Operations Research
Vol. 59, No. 4, July–August 2011, pp. 942–955
ISSN 0030-364X | EISSN 1526-5463 | 11 | 5904 | 0942

http://dx.doi.org/10.1287/opre.1110.0943

Identifying Good Nursing Levels: A Queuing Approach

Natalia Yankovic, Linda V. Green
Columbia Business School, New York, New York 10025 {nyankovic10@gsb.columbia.edu, lvg1@columbia.edu}

Nursing care is arguably the single biggest factor in both the cost of hospital care and patient satisfaction. Inadequate inpatient nursing levels have also been cited as a significant factor in medical errors and emergency room overcrowding. Yet, there is widespread dissatisfaction with the current methods of determining nurse staffing levels, including the most common one of using minimum nurse-to-patient ratios. In this paper, we represent the nursing system as a variable finite-source queuing model. We develop a reliable, tractable, easily parameterized two-dimensional model to approximate the actual interdependent dynamics of bed occupancy levels and demands for nursing. We use this model to show how unit size, nursing intensity, occupancy levels, and unit length-of-stay affect the impact of nursing levels on performance and thus how inflexible nurse-to-patient ratios can lead to either understaffing or overstaffing. The model is also useful for estimating the impact of nurse staffing levels on emergency department overcrowding.

Subject classifications: finite capacity model; queueing model; nurse staffing; emergency room overcrowding; hospital applications.
Area of review: Policy Modeling and Public Sector OR.
History: Received July 2008; revisions received February 2009, July 2009, September 2009; accepted March 2010.

1. Introduction

Maintaining appropriate nurse staffing levels is one of the biggest challenges facing hospitals. Nursing is the largest single component of hospital budgets, typically accounting for over 50% of all costs (Kazahaya 2005), making it an important area for study given the increasing cost of care and pressures from payers to keep prices down.

Furthermore, there is a growing realization of the important role nursing care plays in the quality of health care. Over the last 15 years, evidence has been accumulating that relates higher levels of nurse staffing to lower rates of adverse patient outcomes (Needleman et al. 2002) and a decrease in the likelihood of death (Aiken et al. 2002). It is now recognized by many, including the Institute of Medicine (IOM) and the International Council of Nurses (ICN), that there is a preponderance of evidence establishing a positive relationship between nursing care and quality patient outcomes (IOM 2004, ICN 2006). However, there is still no scientifically based methodology to help nurse managers and hospital administrators efficiently allocate scarce nursing resources to promote quality patient outcomes in their own settings.

Nursing levels are generally determined independently for each clinical unit (sometimes referred to as a ward), which typically varies from 30 to 60 beds. A clinical unit might consist of general medical-surgical beds or correspond to one or more specific hospital services, such as cardiology, infectious diseases, orthopedics, oncology, neurology, pediatrics or obstetrics. For each nursing shift, which generally is either 8 or 12 hours long, there is a nurse manager and a dedicated nursing staff. In some cases, shifts are designed to overlap to better deal with anticipated changes in workload.

Minimum nurse-to-patient ratios are one of the most commonly used methods to establish staffing levels. California's 1999 law AB 394 set minimum nurse-to-patient ratios of 1 to 6 on general medical-surgical wards, and other states are now considering similar proposals (California Department of Health Services (2003), Health Policy Tracking Service 2005). Opponents of mandated ratios (Lang et al. 2004, SHS 2005, Kane 2007) observe that ratios are too inflexible to account for variation in nursing skills and the severity of patients' illnesses. This has been confirmed by a study showing that the mandated staffing ratios implemented in California did not result in the expected patient benefits (White 2006). Other common staffing methods include estimating the total direct productive hours of care per patient day (HPPD) and patient classification systems (PCS). As with mandated nursing ratios, these methods have been criticized for failing to adequately account for differing patient needs (American Nurses Association 1999, Seago 2002).

The American Nursing Association (ANA) and others have called for the development of flexible, patient-centered staffing policies based on multiple variables such as differing patient needs, fluctuations in care needs by day and time, expertise and education of the staff, and other setting characteristics (see, e.g., Lang et al. 2004). Ideally, nurse staffing levels should be based on a quantification of the actual patient needs for nursing and the amount of time associated with these needs, so that nurse services are able to be provided in a timely fashion. This is supported by many adverse patient events that have been linked to inadequate nursing levels, such as failure to rescue, which are clearly time-sensitive (Kane et al. 2007).

This paper makes three major contributions. First, we develop a queuing model that can readily be used by hospital managers to guide nurse staffing decisions. Given the stochastic nature of patient demands and services, the need for a high level of responsiveness, and the flexibility needed to capture varying conditions as described above, queuing methodology seems very well suited for guiding nurse staffing decisions. A clinical unit can be viewed as a finite source queuing system because demands are generated from the inpatients in that unit. However, because of admissions, discharges, and transfers, the number of inpatients varies over a shift, and each of these changes triggers a demand for nursing care. So our queuing model is a unique variable finite source model with two sets of servers: nurses and beds. This model is part of a project with the Hospital for Special Surgery (HSS) in New York City to determine the feasibility and usefulness of using a queuing model to help guide nurse staffing levels in hospitals. Second, we use analytical and computational methods to identify the major factors that affect nurse-to-patient ratios needed to achieve clinically appropriate levels of patient service. Specifically, we show that in addition to unit size and level of nursing intensity, occupancy levels and average length-of-stay could play a significant role in affecting nurse-to-patient ratios that are consistent with timely service. These results demonstrate that fixed nurse-to-patient ratios can result in understaffing or overstaffing, thereby compromising cost and/or quality outcomes. Third, we demonstrate and quantify the impact of nursing levels on backlogs in the emergency department (ED). Although a lack of inpatient beds generally is cited as the primary reason for ED overcrowding, we show that even when the number of beds is sufficient, nurse unavailability can cause major backlogs in the ED.

The rest of the paper is organized as follows. In §2 we review the related literature, and in §3 we develop our basic model and present the major performance metrics. In §4 we demonstrate the reliability of the queuing model by comparing its suggested nurse staffing levels with those of a more realistic simulation model. We then use the queuing model to explore the clinical and patient factors that most affect nurse staffing levels and highlight situations in which the California mandated ratios are likely to be inconsistent with good system performance. Finally, in §5 we summarize and discuss our major results and insights and identify some future research directions.

2. Literature Review

Resource allocation in hospitals has been a subject of operations research studies for many years. Bed occupancy and patient flows within a hospital are examples of early applications of Markov and semi-Markov processes (Young 1965, Kao 1974, Hershey et al. 1981). See Green (2006) for an overview of the use of OR models for capacity planning in hospitals. Given the cost and quality impact of clinical personnel, much of the literature has focused on staffing and scheduling problems. Most of this work has used linear programming models to provide guidance on the scheduling of nurses and other hospital personnel (Miller et al. 1976, Kwak and Lee 1997, Jaumard et al. 1998). Most recently, the work of Wright et al. (2006) uses linear programming to analyze the impact of mandated nurse-to-patient ratios on workforce costs and the ability to provide nurses with "desirable" schedules. All of these have assumed as input a given number of nurses needed for each shift, based on either mandated ratios or one of the other methods mentioned previously.

The use of queuing models to guide staffing decisions in health care facilities has been limited. One exception was described in a study by Green et al. (2006), which used a nonstationary queuing approach to allocate physicians in the emergency department of an urban hospital in order to reduce the fraction of patients who leave before being seen by a physician. de Véricourt and Jennings (2006) is the only previous paper to use a queuing model in the context of nurse staffing. They use a standard finite source multiserver queue to demonstrate that fixed nurse-to-patient ratios as embodied in the 1999 California legislation cannot achieve consistent performance across different unit sizes. Because their model assumes a fixed number of inpatients, it doesn't include the work generated by new arrivals, departures, and transfers of patients. In contrast, the models developed in this paper incorporate the potentially large fluctuations in the census of a clinical unit during a shift, which significantly affect the demand for nursing care (Lapierre et al. 1999, Volpatti et al. 2000).

3. Model Description

We model a specific clinical unit of a hospital for a given fixed staffing interval, which we call a shift, as a queuing system with two types of servers: beds and nurses. The demand for beds comes either from patients waiting in the ED or from elective patients. The demands for nursing care are primarily generated from the current inpatients (e.g., call button requests, medication needs, and monitoring). However, patient movements such as admissions, discharges and transfers change the census level of the unit and also require nurse involvement. For example, a patient admission requires a nurse to move the patient into the bed, set up any necessary monitoring and/or intravenous lines, take vital sign readings, etc. Similarly, once a physician gives a discharge order for a patient, the nurse must be available to provide post-discharge instructions, prescriptions, fill out forms, etc. before the patient can physically leave. So our model must track both inpatient requests and patient movements.

We let B be the number of beds in the unit and N be the number of nurses where $B > N$. We assume that N is fixed during the shift. (While in many cases staffing levels remain constant over a nursing shift, e.g., 8 or 12 hours, in other cases—such as where shifts might overlap—the model would be used for each staffing interval defined as a continuous interval of time during which the nurse staffing level remains constant.) We assume that patients arrive to the unit according to a homogeneous Poisson process. This assumption is very reasonable for units for which most arrivals are unscheduled, such as medical units and obstetrics units (Young 1965). Even in units where most patients are scheduled, such as surgical units, the exact number and timing of patient arrivals into the unit (which typically come from the recovery room) have been found to be very random due to variability, additions, and cancellations in the surgical schedule (Litvak and Long 2000). We assume that patient lengths of stay (LOS) in the clinical unit are exponentially distributed. This assumption is supported by empirical data from a large urban hospital for which the coefficient of variation of LOS was close to one in most clinical units (Green and Nguyen 2001). However, we will relax this assumption when we test the model's estimates against those of a simulation in §4. At any given time t, we assume there is a fixed number of inpatients in the unit, each of which independently generates requests for nursing care with an exponentially distributed time between requests, and that the amount of time a nurse spends on each patient request is exponentially distributed. Although the assumption of exponentially distributed service times is necessary for analytical tractability, it is also well supported by the only study reported in the literature on the use of nurses' time in a medical-surgical unit (Lundgren and Segesten 2001). Although patients are usually assigned to a specific nurse for each shift, we assume, as in common practice, that any available nurse can attend to a patient if the assigned nurse is busy with other patients. We assume that requests are performed on a first-come, first-served basis. (This assumption can be relaxed to allow for priority classes in a heuristic model described in Appendix 4). Finally, we assume that all nurses are equally trained and can perform all requests. This assumption is valid for many hospitals and is also consistent with other nurse staffing methodologies, e.g., nurse-to-patient ratios. We will discuss the implications of this assumption for more complex situations in the last section of the paper.

So our model consists of two queuing systems—the first representing the need for *beds* in the unit and the second the requests for *nurses*. These queues are related because the number of occupied beds will determine the demand for nurses, and nurse availability can influence admissions and discharges, and hence, bed occupancy.

We use the following notation in our analyses:

—B Beds
—N Nurses
—λ_b Arrival rate for the bed system
—μ_b Service rate for the bed system
—λ_n Demand rate for the nurse system
—μ_n Service rate for the nurses.

The complex interaction between nursing workload and bed dynamics would seem to necessitate a four-dimensional state space. To obtain a tractable, two-dimensional state space that captures the essential interaction effects between beds and nurses, we model the system dynamics as follows. We assume, as observed in reality, that a patient arriving to the unit when all nurses are busy but a bed is available will have a bed assigned to him/her and generate a request for nursing care at that time (corresponding to the work associated with a new admission), even though the bed will not be occupied until a nurse is available. We further assume that the arrival rate for the nurse system includes the discharge requests, but that the timing of the physical departure of a patient from his/her bed does not necessarily coincide with the discharge work associated with it. Specifically, we assume that the work associated with discharge might occur in advance of the patient's vacating the bed. This assumption is not unreasonable because some patients remain in their room after being officially discharged; for instance, they must wait for transportation or a relative in order to leave. Based on our work at HSS, we have learned that during this delay, patients often continue to generate demands for nursing; for example, medication administration. So our model assumes that at the time of physical departure, a patient could have an unfulfilled nursing request. However, to capture the impact of nurse unavailability on patient flow and to facilitate our analysis, we assume that discharges of patients who are not being served by a nurse are blocked whenever all nurses are busy. Note that this assumption prevents the hospital LOS for a patient who has been assigned but not yet admitted to a bed from beginning (and therefore also ending) before he/she occupies the bed.

The system described above can be described as a bivariate Markov process with state space (X_b, X_n), where X_b is the number of occupied beds plus the number of patients who are waiting for a bed, and X_n is the number of inpatients being cared for by a nurse plus the number of patients waiting for a nurse. It's important to note that although the bed system has Poisson arrivals and exponential service times, it is not equivalent to an $M/M/c$ system where $c = B$. This is because the times that patients spend in a bed are not independent of one another due to the blocking of discharges when all nurses are busy.

In our fundamental model, we assume that all patients arriving to the bed system when it is full join the queue, and there is no balking or blocking. In reality, many hospitals attempt to block new arrivals to a clinical unit when the number of patients waiting for admission to that unit exceeds some level. This can be done by placing patients in other units that have available beds and/or by going on ambulance diversion, i.e., instructing dispatchers to send ambulances to other hospitals. However, some hospitals have a policy of never going on ambulance diversion, and for those in rural areas it might not be an option. In smaller hospitals, there might be no alternative clinical units to which patients can be admitted. In these cases, and also to study the impact of nurse staffing on ED congestion and bed admissions, the infinite waiting room model presented below is more appropriate. In Appendix 2, we describe a modification of this model that assumes a finite waiting room.

For any given state (i, j), the set of possible successor states and the associated transition rates depend on the "macrostate" as follows:

(a) All beds are occupied, and all patients in the nurse system are being served $(i > B, j < N)$. So arrivals to the bed system join the bed queue, changing the state space to $(i+1, j)$ at rate λ_b because all beds are occupied and no nursing work is generated by this arrival at this time.

Departures from the bed system occur at rate $\mu_b \cdot B$, leading to one of two possible states: If the departing patient was not in the nurse system, the transition is to $(i-1, j+1)$, because the vacated bed allows for the admission of a patient from the bed queue, which generates a new demand for the nurse system. However, if the departing patient was in the nurse queue the resulting state is $(i-1, j)$, because as before, we generate a new nursing job associated with the admission of a patient from the bed queue, but the departure leaves both systems simultaneously. Requests for nursing from inpatients occur with rate $\lambda_n \cdot (B - j)$, changing the state to $(i, j+1)$. Completions of nursing jobs occur at rate $\mu_n \cdot j$ changing the state to $(i, j-1)$.

(b) All beds are occupied, and all nurses are busy $(i \geqslant B, j \geqslant N)$.

Arrivals to the bed system occur as in the previous case. Because all nurses are busy, discharges from the bed system are blocked for those without a nurse. That is, we assume that when all nurses are busy, only the patients who are in service with a nurse will be physically discharged if their length-of-stay is over during their service; this occurs with rate $\mu_b \cdot N$. If $i > B$, a departure allows a patient from the bed queue to reserve an available bed and the transition will be to state $(i-1, j)$. When $i = B$ there is no queue for beds and the departures must be from those currently in service, so the transition is to state $(i-1, j-1)$.

Requests for nursing from inpatients occur with rate $\lambda_n \cdot (B - j)$, changing the state to $(i, j+1)$. Completion of nursing jobs occur at rate $\mu_n \cdot N$, changing the state to $(i, j-1)$.

(c) Both beds and nurses are available $(i < B, j < N)$.

New arrivals are immediately admitted and generate a demand for nursing so that transitions to state $(i+1, j+1)$ occur with rate λ_b. Departures from the bed system can be of the two types described in case (a), but in this case we don't have patients waiting to be admitted. The transition will be to state $(i-1, j)$ if the departing patient was not in the nurse system, or to state $(i-1, j-1)$ otherwise.

Requests for nursing from inpatients occur with rate $\lambda_n \cdot (i-j)$, changing the state to $(i, j+1)$, and nursing job completions occur at rate $\mu_n \cdot j$, changing the state to $(i, j-1)$.

(d) All beds are occupied but there is no queue, and nurses are available $(i = B,\ j < N)$.

Arrivals to the bed system occur as in case (a). Departures from the bed system occur as in case (c). Requests for nursing from inpatients occur with rate $\lambda_n \cdot (B - j)$, changing the state to $(i, j+1)$, and completions of nursing service occur at rate $\mu_n \cdot j$, changing the state to $(i, j-1)$.

(e) Beds are available and all nurses are busy $(i < B, j \geqslant N)$.

Arrivals to the bed system occur as in case (c), and discharges from the bed system are blocked for those without a nurse as in case (b) because all nurses are busy. In this situation we will have departures with rate $\mu_b \cdot N$, and because there is no queue for the bed system and departures must be from those currently in service, the transition is to state $(i-1, j-1)$. Requests for nursing from inpatients occur with rate $\lambda_n \cdot (i-j)$, changing the state to $(i, j+1)$. Completions of nursing jobs occur at rate $\mu_n \cdot N$, changing the state to $(i, j-1)$.

The model described above corresponds to a quasi birth and death process (QBD) with generator Q, which can be partitioned into blocks of size $(B+1) \times (B+1)$ and boundary conditions for all $i < B$. This block structure allows us to check for stability of the system and to solve for the steady-state distribution p_{ij}, where $p_{ij} = \Pr(X_b = i, X_n = j)$, using the ideas of matrix geometric analysis (Neuts 1981). Appendix 1 includes a complete description of the transition matrix and blocks for the case $B = 3$, $N = 2$ as well as the stability conditions.

The steady-state probabilities p_{ij} can be used to derive a variety of performance measures for evaluating the impact of nursing levels on delays for both nurses and beds. Below we describe some of the more commonly used performance measures.

Average bed occupancy rates are widely used and reported to evaluate hospital efficiency (Green 2002). The occupancy level of a clinical unit is simply the bed utilization rate, ρ_b. In our model, this rate is a function not only of the arrival rate λ_b, the service rate μ_b, and the number of beds available B, but also of the probability of discharges being blocked because all nurses are busy, $P(block)$. Other metrics of interest relevant to the bed system include the probability of delay for a bed, the average delay for a bed, and the average number of patients waiting for a bed. The equations for these metrics are presented in Table 1(a)

To identify nurse staffing levels consistent with cost and quality objectives, it is more important to consider performance metrics related to the nursing system. Although other metrics can be readily developed, Table 1(b) presents some commonly used metrics, including average nurse utilization, probability of a delay for nursing care, and average number of patients waiting for a nurse. The derivations appear in Appendix 1.

In Table 1(b), it is important to note the use of the effective arrival rate of nursing requests λ_n^{ef}, which includes both

Table 1. Performance metrics.

(a) Bed system	
Probability discharges are blocked	$P(block) = \sum_{i=N}^{B-1}\sum_{j=N}^{i}\frac{i-N}{i}p_{ij} + \sum_{i=B}^{\infty}\sum_{j=N}^{B}\frac{B-N}{B}p_{ij}$
Bed utilization	$\rho_b = \frac{\lambda_b}{B\cdot\mu_b\cdot(1-P(block))}$
Probability of delay for bed	$P_b(delay) = \sum_{i=N}^{B-1}\sum_{j=N}^{i}p_{ij} + \sum_{i=B}^{\infty}\sum_{j=0}^{B}p_{ij}$
Expected N of patients waiting for bed	$L_q = \sum_{i=B+1}^{\infty}\sum_{j=0}^{B}(i-B)\cdot p_{ij}$
(b) Nurse system	
Nurse utilization	$\rho_n = \frac{\lambda_n^{ef}}{N\cdot\mu_n}$
Arrivals finding system at state (i, j)	$q_{ij} = \frac{\min(i,B)-j}{L_s-L_n}\cdot p_{ij}$
Probability of nurse delay	$P_n(delay) = \sum_{i=N}^{B-1}\sum_{j=N}^{i}q_{ij} + \sum_{i=B}^{\infty}\sum_{j=N}^{B}q_{ij}$

demands from current inpatients and demands due to new admissions and is given by Equation (1):

$$\lambda_n^{ef} = \sum_{i=0}^{B-1}\sum_{j=0}^{i}(\lambda_b + (i-j)\cdot\lambda_n)\cdot p_{ij} + \sum_{j=0}^{B}(B-j)\cdot\lambda_n\cdot p_{Bj} + \sum_{i=B+1}^{\infty}\left[\sum_{j=0}^{N-1}((B-j)\cdot\lambda_n + B\cdot\mu_b)\cdot p_{ij} + \sum_{j=N}^{B}((B-j)\cdot\lambda_n + N\cdot\mu_b)\cdot p_{ij}\right]. \tag{1}$$

As mentioned previously, many hospitals limit the number of patients waiting for admission into a clinical unit either by placing new arrivals "off-service" into a different unit or by going on ambulance diversion. For use in those situations, we consider an alternative model with the same state space and transition dynamics as above, but assuming a finite waiting room. Similar performance metrics can be derived for this situation; see Appendix 2.

4. Results

4.1. Model Validation

As noted previously, our queuing model is an approximation of actual nursing dynamics and does not capture all the elements of a clinical unit that could affect the performance associated with nursing care. To determine its reliability, we tested it against a simulation model that better conforms to the reality of patient, nurse, and bed dynamics.

The major distinction of the simulation is in the modeling of the discharge process, which in the queuing model is assumed to be blocked for patients not with a nurse when all nurses are busy. In the simulation, the discharge of a patient can be handled only by his/her nurse, who is assigned upon admission. In addition, the simulation allows for the inclusion of a bed-cleaning time when a patient leaves the unit. Based on conversations with HSS nursing staff, we also allow the average nursing time associated with admissions and discharges to be different (greater) than that for other nursing activities. The detailed description of the simulation can be found in Appendix 3.

To test the reliability of our queuing model, we examined the nurse staffing levels needed to keep the probability of inpatient delay below a specified threshold over a range of parameter values for unit size, nursing intensity, average length of stay, and bed utilization. We used as our "base case" the specific 42-bed HSS surgical unit that is the focus of our HHS research study. Because the majority of medical/surgical units in most hospitals range between 20 and 60 beds, we varied B accordingly.

Hospitals generally record the average time patients spend in the hospital, or length of stay (ALOS), as well as average occupancy levels. Based on historical data from our HSS study unit, the ALOS is 4.3 days. Based on this and ALOS data from other hospitals (see, e.g., Healthcare Cost and Utilization Project 2008), we varied ALOS from 3 to 8 days.

The average bed occupancy level in HSS is 78%, which is nearly identical to the average daily occupancy level in NYC of 79% (Commission on Health Care Facilities in the 21st Century 2006, final report). We used a set of arrival rates to the bed system λ_b to roughly correspond to different bed occupancy levels. Because actual bed occupancy levels depend upon nurse utilization due to the potential for blocking of discharges when all nurses are busy, we varied the *nominal bed utilization* as defined by

$$\hat{\rho}_b = \frac{\lambda_b}{B\cdot\mu_b}, \tag{2}$$

which is the utilization rate of the bed system without considering the nurses' interaction. We tested for $\hat{\rho}_b \in \{0.65, 0.75, \text{and } 0.85\}$. It is far more challenging to get estimates regarding the nurse system. Inpatient demands for nursing care and the duration of nursing tasks are not generally tracked by hospitals, and HSS is no exception. We relied on the few published sources we have found that report estimates for nursing tasks (Dochterman and Bulechek 2004, Lundgren and Segesten 2001). Based on these, we used an expected time of 15 minutes, i.e., $\mu_n = 4$ in our numerical experiments. For the admission and discharge service times we used uniform distributions ($[12, 60]$ minutes for admissions and $[10, 60]$ minutes for discharges), based on data collected from a nurse focus group conducted at HSS. The time to clean a room after a patient discharge was assumed to be 30 minutes.

To estimate λ_n, the rate at which each inpatient generates a demand for nursing care, we again used the study of Lundgren and Segesten. From their data we computed an average demand rate of 0.38 requests/hour for each inpatient, so in our baseline numerical studies below, we used $\lambda_n = 0.4$. Because some hospital units, such as many surgical units, have more needy patients, we also explored the impact of increasing this parameter value up to 0.5.

Because most hospitals have some finite capacity limit on the number of patients in the ED waiting for beds, we fixed the waiting room size to be 3. However, numerical tests on the size of the waiting room showed that our results were relatively insensitive to this parameter.

Perhaps the most important measure in determining nurse staffing levels is the delay experienced by patients for nursing care. Although there are currently no standards of what constitutes safe delays for nursing care, the development of such standards will be part of our project with HSS. This could take the form, as in call centers, of a maximum fraction of patients who experience a delay of more than a given duration. As described in Appendix 1, the tail distribution for the waiting time conditional on the state of the system is Erlang, so we can easily compute this type of measure. For testing and illustrative purposes, we used probability of delay in our numerical examples.

The total number of combinations of the various parameter settings resulted in an experimental set of 216 cases. For each case, we used the queuing model to determine the minimum number of nurses needed to keep the probability of delay for an inpatient request less than or equal to a specified value. Based on data and conversations with HSS personnel and the widely expressed desire of both patients and nurses to assure timely response to patient needs, we considered two probability of delay targets: 0.05 and 0.10. We then used the simulation model to evaluate whether the nursing level suggested by the queuing model resulted in a probability of delay that met that target performance. To reflect the observation that for most managers the target is not a strict constraint, we counted as "unsuccessful" any case in which the simulation's estimate of probability of delay exceeded the target by more than 10%.

Only 5 of the 216 cases—less than 2.5%—were unsuccessful. All these cases corresponded to a short ALOS, i.e., 3 days, and 4 of the 5 had a nominal bed utilization level of 85%. We hypothesize that in these cases the queuing model slightly underestimates delays because of the frequency of admissions and discharges, which in the simulation require more time on average than regular services. Also, in another 8% of cases the simulation indicated that the target could be met with 1 fewer nurse than suggested by the queuing model. All these latter cases were for fairly high values of ALOS, and most were for very large units, i.e., 60 beds. Our hypothesis is that because large units have many patients, there is a higher number of patient discharges likely to be blocked when all nurses are busy. When ALOS is high, this can result in more patients occupying their beds for a considerable time after the actual length of stay is over, due to the queuing model's discharge assumption, and hence considerably more nursing demands. Thus the queuing model is more likely to overestimate nursing needs in these cases.

To test the robustness of the exponential assumption for length-of-stay, we also ran the simulation assuming a lognormal distribution keeping the coefficient of variation of 1. The results were almost identical, i.e., only 6 of the 216 same cases as before were "unsuccessful."

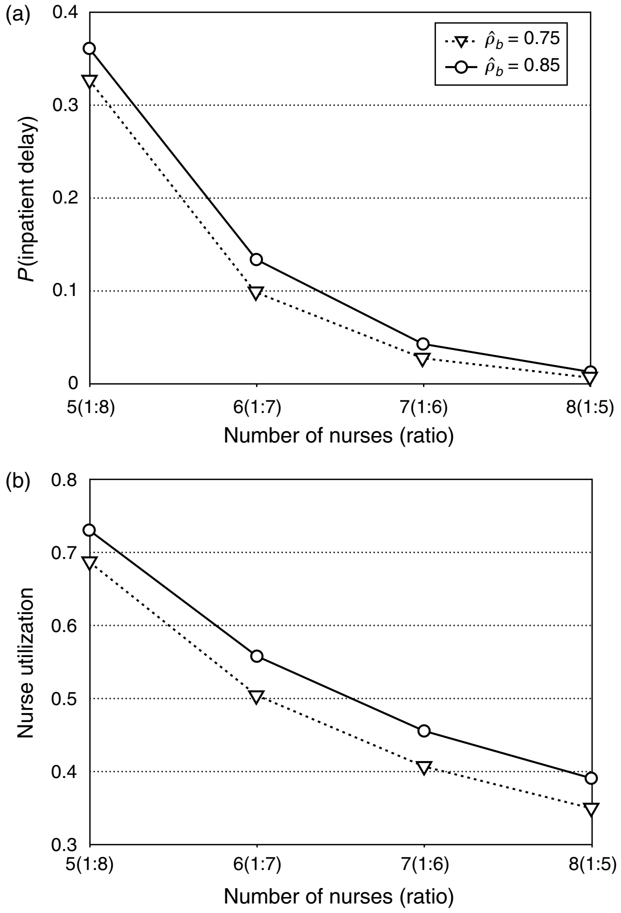
Our results indicate that the queuing model will almost always suggest a staffing level that assures a high level of responsiveness. Given that there are very few nursing units that have on the order of 60 beds, our results indicate that for the vast majority of nursing units, the queuing model's staffing estimates will be very reliable.

4.2. Factors Affecting Nurse Staffing Ratios

In this section, we use the model to identify the major factors that might cause commonly used nurse-to-patient ratios to under- or over-estimate the number of nurses needed to assure good performance. For this purpose, we use as a starting point parameter values from our study unit at HSS. Although HSS is a hospital that specializes in orthopedic surgery, most of the key nursing tasks and practices, such as admission, discharges, medication administration, monitoring, and call button use, are similar to those in other hospitals. Because no one hospital unit is representative of the various conditions that exist across different hospitals and units, we vary the key parameters of our model to identify common characteristics that will likely affect the level of nurse staffing required for a high level of patient responsiveness.

4.2.1. Effect on Inpatient Delay. As mentioned previously, one key element in the quality of patient care is nurse responsiveness. In this subsection we use the probability of inpatient delay as the major performance metric to identify adequate levels of nurse staffing.

Figure 1. Base case performance.



Notes. Base unit. $B = 40$, $WR = 4$, $\hat{\rho}_b = 0.75$, ALOS $= 4.5$, $\lambda_n = 0.4$.

In Figure 1(a) we show the probability of inpatient delay as a function of nurse staffing for our base case unit: $B = 40$ beds, nurse intensity $\lambda_n = 0.4$ requests per hour, average length of stay of 4.5 days, and waiting room $WR = 4$. We show the number of nurses as well as the nurse-to-patient ratio defined as $1{:}M$, where M is the maximum number of patients per nurse. We use two different levels of bed utilization $\hat{\rho}_b \in \{0.75, 0.85\}$ to capture the variation in occupancy levels that occur over the day and over the week.

To be consistent with the minimum nurse-to-patient ratio of 1:6 that was used in the California legislation, which mandates that the minimum ratio is required "at all times," this unit would need to staff 7 nurses. (Recall that actual average bed utilization is higher than nominal bed utilization, so our use of nominal bed utilizations of 0.75 and 0.85 results in a substantial fraction of time that the unit is full.) In the case when nominal bed utilization is 0.85 this level of staffing will translate into a probability of inpatient delay of 4.37%. However, for the case with nominal bed utilization of 0.75 with 6 nurses (nurse-to-patient ratio of 1:7) the probability of inpatient delay is below 10%, which might be considered a reasonable target. Currently, HSS staffs 6 or 7 nurses in the unit under study, depending upon anticipated levels of admissions

Studies investigating nurses' use of time show that direct care accounts for only between 30 and 60% of nurses time (Jinks and Hope 2000, Lundgren and Segesten 2001). Other nurse activities, including indirect care or activities that occur away from the patient (preparing for nurse interventions, medications, and therapies), rounds with the MD, report writing, communication with visitors, and personal activities account for a significant fraction of nurses' time. The numerical results for our model are consistent with those levels of nurse utilization as shown in Figure 1(b).

We now look at some of the factors that are likely to affect the impact of a given level of staffing on delays. In the following analyses, we use the parameter values from our base case and vary only the factor under investigation.

Perhaps the most obvious factor that is likely to affect nursing levels consistent with timely response to patient needs is unit size, because bigger units will benefit from statistical economies of scale. The effect of unit size is illustrated in Table 2, where we show the number of nurses and corresponding nurse-to-patient ratios needed to achieve the targets for inpatient delay for two levels of nominal bed utilization. We can see that varying the number of beds ($B \in \{20, 40, 60\}$) results in very different nurse-to-patient ratios. For a target probability of inpatient delay of 10%, the required ratio goes from 1:5 to 1:7; and for the 5% target the ratio varies even more, going from 1:4 for the smallest unit to 1:7 for the largest.

Another factor likely to affect nurse staffing levels is nursing intensity, i.e., the rate at which patients generate nursing needs. This is illustrated in the results presented in Table 3, in which this parameter varies from 0.35 to 0.5. It is interesting to note that for both delay targets and both levels of bed utilization, an additional nurse is needed when the request rate for nursing care increases roughly 10% from the average level reported in the literature. This seems to strongly indicate that this an important factor to consider in determining nursing levels. More specifically, units with

Table 2. Staffing needed to achieve delay targets—Size effect.

	Number of beds B		
Size effect	20	40	60
		(a)	
P(delay < 0.1)	4(1:5)	6(1:7)	9(1:7)
P(delay < 0.05)	5(1:4)	7(1:6)	9(1:7)
		(b)	
P(delay < 0.1)	4(1:5)	7(1:6)	9(1:7)
P(delay < 0.05)	5(1:4)	7(1:6)	10(1:6)

Notes. WR $= 0.1 \cdot B$, $\hat{\rho}_b = 0.75$, ALOS $= 4.5$, $\lambda_n = 0.4$.
WR $= 0.1 \cdot B$, $\hat{\rho}_b = 0.85$, ALOS $= 4.5$, $\lambda_n = 0.4$.

Table 3. Staffing needed to achieve delay targets—Nursing intensity effect.

	Number of demands λ_n (request/hr.)			
Nurse intensity effect	0.35	0.40	0.45	0.50
	(a)			
P(delay < 0.1)	6(1:7)	6(1:7)	7(1:6)	7(1:6)
P(delay < 0.05)	7(1:6)	7(1:6)	8(1:5)	8(1:5)
	(b)			
P(delay < 0.1)	6(1:7)	6(1:7)	7(1:6)	7(1:6)
P(delay < 0.05)	7(1:6)	7(1:6)	8(1:5)	8(1:5)

Notes. $B = 40$, WR $= 4$, $\hat{\rho}_b = 0.75$, ALOS $= 4.5$.
$B = 40$, WR $= 4$, $\hat{\rho}_b = 0.85$, ALOS $= 4.5$.

patients who have greater nursing needs, e.g., many surgical units, are likely to need higher levels of nurses than other units, and more than suggested by the California mandated minimum ratios if a high level of responsiveness is desired.

We also explore the impact of average length of hospital stay since shorter hospital stays translate into more admissions and discharge events. Our results are summarized in Table 4. Although for most cases ALOS does not seem to affect nursing levels, we do see an effect when ALOS is very low, i.e., 2.5. Given the results of some of our simulation runs, which indicated that our queuing model might occasionally underestimate delays and hence staffing needs when ALOS is very short, we believe that this factor might have a more pervasive impact than indicated by this table.

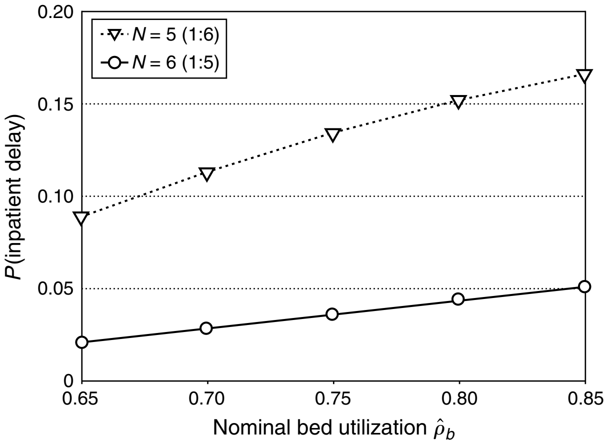
To further illustrate this last point and to demonstrate the potential effect of several factors that are likely to exist concurrently for a single unit, we examine nurse staffing levels for a 30-bed unit with a low ALOS (3 days) and a relatively high rate of demand for nursing care (0.45). In this case, the proposed 1:6 ratio isn't sufficient to meet even the looser 10% probability of delay target if the nominal bed utilization is greater than 0.7. As illustrated in Figure 2, an additional nurse would assure a high level of responsiveness even with high occupancy levels.

Table 4. Staffing needed to achieve delay targets—ALOS effect.

	Average length of stay ALOS$_{\text{(days)}}$			
Average length of stay effect	2.5	4.5	6.5	8.5
	(a)			
P(delay < 0.1)	7(1:6)	6(1:7)	6(1:7)	6(1:7)
P(delay < 0.05)	7(1:6)	7(1:6)	7(1:6)	7(1:6)
	(b)			
P(delay < 0.1)	7(1:6)	7(1:6)	7(1:6)	7(1:6)
P(delay < 0.05)	7(1:6)	7(1:6)	7(1:6)	7(1:6)

Notes. $B = 40$, WR $= 4$, $\hat{\rho}_b = 0.75$, $\lambda_n = 4.5$.
$B = 40$, WR $= 4$, $\hat{\rho}_b = 0.85$, $\lambda_n = 4.5$.

Figure 2. Combined impact of nursing intensity and ALOS effects.



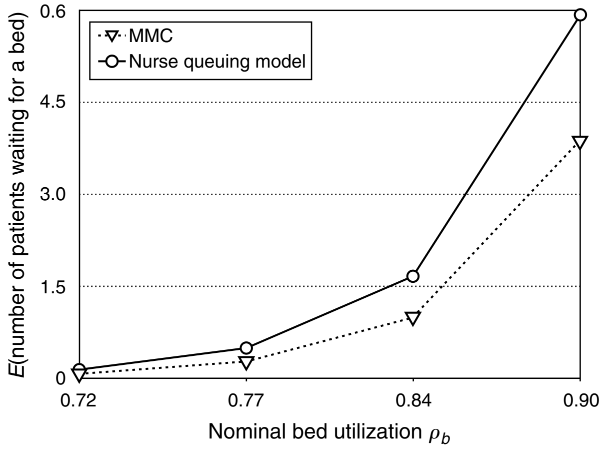
Note. $B = 30$, $WR = 3$, ALOS $= 4.5$, $\lambda_n = 0.45$.

4.2.2. Effect on Emergency Department Congestion.

Our nurse staffing model is also useful for estimating the performance of the bed system, measured either by the expected delay or expected number of patients waiting for a bed. As mentioned previously, our model is not equivalent to an $M/M/c$ queuing system because service times are not independent. The possibility of admission or discharge blocking due to nurse unavailability adds another source of variability that can have a significant adverse affect on system performance.

Figure 3 shows the expected number of patients waiting in the base case unit, assuming no blocking or balking is allowed. We contrast the results from our nurse staffing model using L_q, as defined in Table 1(a), with the results from a standard equivalent $M/M/c$ queuing model with an average utilization equal to the actual bed utilization

Figure 3. Impact of nurse staffing levels on ED overcrowding.



Notes. Base unit. $B = 40$, $WR = 4$, $N = 7$ (1:6), ALOS $= 4.5$, $\lambda_n = 0.4$.

from the nurse staffing model (ρ_b from Table 1(a)) so that both systems have the same workload. We can see that using an $M/M/c$ system without considering the impact of nursing on bed dynamics results in a significant underestimate of the number of patients waiting for a bed and hence delays for beds, especially in cases with high load. These findings are consistent with conversations we have had with ED physicians who report longer waits for beds due to nurses blocking admissions and discharges when units get busy.

It is very interesting to note that even with nursing levels that result in reasonable inpatient delays, i.e., below the 10% target, the expected backlog of patients waiting for a bed can be significantly higher than that predicted from the $M/M/c$ queue. For instance, the probability of inpatient delay when $\rho_b = 0.9$ (in Figure 3) is 5.36%, yet the expected number of patients waiting for a bed for this case is 6, which is 2.01 above the prediction of the $M/M/c$ model. This discrepancy can make the difference between the ED going on diversion or not.

5. Discussion

In this paper, we have presented a two-dimensional queuing model to guide nurse staffing decisions and demonstrated its reliability in identifying good staffing levels across a broad range of parameters corresponding to actual hospital units. From a methodological perspective, this model is unique in its assumption of a finite population of customers which is a random variable dependent on the dynamics of another queuing system. From a practical perspective, this is the first quantitative model that represents the crucial interaction between the nurse and bed systems and can be used by hospital managers to evaluate the impact of nurse staffing decisions on both delays for nursing care and delays for inpatient beds. The model is very flexible and conforms to the guidelines issued by the ANA. Furthermore, the model uses input parameters that correspond to clinical characteristics that hospital managers are familiar with, such as admission rates, ALOS, and nursing intensity. Given the speed with which it can be solved, the queuing model could be easily incorporated into a software package for nurse scheduling, an application that is being implemented by many hospitals. It can also be used to evaluate the benefits of different shift designs to better accommodate changing needs for nursing care over the day by incorporating it into a LAG SIPP approach (see Green et al. 2006).

Numerical results using our model demonstrate the problem in using rigid nurse-to-patient ratios across a broad range of hospital units. Specifically, we have shown that unit size, nursing intensity, bed utilization, and average length of stay will affect nurse staffing levels that are consistent with timely response to patients. The implications of these findings are extremely important from both a patient safety and hospital cost perspective. Using a 1:6 ratio in smaller clinical units with patients who have more intense nursing needs and short lengths of stay—e.g., gynecology and vascular surgery—might result in understaffing, which has been associated with higher levels of medical errors. On the other hand, a 1:6 ratio in units with 40 or more beds and less-needy patients with moderate or long lengths of stay—e.g., many general medicine units—might lead to overstaffing. This is also likely to be true for many units during night and weekend shifts when bed utilization is lower, admissions are less prevalent, and patients generate fewer demands. Based on an average annual salary of $60,000, the use of even one unnecessary nurse on a 12-hour shift for one unit can result in a wasted expense of $300,000 per year.

Our model can also be very useful for hospitals experiencing long ED delays and ambulance diversions, by identifying whether lack of inpatient beds or nursing staff is the bottleneck for a particular nursing unit.

No model is a perfect representation of reality, and this one is no exception. One limitation of our model is the assumption of a homogeneous workforce. In reality, some hospitals use both registered nurses (RNs) and nurses with lower levels of training, e.g., licensed practical nurses (LPNs), who cannot perform all types of interventions. In those cases our model might underestimate actual nursing delays (and hence the number of required nurses) because some requests will have to wait until the appropriate provider becomes available. Another possible complication is that there might be a priority scheme for responding to nursing demands. In preliminary discussions with HSS personnel, we found that although there are no formal priority rules, there are certain patient needs that are attended to ahead of others. Our project will try to identify these as well as determine the importance of incorporating priorities into a future version of the model by modifying our one-dimensional heuristic model.

Our findings clearly point to the need for more accurate data on parameters such as demands for nursing care and times for nursing tasks, as well as the establishment of delay standards. Our current work with HSS involves identifying both existing and potential sources of electronic patient care support data to support queuing-based staffing decisions.

We are hopeful that our model will result in more cost-effective staffing decisions and that as a result of this project, hospitals will develop IT capabilities to track nursing-related data needed to adopt this type of staffing methodology. Given the link between adequate nursing levels and patient safety as well as the imperative for cost efficiency, the adoption of queuing methodology can be an important innovation in hospital management.

Appendix 1. Derivation of Performance Metrics

A.1. Matrix Geometric Structure

Example $B=3$, $N=2$:

	00 10 11 20 21 22	30 31 32 33	40 41 42 43	50 51 52 53	60 61 62 63
00 10 11 20 21 22	B_{00}	B_{01}			
30 31 32 33	B_{10}	B_{11}	A_0		
40 41 42 43		B_{21}	A_1	A_0	
50 51 52 53			A_2	A_1	A_0

where $B_{11}=A_1$, $B_{21}=A_2$. All matrices A_i are of dimension $(B+1)\times(B+1)$.

B_{00} is a matrix of dimension $(B\cdot(B+1))/2 \times (B\cdot(B+1))/2$, B_{01} is $(B\cdot(B+1))/2\times(B+1)$ and B_{10} is $(B+1)\times(B\cdot(B+1))/2$.

$$A_0=\begin{pmatrix}\lambda_b&&&\\&\lambda_b&&\\&&\lambda_b&\\&&&\lambda_b\end{pmatrix},$$

$$A_1=\begin{pmatrix}-(\lambda_b+3\lambda_n+3\mu_b)&3\lambda_n&&\\ \mu_n&-(\lambda_b+2\lambda_n+\mu_n+3\mu_b)&2\lambda_n&\\ &2\mu_n&-(\lambda_b+\lambda_n+2\mu_n+2\mu_b)&\lambda_n\\ &&2\mu_n&-(\lambda_b+2\mu_n+2\mu_b)\end{pmatrix},$$

$$A_2=\begin{pmatrix}0&3\mu_b&0&0\\0&\mu_b&2\mu_b&0\\0&0&2\mu_b&0\\0&0&0&2\mu_b\end{pmatrix}\quad\text{and}\quad A=A_1+A_2+A_3,$$

$$B_{01}=\begin{pmatrix}0&&&\\0&&&\\0&&&\\0&\lambda_b&&\\0&&\lambda_b&\\0&&&\lambda_b\end{pmatrix},$$

$$B_{10}=\begin{pmatrix}0&0&0&3\mu_b&0&0\\0&0&0&\mu_b&2\mu_b&0\\0&0&0&0&2\mu_b&0\\0&0&0&0&0&2\mu_b\end{pmatrix}.$$

B_{00} keeps the same structure as B_{10}, B_{11}, B_{01} and can be decomposed in blocks of increasing dimension.

$$B_{00}=\begin{pmatrix}-\lambda_b&0&\lambda_b&0&0&0\\ \mu_b&-(\lambda_b+\mu+\lambda_n)&\lambda_n&0&\lambda_b&0\\ \mu_b&\mu_n&(-\lambda+\mu+\mu_n)&0&0&\lambda_b\\ 0&2\mu&0&-(\lambda_b+2\mu_b+2\lambda_n)&2\lambda_n&0\\ 0&\mu_b&\mu_b&\mu_n&-(\lambda_b+2\mu_b+\mu_n+\lambda_n)&\lambda_n\\ 0&0&2\mu_b&0&2\mu_n&-(\lambda_b+2\mu_b+2\mu_n)\end{pmatrix},$$

$$B[R]=\begin{pmatrix}B_{00}&B_{01}\\B_{10}&B_{11}+RB_{21}\end{pmatrix},$$

where R is the minimal solution to

$$R^2A_2+RA_1+A_0=0,$$

$$[x_0,x_1]\cdot B[R]=0,$$

$$x_0e+x_1(I-R)^{-1}e=1,$$

$$x_k=x_1R^{k-1}.$$

A.2. Stability Condition

When Q is positive recurrent, we can solve for the steady-state probabilities p_{ij}, which we can use to derive performance measures for evaluating the impact of nursing levels on delays for both nurses and beds. These, in turn, can be used to help hospitals identify nursing levels that are consistent with any given target level of service performance.

We can compute the *traffic* coefficient for the QBD process, γ, to determine if the system is positive recurrent or not (Neuts 1981). If we call $A0$ the transition matrix from states (i,j) to states $(i+1,k)$ (for $j,k=0{:}B$), $A1$ the transition matrix from states (i,j) to states (i,k) (for $j,k=0{:}B$), and $A2$ the transition matrix from states (i,j) to states $(i-1,k)$ (for $j,k=0{:}B$), a necessary and sufficient condition for ensuring that the QBD process will be positive recurrent is given by

$$\gamma=x\cdot A0\cdot e_1-x\cdot A2\cdot e_1<0,$$

where x solves $x \cdot (A0 + A1 + A2) = 0$ and $x \cdot e_1 = 1$, and in this case we will have stability for the bed system.

A.3. Performance Metrics

Bed System:

$$P(block) = \sum_{i=N}^{B-1}\sum_{j=N}^{i}\frac{i-N}{i}p_{ij} + \sum_{i=B}^{\infty}\sum_{j=N}^{B}\frac{B-N}{B}p_{ij}. \tag{A1}$$

The average stay in the unit is given by $\mu_b \cdot (1 - P(block))$. Using Little's law

$$\rho_b = \frac{\lambda_b}{B \cdot \mu_b \cdot (1 - P(block))}. \tag{A2}$$

To compute the probability that an arriving patient must wait for a bed, $P_b(delay)$, or the distribution of the waiting time for a bed, we must consider two different sources of delay. First, an arrival might find all beds occupied $(i \geqslant B)$ and will have to wait until $i - B + 1$ patients are discharged, where the time for each discharge has an exponential distribution with rate $B \cdot \mu_b \cdot (1 - P(block))$. Second, a patient might arrive to the bed system at state $i < B$ and still be delayed because all nurses are busy $(j \geqslant N)$, In this case he/she will have to wait until $j - N + 1$ patients are served, each of which has an exponential distribution with rate $N \cdot \mu_n$. Because patients arrive to the bed system according to a Poisson process, we get

$$P_b(delay) = \sum_{i=N}^{B-1}\sum_{j=N}^{i} p_{ij} + \sum_{i=B}^{\infty}\sum_{j=0}^{B} p_{ij}. \tag{A3}$$

Other standard measures, such as the expected number of patients in queue or inside the unit, can be computed easily. For instance, the expected number of patients waiting for a bed can be computed by

$$L_q = \sum_{i=B+1}^{\infty}\sum_{j=0}^{B}(i-B) \cdot p_{ij}. \tag{A4}$$

Nurse System: To compute the average nurse utilization we need the average effective demand rate for nursing care, λ_n^{ef}, which includes both the demands from current inpatients and the demands due to new admissions.

$$\begin{aligned}\lambda_n^{ef} = {} & \sum_{i=0}^{B-1}\sum_{j=0}^{i}(\lambda_b + (i-j)\cdot\lambda_n)\cdot p_{ij} + \sum_{j=0}^{B}(B-j)\cdot\lambda_n\cdot P_{Bj} \\ & + \sum_{i=B+1}^{\infty}\left[\sum_{j=0}^{N-1}((B-j)\cdot\lambda_n + B\cdot\mu_b)\cdot p_{ij} \right. \\ & \left. + \sum_{j=N}^{B}((B-j)\cdot\lambda_n + N\cdot\mu_b)\cdot p_{ij}\right]. \end{aligned} \tag{A5}$$

In the last half of the above expression, which corresponds to the case when all beds are full, we have to account for nursing jobs due to the admissions of patients from the bed queue which occur when a current inpatient is discharged.

Hence, the nurse utilization ρ_n will be

$$\rho_n = \frac{\lambda_n^{ef}}{N \cdot \mu_n}. \tag{A6}$$

Because the nurse system is a finite source system and therefore doesn't have a stationary Poisson arrival process, PASTA does not apply. So to get patient delay measures we must use the steady state probabilities p_{ij} to derive the set of q_{ij}, state probabilities given an arrival is about to occur (for proof see Gross and Harris 1985).

Hence, the probabilities of an arrival finding the system at state (i, j), for $i = 0, \ldots$ and $j = 0, \ldots \min(i, B)$ can be computed as

$$q_{ij} = \frac{\min(i, B) - j}{L_s - L_n} \cdot p_{ij}, \tag{A7}$$

where L_n is the expected number of patients in the nurse system $(L_n = \sum_{i,j} j \cdot p_{ij})$, and L_s is the expected number of inpatients $(L_s = \sum_{i,j} \min(i, B) \cdot p_{ij})$.

$$P_n(delay) = \sum_{i=N}^{B-1}\sum_{j=N}^{i} q_{ij} + \sum_{i=B}^{\infty}\sum_{j=N}^{B} q_{ij}. \tag{A8}$$

We can use the probabilities q_{ij} to compute the distribution of the waiting time for a delayed inpatient. A patient arriving to the nurse system at state (i, j) with $(j \geqslant N)$ will have to wait until $(j - N + 1)$ patients finish their services $j \geqslant N$ and his/her waiting time will follow an Erlang$((j - N + 1), N \cdot \mu_n)$ distribution.

Appendix 2. Finite Capacity Model

We consider an alternative model with the same state space and transition dynamics as the ones described in §3, but assuming a finite waiting room, $WR \geqslant 1$, so that $i \leqslant B + WR$.

For any given state (i, j), the set of possible successor states and the associated transition rates depend on the "macrostate" as in the infinite model. The only difference is when the bed system reaches capacity, $i = B + WR$, all arrivals to the bed system are "lost."

The steady-state probabilities for this model, p_{ij}, can be directly computed by solving the system of linear equations, or as in the previous case, we can take advantage of the special structure of the transition matrix. Akar and Sohraby (1997) present an unified approach for solving infinite and finite QBD processes using the ideas of the matrix geometric solutions without increasing complexity.

As in the infinite capacity model, the performance measures of interest can be computed from these steady-state probabilities. The utilization of the bed system will, as before, be affected by the interaction with the nurse system due to the potential blocking of discharges. However, the average arrival rate is no longer λ_b because of the probability of finding the waiting room full, so the bed utilization is given by

$$\rho_b = \frac{\lambda_b \cdot (1 - P(full))}{B \cdot \mu_b \cdot (1 - P(block))}, \tag{B1}$$

where

$$P(full) = \sum_{j=0}^{B} P_{(B+WR)j}. \tag{B2}$$

As in the infinite case, we can compute the effective arrival rate for the nurse system and its utilization rate:

$$\begin{aligned}\lambda_n^{ef} &= \sum_{i=0}^{B-1}\sum_{j=0}^{i}(\lambda_b + (i-j)\cdot\lambda_n)\cdot p_{ij} + \sum_{j=0}^{B}(B-j)\cdot\lambda_n\cdot P_{Bj}\\ &\quad + \sum_{i=B+1}^{B+WR}\left[\sum_{j=0}^{N-1}((B-j)\cdot\lambda_n + B\cdot\mu_b)\cdot p_{ij}\right.\\ &\quad\left. + \sum_{j=N}^{B}((B-j)\cdot\lambda_n + N\cdot\mu_b)\cdot p_{ij}\right].\end{aligned} \tag{B3}$$

Hence, the nurse utilization ρ_n will be

$$\rho_n = \frac{\lambda_n^{ef}}{N \cdot \mu_n}. \tag{B4}$$

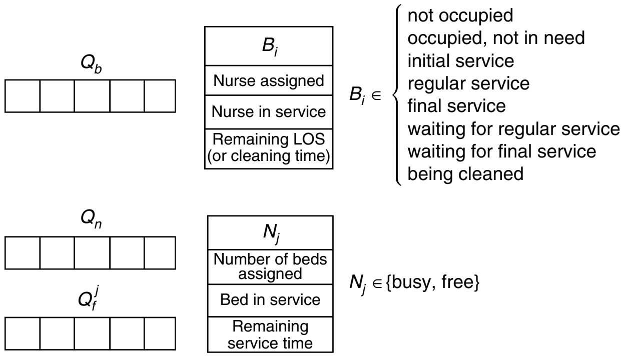
For computing the probability of delay and the distribution of the waiting time for an inpatient demand for nursing care, we need to use the normalization factors associated with the finite source, using the same approach as for the infinite case presented in Equation (A7).

Appendix 3. Description of the Simulation

This discrete event simulation shares most of the assumptions of our queuing model. We assume we have one unit of B beds and N nurses, which are fixed during the simulation in order to capture the dynamics occurring during a shift. Arrivals to the bed system follow a Poisson process with rate λ_b, and an admission can occur only if both a bed and a nurse are available. Each patient on the unit generates nursing requests with an interarrival time that is exponentially distributed with mean $1/\lambda_n$, and the service times for requests are assumed to follow an exponential distribution with mean $1/\mu_n$. As in the queuing model, the patient stay in the unit begins at the moment the nurse starts the admission process, and it follows an exponential random variable with mean $1/\mu_b$

Based on our observations and conversations with nurses at HSS and discussions with hospital personnel in other hospitals, we divide nurse patient interactions into three categories: initial service (admission), regular service, and final service (discharge). We assume that when there are several nursing jobs waiting for service, priority is given to regular requests (e.g., medication administration and call button requests), then discharge work, and finally new admissions. Each patient has a nurse assigned to him/her upon admission who, if available, responds to that patient's needs. If the assigned nurse is busy at the time of the request, another nurse, if available, responds. However, only the assigned nurse can process the patient's discharge.

Figure C.1. Simulation of a hospital unit.



Note. Space state description.

As mentioned above, we assume that each time a patient physically leaves the unit, a bed cleaning time begins. This cleaning time is uniformly distributed between $[a_c, b_c]$.

The events that can trigger the clock to advance are the following: new arrivals to the bed system, requests for normal service, requests for final service, end of initial service, end of regular service, end of final service, and end of cleaning time.

The space state needed for tracking all the interactions is summarized in Figure C.1. Patients *in bed but not in need* or *waiting for final service* generate regular nursing requests. If all nurses are busy the jobs join the nurse queue. A request for final service occurs when the remaining time in the unit of a patient hits zero. If this happens in the middle of a regular service, it is delayed until the current service finishes. If the assigned nurse is busy at the time of the discharge request, the request for final service joins the patient's assigned nurse queue.

Initial and final services are assumed to have a different distribution than regular nursing requests. Based on data collected from nurse focus groups at HSS, we assume their durations follow uniform distributions ranging between $[a_i, b_i]$ and $[a_f, b_f]$.

The simulation was programmed in Matlab 7.4 and the average time per run, using 20 beds, was 2.238 minutes using an IMac, with an Intel Core 2 Duo processor 2.64 Ghz and 2 GB 667 Mhz RAM. This is in contrast to an average run time of 0.403 seconds for the queuing model.

Appendix 4. Heuristic

Although the two-dimensional models discussed above are tractable and easy to solve numerically, it would be preferable to have a simpler model that focuses on the nursing system to make it easier for hospital mangers to use on a regular basis. For this reason, we developed and tested a one-dimensional modified finite source queuing system ($M/M/N/B$ mod). To do this, we assume a fixed number of occupied beds $\bar{B} = E$[occupied beds] that generate nursing requests at a rate λ_n each, plus an outside source of requests that arrive at rate λ_b, representing the workload associated with admissions.

$$\lambda_i = \begin{cases} (\bar{B} - i) \cdot \lambda_n + \lambda_b & i < \bar{B} \\ \lambda_b & i \geqslant \bar{B} \end{cases}, \tag{D1}$$

$$\mu_i = \min(i, N) \cdot \mu_n. \tag{D2}$$

The question is how to compute the expected number of occupied beds. If we assume there is no blocking from all nurses being busy, we have

$$\bar{B} = \hat{\rho}_b \cdot B, \tag{D3}$$

which gives us a lower bound for the actual number when the interaction between nurses and discharges is considered. This is not a bad assumption if the nurse system is lightly loaded but becomes a terrible one when bed utilization approaches 1, as illustrated in the graphs in the previous sections. In those situations it appears better to use $\bar{B} = B$ because in practice most of the time we will see a full unit.

We propose a two-step heuristic, where we start with a lightly loaded system, assuming the fixed source $\bar{B}$ as in Equation (D3). Then we adjust $\bar{B}$ using the probability of finding all nurses busy in this modified model.

The heuristic works as follows:

(a) Define $\bar{B}^1 = \hat{\rho}_b \cdot B$.

(b) Solve the steady-state probabilities for the death-and-birth process defined by the transition rates of (D1) and (D2). We call these probabilities p_i^{h1}.

(c) Compute the probability of having all nurses busy, where departures would have been blocked in the coupled model.

$$P^{h1}(block) = \sum_{j=N}^{\infty} p_i^{h1}. \tag{D4}$$

(d) Compute an approximation for the utilization of the bed system, using the logic of Equation (A2).

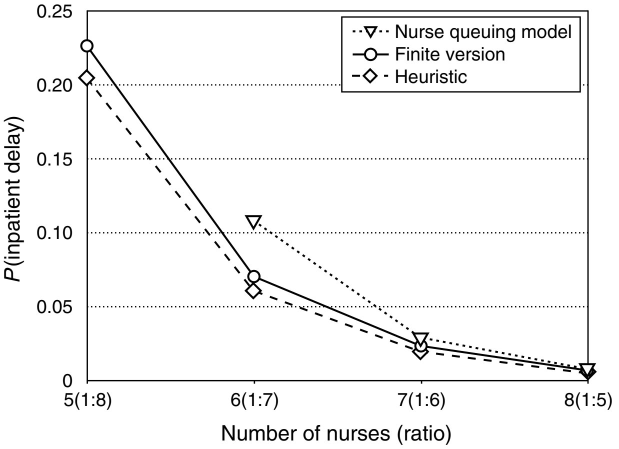
$$\rho_b^{h1} = \min\left\{1, \frac{\lambda_b}{B \cdot \mu_b \cdot (1 - P^{h1}(block))}\right\}. \tag{D5}$$

(e) Define $\bar{B}^2 = \rho_b^{h1} \cdot B$.

(f) Solve the steady-state probabilities using this new fixed source. We call these probabilities p_i^{h2}.

The procedure proposed involves twice solving a one-dimensional queuing model with a transition matrix of size $(\bar{B} + 1) \times (\bar{B} + 1)$ and can be easily implemented.

Figure D.1. Heuristic performance.



Notes. Base unit. $B = 40$, $WR = 4$, $\hat{\rho}_b = 0.75$, ALOS $= 4.5$, $\lambda_n = 0.4$.

The proposed heuristic allows the approximation of the key performance indicators needed to guide nurse staffing decisions. For instance, inpatient delay can be approximated by the probability of having all nurses busy, and from this modified model we can compute nurse utilization and even tail probabilities.

Although the heuristic considers only the workload for nurses, we can still use it to say something about the bed system. Bed utilization can be approximated by Equation (D5), and the degree to which it differs from the nominal bed utilization $\hat{\rho}_b$ provides important information on the actual bed congestion.

Figure D.1 presents the approximated nurse utilization and probability of inpatient delay from the heuristic together with the results for the infinite and finite models.

We can see that for both performance indicators, the heuristic does a very good job of approximating the two-dimensional model.

Because the heuristic is one-dimensional, it can be modified to incorporate a priority service discipline so that staffing levels can be determined on the basis of virtually immediate response to emergent patient needs.

References

Aiken, L. A., S. P. Clarke, D. M. Sloane, J. A. Sochalski, J. H. Siber. 2002. Hospital nurse staffing and patient mortality, nurse burnout, and job dissatisfaction. *J. Amer. Medical Assoc.* **288**(16) 1987–1993.

Akar, N., K. Sohraby. 1997. Finite and infinite qbd chains: A simple and unifying algorithmic approach. *Proc. INFOCOM '97.*

American Nursing Association, ANA. 1999. Principles for nurse staffing. Accessed June 18, 2007, http://www.nursingworld.org/readroom/stffprnc.htm.

California Department on Health Services. 2003. Final statements of reasons, on hospital nurse staff ratios and quality of care. Report on AB394.

Commission on Health Care Facilities in the 21st Century. 2006. Final report. Accessed July 25, 2007, http://www.nyhealthcarecommission.org/docs/final/commissionfinalreport.pdf.

de Véricourt, F., O. B. Jennings. 2006. Nurse-to-patient ratios in hospital staffing: A queuing perspective. ESMT Working Paper 08-005, Duke University, Durham, NC.

Dochterman, J., G. Bulechek. 2004. *Nursing Interventions Classification* (*NIC*), 4th ed. Mosby, St. Louis.

Green, L. V. 2002. How many hospital beds? *Inquiry* **39**(4) 400–412.

Green, L. V. 2006. Patient flow: Reducing delay in healthcare delivery. R. W. Hall, ed. *Queueing Analysis in Healthcare*. Springer, New York.

Green, L. V., P. J. Kolesar. 2004. Improving emergency responsiveness with management science. *Management Sci.* **50**(8) 1001–1014.

Green, L. V., V. Nguyen. 2001. Strategies for cutting hospital beds: The impact on patient service. *HSR: Health Services Res.* **36**(2) 421–442.

Green, L. V., J. Soares, J. F. Giglio, R. A. Green. 2006. Using queueing theory to increase the effectiveness of emergency department provider staffing. *Acad. Emergency Medicine* **13**(1) 61–68.

Gross, D., C. M. Harris. 1985. *Fundamentals of Queueing Theory*. John Wiley & Sons, New York.

Healthcare Cost and Utilization Project. 2008. HCUP facts and figures, 2006. Statistics on hospital based care in the United States. Healthcare Cost and Utilization Project (HCUP). Agency for Healthcare Research and Quality (US), Rockville, MD.

Health Policy Tracking Service. 2005. 2005 state health care priorities survey report. Netscan iPublishing Inc., Falls Church, VA.

Hershey, J. C., E. N. Weiss, M. A. Cohen. 1981. A stochastic service network model with application to hospital facilities. *Oper. Res.* **29**(1) 1–22.

Institute of Medicine of the National Academies, IOM. 2001. *Crossing the Quality Chasm: A New Health System for the 21st Century*. The National Academy Press, Washington, DC.

Institute of Medicine of the National Academies, IOM. 2004. *Keeping Patients Safe—Transforming the Work Environment of Nurses*. The National Academy Press, Washington, DC.

International Council of Nurses, ICN. 2006. The global nursing shortage: Priority areas for intervention. International Council of Nurses, Geneva.

Jaumard, B., F. Semet, T. Vovor. 1998. A generalized linear programming model for nurse scheduling. *Eur. J. Oper. Res.* **107**(1) 1–18.

Jinks, A. M., P. Hope. 2000. What do nurses do? An observational survey of the activities of nurses on acute surgical and rehabilitation wards. *J. Nursing Management* **8**(5) 273–279.

Kane, R. L., T. A. Shamilyan, C. Mueller, S. Duval, T. J. Wilt. 2007. The association of registered nurse staffing levels and patient outcomes: Systematic review and meta-analysis. *Medical Care* **45**(12) 1195–1204.

Kao, E. P. C. 1974. Modeling the movement of coronary patients within a hospital by semi-Markov processes. *Oper. Res.* **22**(4) 683–699.

Kazahaya, G. 2005. Harnessing technology to redesign labor cost management reports. *Healthcare Financial Management* **59**(4) 94–100.

Kwak, N. K., C. Lee. 1997. A linear goal programming model for human resource allocation in a health-care organization. *J. Medical Systems* **21**(3) 129–140.

Lang, T. A., M. Hodge, V. Olson, P. S. Romano, R. L. Kravitz. 2004. Nurse-patient ratios: A systematic review on the effects of nurse staffing on patient, nurse employee, and hospital outcomes. *J. Nursing Admin.* **34**(7–8) 326–337.

Lapierre, S. D., D. Goldsman, R. Cochran, J. DuBow. 1999. Bed allocation techniques based on census data. *Socio-Econom. Planning Sci.* **33**(1) 25–38.

Litvak, E., M. C. Long. 2000. Cost and quality under managed care: Irreconcilable differences? *Amer. J. Managed Care* **6**(3) 305–312.

Lundgren, S., K. Segesten. 2001. Nurses' use of time in a medical-surgical ward with all-RN staffing. *J. Nursing Management* **9**(1) 13–20.

Miller, H. E., W. P. Pierskalla, G. J. Rath. 1976. Nurse scheduling using mathematical programming. *Oper. Res.* **24**(5) 857–870.

Needleman, J., P. Buerhaus, S. Mattke, M. Stewart, K. Zelevinsky. 2002. Nurse-staffing levels and the quality of care in hospitals. *New England J. Medicine* **346**(22) 1715–1722.

Neuts, M. F. 1981. *Matrix-Geometric Solutions in Stochastic Models*. The Johns Hopkins University Press, Baltimore.

Seago, J. A. 2002. A comparison of two patient classification instruments in an acute care hospital. *J. Nursing Admin.* **32**(5) 243–249.

Society for Health Systems, SHS. 2005. Position statement on mandated nursing ratios. Accessed June 18, 2007, http://www.iienet2.org/uploadedFiles/SHS/Resource_Library/Details/positionPaper.pdf.

Volpatti, C., M. Leathley, K. R. Walley, P. M. Dodek. 2000. Time-weighted nursing demand is a better predictor than midnight census of nursing supply in an intensive care unit. *J. Critical Care* **15**(4) 147–150.

White, K. M. 2006. Policy spotlight: Staffing plans and ratios. *J. Nursing Management* **37**(4) 1195–1204.

Wright, P. D., K. M. Bretthauer, M. J. Côté. 2006. Reexamining the nurse scheduling problem: Staffing ratios and nursing shortage. *Decision Sci.* **37**(1) 39–70.

Young, J. P. 1965. Stabilization of inpatient bed occupancy through control of admissions. *Hospitals: J. Amer. Hospital Assoc.* **39**(19) 41–48.