Anesthesiology 2004; 100:1271–6

Queuing Theory Accurately Models the Need for Critical Care Resources

Michael L. McManus, M.D., M.P.H., Michael C. Long, M.D.,† Abbot Cooper,‡ Eugene Litvak, Ph.D.§*

Background: **Allocation of scarce resources presents an increasing challenge to hospital administrators and health policy makers. Intensive care units can present bottlenecks within busy hospitals, but their expansion is costly and difficult to gauge. Although mathematical tools have been suggested for determining the proper number of intensive care beds necessary to serve a given demand, the performance of such models has not been prospectively evaluated over significant periods.**

Methods: **The authors prospectively collected 2 years' admission, discharge, and turn-away data in a busy, urban intensive care unit. Using queuing theory, they then constructed a mathematical model of patient flow, compared predictions from the model to observed performance of the unit, and explored the sensitivity of the model to changes in unit size.**

Results: **The queuing model proved to be very accurate, with predicted admission turn-away rates correlating highly with those actually observed (correlation coefficient = 0.89). The model was useful in predicting both monthly responsiveness to changing demand (mean monthly difference between observed and predicted values, 0.4 ± 2.3%; range, 0–13%) and the overall 2-yr turn-away rate for the unit (21% *vs.* 22%). Both in practice and in simulation, turn-away rates increased exponentially when utilization exceeded 80–85%. Sensitivity analysis using the model revealed rapid and severe degradation of system performance with even the small changes in bed availability that might result from sudden staffing shortages or admission of patients with very long stays.**

Conclusions: **The stochastic nature of patient flow may falsely lead health planners to underestimate resource needs in busy intensive care units. Although the nature of arrivals for intensive care deserves further study, when demand is random, queuing theory provides an accurate means of determining the appropriate supply of beds.**

IN the United States, after more than a decade of healthcare restructuring, the number of hospitals continues to decline.[1] In some regions of the country, this has produced serious overcrowding, particularly in emergency departments[2–4] and intensive care units (ICUs).[5] Although there may be growing recognition that mortality is increased among patients to whom admission to crowded ICUs is refused,[6] there is incomplete understanding of the limits of the downsizing process and no consensus as to the number of ICU beds necessary to serve a given population.[7] Nevertheless, ICUs are among the most complex and expensive of all medical resources, and hospital administrators are challenged to meet the demand for intensive care services with an appropriate capacity.

Queuing theory is used widely in engineering and industry for analysis and modeling of processes that involve waiting lines.[8] In appropriate systems, it enables managers to calculate the optimal supply of fixed resources necessary to meet a variable demand. In the past, attempts have been made to apply queuing analysis to a variety of hospital activities, including cardiac care units,[9] obstetric services,[10] operating rooms,[11,12] and emergency departments,[13] as a means of directing the allocation of increasingly scarce resources. More recently, health policy investigators have also sought to apply these techniques more widely across entire healthcare systems.[14–16] Unfortunately, most proposed queuing models lack real-world validation[17] and, perhaps for this reason, have yet to be embraced by physicians and hospital administrators. Therefore, to explore the utility and implications of queuing theory as it relates to the supply and demand for critical care services, we sought to validate a simple queuing model in a busy ICU.

Materials and Methods

We studied all admissions to the medical–surgical ICU of a large, urban children's hospital during a 2-yr period. The 18-bed unit provides all manner of noncardiac intensive care services and, in addition to local emergencies, serves a large regional, national, and international referral population. During periods of high demand, external requests for transfer are diverted to other institutions in the region, whereas internal overflow is accommodated in off-service care sites, such as the PACU or available beds in a separate, specialized cardiac ICU. Data were collected prospectively as part of the unit's patient care database and are analyzed here for frequency of admission requests, durations of stay, and crowding.

Queuing analysis is dependent on accurate measurement of three variables: arrival rate, service time, and the number of servers in the system. We therefore collected data with special attention to the corresponding hospital

* Department of Anesthesia and the Multidisciplinary Intensive Care Unit, Children's Hospital Boston, and Associate Professor, Harvard Medical School. † Senior Anesthetist, Massachusetts General Hospital, and Adjunct Associate Professor, Boston University School of Management. ‡ Senior Analyst, § Professor of Operations Management and Director, Boston University Health Policy Institute Program on Variability.

Received from the Department of Anesthesia, Pain and Perioperative Medicine, Children's Hospital, Boston, Massachusetts, and the Health Policy Institute, Boston University, Boston, Massachusetts. Submitted for publication May 28, 2003. Accepted for publication November 15, 2003. Support was provided solely from institutional and/or departmental sources.

Address reprint requests to Dr. McManus: Children's Hospital, 300 Longwood Avenue, Boston, Massachusetts 02115. Address electronic mail to: michael.mcmanus@childrens.harvard.edu. Individual article reprints may be purchased through the Journal Web site, www.anesthesiology.org.

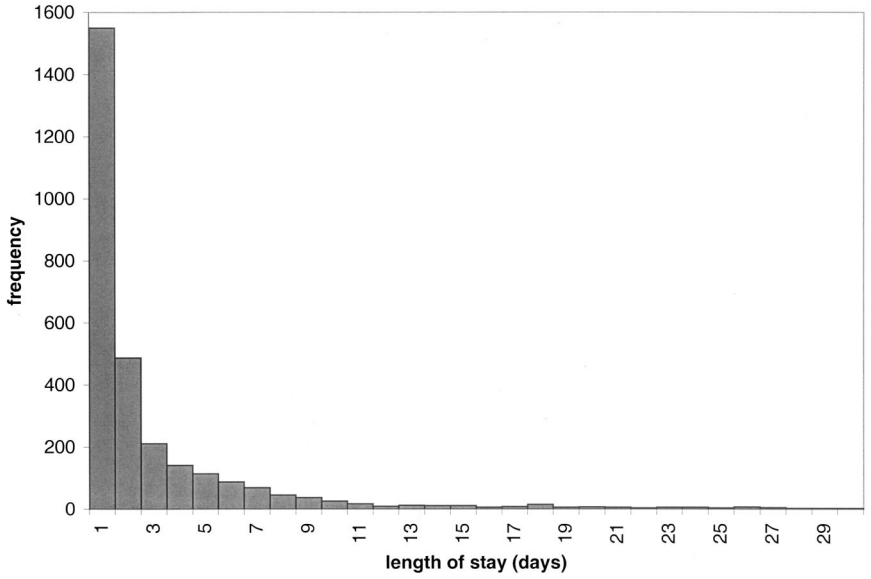

Fig. 1. Distribution of intensive care unit durations of stay over a 2-yr period. Data include all stays of 30 or fewer days and are described by the equation $y = 227.95e^{-0.1662x}$ with $R^2 = 0.8208$.

variables (admission rate, duration of stay, and number of available beds) prospectively. For purposes of modeling, all patients referred for ("requesting") admission were considered arrivals. Durations of stay were calculated as (discharge date) − (admission date), with all admissions assigned a minimum of 1 day.

A computer simulation model of ICU flow was then constructed using spreadsheet software (Excel 2000®; Microsoft Corporation, Redmond, WA) and standard queuing formulae.[18] The ICU was modeled as a multichannel, single-stage system of identical parallel servers that process randomly patterned arrivals according to exponentially distributed service times. Each ICU bed was treated as one server and a "first come, first served" queuing discipline was assumed. It was further understood that no waiting line was possible for these critically ill patients and, therefore, the probability of waiting equals the probability of rejection. Such a system has been suggested by others as an appropriate construct for evaluating turn-away probabilities and, in the queuing literature, is denoted as M/M/c/c (shorthand notation for systems involving Markovian interarrival times, which are modeled as a Poisson process, Markovian service times, *c* servers, and *c* spaces in the system).[16] Readers unfamiliar with queuing theory may find introductions, useful tutorials, and downloadable software suitable for duplicating this work on numerous Internet sites such as those listed in the appendix.

With observed monthly admission rates, available beds, and stay durations as inputs, monthly utilizations and rejection probabilities were calculated using queuing theory. Summary calculations over the 2-yr period were also completed. For purposes of monthly analysis, any bed occupied for more than 1 continuous month by the same patient was treated as a bed lost to the system. During the occupied month, therefore, the number of servers in the model was reduced, and the corresponding admission days were not included in that month's duration-of-stay calculations. Associated admission days carrying over into contiguous months were treated as separate admissions when calculating those months' average durations of stay. Patients to whom admission was refused and who were transported to another hospital or those diverted to an alternative care site within the hospital (*e.g.*, PACU or specialty ICU) because a bed was unavailable in the primary unit were considered "rejected" or "turn aways." Observed rejection rates were calculated as: (no. of patients refused + no. of patients diverted)/total no. of patients requesting admission. During portions of the observation period when the practice was to always maintain one open bed in the primary unit for new in-house emergencies, the total number of available servers in the model was decreased by 1. Very brief (< 1 day) bed closures due to staffing shortfalls were neglected.

Statistical Analysis

The queuing model selected assumes that daily admission rates (arrivals) follow a Poisson distribution (coefficient of variation = 1) and that durations of stay (service times) are either constant or follow an exponential distribution. Others have shown that the arrival rate of patients to ICUs follows a Poisson distribution,[7,19] and this behavior was confirmed in data here both by coefficient of variation (1.1) and Kolmogorov-Smirnov test (with fit accuracy of 0.0003 and α of 0.05, a Poisson distribution is not rejected, $P = 0.262$, using Statfit®; Geer Mountain Software Corporation, South Kent, CT). As illustrated in figure 1, durations of stay were found to follow an exponential distribution. Validity of the queuing model was assessed using a correlated inspection approach[20] with agreement between observed turn-away rates and those predicted by the model assessed

Table 1. Monthly Intensive Care Unit Service Parameters over 2 Years

Month	Admissions	Beds Available for New Admissions	Average Duration of Stay, days	Utilization, %	Ambulance Diversions	Off-service Diversions	Total Rejections
1	145	17	3.2	78	0	24	24
2	155	17	3.7	88	0	35	35
3	156	17	2.8	76	0	24	24
4	138	17	4.6	89	0	22	22
5	139	17	3.7	82	0	12	12
6	145	18	3.5	81	0	20	20
7	166	18	3.3	83	0	24	24
8	144	18	3.2	76	0	4	4
9	141	18	3.5	80	0	10	10
10	147	18	3.3	78	2	11	13
11	170	18	3.5	86	6	28	34
12	163	18	2.5	71	2	10	12
13	192	17	2.8	84	0	36	36
14	166	17	3.8	89	14	34	48
15	152	15	4.4	91	14	64	78
16	143	13	4.0	90	5	54	59
17	145	15	4.5	90	15	45	60
18	155	16	3.3	85	12	36	48
19	152	15	3.5	86	4	36	40
20	160	17	3.9	88	13	31	44
21	148	17	2.4	69	3	18	21
22	159	17	3.2	82	3	20	23
23	142	15	3.7	86	5	35	40
24	158	16	3.6	86	8	31	39
Total	3,680		3.5	83	106	664	770

via linear regression, paired *t* test, and standard residual analysis (SPSS software; Chicago, IL).

Results

There were 3,786 requests for admission during the period studied. Of these, 3,680 were admitted to the hospital and 106 were diverted to other institutions. Of admissions, 1,374 were patients requiring care for medical illnesses and 2,306 required management of issues related to surgery. Of surgical patients, 2,131 were admitted for care after scheduled procedures whereas 175 were admitted after emergency procedures or unanticipated intraoperative events. Overall service parameters for the study period are presented in table 1.

Monthly average admission request rates ranged from 4.6 to 6.2 patients/day. Individual durations of stay ranged from 1 to 190 days. Monthly average durations of stay ranged from 2.4 to 5.5 days. Seventeen patients had durations of stay greater than 45 days, with each occupying a bed for more than 1 calendar month. In addition, during short periods, up to two beds were closed for administrative reasons. In the 18-bed unit, then, the actual number of available beds ranged from 13 to 18 (mean = 17).

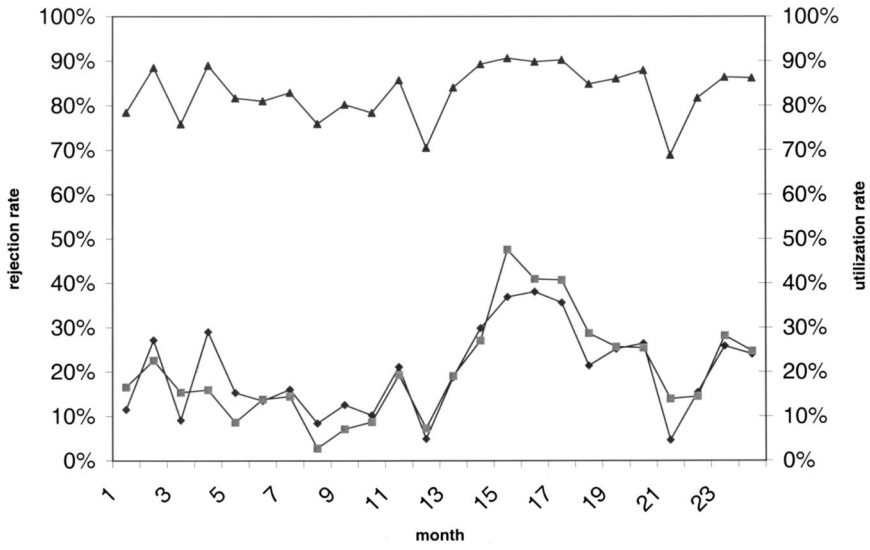

Fig. 2. Monthly intensive care unit utilization and rejection rates. *Diamonds* = monthly rejection rates predicted by the queuing model; *squares* = percent of total admission requests that could not be accommodated; *triangles* = percent utilization of unit resources during each month.

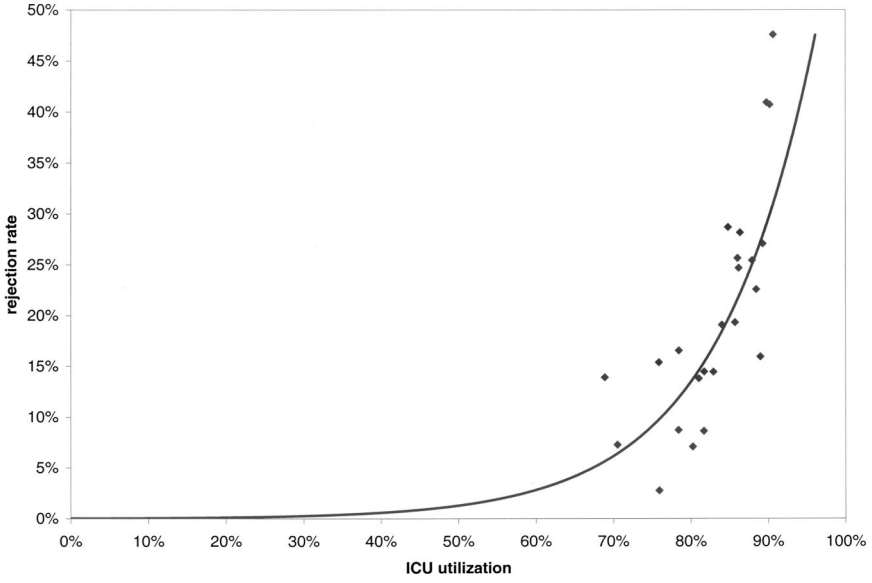

Fig. 3. Intensive care unit (ICU) rejection rate as a function of utilization. The least squares curve fit to the data is described by the equation $y = 0.0003e^{7.8221\times}$ with $r^2 = 0.53$.

Observed monthly turn-away rates varied widely, ranging from 3 to 47% (fig. 2). Over the observation period, turn-away rates corresponded closely to calculated utilization and were accurately predicted by the queuing model (correlation coefficient = 0.897; $P < 0.001$). Overall, the mean difference between observed and predicted values was 0.4% (95% paired *t* confidence interval = ±2.3%) with a maximum difference of 13% and minimum of 0. Residual and normal probability plots (not shown) contained no significant outliers or systematic deviations, while the plot of residuals *versus* predicted values disclosed no nonlinear dependences. For the entire 2-yr period, the observed overall turn-away rate was 21%, and that predicted by the model (using 2-yr average duration of stay and overall average admission rates as inputs) was 22%.

In practice, it was observed that when utilization increased above 80–85%, blocking rates (hospital diversions + off service transfers) increased abruptly (fig. 2). At the highest utilization rate (91%), nearly one half (48%) of all requests for admission could not be accommodated. As illustrated in figure 3, the observed rejection rate was best viewed as an exponential function of utilization. This behavior is consistent with predictions from queuing theory and is widely appreciated as a general property of systems involving waiting lines.[21]

Sensitivity analysis using the model illustrates the impact of bed closure or patients with very long stay durations on the responsiveness of ICUs running near capacity. Using data from a representative month as inputs, an average admission rate of 5.7 patients/day and a 3.5-day average duration of stay yielded a predicted utilization rate of 86%. The associated predicted rejection rate (21%) agreed well with the observed rejection

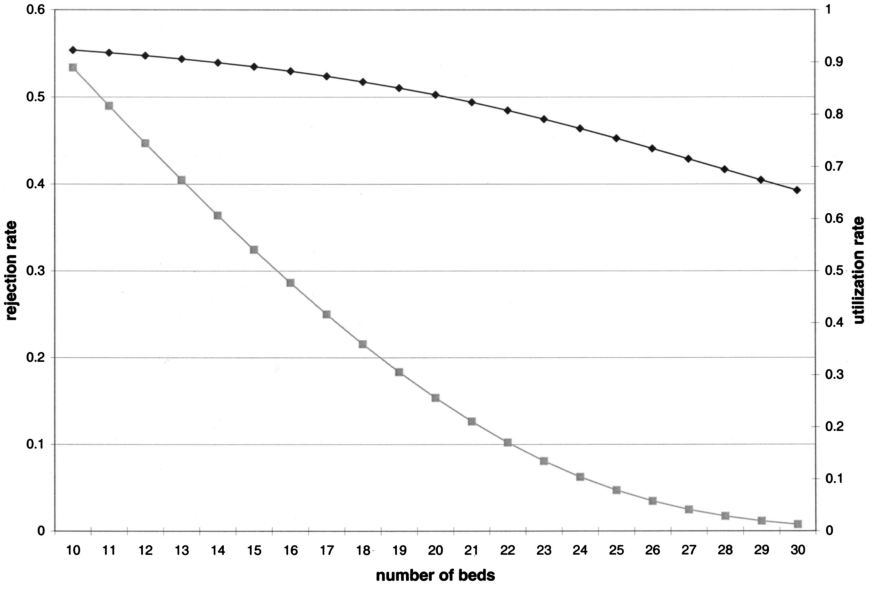

Fig. 4. Sensitivity of rejection rate to the number of available intensive care beds. *Diamonds* = utilization; *squares* = rejection rate. As utilization increases above approximately 85%, further increases are accompanied by large increases in rejection rate. At high utilization rates, loss of even a few available beds markedly increases rejections.

rate (19%), and the corresponding monthly utilization was similar to the overall average observed during the 2 yr studied. Given these routine parameters, figure 4 illustrates the tradeoff between utilization and turn-away rate as the number of available beds is varied (whether by staffing changes or the presence of patients with long stays). Because utilization rates approach 100% asymptotically while rejection rates increase exponentially, small gains in utilization are accompanied by rapid degradation of the ability to handle new admissions. This graphically illustrates the dilemma facing many ICUs: Units financially forced to high utilization must increasingly reject new admissions.

Discussion

This 2-yr experience illustrates that queuing theory may be used to accurately model ICU bed utilization in a large unit operating at or near capacity. Here, the correlation between observed and predicted turn-away rates was extremely high, particularly when noting that day-to-day variations in bed availability (due to patient flow, temporary staffing issues, or the special bed requirements of individual patients) were not considered. To our knowledge, this is the largest experience comparing prospectively acquired data from a functioning ICU with the behavior predicted by a stochastic model.

Although the findings here may be generalized to similar units facing similar demand patterns, they are not necessarily applicable to smaller units, units operating below capacity, or units containing specialized subunits. For example, the results here may significantly underestimate stresses on smaller units because, for a given utilization rate, rejection rates are higher in smaller than in larger service systems.[22] Similarly, the sensitivity analysis provided in figure 3 describes functioning to be expected under arrival rate and service time patterns similar to those observed in our unit. Units with significantly different patient flow patterns might behave differently.

Despite the above limitations, this analysis holds at least four practical implications. First, it clearly demonstrates that the realistic capacity of an ICU is significantly overestimated by measures that fail to account for the variability of demand. Because patient arrivals are random, occupancy rates are more appropriately discussed in terms of probabilities. As demonstrated here, amid a fixed number of available beds, these probabilities are mathematically determined by duration of stay and arrival rate. Common measures of utilization, such as daily census and average occupancy, fail to capture flow-related stresses in the system and mask the reality that patients may frequently be denied access even if the unit seems less than "full."

A corollary to the above observation is that when utilization is maintained at high levels, there is increasing probability that patients will be rejected from the system. As the data show, for a typical range of stay durations and arrival rates, lower utilization necessarily produces lower rejection rates, and higher utilization produces higher rejection rates. In the past, conventional wisdom has held that average occupancy targets of 85% may be considered optimal.[7] The findings here are consistent with this because utilization above 85% was associated with rapidly increasing rejection rates. However, averages may be misleading because seemingly acceptable average utilization of 83% may mask prolonged periods of higher utilization wherein rejection rates might be unacceptable. Therefore, for a system to respond adequately to natural peaks in demand, true continuous utilization must be limited, and a predictable number of empty beds must always be maintained in readiness. Although not the subject here, the associated cost of this readiness could be calculated using queuing theory and fairly assigned to benefiting stakeholders.

Third, the queuing model shows the exquisite sensitivity of "bed crises" to sudden staffing shortfalls or the presence of patients with extremely long durations of stay. Because both conditions effectively lower the number of available "servers," they rapidly degrade the performance of the system. For this reason, analyses that rely on simple duration of stay averages but do not appropriately adjust the number of available servers may tend to overestimate the performance of the system. However, as demonstrated here, if server number is accurately accounted for, queuing theory may be useful in making decisions regarding staffing costs and construction of step-down units.

Finally, to the extent that ICU resources are expensive and often saturated, it is important to reconsider the nature of patient arrival patterns. Here, overall arrivals rates were found to be random, and this randomness permitted successful application of a standard stochastic model. However, it is puzzling that this is so when the majority of admissions resulted from scheduled surgical procedures. Although the utilization and rejection relations described above are mathematical consequences of variability within the system, operations management teaches that lower rejection rates should be anticipated if this variability can be reduced. Therefore, more effective management of the elective surgery scheduling process could produce a much smoother demand pattern and, as a result, increase the effective capacity of busy units. Sources of variability may be classified as natural when they result from uncontrollable variations in disease prevalence, severity, or responsiveness and may be classified as artificial when they result from controllable variations in the manner by which we choose to deliver care.[23] Here, a substantial amount of artificial variability can be inferred because the unit modeled precisely as a random process despite the presence of substantial schedulable patient flow. In separate studies, we have

attempted to estimate the impact of uncontrolled patient flow variability on access to intensive care.[24]

Traditionally, regional requirements for ICU beds have been determined by historical experience and population estimates.[25] However, in a market-driven or otherwise financially austere environment, such determinations are increasingly based on average census figures and occupancy rates. When shortfalls arise or disaster responses are planned, it may be difficult for legislators, health planners, and hospital executives to grasp the true capacity of an intensive care delivery system. Findings here suggest that queuing theory represents a simple and reasonable "first approach" to analysis of ICU capacity until more sophisticated and robust models become available.

References

1. American Hospital Association: Hospital Statistics, 2002 edition. Chicago, Health Forum, LLC2002
2. Derlet RW, Richards JR, Kravitz RL: Frequent Overcrowding in U.S. emergency departments. Acad Emerg Med 2001; 8:151–5
3. McCabe JB: Emergency department overcrowding: A national crisis. Acad Med 2001; 76:672–4
4. Schull MJ, Szalai J-P, Schwartz B, Redelmeier DA: Emergency department overcrowding following systematic hospital restructuring: Trends at twenty hospitals over ten years. Acad Emerg Med 2001; 8:1037–43
5. Nelson M, Waldrop RD, Jones J, Randall Z: Critical care provided in an urban emergency department. Am J Emerg Med 1998; 16:56–9
6. Metcalfe MA, Sloggett A, McPherson K: Mortality among appropriately referred patients refused admission to intensive-care units. Lancet 1997; 350: 7–11
7. Green LV: How many hospital beds? Inquiry 2002; 39:400–12
8. Duckworth WE: Operational Research Techniques. London, Methuen & Co., 1962
9. Cooper JK, Corcoran TM: Estimating bed needs by means of queuing theory. N Engl J Med 1974; 291:404–5
10. Milliken RA, Rosenberg L, Milliken GM: A queuing theory model for the prediction of delivery room utilization. Am J Obstet Gynecol 1972; 114:691–9
11. Taylor TH, Jennings AM, Nightingale DA, Barber B, Leivers D, Styles M, Magner J: A study of anesthetic emergency work: I. The method of study and introduction to queuing theory. Br J Anaesth 1969; 41:70–5
12. Tucker JB, Barone JE, Cecere J, Blabey RG, Rha CK: Using queueing theory to determine operating room staffing needs. J Trauma 1999; 46:71–9
13. Scott DW, Factor LE, Gorry GA: Predicting the response time of an urban ambulance system. Health Serv Res 1978; 13:404–17
14. el-Darzi E, Vasilakis C, Chaussalet T, Millard PH: A simulation modeling approach to evaluating length of stay, occupancy, emptiness and bed blocking in a hospital geriatric department. Health Care Manag Sci 1998; 1:143–9
15. Bagust A, Place M, Posnett JW: Dynamics of bed use in accommodating emergency admissions: Stochastic simulation model. BMJ 1999; 319:155–8
16. Mulligan JG: The stochastic determinants of hospital-bed supply. J Health Econ 1985; 4:177–85
17. Costa AX, Ridley SA, Shahani AK, Harper PR, De Senna V, Nielsen MS: Mathematical modelling and simulation for planning critical care capacity. Anaesthesia 2003; 58:320–7
18. Gross D, Harris CM: Fundamentals of Queueing Theory, 3rd edition. Indianapolis, Wiley & Sons, 1998
19. Milne E, Whitty P: Calculation of the need for paediatric intensive care beds. Arch Dis Child 1995; 73:505–7
20. Law AM, Kelton WD: Simulation Modeling and Analysis, 3rd edition. Boston, McGraw-Hill, 2000
21. Hillier F, Lieberman G: Introduction to Operations Research, 6th edition. Boston, McGraw-Hill, 1995
22. Whitt W: Understanding the efficiency of multi-server service systems. Management Sci 1992; 38:708–23
23. Litvak E, Long MC: Cost and quality under managed care: Irreconcilable differences? Am J Manag Care 2000; 6:305–12
24. McManus ML, Long MC, Cooper AB, Mandell J, Berwick DM, Pagano M, Litvak E: Variability in surgical caseload and access to intensive care services. Anesthesiology 2003; 98:1491–6
25. Schwartz S, Cullen DJ: How many intensive care beds does your hospital need? Crit Care Med 1981; 9:625–9

Appendix: Selected Queuing Resources Available on the Internet

1. Ferrier A: An Introduction to Queueing Theory. 1999. Available at: http://www.new-destiny.co.uk/andrew/past_work/queueing_theory/Andy/. Accessed October 10, 2003
2. Slater T: The Queueing Theory Tutor. 2000. Available at: http://www.dcs.ed.ac.uk/home/jeh/Simjava/queueing/. Accessed October 10, 2003
3. Ingolfsson A, Gallop F: Queueing ToolPak 3.0. 2002. Available at: http://www.bus.ualberta.ca/aingolfsson/QTP/. Accessed October 10, 2003
4. Kamath M: A Software Package for Rapid Analysis of Queueing Systems [RAQS]. 2001. Available at: http://www.okstate.edu/cocim/raqs/. Accessed October 10, 2003