For example, the policy at Columbia Presbyterian Hospital in New York City is to go on diversion when 15 or more patients are delayed in being admitted from the ED for lack of an appropriate bed. Other less dramatic results of bed unavailability include: patients being placed "off service" (e.g., a cardiac patient placed in a neurology unit); urgent patients bumping less critically sick patients from intensive care units to "step down units" (with less technical and nursing support); and early discharge of patients to make room for new admissions. Bed unavailability also can lead to holding patients in upstream areas such as the surgical area, where long delays also may result in backups for the operating room (which is often a bottleneck area) causing the postponement or cancellation of surgical procedures. All of these situations have the potential for adverse financial as well as clinical effects (Cohen, Hershey, and Weiss 1980; Sarasin et al. 1996; Morris et al. 1999). And, of course, if patients experience considerable delays for a bed and/or are placed in inappropriate nursing units, patient satisfaction ratings may suffer, an increasingly important concern in the current competitive environment.

From the preceding discussion, it is clear that many fundamental factors must be considered in determining the number of beds that a specific hospital or hospital service should have, and, therefore, whether an "excess" exists. Additional factors also must be considered: hospital location, demographics, and forecasts and patterns of utilization for various services. A different model for determining hospital capacity needs and policies is necessary—one that incorporates these factors. Other types of service organizations, such as telecommunications, airlines, and police, face similar capacity decisions. Typically, their decision making begins with an evaluation of the trade-off between cost and the length and likelihood of a customer's delay for service. In such organizations, this evaluation is facilitated by the use of a queueing model that estimates the impact of a given capacity level on customer delays for service. This generally results in a target average delay or a target probability of losing customers. The utilization level is then a by-product of the analysis, *not* the target itself. The next section starts with the premise that a fundamental mission of a hospital is to provide appropriate medical care and that timeliness is a critical dimension of care. Therefore, capacity decisions should be based primarily on clinically appropriate standards for bed availability.

Data and Modeling Assumptions

Evaluating bed capacity based on a target probability of bed availability or other measure of delay can lead to very different conclusions than would be reached from the use of a target occupancy level. The analyses reported here are based on 1997 data for obstetrics and intensive care units obtained from institutional cost reports (ICRs) for New York state hospitals. These units were chosen because: 1) the vast majority of patients needing these facilities are "urgent" or "emergent" (i.e., must be treated quickly), and hence adequate capacity is particularly important; 2) patients in these units generally cannot be placed off-service; and 3) the ICRs contain separate data for these types of units. The latter two factors make it possible to analyze these units independently of other units in the hospital. The data include number of discharges, ALOS, and average occupancy levels for each hospital.

The analyses use an M/M/s queueing model to estimate delays (Gross and Harris 1985). Due to the robustness of its assumptions and its ease of use, this type of model is used extensively for capacity planning in a very broad variety of service industries. This model assumes a single queue with unlimited capacity that feeds into s identical servers (beds). Arrivals (patient demands for beds) occur according to a time-homogeneous Poisson process with rate λ; the service duration (LOS) has an exponential distribution with mean $1/\mu$. (These assumptions are often called Markovian, hence the use of the two "M's" in the notation used for the model.) Many real arrival and demand processes have been shown empirically to be well approximated by a Poisson process. Among these are demands for emergency services such as police, fire, and ambulance; arrivals to banks and other retail establishments; and arrivals of telephone calls to customer service call centers. Consequently, the Poisson process is the most commonly used arrival process in modeling service systems.

An important characteristic of the exponential distribution is that the mean equals the standard deviation. Another way of saying this is that the coefficient of variation, which is defined as the ratio of the standard deviation to the mean, equals one. The performance that is predicted by the M/M/s model is fairly insensitive to the exponential assumption provided the coefficient of variation of service times is close to one, a characteristic which is found in many real service systems. In general, the greater the coefficient of variation of the service time, the worse the performance of the system.

One advantage of using this model is that given an arrival rate, an average service duration, and the number of servers, closed form expressions for performance measures such as the probability of a positive delay or the mean delay can be obtained easily. The delay is measured from the time of the demand for service (i.e., request for a bed) to the time at which service begins (i.e., a bed is available).

There are many possible performance measures that can be used to determine the efficiency and effectiveness of a service system. In emergency systems such as a hospital emergency room, the most common measure of service performance is the probability that an arrival has any wait. This measure is called probability of delay. If we define p_n to be the steady-state probability that there are n customers in the system, then the probability of delay, p_D, is given by:

$$p_D = 1 - \sum_{n=0}^{s-1} p_n.$$

Server utilization (the average fraction of servers busy), denoted by ρ, usually is considered a measure of system efficiency and is referred to as average occupancy level in the hospital context. It is given as:

$$\rho = \lambda/s\mu.$$

Another common performance measure is the expected wait in queue until a server (bed) is available, W_q. This is given by:

$$W_q = p_D/[(1 - \rho)s\mu].$$

It is important to note that probability of delay and expected delay, as well as other critical measures of customer performance, increase at an increasing rate with server utilization. This not only implies the intuitive notion that higher hospital occupancy levels result in longer delays for beds, but, perhaps nonintuitively, that relatively small increases in occupany levels can result in very large increases in delays, particularly at "critical" levels. Furthermore, the smaller the hospital (or more accurately, the nursing unit), the lower this critical level will be. Thus, large hospital units can operate at higher occupancy levels than small ones and achieve the same delay levels.

Though more complex queueing models have been and should be used to more accurately estimate capacity requirements in specific hospital settings (Hershey, Weiss, and Cohen 1981; Dumas 1984, 1985; Vassilacopoulos 1985; LaPierre et al. 1999; and Green and Nguyen 2001), this simple model was chosen for several reasons. The first is tractability, since the analysis of hundreds of different units requires a model that can be solved quickly and requires only the data which is publicly available. Second, the model's assumption of Poisson arrivals is very reasonable for both obstetrics and intensive care units since the vast majority of patients in these units are unscheduled. Finally, as discussed subsequently, while other assumptions of the model may not be as good, exogenous factors and evidence from other studies indicate that the following analyses generally err on the side of underestimating the true likelihood of bed unavailability. Thus, to the extent that the model's estimates deviate from reality, they do so in a way that support the resulting conclusions that target occupancy levels often underestimate the number of beds needed and that many units have insuffiicient capacity to meet reasonable service performance standards. This is discussed in more detail later.

One questionable assumption of the M/M/s model is that when an arriving patient finds all beds full, the patient must wait until one is available. From the previous discussion, this is clearly not always the case, but is useful here for several reasons. One is that unlike other patient types, those needing an obstetrics or intensive unit bed usually cannot be placed in another type of unit and so often do wait (Green and Nguyen 2001). Also, lack of data and consistent policies would make it impossible to accurately model all of the potential consequences of bed unavailability, particularly for so many different hospitals. Most importantly, the philosophy underlying the analyses is that from a planning perspective, there should be sufficient beds in a unit to assure a given level of availability without off-service placements, bumping, and early discharges.

An Example: Obstetrics

Unlike most other hospital services, such as neurology or cardiology, obstetrics generally operates entirely independently of other services. It is also a service for which the use of a standard M/M/s queueing model is quite good: most obstetrics patients are unscheduled, and in studies of unscheduled hospital admissions (Young 1965) the assumption of Poisson arrivals has been shown to be reasonable; also, the coefficient of variation (CV) of LOS is typically very close to 1.0 (Green and Nguyen 2001).

To illustrate the use of the M/M/s in this context, consider the obstetrics unit of Beth Israel Deaconess Medical Center in Boston (Green and Nguyen 2001). Traditionally, the service is organized with patients moving from labor room to delivery room to recovery room and finally to a postpartum bed. In 1996, the unit had 56 postpartum beds with an ALOS of 2.9 days. The coefficient of variation of 1.04 makes the assumption of exponential service times an excellent one. The average daily arrival rate was 14.8 patients, resulting in an average occupancy level of about 76.5%. Using these data, the M/M/s model estimates that 4% of patients were delayed in the recovery area waiting for a postpartum bed. However, if Beth Israel operated at its target occupancy level of 85%, this probability would rise to more than 16%, and the average delay for waiting patients would be more than eight hours. Using this model, we also can calculate tail probabilities, such as the probability that a patient waits more than two hours, which in this example would be about 22%.

Figure 1 shows the distribution of average occupancy rates for 148 obstetrics units in New York state for 1997. These data, representing nearly all obstetrics units in New York, were obtained from ICRs, and unlike most other published data reflect staffed beds rather than certified beds. The graph shows that many maternity units did have low average occupancy levels. (To some extent, the data show that larger units tended to have higher average occupancies, as would be suggested by queueing theory. However, this pattern reversed for units with more than 50 beds, perhaps because many of these hospitals actually have more than one obstetrics unit, as discussed subsequently.) Since obstetrics patients are generally considered emergent, the American College of Obstetrics and Gynecology (ACOG) has recommended that target occupancy levels for maternity units be 75% (Freeman and Poland 1997), which is considerably lower than the commonly used target of 85%. However, the overall average occupancy level for the study hospitals was only 60%, which, based on the ACOG standard, would imply significant excess capacity. Applying this 75% standard to the 1997 data, 117 of the 148 New York state hospitals had excess beds, while 27 had insufficient beds.

However, if one considers a bed delay target

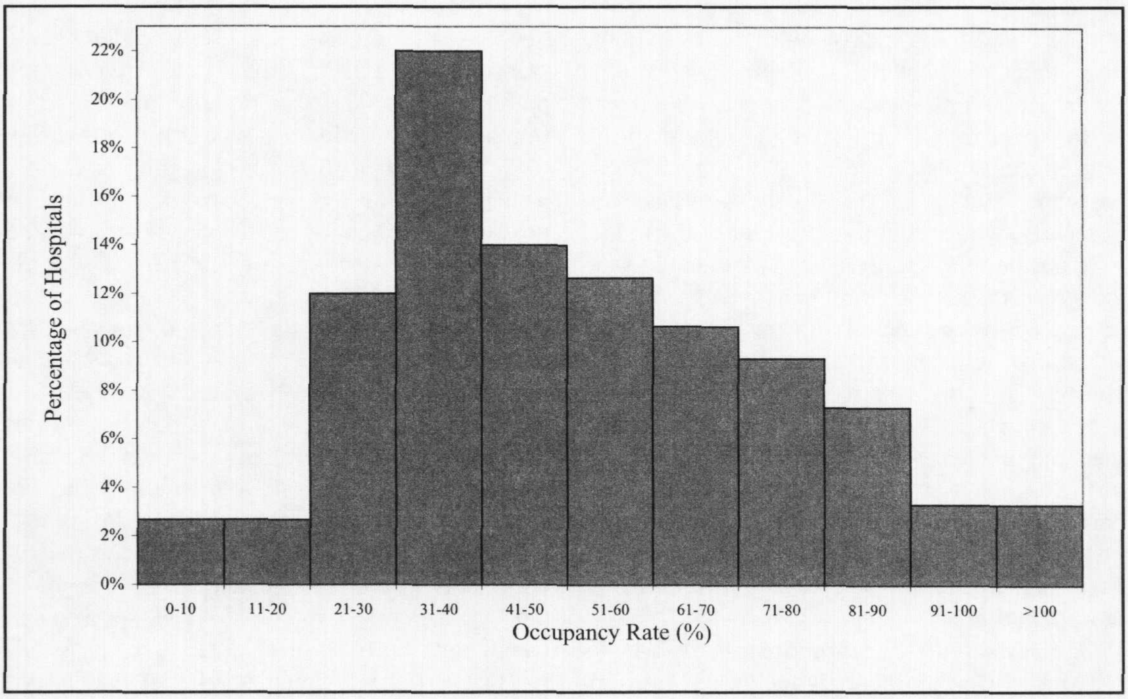

Figure 1. Average occupancy rates of New York state maternity units, 1997

as a more appropriate measure of capacity needs, the conclusions can be quite different. Now the number of beds in each unit becomes a major factor since, for a given occupancy level, delays increase as unit size decreases. While obstetrics units usually are not the smallest units in a hospital, there are many small hospitals, particularly in rural areas, and the units in these facilities may contain only five to 10 beds. Of the New York state hospitals considered here, more than 50% had maternity units with 25 or fewer beds.

In the M/M/s model, probability of delay is a function of only two parameters: s and ρ, which in our context are number of beds and occupancy level. Each of the three curves shown in Figure 2 represents a specific probability of delay as a function of these two variables as generated by the model. Thus, using the unit size and occupancy level reported on the ICR report for a given maternity unit, we can determine from this figure if the probability of delay meets or exceeds any one of these targets. For example, if a maternity unit has 15 beds and an occupancy level of 45%, it would fall below all three curves and hence have a probability of delay less than .01 or 1%, meeting all three targets. Doing this for every hospital in the database, 30 hospitals had insufficient capacity based on even the most slack delay target of 10%. (It is interesting to note that two of the hospitals that would be considered overutilized under the 75% occupancy standard had sufficient capacity under this delay standard.) Tightening the probability of delay target to 5% yields 48 obstetrics units that do not meet this standard; adopting a maximum probability of delay of 1% as was suggested in the only publication identified as containing a delay standard for obstetrics beds (Schneider 1981), then 59, or 40%, of all New York state maternity units can be deemed to have insufficient capacity.

How many hospitals in New York state have maternity units large enough to achieve the ACOG-suggested 75% occupancy level and also meet a specified probability of delay standard? Using Figure 2, we see that for a 10% target, an obstetrics unit would need to have at least 28 beds, a size that exists in only 40% of the state hospitals. For a 5% standard, the minimum number of beds needed is 41, a size achieved in only 14% of the hospitals; for a 1% standard, at

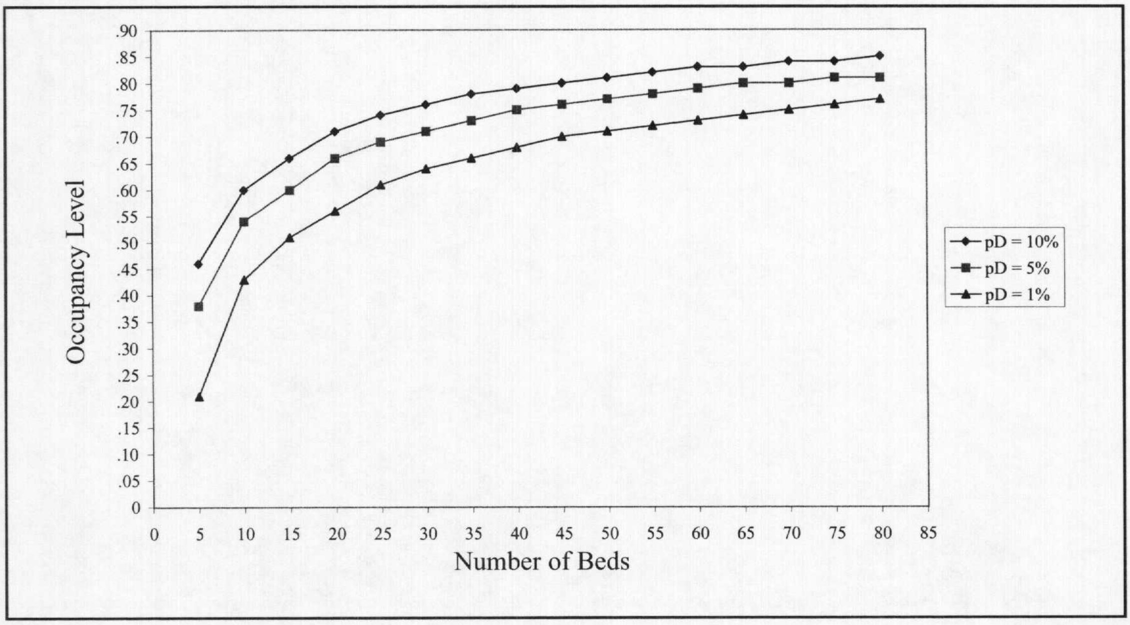

Figure 2. Probability delay (PD) by occupancy and size of maternity units

least 67 beds are needed, leaving only three of the 148, or 2%, of the hospitals of sufficient size.

These estimates are likely to be conservative for several reasons. First, 19 of the hospitals represented in Figure 1 have multiple facilities as a result of mergers or other affiliation agreements. In at least some of these, the number of maternity beds reported is the sum of the beds in two or more geographically distinct hospitals. Therefore, the actual unit sizes are smaller, and hence lower utilizations would be needed to achieve the given delay targets. In addition, day of week and seasonal fluctuations in demand are common, which means that actual delays will be higher than those predicted by a model which assumes a constant rate of demand (Green and Nguyen 2001). For example, at Beth Israel Deaconess, the occupancy level rises to approximately 88% in July, boosting the estimated probability of delay to almost 25% from a low of nearly zero in January.

What are the possible consequences of congestion? First, it is important to note that the obstetrics beds depicted in Figure 1 are primarily postpartum beds. While patients in some hospitals remain in the same bed through labor, delivery, recovery, and postpartum, in most maternity units, as in Beth Israel Deaconess, there are separate areas for some or all of these stages of birth. Therefore, a delay for an obstetrics bed often means that a postpartum patient will remain in a recovery bed longer than necessary. This, of course, may cause a backup in the labor and delivery areas so that newly arriving patients may have to wait on gurneys in hallways or in the emergency room. Some hospitals have overflow beds in a nearby unit that is opened (staffed) when all regular beds are full. (This is likely the case for the five hospitals that reported average occupancy levels exceeding 100%.) In some hospitals, congestion results in some patients being discharged earlier than normal. While these effects of congestion likely pose no medical threat for most patients who experience normal births, there could be adverse clinical consequences in cases in which there were complications. In particular, whether patients are placed in hallways or overflow units, the nursing staff is likely to be severely strained, thereby limiting the quantity and quality of personal attention. Even if a hospital is able to obtain additional staffing, it is usually by using agency nurses who are more expensive and not as familiar with the physical or operating environment, thereby jeopardizing quality of patient care. In addition, telemetry devices, such as fetal monitors that are usually in labor and deliv-

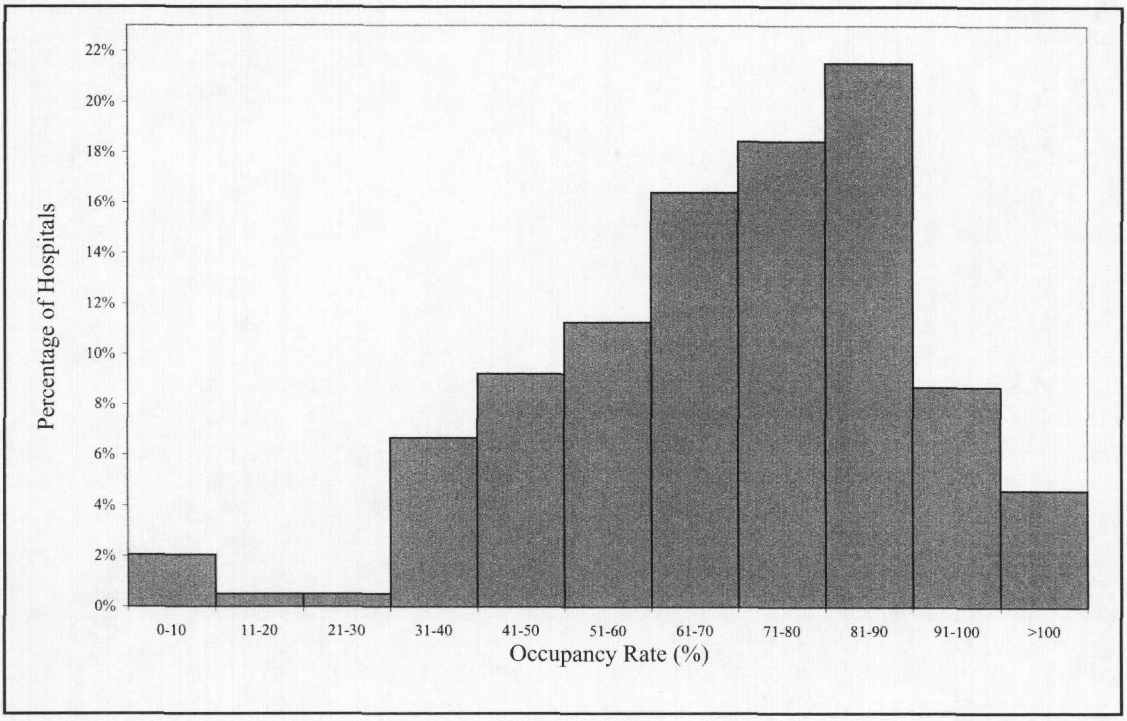

Figure 3. Average occupancy rates of New York state intensive care units, 1997

ery rooms, may be unavailable in other locations, thus compromising the ability to monitor vital body functions of both mother and baby. Again, it is worth noting that such results of congestion may negatively affect patients' perceptions of service quality.

Based on all of this, inferring the number of "excess" obstetrics beds based on the 75% occupancy standard is likely to lead to a significant overestimate. More importantly, using a 75% occupancy target to plan capacity may have adverse consequences for maternity patients. Several hospitals represented in Figure 1 have reduced the size of their obstetrics units in the past several years, and based on conversations with hospital administrators, often did so because utilization fell below 75%. Given increasing cost pressures, there is little doubt that others will follow suit.

An Example: Intensive Care

Although admissions to hospitals are decreasing due to pressure from managed care to reduce hospital use and the growing number of procedures that can be done on an outpatient basis, patients who are critically ill still need the facilities of a hospital. In particular, such patients often need the resources of an intensive care unit (ICU). However, ICUs are usually the most expensive units in the hospital due to both the technology used and staff needed. The full per-day cost in an ICU is about three to five times as much as in a regular inpatient unit (Groeger et al. 1992). ICUs tend to be quite small and are used only for patients who need the intense monitoring provided by these units. Figure 3 shows the distribution of average occupancy levels for ICUs in New York state in 1997. It is important to keep in mind that the average size of the units in this sample was only 15 beds and the mode was 10 beds. There are no occupancy standards for ICUs, but the data show an average occupancy of 75%. It is interesting to note that with the exception of extremely small units (five or fewer beds), which had an average occupancy of 47%, occupancy levels did not vary systematically with size. Given the overall 85% rule-of-thumb for occupancy levels, it might appear that these units were not optimally utilized. However, employing an M/M/s model to estimate delays reveals a very different picture. Since probability of delay depends only on size and server utilization, we can again use Figure 2 to estimate the number of beds needed to meet our various delay targets and the resulting occupancy levels.

Adopting the standard of a maximum probability of delay of 10%, 112 of the 194 ICUs, or about 58%, were overutilized. If that target is reduced to 5%, 143 or 74% of the units were too small to handle their experienced workloads; for a 1% target, 175 or over 90% were of inadequate size. As with the obstetrics units, these estimates are likely to be conservative for several reasons. First, in an analysis of intensive care units at Beth Israel Deaconess, the coefficient of variation of LOS ranged from 1.1 to 1.6 (Green and Nguyen 2001), suggesting that the M/M/s assumption of exponential service times leads to underestimates of actual congestion. Second, as stated before, several of the hospitals have multiple divisions or locations, and the reported units are sometimes the sum of two or more smaller units. Another reason is that in many larger hospitals there are several types of ICUs (e.g., medical, surgical, neurological, and cardiac). Therefore, some of the unit sizes reported in this data likely represent the combined size of smaller, specialized units.

It is also important to note that the average reported length of stay in these units was almost 18 days, so that if a patient experienced a delay, it could be quite long. One possible consequence of long delays for critical care beds is ambulance diversions, as reported in recent news accounts (Goldberg 2000; *New York Times* 2002). However, when possible, a current occupant may be "bumped" from an ICU bed earlier than planned and transferred to another unit. (Friedman and Steiner [1999], using data from Massachussets and Florida, found that patients in hospitals with the most constrained supply of ICU beds relative to demand received fewer ICU resources: 28% shorter LOS in Massachusetts and 56% fewer ICU-associated services in Florida.) The receiving unit may be another type of ICU or a "step-down" unit, where there are telemetry beds but the nursing level is lower and less skilled than in the original ICU. Another possible scenario is that a "bumped" patient may be placed in a nontelemetry bed that is "jury-rigged" with additional equipment and staffing to mimic as closely as possible the care administered in the ICU. (One or more of these situations are likely the reason that seven hospitals reported average occupancy levels exceeding 100%.) Many of these scenarios are likely to be suboptimal from both the patient and hospital perspective.

Conclusions and Discussion

The preceding examples illustrate how the current definition, and hence estimates, of excess hospital capacity may be misleading and potentially dangerous. Similar arguments and analyses can be based on data from even larger hospital units such as those in general medicine-surgery. For example, the overall 1997 reported average occupancy level for medical-surgical beds in New York City was 69.6%, which might be considered low given that most New York City hospitals have hundreds of such beds. However, in most hospitals, beds are not all interchangeable. Rather, patients are assigned beds according to clinical service and sometimes even by subspeciality, particularly in academic medical centers, which often have the highest occupancy levels (Heisler 2000).

Of course, any reasonable analysis of required capacity should consider the factors that may affect the demand and use of the service in the future. While many believe that admissions to hospitals and ALOS will continue to fall, resulting in even lower occupancy levels, this assumption could be dangerously wrong. For example, New York City experienced a severe and protracted citywide shortage of inpatient hospital beds in 1987/88 (Myers, Fox, and Vladeck 1990) due to an unexpected 18% growth in admissions as a result of the AIDS epidemic and increased drug abuse. During this period, ambulances were routinely turned away from full hospitals and urgently sick patients experienced delays of days waiting for an open bed. During the two years prior to this hospital crisis, there was a 9% decline in capacity that was due largely to new state regulations linking Medicaid reimbursement to occupancy levels (which were regulated to be at least 85%) so as to reduce the number of "excess" beds in the city. There is no reason to believe that other unanticipated disease outbreaks or increases in risky behaviors will not occur in the future.

In addition, at high levels of system utilization, queueing delays are extremely sensitive to even temporary increases in arrival rates. For example, as noted previously, obstetrics units often experience increased arrivals in the summer. In the case of Beth Israel Deaconess cited earlier, this means that the average occupancy level in July rises to 88%. If, on a given day, the arrival rate were 10% higher than this, the probability of delay as predicted by the M/M/s model would rise to over 65% with very long delays for those patients experiencing a delay. Understanding the potential for extremely long delays at critically high occupancy levels is even more important since the events of September 11, 2001, which have led to pressures on hospitals to maintain "surge capacity."

Of course, the level at which occupancy becomes "critical" is dependent on the size of a hospital unit. While for large hospitals this level might start at 80% or 85%, for small hospitals it could be as low as 45% for certain units. This implies that rural hospitals are likely to need even more "surge capacity" proportionally, given their generally small sizes and lack of proximity to other facilities that can accommodate overflow in an emergency. One option, as mentioned in the discussion of obstetrics beds, is to keep "overflow" beds that generally are unstaffed but may be staffed for use when admission spikes occur. This, of course, is dependent on the ability of the hospital to obtain additional nursing staff on short notice.

Furthermore, it is important to note that the median age in the United States is increasing. Given that older people have a higher likelihood of being hospitalized and of experiencing longer lengths of stay once they are in the hospital, it seems unlikely that the current trends in admissions and ALOS are likely to last for a long time. Indeed, the latest statistics from New York City show slight increases in admissions and a tapering off of ALOS reductions for the last two years (Heisler 2000). Hospitals need to develop forecasts for admissions and ALOS for each clinical area based on changing demographics, as well as new technologies and clinical management methodologies.

So are hospitals doomed to operate at occupancy levels that will result in financial losses? Not necessarily. First, it is informative to look at occupancy levels in the for-profit sector. In 1998, the average occupancy level for all community and government hospitals was 63.9%; for investor-owned hospitals it was only 53.2% (American Hospital Association 2000). Although there are many factors that may account for this discrepancy, including average size of hospitals, this observation indicates that financial health is not necessarily dependent on high occupancy levels. In particular, nursing units with a higher degree of control over admissions can operate at higher occupancy levels without incurring unacceptable delays for beds. More generally, though the analyses in this paper focused on nursing units with too few beds, other units, perhaps in the same hospital, may have too many. By using more sophisticated models to better identify the needs of each unit, hospitals can improve bed availability by reallocating beds from some clinical areas to others.

Second, there are ways for hospitals to achieve greater efficiencies in bed utilization through the use of more flexible nursing units, identification and better capacity management of bottleneck areas, and appropriate sizing and staffing of support services such as laboratories and radiology. Some investments—such as cross training nurses and increased use of telemetry—may be needed to realize some of these improvements. However, the resulting savings due to economies of scale, decreased ALOS, and fewer transfers among units likely would make these investments financially worthwhile as well as increase service levels and patient satisfaction. There are almost certainly other opportunities for increasing operational efficiency as well. One example is in the management of nurse staffing levels. Though most hospitals try to adjust the level of nurses across the day and week to account for changes in census, many (if not most) do not use any formal optimization models and, as a result, wind up with high costs for overtime and agency nurses that likely could be reduced. The author's personal experience dealing with nurse staffing in a large hospital, as well as conversations with administrators in other hospitals, supports this hypothesis.

In summary, hospital executives and government officials need to be better informed about the factors affecting the trade-off between utilization and the ability to provide an appropriate bed in a timely fashion. These include nursing unit sizes, the variability and time-dependent patterns of demands for beds, and bed allocation policies. Most importantly, capacity planning and management should be driven primarily by clinical and service performance standards, not target occupancy levels. Without such standards, it is impossible to make any real determination about what is the "right" number of beds for a given nursing unit or hospital. In order to assure quality care and service, policymakers and hospital executives must collect and track data on critical service performance indicators, such as probability and lengths of waits for beds, ambulance diversions, the frequency of patient bumping, and the fraction of patient days spent in an inappropriate unit. Of course, evaluations of bed capacity requirements also are related to the levels and utilization of other health care resources such as physicians, nurses, and various types of technology. Comprehensive models are needed to assess cost-benefit trade-offs and identify opportunities for increased efficiency and effectiveness, and all of these issues must be analyzed in the context of an increasingly complex and dynamic health care environment.

Notes

[1] This was based on 1997 data obtained from Beth Israel Deaconess and based on the former Beth Israel Hospital only.

[2] This was based on 2000 data obtained from Maimonides Hospital.

[3] Telemetry beds are those that are equipped with electronic monitoring of vital functions.

References

American Hospital Association. 2000. *Hospital Statistics 2000.* Chicago, Ill.: American Hospital Association.

Baker, L., C. Phibbs, J. Reynolds, and D. Supina. 2000. Within-Year Variation in Hospital Utilization and Its Implications for Hospital Costs. Unpublished.

Barro, J.R., and D.M. Cutler. 1997. Consolidation in the Medical Care Marketplace: A Case Study from Massachusetts. NBER Working Paper 5957. Cambridge, Mass.: National Bureau of Economic Research.

Billings, J., S. Kaplan, and T. Mijanovich. 1996. Projecting Hospital Utilization and Bed Need in New York City for the Year 2000. *HRP Reports.* New York University.

Brecher, C., and S. Speizio. 1995. *Privatization and Public Hospitals.* New York: Twentieth Century Fund Press.

Cohen, M.A., J.C. Hershey, and E.N. Weiss. 1980. Analysis of Capacity Decisions for Progressive Patient Care Hospital Facilities. *Health Services Research* 14: 145–160.

Dumas, M.B. 1984. Simulation Modeling for Hospital Bed Planning. *Simulation* 43: 69–78.

———. 1985. Hospital Bed Utilization: An Implemented Simulation Approach to Adjusting and Maintaining Appropriate Levels. *Health Services Research* 20: 43–61.

Freeman, R.K., and R.L. Poland. 1997. *Guidelines for Perinatal Care,* 4th ed. Washington, D.C.: American College of Obstetricians and Gynecologists.

Friedman, B., and C. Steiner. 1999. Does Managed Care Affect the Supply and Use of ICU Services? *Inquiry* 36(1): 68–77.

Goldberg, C. 2000. Emergency Crews Worry as Hospitals Say, "No Vacancy." *New York Times* (Dec. 17): 39.

Green, L.V., P. J. Kolesar, and A. Svoronos. 1991. Some Effects of Nonstationarity on Multi-server Markovian Queueing Systems. *Operations Research* 39: 502–511.

Green, L.V., and V. Nguyen. 2001. Strategies for Cutting Hospital Beds: The Impact on Patient Service. *Health Services Research* 36: 421–442.

Groeger, J.S., K.K. Guntupalli, M. Strosberg, N. Halpern, R.C. Raphaely, F. Cerra, and W. Kaye. 1992. Descriptive Analysis of Critical Care Units in the United States. *Critical Care Medicine* 21: 279–291.

Gross, D., and C.M. Harris. 1985. *Fundamentals of Queueing Theory,* 2nd ed. New York: John Wiley & Sons, Inc.

Health Care Financing Administration (HCFA). 2001. *U.S. National Health Accounts.* Table 1, Office of the Actuary, National Health Statistics Group. Rockville, Md.: HCFA.

Heisler, T. 2000. *Health Care Annual.* New York: United Hospital Fund of New York.

Hershey, J.D., E.N. Weiss, and M.A. Cohen. 1981. A Stochastic Service Network Model with Application to Hospital Facilities. *Operations Research* 29: 1–22.

LaPierre, S.D., D. Goldsman, R. Cochran, and J. Dubow. 1999. Bed Allocation Techniques Based on Census Data. *Socio-Economic Planning Sciences* 33: 25–38.

McClure, W. 1976. *Reducing Excess Hospital Capacity.* Excelsior, Minn.: Bureau of Health Planning.

Morris, D.L., W.D. Rosamond, A.R. Hinn, et al.

1999. Time Delays in Accessing Stroke Care in the Emergency Department. *Academic Emergency Medicine* 6: 218–223.

Myers, L.P., K.S. Fox, and B.C. Vladeck. 1990. Health Services Research in a Quick and Dirty World: The New York City Hospital Occupancy Crisis. *Health Services Research* 25: 739–755.

New York State Department of Health. 1993. *Acute Care Bed Need Methodology Background for the Derivation of 1996 Adult and Pediatric Bed Need.* Albany: Bureau of Health Facility Planning.

New York Times. 2002. 1 in 3 Hospitals Say They Divert Ambulances. (April 9).

Pasley, B.H., R.J. Lagoe, and N.O. Marshall. 1995. Excess Acute Care Bed Capacity and Its Causes: The Experience of New York State. *Health Services Research* 30 (1): 115–131.

Pendergast, J.F., and W. B. Vogel. 1988. A Multistage Model of Hospital Bed Requirements. *Health Services Research* 23 (3): 381–399.

Sarasin, F.P., M. Louis-Simonet, J.M. Gaspoz, et al. 1996. Detecting Acute Thoracic Aortic Dissection in the Emergency Department: Time Constraints and Choice of the Optimal Diagnostic Test. *Annals of Emergency Medicine* 28: 278–288.

Schneider, D. 1981. A Methodology for the Analysis of Comparability of Services and Financial Impact of Closure of Obstetrics Services. *Medical Care* 19: 395–409.

Shute, N., and M.B. Marcus. 2001. Code Blue: Crisis in the ER. *U.S. News & World Report* (September 10): 55–61.

Vassilacopoulos, G. 1985. A Simulation Model for Bed Allocation to Hospital Inpatient Departments. *Simulation* 45: 233–241.

Whitt, W. 1992. Understanding the Efficiency of Multi-server Service Systems. *Management Science* 38: 708–723.

Young, J.P. 1965. Stabilization of Inpatient Bed Occupancy through Control of Admissions. *Journal of the American Hospital Association* 39: 41–48.