Strategies for Cutting Hospital Beds: The Impact on Patient Service

Linda V. Green and Vien Nguyen

Objective. To develop insights on the impact of size, average length of stay, variability, and organization of clinical services on the relationship between occupancy rates and delays for beds.

Data Sources. The primary data source was Beth Israel Deaconess Medical Center in Boston. Secondary data were obtained from the United Hospital Fund of New York reflecting data from about 150 hospitals.

Study Design. Data from Beth Israel Deaconess on discharges and length of stay were analyzed and fit into appropriate queueing models to generate tables and graphs illustrating the relationship between the variables mentioned above and the relationship *between occupancy levels and delays. In addition, specific issues of current concern* to hospital administrators were analyzed, including the impact of consolidation of clinical services and utilizing hospital beds uniformly across seven days a week rather than five.

Principal Findings. Using target occupancy levels as the primary determinant of bed capacity is inadequate and may lead to excessive delays for beds. Also, attempts to reduce hospital beds by consolidation of different clinical services into single nursing units may be counterproductive.

Conclusions. More sophisticated methodologies are needed to support decisions that involve bed capacity and organization in order to understand the impact on patient service.

Key Words. Hospital capacity planning, occupancy levels, bed delays, clinical consolidation

The largest single source of health care costs is hospitals, which account for close to 40 percent of all health care expenditures. 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 have engaged in various cost-cutting efforts, including downsizing, mergers, consolidation of small services, decreasing average length of stay (ALOS), and establishing "clinical pathways," which standardize clinical care protocols and hence reduce the variability of length of stay (LOS). All of these strategies involve a reassessment of the number of beds needed to serve the hospital's target population. Indeed, much of the current activity is due to the widespread perception that, given decreasing ALOS and fewer inpatient admissions, there are "too many" hospital beds (Pasley, Lagoe, and Marshall 1995).

In determining the number and organization of hospital beds, managers must consider many factors, including costs, the likelihood and length of patient backups in the emergency room, the probability of turning patients away, waits for elective patients, and the medical and satisfaction consequences of placing a patient in an inappropriate unit (e.g., putting a cardiac patient in a noncardiac-care unit). In many other service systems, such as telecommunications, airlines, and police (Brigandi et al. 1994; Brusco, Jacobs, Bongiorno, et al. 1995; Taylor and Huxley 1989), similar capacity decisions are framed in terms of the trade-off between cost and the length and likelihood of a customer's delay for service. In these organizations, a target average delay or a target probability of losing customers is chosen, and a queueing model is used to determine the minimum capacity needed to meet that target. Hospital managers, on the other hand, generally employ a simpler approach in planning bed capacity, relying primarily on target occupancy levels. Target occupancy levels, which may vary by clinical service within a given hospital, are assumed to reflect capacity levels that achieve an appropriate balance of cost and patient delays. Yet, the marginal cost of a bed is unclear, and a given occupancy level may result in very short or very long delays depending on other factors such as size and the variability in demand and length of stay.

Although the impact on the medical outcome of a patient's delay for an appropriate bed is difficult to measure and clearly depends upon the specific medical condition, hospitals do recognize the adverse consequences of delays. Admissions are classified as emergent, urgent, or elective according to what is deemed a maximum tolerable delay. Unavailability of beds may require placement of patients in less appropriate units, often compromising the quality of care and possibly resulting in increased costs, for example, when additional

Address correspondence to Linda V. Green, Ph.D., Armand G. Erpf Professor of Business, Graduate School of Business, Columbia University, 423 Uris Hall, 3022 Broadway, New York, NY 10027-6902. Vien Nguyen, Ph.D. is Vice President, Morgan Stanley Dean Witter & Co., New York, NY. This article, submitted to *Health Services Research* on April 12, 1999, was revised and accepted for publication on February 4, 2000.

staffing is needed. Moreover, backups in emergency rooms, surgical recovery rooms, and labor/delivery rooms impede the ability to treat new patients and may result in lost revenues from the hospital going "on diversion," that is, sending patients to other hospitals. Finally, in an increasingly competitive environment, delays will likely become more important in consumers' evaluation of hospitals. Yet, explicit standards for acceptable levels of delay do not generally exist for specific patient categories, and actual delays are not systematically recorded nor explicitly used in evaluating the impact of cost-cutting measures.

The major purpose of this article is to apply a queueing model approach to the hospital bed planning issue to gain insights on the potential impact of cost-cutting strategies on patients' delays for beds. Using this approach, we also identify those factors that have the greatest impact on the trade-off between hospital occupancy levels and delays. Our analyses are based on data obtained from a major New England hospital but are designed to provide more general insights. Specifically, we (1) explore the conditions under which downsizing or increasing admissions to achieve a given target occupancy level may result in undesirable service performance; (2) examine the impact of consolidating various hospital services into single managerial units in order to increase bed flexibility; (3) provide insight on the effectiveness of reducing ALOS or reducing variability in LOS; and (4) illustrate the danger of ignoring weekly and yearly patterns of demand variability when determining bed capacities. Overall, these analyses reveal the shortcomings of using target occupancy levels for capacity planning in hospitals and the need for more sophisticated decision support systems.

Background and Related Literature

Although there is extensive literature related to the management of hospital resources, very little of it addresses the types of issues described above. In the field of operations research, capacity planning in hospitals has been the subject of several studies (Hershey, Weiss, and Cohen 1981; Dumas 1984, 1985; Vassilacopoulos 1985; Worthington 1987). However, these have focused on the detailed assessment of allocation and scheduling rules or on the development of models for specific facilities. In the economics literature, Graham and Cowing (1997) studied the determinants of hospital reserve margins—beds in excess of average patient census. Their results indicate that larger hospitals and hospitals with more diverse clinical services have smaller reserve margins, indicating both economies of scale and scope. Research on the cost of empty hospital beds has been the subject of several studies, including Schwartz and Joskow (1980), Friedman and Pauly (1981, 1983), Pauly and Wilson (1986), and, more recently, Gaynor and Anderson (1995) and Keeler and Ying (1996). Much of this work is framed as estimating the cost of excess capacity, which is defined as the average unutilized capacity. Yet, uncertain demands and lengths of stay require that hospitals operate at less than 100 percent utilization in order not to turn away a large percentage of patients, and therefore it is unclear at what occupancy level a hospital truly has excess capacity, that is, more than needed to provide a target level of service. Gaynor and Anderson (1995) explicitly incorporate this concept of capacity choice being dependent on a target probability of turning away patients in their calculation of the cost of an empty bed.

Target occupancy levels were originally developed at the federal government level in the 1970s to control accelerating health care costs (McClure 1976). These occupancy targets were the result of analytical modeling for "typical" hospitals in various size categories and were based on "acceptable" delays. Historically, the most commonly used occupancy target has been 85 percent, and current estimates of the number of "excess" beds are usually based on this "optimal" occupancy figure (Brecher and Speizio 1995, p. 55). (The current average occupancy rate for nonprofit hospitals is about 63 percent; AHA 1996.) In recent years, some managers of large hospitals have used target levels greater than the traditional 85 percent because of increased financial pressures. Research from queueing theory (Whitt 1992) and economics (Lynk 1995) support the conclusion that larger clinical units can achieve higher utilization levels than smaller ones in trying to achieve a given patient delay objective. While Pauly and Wilson (1986) found no explicit relationship between occupancy level and cost, the more recent work of Keeler and Ying (1996) found that increasing utilization of beds lowers costs.

PROBLEMS IN USING OCCUPANCY LEVELS FOR CAPACITY PLANNING

It is important to note that reported occupancy levels—defined as the ratio of the average daily census (ADC) to the number of inpatient hospital beds—may be inaccurate or misleading for the following reasons. (1) Published occupancy levels are based on the number of *certified* beds, that is, approved by the state or, alternatively, reported by hospitals to the state department of health. However, certified beds are often taken out of service (not staffed) when use decreases to cut costs or for reasons of maintenance, construction, patient isolation, or staff shortages. (2) Occupancy is measured based on the "midnight census," used for billing purposes, which generally reflects the lowest occupancy level of the day. (3) Reported occupancy levels are yearly averages and hence do not reflect significantly higher levels that may exist for extensive periods of time due to seasonal effects and significant disparities between weekdays and weekends (when few procedures are scheduled).

From the consumer perspective, a target occupancy level does not necessarily correspond to a desired level of service, for example, the waiting time for a specific bed type. In the following analyses, we will assume that both fixed and variable hospital bed costs are known and hence focus on the impact of various factors on the relationship between occupancy and service level.

METHODOLOGY

Most of our 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. The model assumes arrivals (patient demands for beds) occur according to a Poisson process and that the service duration (*LOS*) has an exponential distribution. The number of servers (beds), s, can be varied to determine the impact on patients' delays for beds. 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 easily obtained. 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). In our context, the time of the demand would depend on the patient type and whether the patient is arriving from outside the hospital or being transferred from within. As in other hospitals, most admissions (65 percent) to Beth Israel are unscheduled and fall into two categories. Emergent patients are admitted through the emergency room and usually require a bed almost immediately. Urgent patients are those admitted from the outside and require a bed within a day or less. The demand epoch for these patients, as well as some elective patients, is the time at which the emergency room physician or referring physician requests that the patient be admitted as an inpatient. For most elective admissions, usually surgical patients, the demand epoch generally corresponds to the time at which the patient comes out of the recovery room or intensive care unit after surgery. The Poisson arrival assumption for the unscheduled patients is very reasonable based on prior studies of unscheduled arrivals to hospitals (Young 1965). As for surgical units, which usually have a substantial fraction of scheduled patients, the Poisson assumption may result in overestimates of delays, which are likely to be offset by other factors. In most of the clinical services we studied, the coefficient of variation (CV) of LOS was very close to 1.0, so the assumption of exponential service times is good. In the services for which this is not a good assumption, we use an *M/G/s* approximation (since there are no exact closed form expressions for delays in this case) in which service times are assumed to have an arbitrary distribution and the delay is dependent on both the mean and standard deviation (Hokstad 1978). To study consolidation of services, we also use a variant of the *M/M/s* model in which one class of patients has priority for service over the other (Cobham 1954). Delay estimates for the analyses that assume time dependent arrival rates are based on numerical integration of the set of differential equations that descirbe the system dynamics. [See Green, Kolesar, and Svoronos (1991) for more detail.] In all cases, the models assume that patients are not assigned a bed in an alternative unit or are turned away if delays get long. In reality, this may occur if other units have spare capacity or if the hospital goes on diversion. Another actual possiblility for handling long delays in some units is staffing noncertified beds.

DESCRIPTION OF THE DATA

The data used in our analyses were obtained from Beth Israel Deaconess Hospital in Boston, which was officially created in the fall of 1996 as the result of the merger of two Harvard-affiliated hospitals. At the time of this analysis, clinical consolidation was being planned but had not yet occurred. Therefore, these data reflect only the operations of the former Beth Israel Hospital.

Discharge length of stay reports were obtained for all patients over the three-year period beginning October 1, 1993 and ending September 30, 1996. They provide the distribution of LOS for all patients discharged from the hospital each month, categorized by the hospital service from which the patient is discharged. A service corresponds to a clinical department comprised of physicians in a given medical specialty. There are 20 hospital services at Beth Israel. Unfortunately, admissions data are collected not by service but by physical location and type of admission, for example, elective, urgent, emergency. There is no simple way to connect admissions information to services, and patients are sometimes transferred among units and services during their stay in the hospital. After conferring with hospital personnel, we determined that for our purposes, discharge data would be a reasonable surrogate for arrivals where necessary.

The discharge LOS data is the total LOS in the hospital and may include time in multiple services. LOS for each individual service is not kept. Since the number of transfers in and out of most services is relatively small, we were assured that this data would generally provide reasonable estimates for LOS per service.

We also obtained admissions data categorized by day of the week for a six-month period in 1997 to explore the impact of weekly patterns of demand variability. At Beth Israel, bed assignments are made by a central admitting department to a specific nursing unit based primarily on the clinical service for which the patient is admitted. A nursing unit corresponds to a specific physical location with a dedicated nursing staff headed by a nurse manager. Although nursing units usually correspond to a single clinical service, some nursing units consist of multiple services that were consolidated to achieve greater bed flexibility.

To determine the generalizability of our findings concerning obstetrics, we also examined 1998 ALOS data for obstetrics units from the roughly 150 hospitals surveyed by the United Hospital Fund of New York (Heisler and Cantor 1998).

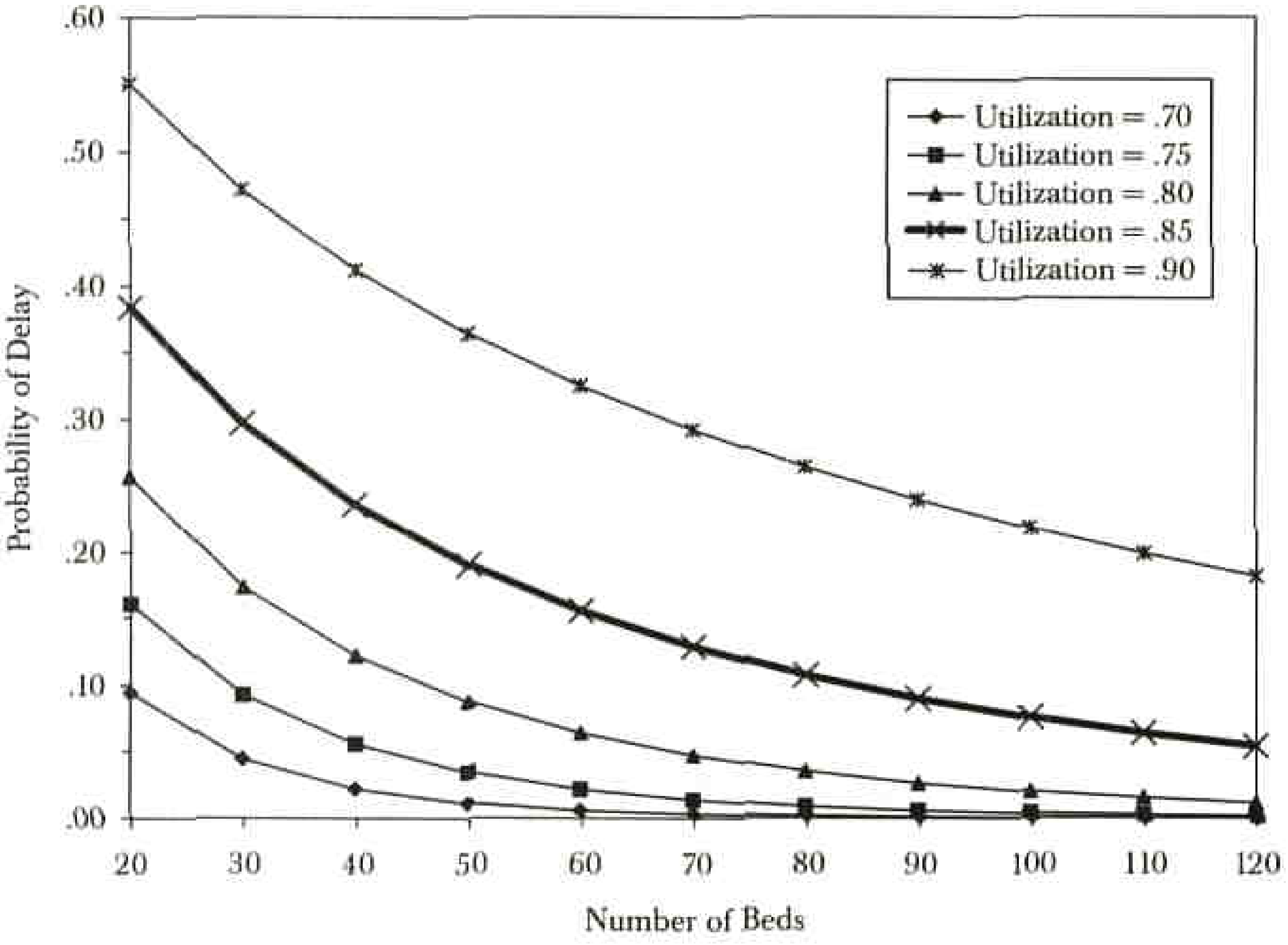
IMPACT OF SIZE AND URGENCY ON OPTIMAL OCCUPANCY LEVELS

We studied two types of services—obstetrics and surgery—to better understand the relationship of size, occupancy levels, and patient delays in hospitals.

Obstetrics

We used Beth Israel's ALOS of 2.9 days and looked at the effect of unit size and occupancy levels on delays. Figure 1 shows probability of delay (p_D) as a function of the number of beds for utilization levels ranging from 0.70 to 0.90. We chose probability of delay as our prime measure of service performance

Figure 1: Obstetrics, ALOS = 2.9



because obstetrics patients are classified as emergent, that is, requiring an immediate bed. (Probability of delay is the standard measure used in service systems where customers have low tolerance for any delays, particularly emergency systems.) Several interesting observations can be made. First, note that the occupancy level recommended by the American College of Obstetrics and Gynecology (ACOG) is 75 percent (Freeman and Poland 1992). Looking at the 0.75 curve, we see that if it is desirable to keep p_D under 0.10, this could only be achieved in obstetric units larger than 30 beds. Yet smaller hospitals, often in rural areas, may have far fewer beds. On the other hand, a hospital such as Beth Israel, which has about 56 obstetrics beds, could increase its occupancy to 80 percent and still stay below this target. At their target occupancy of 85 percent the model estimates that about 16 percent of patients will not get a bed when needed. It is important to note that the authors have not come across any official or operational standard regarding patient delays for obstetrical beds. At least one article on analyzing the need for obstetrical beds used a p_D target of 0.01 (Schneider 1981), implying that such units would have to be quite large to operate at high utilizations or, conversely, have to settle for low occupancy levels. Another interesting performance measure is the expected delay for those patients who have a positive delay ($ED/D > 0$). We calculated this statistic as well and found that at an 85 percent occupancy level, $ED/D > 0$ for Beth Israel, would be about 0.35 days or over eight hours. This is clearly important information for hospital administrators in making capacity decisions.

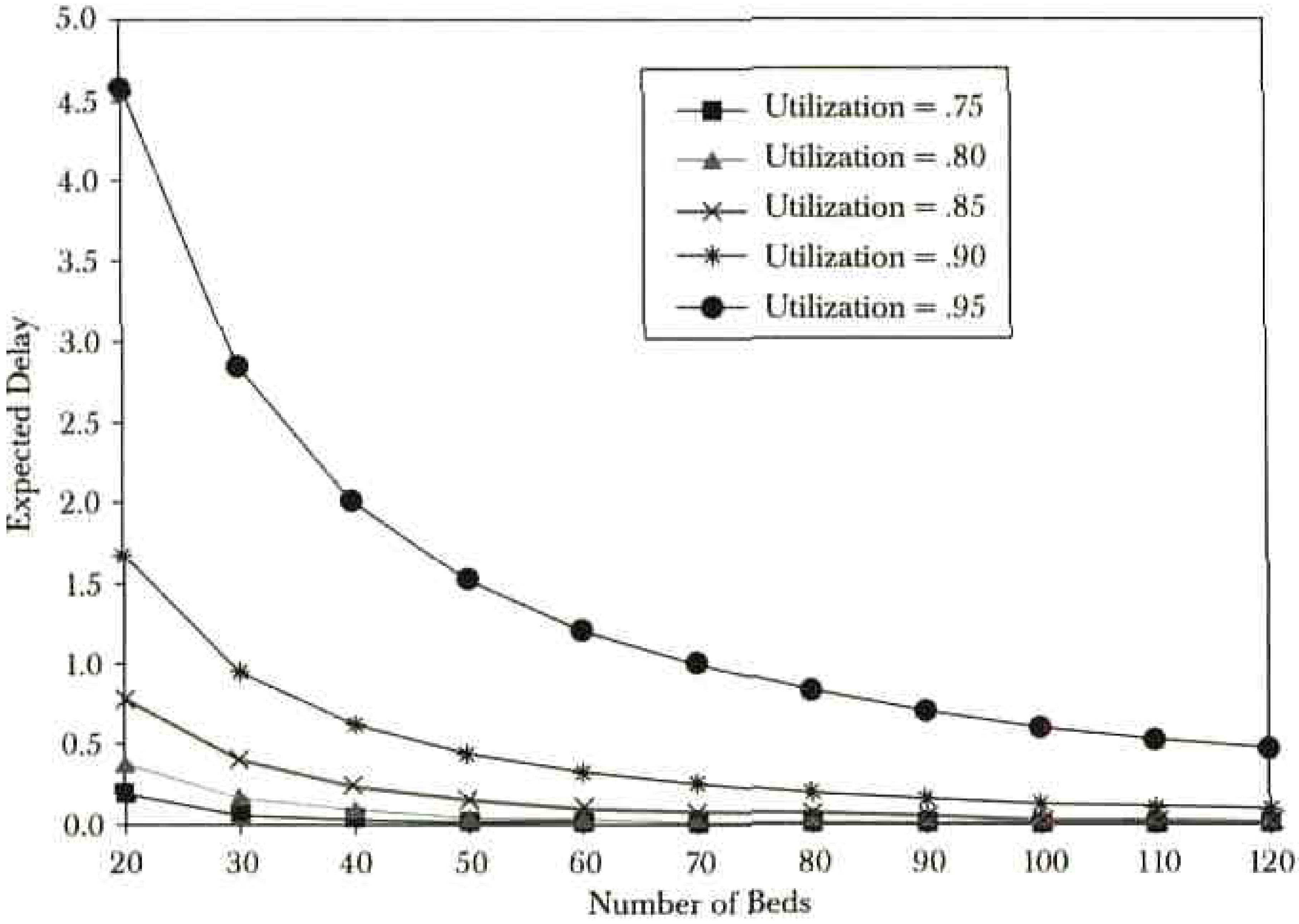
The above findings are quite generalizable because the only fixed parameter in the analyses is ALOS, which varies very little for obstetrics across hospitals (Heisler and Cantor 1998). Of course, these results are meant to be illustrative. For example, more reliable planning estimates could be obtained by collecting LOS data on labor and delivery rooms, if distinct from postpartum beds, and by looking at seasonality effects and peak demand periods, which we examine in the Changes in Demand Rate section.

Surgery

About half of general surgery patients are elective, or scheduled. The other half are either urgent, meaning they must be admitted within 24 hours, or emergent. Administrators at Beth Israel have tried to keep delays for surgical beds down to an average of one day, so we choose expected delay (ED) as our primary measure for our analysis.

Figure 2 plots ED in days against number of beds for utilization levels ranging from 0.75 to 0.95. The most interesting observation here is that for a unit of comparable size to the obstetrics service, occupancy levels can be much higher and still meet acceptable standards. For example, for 56 beds and a target ED of one day, occupancy levels would be over 90 percent. This is consistent with the actual delays and occupancies reported to us by the hospital. Even with fewer than 20 beds, this type of unit could operate at over the standard 85 percent level. This, of course, reflects the difference in the tolerance for delays between surgery and obstetrics. Again, this is only illustrative. The use of the $M/M/s$ here may overestimate delays because it assumes all arrivals are unscheduled. Also, because of the more discretionary nature of surgical admissions, there is a distinct day-of-week pattern to arrivals, with Sundays' arrival rate less than 40 percent of the average and Tuesdays' peaking at about 30 percent higher than average. We examine these latter effects in the Changes in Demand Rate section.

Figure 2: Surgery, ALOS = 5.9



IMPACT OF CONSOLIDATING DISPARATE SERVICES ON BED REQUIREMENTS

As in other hospitals, Beth Israel has several nursing units that consist of two or more small clinical services. One of these is the consolidated cardiac and thoracic surgery unit. Many hospitals have such a unit because thoracic patients are relatively few and require similar nursing skills as cardiac patients. The average arrival rate of cardiac patients in Beth Israel is 1.91 bed requests per day versus 0.42 for thoracic patients. Cardiac patients also stay, on average, more than twice as long with an ALOS of 7.7 days versus 3.8 for thoracic patients.

Table 1, panel A shows the number of beds required to meet several performance targets by each of the two services operating independently as well as in a combined unit. Most of these surgeries are elective. However, delays of more than one or two days are problematic because they may result in backups in the emergency room or other units.

Table 1: Cardiac and Thoracic Surgery

A. Number of beds needed to meet service targets

Target	*Cardiac*		*Thoracic*		*Combined*	
Maximum ED (Days)	*No. Beds*	*Utilization*	*No. Beds*	*Utilization*	*No. Beds*	*Utilization*
0.5	19	0.84	4	0.40	22	0.81
1	19	0.84	3	0.53	21	0.85
2	18	0.88	3	0.53	20	0.89
3	18	0.88	3	0.53	20	0.89

B. Delays when priority given to cardiac patients

		ED (Days)		
Number of Beds	*Cardiac*	*Thoracic*	*Overall*	*Utilization*
23	0.17	0.77	0.28	0.78
22	0.28	1.53	0.50	0.81
21	0.47	3.20	0.96	0.85
20	0.77	7.49	1.98	0.89

The results show that for each delay target, the combined unit results in a savings of only one bed out of a total of about 20 beds. For example, 22 (19 cardiac and three thoracic) beds would be needed to achieve Beth Israel's ED target of less than one day if the services were operated independently, while 21 beds are necessary if they are combined. However, this assumes that the admissions policy is the same for all patients. Yet in this hospital, as in others, cardiac patients have priority over thoracic patients. Incorporating this into our analysis yields different results, which appear in Table 1, panel B. Focusing on a target of less than one day, we see again that 21 beds is the minimum that produces this result. However, the resulting ED for the low-priority thoracic patients is now more than three days. This long delay is due to the fact that thoracic patients represent less than 20 percent of the total arrivals and thus will often be bumped in queue by the far more prevalent cardiac patients. Even worse, this predicted ED for thoracic patients of 3.2 days is actually an underestimate. This is because the model assumes the same (weighted) average service time for both customer classes, while in reality the higher priority cardiac patients have much *longer* stays resulting in even longer delays than predicted for the thoracic patients. If one bed is added, the resulting delay for thoracic patients goes down to 1.5 days, a more reasonable level, but there will be no savings over operating the units separately. And to maintain a maximum ED of one day for each patient group, the combined unit would actually require one more bed than the separate units.

Therefore, the increased efficiency in terms of reduced beds (and thus higher occupancy level) is at best small and may actually be nonexistent. We do not mean to suggest that there is no other advantage to combining these services. Clearly, a unit of just three beds is likely to be inefficient from a physical space and overhead perspective. Another alternative is to operate the two services in one unit but not treat the beds as completely interchangeble, that is, reserving a number of beds for thoracic patients.

The above analysis suggests that hospital managers may be misguided in their thinking about consolidation of clinical services. However, looking at another unit in Beth Israel yields somewhat different results. The SVGEG unit consists of five services: surgery (SURG), vascular surgery (VSUR), gynecology (GYN), otorhinolaryngology (ENT), and urology (GU). Table 2 shows the admissions and LOS data for each service as well as the number of beds required for two delay targets for each service operating independently as well as for the consolidated unit. Note that the consolidation savings in number of beds needed to meet a one- or two-day ED service target is six and three beds, respectively. Why are the relative savings greater in this case than for combining cardiac and thoracic surgery? Primarily because much of the savings is the result of combining the four small (and hence relatively inefficient) services, that is, all but general surgery. If these four services are managed independently, the number of beds needed to meet expected delay targets of less than one or two days is 29 and 26, respectively, while the combined unit would need 26 or 25. These analyses illustrate that combining

Table 2: Miscellaneous Surgical Services: Number of Beds Needed to Meet Service Targets

				Number of Beds	
Unit	*Arrival Rates*	*ALOS*	*CV*	*ED ≥ 1*	*ED ≥ 2*
SURG	4.40	6.0	1.06	31	30
VSUR	1.34	5.9	1.10	11	10
GYN	2.58	2.5	0.74	8	8
ENT	0.24	3.8	1.42	3	2
GU	1.46	3.4	0.94	7	6
			Totals	60	56
SVGEG	10.02	4.8	1.42	54	53

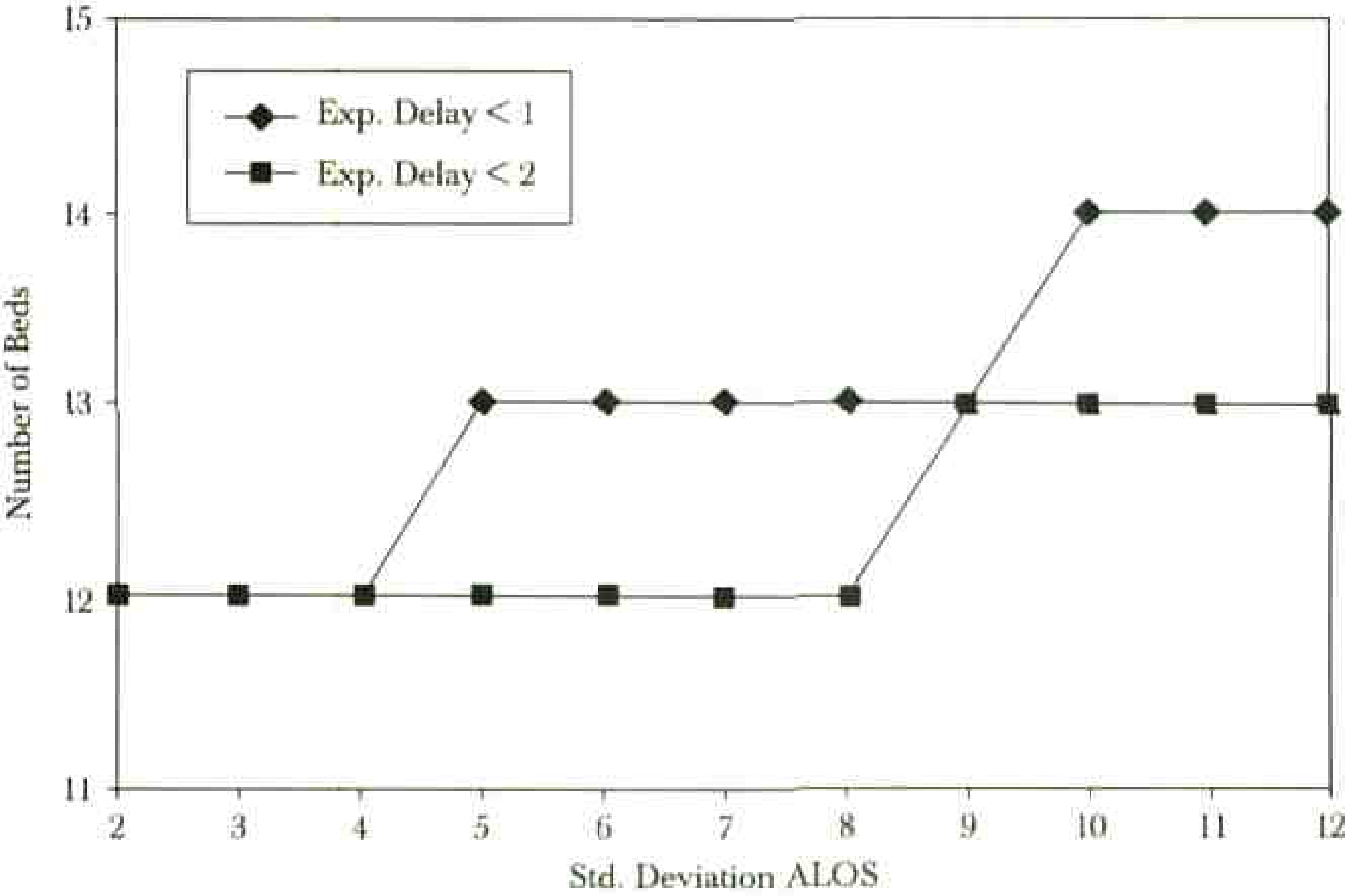
clinical services may result in substantial bed savings when several small services are combined, but they also show that the impact of any priorities must be considered.

REDUCING LOS: MEAN VERSUS VARIABILITY

Hospitals have experienced steadily decreasing ALOS for years largely as a result of discharging patients sooner, while more recently there has been increased focus on reducing the variability of LOS using critical pathways, controlling admissions based on certain demographic or socioeconomic factors, or both. This raises an interesting question: What is the relative impact of reducing the variability of hospital stay versus the mean?

We chose the neurosurgery unit to explore this issue because it has the highest CV of LOS—about 2.0. We calculated the number of beds needed to meet service performance targets of ED less than one and two days as a function of the standard deviation of LOS. We used the empirically derived arrival rate of 1.7 patients per day and an ALOS of 5.8 days. The results, shown in Figure 3, illustrate the relative insensitivity of number of required

Figure 3: Neurosurgery, ALOS = 5.8

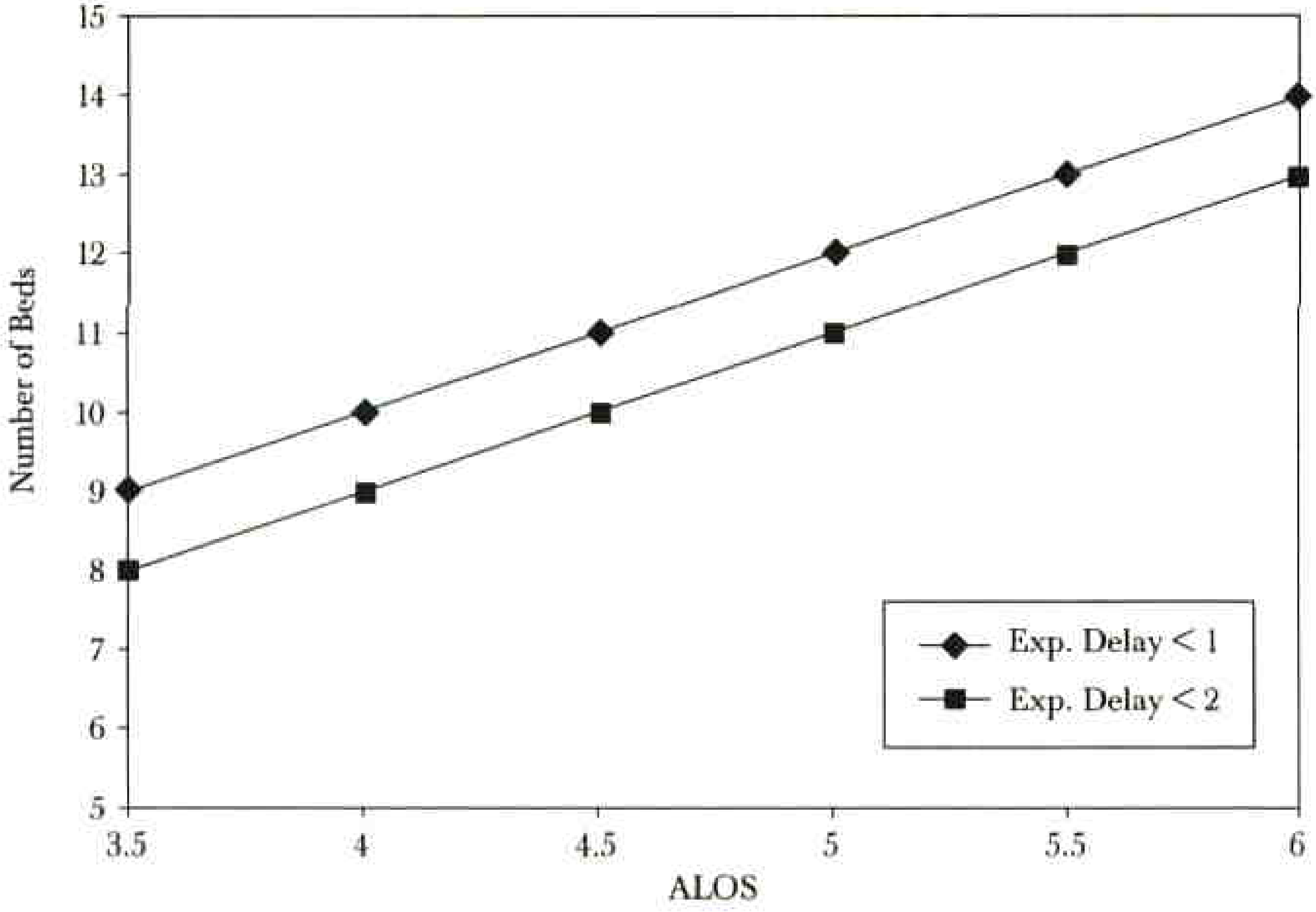


beds to standard deviation of LOS. For example, for a range of standard deviation corresponding to coefficients of variation of less than 0.4 up to more than 2, the number of beds needed to meet an ED of less than two days changes by only one bed—from 12 to 13 beds. To contrast this with reducing the mean, Figure 4 shows the effect of changing the ALOS while keeping the CV constant at 2.0. Every reduction of half a day results in a savings of a bed.

Based on these results and using a target of expected delay of less than one day, we can quantify the impact on number of required beds resulting from reducing the mean versus the standard deviation of LOS: A reduction in ALOS of about 10 percent would result in a one bed savings while the same relative reduction in the standard deviation would not result in any savings; a reduction of 20 percent in ALOS would save two beds while the same reduction in standard deviation would save only one; and a 50 percent reduction in ALOS would result in over a 40 percent reduction in beds while the same reduction in standard deviation would still only decrease the number of beds required by one.

The above analysis confirms the general belief that the potential benefits from efforts focused on eliminating unnecessary time spent in the hospital

Figure 4: Neurosurgery, CV = 2.0



are significant. It also indicates that activities aimed more at standardization of LOS are probably not as worthwhile from a bed utilization perspective, although, of course, they may have other benefits. From a queueing theory perspective, it is well known that delays are relatively insensitive to small changes in standard deviation, and, hence, there is good reason to believe that this finding is generalizable to other units and hospitals.

CHANGES IN DEMAND RATE

Conversations with hospital administrators at several hospitals revealed that many clinical services experience some degree of seasonality in admissions due to, for example, a tendency not to schedule elective procedures during vacation and holiday seasons. Another common example is obstetrics. Our data from Beth Israel confirmed what had been related to us by several hospitals—there are more births in summer than winter. At Beth Israel, admissions range from about 12.3 per day in January to almost 17 per day in July.

Figure 5: Obstetrics, Beds = 56

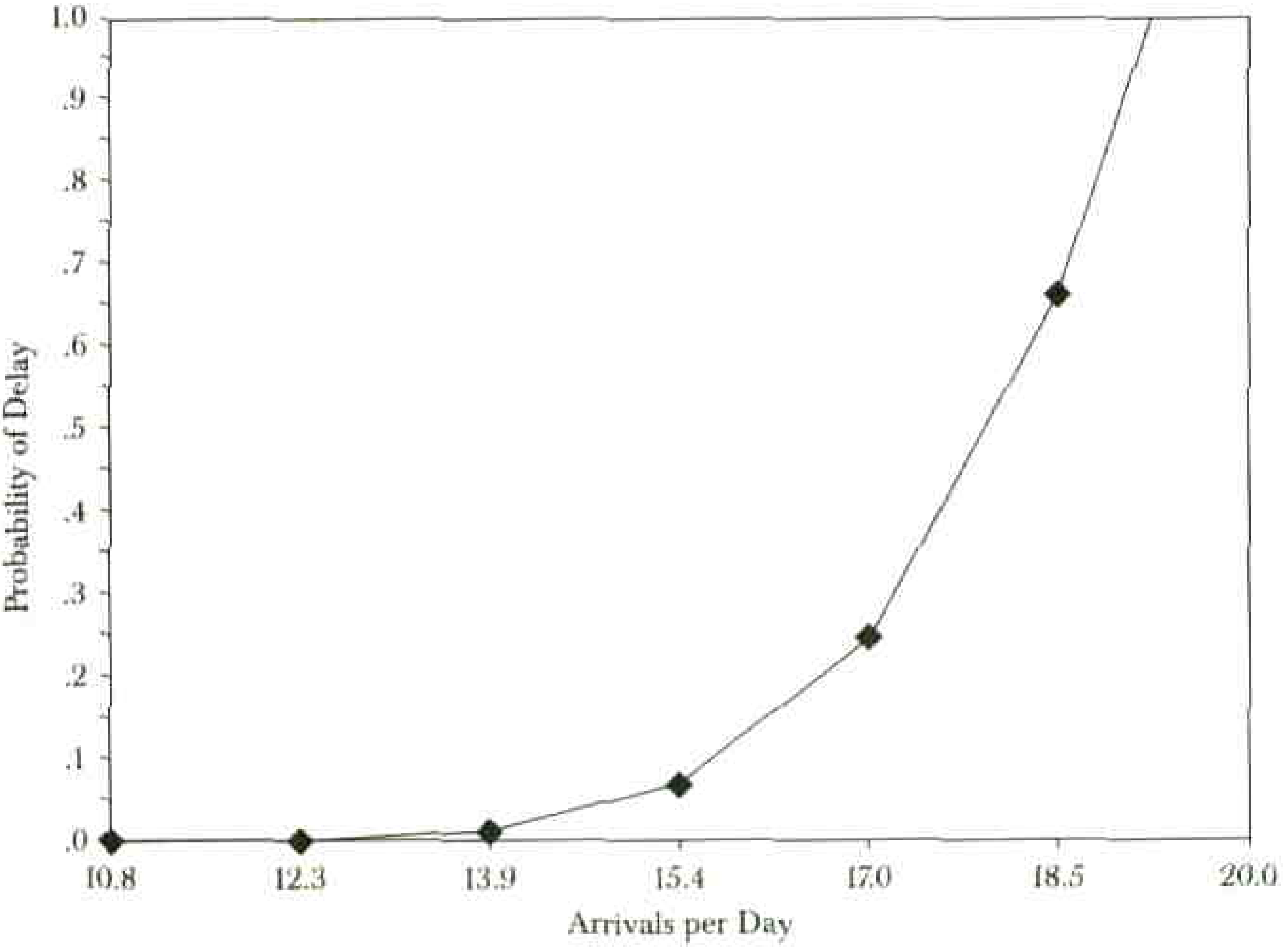


Figure 5 shows the probability of delay for an obstetrics bed as a function of the arrival rate using the Beth Israel ALOS of 2.9 days and 56 beds. The graph indicates that while probability of delay will be close to zero in January, in July it reaches almost 25 percent. Moreover, the curve rises very steeply above this arrival rate level. If there should be a period of time in July when arrivals increase, say, an additional 10 percent, probability of delay would increase dramatically to over 65 percent. This is because at July levels, average occupancy levels are about 88 percent, so even relatively small percentage increases raise utilization into a precariously high range. (This "elbow" in the delay curve is the result of the well-known result that as utilization approaches 100 percent, the probability of delay in a queueing system approaches 1.) This degree of disparity between low and high seasons suggests that, if possible, the number of obstetrics beds be adjusted over the year. For example, Figure 5 suggests that an additional eight beds would be needed to keep probability of delay in the 5 percent range during peak demand months. Such adjustments could be made in a hospital by the use of "swing" beds, which are employed at many large hospitals during peak demand times, or by staffing fewer beds during slower times by scheduling vacations accordingly.

Hospitals also have significant disparities in admissions rates across the week. In particular, admissions drop significantly on weekends when virtually no elective procedures are scheduled. Several hospitals have considered whether it would be worthwhile to try to operate in a true seven-day-a-week mode. Although it may cause increased difficulty in staffing for weekends, the potential benefits of "smoothing" demand over seven days include operating with fewer beds or increasing admissions without an adverse impact on delays for beds. To provide insight, we looked at the surgical intensive care unit (SIC), which has a preponderance of elective patients and for which daily admissions data were available.

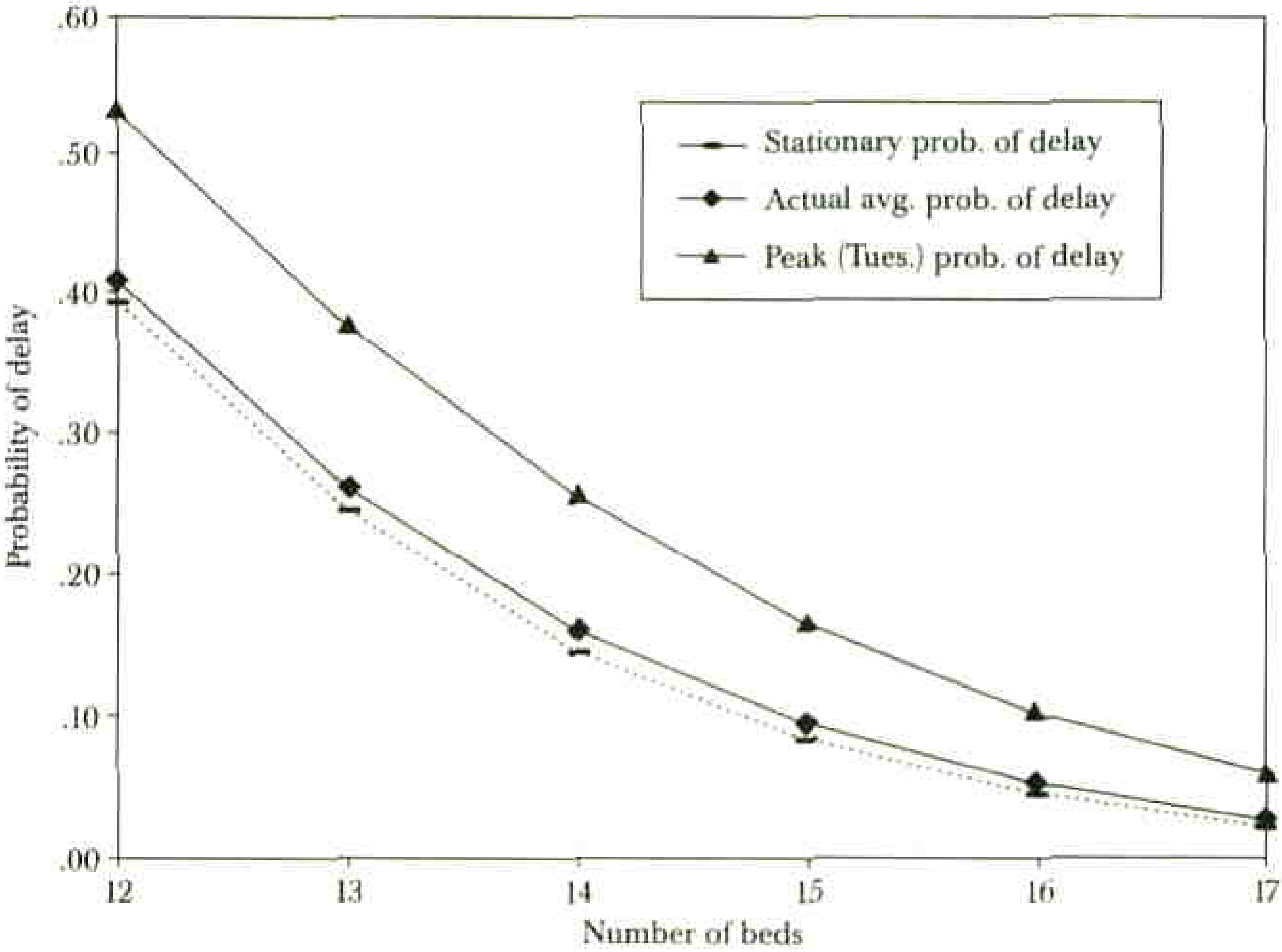
Table 3: Surgical Intensive Care—Admissions

Day	*Admissions/Day*
Sunday	1.44
Monday	3.36
Tuesday	4.42
Wednesday	3.59
Thursday	3.92
Friday	4.40
Saturday	2.21
Average	3.34

The average daily admissions rate over the year is 3.34, with Sundays having the lowest level at 1.44 per day and Tuesdays the highest at 4.42 per day. We compared the actual delays, as estimated from a numerical solution of the differential equations, assuming exponential LOS and using the admissions data in Table 3 [see Green, Kolesar, and Svoronos (1991) for details on the solution methodology] to those that would result if the average daily admissions rate was constant across the days of the week. Many patients have an LOS in intensive care of less than one day, which is recorded by Beth Israel in the category "less than 24 hours." Based on the assumption that these stays of less than one day were half a day, we calculated an ALOS of 3.05 days.

Figure 6 plots probability of delay versus number of beds for both the actual and time-stationary cases. Figure 6 shows that although, as expected, the stationary case results in lower delay probabilities, the difference is not dramatic. However, the weekly average probability of delay is not a good representation of actual congestion on the heavy demand days, Tuesdays. Therefore, we also plotted the peak probability of delay curve on Figure 6. Here, the differences are more significant. For example, if the performance

Figure 6: Surgical Intensive Care: Day-of-Week Admissions Variability



standard was to keep the daily probability of delay below 10 percent, at least 17 beds would be needed with the current pattern of daily variability in demand resulting in less than 60 percent utilization. If demand was spread evenly over the week, 15 beds would be sufficient, resulting in a 68 percent occupancy.

SUMMARY AND CONCLUSIONS

Although queueing and other operations research models are routinely used by a great number of service organizations in various industries to evaluate trade-offs regarding efficiency and service, public policy guidelines as well as the dozens of hospital managers to whom we have spoken continue to rely primarily on occupancy levels in determining capacity. In this article, we have shown that this approach is generally flawed. Using the data of a major urban hospital and assuming reasonable levels of patient delays, we have also provided insight on the efficacy of some strategies for cutting costs by reducing beds (or alternatively, increasing revenues by increasing admissions). Specifically, we have demonstrated the following.

- The standard practice of using a target occupancy level of 85 percent will result in unacceptable delays in clinical units that are small, have a high percentage of urgent admissions, or both. Conversely, large units that are used primarily for elective patients may be able to achieve occupancy levels above 90 percent without serious impact on patient service. Use of the appropriate measure of delay is important in determining bed needs.
- The common belief that consolidation of small clinical units can save beds without compromising patients' delays may be incorrect and depends on the relative demands, LOSs, and admissions priorities of the individual patient classes.
- Reducing ALOS has far more potential to reduce required capacity than reducing LOS variability.
- Increases in arrival rates due to seasonality or unexpected increases in the incidence of a disease can seriously compromise the timely availability of beds when capacity is based on maintaining a high average occupancy level. In the case of seasonality, which is predictable, this implies that capacity planning in some units should consider peak demand periods. Efficiencies could be achieved by resource sharing with other units in the hospital or other hospitals in the community during high demand periods by taking beds out of service (i.e., reduce staffing) during low demand periods, or both. In general, the sensitivity of delays to small

increases in the arrival rate at high utilization levels should be considered in all hospital capacity planning. Failure to do so can have dramatic consequences. An example of this was the New York City hospital occupancy crisis in 1987/88 (Myers, Fox, and Vladeck 1990), during which there was a severe and protracted city-wide shortage of inpatient hospital beds and which resulted in ambulances routinely being turned away from full hospitals and delays of days for urgently sick patients waiting for an open bed. Subsequent analysis showed that the crisis was the result of the simultaneous 9 percent decline in capacity between 1985 and 1987 that was largely due to new state regulations linking Medicaid reimbursement to occupancy levels (which were regulated to be 85 percent) and of an unanticipated 18 percent growth in admissions largely due to a rise in AIDS and drug abuse.

- While the phemomenon of not scheduling elective procedures on weekends and holiday periods has little effect on overall average delays, it may significantly increase delays for beds on resulting high-demand weekdays. Although hospitals can and do reduce staffing levels on weekends, the savings may be more than offset by the need to have additional beds to provide a consistently good level of care. This is particularly relevant for units like intensive care where the combination of day-of-week fluctuations and small size will result in a need for very low average occupancy levels to keep the availability of beds for critically ill patients sufficiently high.
- Based on our experience with Beth Israel and our conversations with other hospitals, the operational data necessary to accurately model hospital units for the purpose of evaluating capacity decisions and patient delays appears not to be routinely captured by the hospitals' information systems. This includes arrival rates and patterns by clinical service, ALOS by unit, and even the actual number of beds in service by unit.

Although the specific quantitative results are limited by virtue of their dependence on data from one hospital, we believe that the above qualitative findings are fairly generalizable for several reasons. First, Beth Israel is probably quite typical of large urban teaching hospitals, which as a class provide a very large percentage of all hospital care. Second, we varied several of the factors such as size to represent other hospital situations, and parameters such as ALOS are often similar across hospitals for certain units such as obstetrics. Finally, most of the findings are the result of structural properties of queueing systems. Further research is needed to validate and refine these findings and to provide additional insights concerning the dynamics and interrelationships among various areas of the hospital or among a group of affiliated hospitals.

Another important issue to consider in capacity planning is the relationship between size and quality of care. Several research studies have found that hospitals that handle larger volumes of patients with certain diagnoses or procedures have better medical outcomes, such as lower mortality, reduced readmission rates, and lower lengths of stay (Luft et al. 1990; Luft, Hunt, and Maerki 1987; Phillips, Luft, and Ritchie 1995; Yao and Yao 1999). Equally important is better information concerning cost structures and revenue characteristics and how these affect capacity and resource allocation decisions.

As competitive factors as well as public policy concerns create increasing focus on quality and service issues in health care, it will become imperative for hospital managers and government officials to understand these issues and others affecting efficiency and effectiveness when evaluating decisions involving the capacity and organization of hospitals. Of immediate concern is the evaluation and implementation of the increasing number of merger and consolidation decisions from an operational perspective. These decisions have major consequences for the welfare of populations as well as the financial health of health care providers and therefore call for much more careful analyses than have been done to date.

ACKNOWLEDGMENTS

The authors would like to thank Michael Littman, Bob Norton, and Susan Stengrevics of Beth Israel Hospital for all of their help in providing the data and information needed to perform the analyses.

REFERENCES

American Hospital Association (AHA). 1996. *1996/1997 Hospital Statistics*. AHA: Chicago.

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

Brigandi, A. J., D. R. Dargon, M. J.Sheehan, and T. Spencer III. 1994. "AT&T's Call Processing Simulator (CAPS) Operational Design for Inbound Call Centers." *Interfaces* 24 (1): 6–28.

Brusco, M. J., L. W. Jacobs, R. J. Bongiorno, D. V. Lyons, and B. Tang. 1995. "Improving Personnel Scheduling at Airline Stations." *Operations Research* 43 (5): 741–51.

Cobham, A. 1954. "Priority Assignment in Waiting Line Problems." *Operations Research* 2: 70–76.

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 (1): 43–61.

Freeman, R. K., and R. L. Poland. 1992. *Guidelines for Perinatal Care*, 3rd Edition, p. 14. Chicago: American College of Obstetricians and Gynecologists.

Friedman, B., and M. Pauly. 1981. "Cost Functions for a Service Firm with Variable Quality and Stochastic Demand." *Review of Economics and Statistics* 63 (November): 610–24.

———. 1983. "A New Approach to Hospital Cost Function and Some Issues in Revenue Regulation." *Health Care Financing Review* 4 (March): 105–14.

Gaynor, M., and G. F. Anderson. 1995. "Uncertain Demand, the Structure of Hospital Costs, and the Cost of Empty Hospital Beds." *Journal of Health Economics* 14: 291–317.

Graham, G. G., and T. G. Cowing. 1997. "Hospital Reserve Margins: Structural Determinants and Policy Implications Using Cross-Section Data." *Southern Economic Journal* 63 (3): 692–709.

Green, L. V., P. J. Kolesar, and A. Svoronos. 1991. "Some Effects of Nonstationarity on Multi-Server Markovian Queueing Systems." *Operations Research* 39 (3): 502–11.

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

Heisler, T., and J. C. Cantor. 1998. *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): 1–22.

Hokstad, P. 1978. "Approximations for the M/G/m Queue." *Operations Research* 26 (3): 510–23.

Keeler, T. E., and J. S. Ying. 1996. "Hospital Costs and Excess Bed Capacity: A Statistical Analysis." *Review of Economics and Statistics* 78 (3): 470–81.

Luft, H. S., D. Garnick, D. H. Mark, and S. J. McPhee. 1990. *Hospital Volume, Physician Volume, and Patient Outcomes: Assessing the Evidence*. Chicago: Health Administration Press.

Luft, H. S., S. S. Hunt, and S. C. Maerki. 1987. "The Volume-Outcome Relationship: Practice Makes Perfect or Selective Referral Patterns?" *Health Services Research* 22 (2): 157–82.

Lynk, W. J. 1995. "The Creation of Economic Efficiencies in Hospital Mergers." *Journal of Health Economics* 14: 507–30.

McClure, W. 1976. *Reducing Excess Hospital Capacity*. Washington, DC: Bureau of Health Planning and Resources Development, Department of Health, Education and Welfare.

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 (5): 739–55.

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, Part I): 115–31.

Pauly, M. V., and P. Wilson. 1986. "Hospital Output Forecast and the Cost of Empty Hospital Beds." *Health Services Research* 21 (3): 403–28.

Phillips, K. A., H. S. Luft, and J. Ritchie. 1995. "The Association of Hospital Volumes with Adverse Outcomes, Length of Stay and Charges in California for Percutaneous Transluminal Coronary Angioplasty (PTCA)." *Medical Care* 33 (5): 502–14.

Schneider, D. 1981. "A Methodology for the Analysis of Comparability of Services and Financial Impact of Closure of Obstetrics Services." *Medical Care* 19 (4): 393–409.

Schwartz, W., and P. Joskow. 1980. "Duplicated Hospital Facilities: How Much Can We Save by Consolidating Them?" *New England Journal of Medicine* 303 (25): 1449–57.

Taylor, P. E., and S. J. Huxley. 1989. "A Break from Tradition for the San Francisco Police: Patrol Officer Scheduling Using an Optimization-Based Decision Support System." *Interfaces* 19 (1): 4–24.

Vassilacopoulos, G. 1985. "A Simulation Model for Bed Allocation to Hospital Inpatient Departments." *Simulation* 45 (5): 233–41.

Whitt, W. 1992. "Understanding the Efficiency of Multi-Server Service Systems." *Management Science* 38 (5): 708–23.

Worthington, D. J. 1987. "Queueing Models for Hospital Waiting Lists." *Journal of the Operational Research Society* 38: 413–22.

Yao, S., and G. Lu-Yao. 1999. "Population-Based Study of Relationships Between Hospital Volume of Prostatectomies, Patient Outcomes, and Length of Hospital Stay." *Journal of the National Cancer Institute* 91 (22): 1950–56.

Young, J. P. 1965. "Stabilization of Inpatient Bed Occupancy Through Control of Admissions." *Hospitals: Journal of the American Hospital Association* 39 (19): 41–48.