3-2009

OR Practice—Efficient Short-Term Allocation and Reallocation of Patients to Floors of a Hospital During Demand Surges

Steven M. Thompson
University of Richmond, sthomps3@richmond.edu

Manuel Nunez

Robert Garfinkel

Matthew D. Dean

Recommended Citation
Thompson, Stephen, Manuel Nunez, Robert Garfinkel, and Matthew D. Dean. "OR Practice--Efficient Short-Term Allocation and Reallocation of Patients to Floors of a Hospital During Demand Surges." *Operations Research* 57, no. 2 (March/April 2009): 261-73. doi:10.1287/opre.1080.0584.

Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

OR Practice—Efficient Short-Term Allocation and Reallocation of Patients to Floors of a Hospital During Demand Surges

Steven Thompson, Manuel Nunez, Robert Garfinkel, Matthew D. Dean,

To cite this article:
Steven Thompson, Manuel Nunez, Robert Garfinkel, Matthew D. Dean, (2009) OR Practice—Efficient Short-Term Allocation and Reallocation of Patients to Floors of a Hospital During Demand Surges. Operations Research 57(2):261-273. http://dx.doi.org/10.1287/opre.1080.0584

OPERATIONS RESEARCH
Vol. 57, No. 2, March–April 2009, pp. 261–273
ISSN 0030-364X | EISSN 1526-5463 | 09 | 5702 | 0261

DOI 10.1287/opre.1080.0584

OR PRACTICE

Efficient Short-Term Allocation and Reallocation of Patients to Floors of a Hospital During Demand Surges

Steven Thompson
Robins School of Business, University of Richmond, Richmond, Virginia 23173, sthomps3@richmond.edu

Manuel Nunez, Robert Garfinkel
School of Business, University of Connecticut, Storrs, Connecticut 06269-1041
{mnunez@business.uconn.edu, rgarfinkel@business.uconn.edu}

Matthew D. Dean
College of Business, University of New Orleans, New Orleans, Louisiana 70148, mddean@uno.edu

Many hospitals face the problem of insufficient capacity to meet demand for inpatient beds, especially during demand surges. This results in quality degradation of patient care due to large delays from admission time to the hospital until arrival at a floor. In addition, there is loss of revenue because of the inability to provide service to potential patients. A solution to the problem is to proactively transfer patients between floors in anticipation of a demand surge. Optimal reallocation poses an extraordinarily complex problem that can be modeled as a finite-horizon Markov decision process. Based on the optimization model, a decision-support system has been developed and implemented at Windham Hospital in Willimantic, Connecticut. Projections from an initial trial period indicate very significant financial gains of about 1% of their total revenue, with no negative impact on any standard quality of care or staffing effectiveness indicators. In addition, the hospital showed a marked improvement in quality of care because of a resulting decrease of almost 50% in the average time that an admitted patient has to wait from admission until being transferred to a floor.

Subject classifications: hospitals; health care; dynamic programming; Markov decision processes; decision-support systems.
Area of review: OR Practice.
History: Received July 2005; revisions received May 2006, March 2007, August 2007; accepted December 2007. Published online in *Articles in Advance* January 5, 2009.

1. Introduction

Over the past 25 years, United States hospitals have been subjected to significant transformations of the operating landscape. The large-scale penetration of health maintenance organizations in the early 1980s, the Emergency Medical Treatment and Labor Act in 1986, and Medicare reform associated with the Balanced Budget Act of 1997, have limited the ability of hospitals to turn away patients that are unable to pay for services, and have placed limits on the amount hospitals are able to collect for the services they provide. These changes have forced hospitals to improve operating efficiency (Krein and Casey 1998). As a result, hospitals have aggressively reduced inefficiencies by cutting staff, managing length of stay, and finding innovative ways to reduce incidental costs.

One side effect of these cost-containment efforts has been a reduction in the number of inpatient beds for which hospitals maintain staff and to which patients can be admitted. The United States lost a total of 100,000 hospital beds, including 7,800 intensive care beds between 1990 and 1999 (Kellerman 2001). The loss of inpatient beds has left many hospitals with minimal surge capacity to handle spikes in demand. The result is that beds are often in short supply, and the problem is expected to get worse (Dodge 2001, Solberg et al. 2003, Wilson 2001).

Emergency department (ED) overcrowding is another undesirable result of insufficient floor capacity. In a March 2003 report (GAO 2003), the United States General Accounting Office found that the factor most commonly associated with crowding was the inability to move emergency patients to inpatient beds, once a decision had been made to admit them as hospital patients, rather than to treat and release them. The impact on ED management is clear. As the number of "boarding" patients who are waiting for beds on their assigned floors increases, the resources available to treat other ED patients are reduced, and eventually the department ceases to function effectively. Often hospitals

will then attempt to divert ambulances to other facilities until some of the boarded patients are transferred out of the ED, yielding a very significant negative outcome for patients and the hospital. For patients, the effect of diversion is a longer ambulance ride, which prolongs the delay in receiving medical treatment. For the hospital, each ambulance that is diverted represents lost revenue. In addition, when the ED becomes overcrowded, patients with minor complaints tend to leave the hospital before being seen by a physician. In these cases, the hospital has missed the opportunity to provide a service for which payment would otherwise have been rendered.

An example of a hospital facing the problems delineated above is Windham Community Memorial Hospital (WCMH), located in Willimantic, Connecticut, and servicing a surrounding community of approximately 100,000 people. Of interest here was that in 2004, when this study was initiated, WCMH would often begin to experience capacity-related problems that resulted in patients being boarded in the ED for the short term, e.g., 4–6 hours, well before all inpatient beds were utilized. As a result, delays were incurred when a patient had to be transferred from a given floor to make room for a new patient that could only be assigned to that floor. These "last-minute" transfers are often done under duress during critical "crunch" periods and are generally undesirable from the standpoints of patient flow and quality of care.

The "bed manager" (decision maker) determines the initial assignments of patients to floors. Another task is to determine when, and if, it is necessary to transfer patients from one floor to another, even in noncrunch periods. While it is generally undesirable to transfer patients after admission, it was determined that there was a critical need at WCMH for an optimality-based decision-support system (DSS) for the bed manager that would allow for preemptive (prior to the occurrence of a demand surge) transfers of patients between floors, and for the assignment of patients to floors based partially on capacity considerations. For the former, in-house patients are transferred for the purpose of capacity reallocation (proactive transfer), as opposed to as a "last-minute" immediate response to make room for newly admitted patients (reactive transfer). For the latter, even when beds are available on the "ideal" (based strictly on floor specialization) floor, newly admitted patients may be assigned to feasible "alternate" floors.

We modeled the problem of finding an optimal capacity utilization strategy based on patient allocation as a multidimensional, discrete-time, finite-horizon Markov decision process. The model has been integrated into a DSS that has been implemented and, based on an initial trial period, is projected to result in very significant financial gains of about \$600,000 per year, or 1% of total revenue. No negative impact resulted on any standard quality of care or staffing effectiveness indicators. In addition, there was a marked improvement in quality of care because of a resulting decrease of almost 50% in the average time that an admitted patient has to wait from admission until being transferred to a floor. Based on this success, WCMH decided to create an "operations manager" position, to be filled by an individual who will work with the system and will also identify other opportunities to improve patient flow and hospital efficiency.

1.1. Overview

Many hospitals maintain a myopic strategy of assigning admitted patients to their "ideal" floors, based on diagnosis, as long as there are available beds on those floors. Although this strategy works well in many cases, when capacity is limited it can result in patient flow bottlenecks that have a number of negative quality of care and financial implications.

In hospital settings, the general overall problem is quite complex. Departures must be considered as well as arrivals, and the number of floors and patient categories could be large. Although techniques have been developed for solving stochastic sequential decision problems, the basic problem presented here is challenging for a few reasons:

1. There generally are a large number of different patient categories and floors, depending on the size of the hospital and the range of diagnoses it can treat.
2. The number of possible actions/policies to consider is very large.
3. Random events such as patient arrivals and departures depend on patient category, type of floor, and the current state of the hospital, and are often nonhomogenous with respect to time of day, day of week, and season of the year.
4. The amount of time available to reach a decision is fairly short, i.e., typically less than five minutes.

In this paper, we develop and implement a solution methodology that addresses those issues. The remainder of this paper is organized as follows. First, we define the main problem under consideration in §2. In §3, we model the problem as a finite-horizon Markov decision process (MDP) and study the computational complexity of finding an optimal solution to the MDP and conclude that traditional approaches are not practical for this problem. In §3.7, we present an original approximation methodology that relies on event aggregation to find an approximate decision rule solution to the MDP. Section 4 deals with several issues related to the implementation of our approximation methodology in general, and we describe the "rolling-horizon" algorithm used to implement a DSS based on the MDP model. Section 5 describes our computational experience, and §6 presents the details of the implementation of the DSS at WCMH, including anecdotal experience, managerial insights, and an analysis of the impact on the hospital.

1.2. Literature Review

The related problems of hospital and ED crowding have been addressed from clinical and managerial viewpoints

by health care researchers. Medical researchers have found that diverting ambulances significantly lessens the availability of ambulances for patients in need of medical treatment (Eckstein and Chan 2004). This diversion of ambulances has been found to be primarily due to holding admitted patients in the ED (Schull et al. 2003). The quality of care and financial impact of holding patients in the ED have been found to be significant for patients with chest pain (Bayley et al. 2005) and those in need of thrombolytic therapy (Schull et al. 2004).

Efforts to alleviate the ED crowding problem have largely focused on reducing the amount of time required to provide emergency services. Examples include bedside registration (Parker 2004) and applications of simulation and queueing theory to identify and eliminate bottlenecks within the ED (Litvak et al. 2001). Other strategies designed to move admitted patients out of the ED as quickly as possible include faxing patient condition reports to the floors (Caissie 2004) and admitting patients to hallways (Derlet and Richards 2000). These efforts, while achieving some success in ensuring that patients in need of emergency medical treatment are able to receive it, have had little effect on the waiting time before moving admitted patients from the ED to the floors.

The OR/MS literature contains a long history of research devoted to solving assignment/allocation problems in the health care and emergency services settings. For a summary of applications of OR/MS to problems found in the emergency services setting, see Green and Kolesar (2004). In addition, a history of the applications of OR techniques to problems in the health care industry can be found in Flagle (2002). The application of OR techniques to health care problems also received considerable attention in Brandeau et al. (2004). Nevertheless, none of these models can be directly used in our situation due to characteristics of the problem that are formalized later in the paper.

There has also been a substantial amount of work addressing the fundamental problem of assigning different "jobs" to different "stations" when the number of jobs (patients in our case) and/or the amount of resources (beds) available to complete each job is unknown. In general, when the randomness arises from either the number of jobs or the amount of resources available, but not both, the problem is typically modeled as a stochastic generalized assignment problem (SGAP), e.g., see Albareda and Fernandez (2000) and Mine et al. (1983). To the best of our knowledge, there are no known exact results or algorithms for the SGAP.

When both the number of jobs and the amount of resources available are random, queueing theory is typically used. Queueing theory has been extensively researched and applied in health care scenarios such as clinics (Cox et al. 1985), hospital appointment systems (Jackson et al. 1964), and operating room staffing (Tucker et al. 1999). In these cases, the stochastic processes and the problem structures were amenable to established queueing models. Specifically, the arrival of all patients was assumed to follow a Poisson process, all patients were treated as identical in terms of service time, and no distinctions were made in terms of the clinical requirements of the patients. Recent work on the scheduling of emergent and elective surgeries (Gerchak et al. 1996) and emergent and elective radiology cases (Green et al. 2006) are related to the problem we address in that multiple classes of demand are considered. Despite some similarities, our problem incorporates many other characteristics that render inappropriate the modeling assumptions made in those references. It is pointed out by Green and Nguyen (2001) that to understand the impact on patient service, more sophisticated methodologies are needed to support decisions that involve bed capacity and organization. The goal of this paper is to develop a method to improve the efficiency of capacity allocation.

2. Problem Definition

The number of floors in a hospital is very small compared to the possible number of characteristics of the patients that it serves. There exists a standard grouping terminology called "diagnostic related group" (DRG) that is based on diagnosis upon admission. However, there are over 500 possible DRGs, so that using them as the basis for optimization models results in inordinately large problems. Instead, we partition the set of DRGs into "categories" according to three parameters: (1) the set of floors that can treat the patients; (2) the "ideal" (to be expanded on later in the paper) floor for the DRG; and (3) the expected lengths of stay of the patients. Then, within a category DRGs must be identical in the first two parameters, and very similar in the third. In the WCMH implementation, we used 12 patient categories because that was the minimum number that satisfied the above criteria. All patients are considered to have arrived to the system as soon as they are admitted to the hospital. Once admitted, patients wait to be assigned and then transported to floors.

In our model, the system state at any point in time is represented, for each patient category, by the number waiting and the number being cared for on each floor. The patient interarrival times to the system, and the number of patient departures during a given time interval from the hospital, are random and dependent on the patient category, floor assigned, and the time dimension. The parameters of the arrival and departure distributions are based on actual historical data at the hospital rather than following any predetermined distributions. Patient interarrival times are independent of the number of departures.

The bed manager observes the state of the hospital after a fixed-length time interval (period), records the number of patient arrivals and departures during the period, and makes assignment and transfer decisions. Because of capacity constraints, some admitted patients may not be assignable at that time, and would have to wait until the next period before assignment to any floor. The problem is to determine the best patient assignment and transfer decision in

terms of immediate assignment rewards and transfer costs incurred from making the decision, and in terms of long term future expected rewards and costs.

2.1. Staff and Quality of Care

In any hospital, one measure of quality of care is based on predefined maximum ratios of patients-to-staff and patients-to-licensed-staff. Staff arrivals and departures exhibit much less variability than those associated with patients and are assumed known in advance. Staff members (e.g., registered nurse, nurse's aide, etc.) can always be deterministically assigned so as to satisfy these ratios, which is possible because the number and type of staff of the hospital are already set up to correspond to the bed capacity on each of the floors. In the short term, it is always true that safe staffing can be met, even if a manager or supervisor must step in. Furthermore, as a general rule the hiring strategy of a hospital is to obtain enough staff to run a floor at maximum capacity if needed. In the case of a long-term inability to maintain full staffing levels (injuries, vacations, etc.), the ratios can be satisfied by adjusting the capacities of the floors. For example, if the bed manager knew that for several weeks she would not be able to staff more than four nurses on a floor where the ratio is one nurse for every two patients, she could temporarily reduce the capacity of the floor to eight beds from its present level during that period. Hence, as long as patients are on the floors, proper care will be provided and we do not explicitly consider staff allocation as part of our model.

3. Model Formulation and Notation

We model the allocation problem as a finite-horizon, discrete-time, stationary MDP. Time is split into fixed-length intervals (periods). The state of the process is observed at the beginning of a period, and a decision is chosen from a finite set of possible decisions. An immediate cost is incurred depending on the state and decision, which determine the transition probabilities for the next state. That state is realized at the end of the period, the process state is updated, and the procedure repeats.

3.1. Notation Summary

We give the essential notation for various sections of the paper:

Parameters

- c_j: maximum capacity of floor j;
- F: index set of floors;
- Q: index set of patient categories;
- F_i: index set of feasible floors for patients of category i ($F_i \subset F$);
- a_{ij}: reward from assigning a category i patient to floor j;
- b_{ijk}: cost of a category i patient transfer from floor j to floor k.

State Space

- m: number of time periods in the process;
- M: index set of the time periods ($M = \{1, \ldots, m\}$);
- t: time period index ($t \in M$);
- x_{ij}: number of category i patients on floor j;
- x_{i0}: number of category i admitted patients not yet on any floor;
- X: matrix of x_{ij} values;
- S: state of the process ($S := [X, t]$).

Random Variables

- g_{it}: number of category i patient arrivals during period t;
- d_{ijt}: number of category i patient departures from floor j during period t;
- G: matrix of g_{it} variables;
- D: array of d_{ijt} variables.

Decision Variables

- y_{i0j}: number of category i nonassigned patients to be assigned to floor j;
- y_{ijk}: number of category i patients to be transferred from floor j to floor k;
- Y: array of decision variables;
- $\mathcal{Y}(S)$: set of feasible decisions for a given state S;
- $C(Y)$: total cost associated with decision Y.

Objective Values

- $V_n(S)$: minimum expected n-stage cost in state S;
- $\tilde{V}_n(S)$: approximation based on simulation;
- $\hat{V}_n(S)$: approximation based on simulation and simplified state space.

3.2. Process State

We consider a decision process with a finite time horizon divided into m time periods of constant length. The transition probabilities of the embedded Markov chain might vary from time period to time period, but they cycle with respect to m. That is, the transition probability from state A to state B is the same during periods $t, t + m, t + 2m$, etc. Hence, we have a stationary MDP where the period index t is part of the state definition. To capture the intraday, intraweek, and seasonal fluctuations in patient arrival and departure rates, we must allow for this m-period time frame to be very large, perhaps spanning an entire year. The state of the process then depends on the number of patients on the floors, the number of patients admitted but not yet on any floor, and the time. Thus, the state of the process is $S := [X, t]$.

3.3. Constraints on Decision Variables

As indicated earlier, there will generally be restrictions on the floors to which different categories of patients can be assigned. Among feasible floors it is generally more desirable to assign a patient of a given category to one floor than another. The preference ranking of multiple feasible floors for a given patient category is achieved by setting assignment costs, a topic that is discussed in §3.4. All floors other than the unique lowest cost ("ideal") floor are called "alternate."

For a state $S = [X, t]$, decision variables in Y must satisfy:

- *Patients Waiting*:

$$\sum_{j\in F} y_{i0j} \leqslant x_{i0}, \quad i \in Q. \tag{1}$$

- *Floor Capacity*:

$$\sum_{i\in Q}\left(x_{ik} + y_{i0k} + \sum_{j\in F} y_{ijk} - \sum_{j\in F} y_{ikj}\right) \leqslant c_k, \quad k \in F. \tag{2}$$

- *Nonnegativity*:

$$y_{i0k} \geqslant 0, \quad y_{ijk} \geqslant 0, \quad i \in Q,\ j, k \in F. \tag{3}$$

- *Patient Restrictions*:

$$y_{i0k} = 0, \quad y_{ijk} = 0, \quad i \in Q,\ j \in F,\ k \in F\backslash F_i. \tag{4}$$

The set of decision rules, formalized in §4.3, satisfying (1)–(4) for a state S, are called "feasible," and are denoted by $\mathcal{Y}(S)$.

3.4. Stage Cost

Once a feasible decision Y has been made at the beginning of a period, there are a number of resulting expected costs and rewards. First, there is an expected positive reward a_{ij} to the hospital of assigning a category i patient to floor j. For this work, a_{ij} is the expected financial gain from treating a patient of category i scaled to reflect the desirability of floor j. The nonnegative parameter b_{ijk} represents the intrinsic undesirability of the transfer of a category i patient from floor j to floor k. Then, the stage cost associated with Y is

$$C(Y) := \sum_{i\in Q}\sum_{k\in F}\left(-a_{ik}y_{i0k} + \sum_{j\in F} b_{ijk}y_{ijk}\right). \tag{5}$$

3.5. Transition Probabilities

For every patient category, there are two possible random events—new arrivals and departures from floors. Both event types depend on the time period. It is natural to assume that entries of $G := [g_{it}]$ are statistically independent of each other. Similarly, the entries of $D := [d_{ijt}]$ are assumed statistically independent of each other and of the entries of G, but are clearly dependent on the state matrix X.

Let $S := [X, t]$ be the current state, $Y \in \mathcal{Y}(S)$ the chosen decision array, and $\hat{S} := [\hat{X}, \hat{t}]$ the state after arrivals and departures occur. State $\hat{S}$ is updated as follows:

$$\hat{x}_{i0} = x_{i0} + g_{it} - \sum_{k\in F} y_{i0k}, \quad i \in Q,\ t \in M, \tag{6}$$

$$\hat{x}_{ij} = x_{ij} - d_{ijt} + y_{i0j} + \sum_{k\in F}(y_{ikj} - y_{ijk}), \quad i \in Q,\ j \in F,\ t \in M, \tag{7}$$

$$\hat{t} = \begin{cases} 1 & \text{if } t = m, \\ t+1 & \text{if } t < m. \end{cases} \tag{8}$$

The transition probability from S to $\hat{S}$, given decision Y, is

$$P_{S\hat{S}}(Y) = \prod_{i\in Q}\mathcal{P}\left\{g_{it} = \hat{x}_{i0} - x_{i0} + \sum_{k\in F} y_{i0k}\right\} \cdot \prod_{i\in Q}\prod_{j\in F}\mathcal{P}\left\{d_{ijt} = -\hat{x}_{ij} + x_{ij} + y_{i0j} + \sum_{k\in F}(y_{ikj} - y_{ijk})\right\}. \tag{9}$$

Here the values of g_{it} and d_{ijt} in (9) are the possible numbers of arrivals and departures that respectively must have occurred to permit the transition from the state S to the state $\hat{S}$, given that the decision Y has been made.

3.6. Objective Function

Our objective is to find a decision rule that minimizes the sum of the immediate cost and the subsequent expected costs that result from the future evolution of the process. Let $V_n(S)$ be the minimum expected cost for an n-stage problem (evolving for n time periods) that starts in state S. We have

$$V_1(S) = \min_{Y\in\mathcal{Y}(S)} C(Y), \tag{10}$$

$$V_n(S) = \min_{Y\in\mathcal{Y}(S)}\left\{C(Y) + \sum_{\hat{S}} P_{S\hat{S}}(Y)V_{n-1}(\hat{S})\right\}, \quad n > 1. \tag{11}$$

With respect to the complexity of solving $V_n(S)$ for a given state $S = [X, t]$, note that for $n = 1$ the computation of $V_1(S)$ reduces to solving a single-commodity minimum cost flow problem without directed cycles of negative cost and can be readily solved in polynomial time on the sizes of Q and F as a linear program subject to constraints (1)–(4). On the other hand, the complexity of computing $V_n(S)$ for $n > 1$ depends on the size of the decision space $\mathcal{Y}(S)$ and the number of states $\hat{S}$ in (11) which in some cases can be infinite or finite but with exponential growth in the size of Q and the number of stages. Hence, as n grows, the exact computation of $V_n(S)$ becomes impractical for traditional methods such as dynamic programming. In the online supplement, we elaborate on the complexity issues. An electronic companion to this paper is available as part of the online version that can be found at http://or.journal.informs.org/.

3.7. Approximation Methodology

To approximate $V_n(S)$, we use random sampling to estimate the expectation term

$$\sum_{\hat{S}} P_{S\hat{S}}(Y)V_{n-1}(\hat{S}).$$

Concretely, let $(G_1, D_1), \ldots, (G_\nu, D_\nu)$ be a sample of i.i.d. random variables for a given state S. Then, we approximate $V_n(S)$ using $\widetilde{V}_n(S)$, where

$$\widetilde{V}_1(S) := \min_{Y\in\mathcal{Y}(S)} C(Y) = V_1(S), \tag{12}$$

$$\widetilde{V}_n(S) := \min_{Y\in\mathcal{Y}(S)}\left\{C(Y) + \frac{1}{\nu}\sum_{i=1}^{\nu}\widetilde{V}_{n-1}(\hat{S}_i)\right\}, \quad n > 1, \tag{13}$$

and $\hat{S}_1, \ldots, \hat{S}_\nu$ are the corresponding updated states.

Before we study the convergence of $\widetilde{V}_n(S)$, we introduce some additional notation. Consider the following definitions:

$$A_i := \max_{j\in F}\{a_{ij}\}, \quad i \in Q, \tag{14}$$

$$\hat{g}_{it} := E[g_{it}], \quad i \in Q,\ t \in M, \tag{15}$$

$$\mu_{it} := E[\max\{g_{it}, \hat{g}_{it}\}], \quad i \in Q,\ t \in M, \tag{16}$$

$$\sigma_{it}^2 := \mathrm{Var}[\max\{g_{it}, \hat{g}_{it}\}], \quad i \in Q,\ t \in M. \tag{17}$$

Let

$$s(t) := 1 + (t-1) \bmod m \quad \text{for all } t \in M$$

denote the successor function modulo m. Let $s^0(t) := t$, $t \in M$, denote the identity function on M. For $\tau > 0$, we introduce the notation

$$s^\tau(t) = s(s^{\tau-1}(t)).$$

Using this notation, we state the following bounds on $V_n(S)$. (Proofs are found in the online supplement.)

Proposition 1. *The following bounds hold for all states* $S = [X, t]$ *and* $n \geqslant 1$:

$$-\sum_{i\in Q} A_i x_{i0} - \sum_{i\in Q}\sum_{\tau=0}^{n-2} A_i \hat{g}_{is^\tau(t)} \leqslant V_n(S) \leqslant 0.$$

A similar argument to the proof of Proposition 1 can be used to show analogous bounds for $\widetilde{V}_n(S)$ in the following proposition.

Proposition 2. *The following bounds hold for all states* $S = [X, t]$ *and* $n \geqslant 1$:

$$-\sum_{i\in Q} A_i x_{i0} - \sum_{i\in Q}\sum_{\tau=0}^{n-2}\sum_{l=1}^{\nu} \frac{1}{\nu^{\tau+1}} A_i g^l_{is^\tau(t)} \leqslant \widetilde{V}_n(S) \leqslant 0,$$

where $g^l_{is^\tau(t)}$, $l = 1, \ldots, \nu$, *are the realizations of the* ν *random variables in each of the samples corresponding to each period* $t, s(t), \ldots, s^{n-2}(t)$, *and each category* $i \in Q$.

Using Proposition 1, we establish the following result concerning the convergence of $\widetilde{V}_n(S)$ to $V_n(S)$ as the sample size ν increases.

Proposition 3. *Let* $\epsilon > 0$, $0 < \alpha < 1$, *and* $S = [X, t]$ *be a given state. Then,*

$$\mathcal{P}\{|V_n(S) - \widetilde{V}_n(S)| < \epsilon\} \geqslant 1 - \alpha \tag{18}$$

for all

$$\nu \geqslant \frac{1}{\alpha\epsilon^2}\left(\left(\sum_{i\in Q} A_i x_{i0} + \sum_{i\in Q}\sum_{\tau=1}^{n-1} A_i \hat{g}_{is^\tau(t)} + \sum_{i\in Q} A_i \mu_{it}\right)^2 + \sum_{i\in Q} A_i^2 \sigma_{it}^2\right). \tag{19}$$

One important implication of Proposition 3 is that the sample size needed to achieve good approximations is independent of the decisions that are made. The result also implies that $\widetilde{V}_n(S)$ converges in probability to $V_n(S)$ as $\nu \to \infty$ and so, $\widetilde{V}_n(S)$ is a consistent estimator of $V_n(S)$. On the other hand, the term $1/\alpha$ in the bound (19) that originates from using Chebyshev's inequality in the proof of Proposition 3 yields very large values of the sample size ν to obtain reasonable approximations. The bound can be improved by using the central limit theorem because for large ν we can easily modify the proof of the proposition to obtain an approximate bound

$$\nu \geqslant \frac{z^2_{\alpha/2}}{\epsilon^2}\left(\left(\sum_{i\in Q} A_i x_{i0} + \sum_{i\in Q}\sum_{\tau=1}^{n-1} A_i \hat{g}_{is^\tau(t)} + \sum_{i\in Q} A_i \mu_{it}\right)^2 + \sum_{i\in Q} A_i^2 \sigma_{it}^2\right),$$

where $z_{\alpha/2}$ satisfies the equation $\Phi(z) = 1 - \alpha/2$ and Φ is the distribution function of a standard normal random variable.

4. DSS Specification

We provide details about how the solution algorithm described above was incorporated into a DSS. In particular, we indicate how to simplify the decision space under the conditions of a typical hospital, how to implement a "rolling-horizon" algorithm to provide timely recommendations to the bed manager, and provide details on the computation of probabilities for the random variables considered in the hospital model.

4.1. Probability Computation

The number g_{it} of category i patient arrivals for a given state $S = [X, t]$ is assumed to have an aggregated Poisson distribution with parameter λ_{it}. Note that λ_{it} is the average number of category i patients who arrive during the time interval between periods t and $s(t)$, where as before $s(t) = 1 + (t-1) \bmod m$. To compute λ_{it}, we partition the interval between periods t and $s(t)$ into η subintervals. We assume that the arrivals during each subinterval follow a Poisson distribution with rate $\lambda_{it}^{(j)}$ for $j = 1, \ldots, \eta$. Hence,

$$\lambda_{it} = \sum_{k=1}^{\eta} \lambda_{it}^{(j)}.$$

The use of such an aggregated Poisson distribution for patient arrivals fits the observed behavior at WCMH very well.

For departure variables d_{ijt} given state S, we assume that each variable follows a binomial distribution with parameter pair (p_{ijt}, x_{ij}). The parameter p_{ijt} denotes the probability of a category i patient departure from floor j during period t. To compute the parameter p_{ijt}, we use a method similar to that used for the arrivals. We partition the interval between periods t and $s(t)$ into η subintervals. We denote

by $p_{ijt}^{(k)}$ the probability of one departure during the kth subinterval, and then we set

$$p_{ijt} := 1 - \prod_{k=1}^{\eta} (1 - p_{ijt}^{(k)}).$$

All of the parameters are readily estimated from historical data and/or forecasting methods. Moreover, note that in this case

$$\hat{g}_{it} = \lambda_{it},$$
$$\mu_{it} \leqslant 2\lambda_{it},$$
$$\sigma_{it}^2 \leqslant \lambda_{it}(3\lambda_{it} + 1),$$

for all $i \in Q$ and $t \in M$. Hence, based on bound (19) from Proposition 3, we obtain the following corollary.

COROLLARY 4. *Let $\epsilon > 0$, $0 < \alpha < 1$, and $S = [X, t]$ be a given state. Then,*

$$\mathcal{P}\{|V_n(S) - \widetilde{V}_n(S)| < \epsilon\} \geqslant 1 - \alpha \tag{20}$$

for all

$$\nu \geqslant \frac{1}{\alpha\epsilon^2}\left(\left(\sum_{i\in Q} A_i x_{i0} + \sum_{i\in Q}\sum_{\tau=1}^{n-1} A_i \lambda_{is^\tau(t)} + 2\sum_{i\in Q} A_i \lambda_{it}\right)^2 + \sum_{i\in Q} A_i^2 \lambda_{it}(3\lambda_{it} + 1)\right). \tag{21}$$

Bound (21) can be easily computed from the data and problem parameters corresponding to the WCMH case. As remarked before, bound (21) can be improved to

$$\nu \geqslant \frac{z_{\alpha/2}^2}{\epsilon^2}\left(\left(\sum_{iQ} A_i x_{i0} + \sum_{i\in Q}\sum_{\tau=1}^{n-1} A_i \lambda_{is^\tau(t)} + 2\sum_{i\in Q} A_i \lambda_{it}\right)^2 + \sum_{i\in Q} A_i^2 \lambda_{it}(3\lambda_{it} + 1)\right)$$

for large ν. For example, using this bound with the data provided by WCMH, we can show that the relative error between $V_2(S)$ and $\widetilde{V}_2(S)$ is no more than 34% with approximately 99% probability for a sample size of $\nu = 500$. Also, for the same sample size, the relative error between $V_2(S)$ and $\widetilde{V}_2(S)$ is no more than 17% with approximately 80% probability.

4.2. Real-Time DSS

The real-time DSS operates by solving the approximation $\widehat{V}_2(S)$ to the dynamic programming formulation (10)–(11) on a "rolling-horizon" basis, according to the following algorithm. Given state S and sample size $\nu \geqslant 1$, we generate a random sample $(G_1, D_1), \ldots, (G_\nu, D_\nu)$. Then, we choose any Y^* such that

$$Y^* \in \arg\min_{Y\in\mathcal{Y}(S)}\left\{C(Y) + \frac{1}{\nu}\sum_{i=1}^{\nu} \widetilde{V}_1(\hat{S}_i)\right\}.$$

The decision is applied in real time to the current state S. Next, the decision maker waits for an interval of time during which realizations of G and D are observed. At the end of the interval, the state is further updated by adding the changes that occurred corresponding to (G, D) to yield a new state $\hat{S}$. Then, the algorithm is reapplied to the updated state $\hat{S}$.

In our computations, the interval between periods t and $s(t)$ is eight hours long. In computing the distributions for G and D, we partition the eight-hour interval into $\eta = 32$ subintervals of 15 minutes each. Using our real-time algorithm every 15 minutes, a decision Y is computed and applied. Time intervals of 15 minutes were chosen to strike a balance between providing the DSS enough time to arrive at a "good" recommendation, while avoiding long periods of computational time that would introduce another bottleneck to the patient flow process. In addition, eight hours is a sufficiently long planning horizon to allow for a reasonable probability of random events occurring, while short enough to keep the process under control.

4.3. Further Simplification Using Decision Rules

In using $\widetilde{V}_n(S)$ to approximate $V_n(S)$, one advantage is that the term

$$\sum_{j=1}^{\nu} \widetilde{V}_{n-1}(\hat{S}_j)$$

is easier to compute when the number of scenarios $\hat{S}_j$ is not very large. However, the computational complexity of the algorithm will depend on the size of the decision space $\mathcal{Y}(S)$. Using (1), there are at least

$$\prod_{i\in Q}\binom{x_{i0} + |F\backslash F_i|}{|F\backslash F_i|}$$

feasible solutions to consider, which could be large even for small x_{i0}.

In a typical hospital like WCMH, the following can be assumed to be true:

1. If floor $j^* \in F_i$ is ideal for category i patients, then a_{ij^*} is significantly larger than a_{lj^*} for any category l patient for which floor j^* is alternate.

2. Under the same conditions, a_{ij^*} is also significantly larger than the cost b_{lj^*k} of transferring out a category l patient for which floor j^* is alternate.

Under those assumptions, for each $i \in Q$ there exists a unique index $j_i \in F_i$ such that

$$A_i = a_{ij_i},$$

where A_i is as defined in (14). Floor j_i is the ideal floor for category i patients. In addition, we can partition the set of categories into subsets Q_j for all $j \in F$. Each subset Q_j corresponds to the categories for which floor j is ideal; that is,

$$Q_j := \{i \in Q\colon j_i = j\}. \tag{22}$$

Using this partition, we are able to prove the following statement.

PROPOSITION 5. *Let α be such that $0 < \alpha < 1$ and*

$$\theta_j := (1-(1-\alpha)^{1/(\nu|F|)})^{-1} \max_{t\in M}\left\{\sum_{i\in Q_j} \hat{g}_{it}\right\}$$

for all $j \in F$. If state $S = [X, t]$ is such that

$$c_j - \sum_{i\in Q} x_{ij} - \sum_{i\in Q_j} x_{i0} \geqslant (n-1)\theta_j \tag{23}$$

for all $j \in F$, then with probability at least

$$(1-\alpha)^{(\nu^{n-1}-1)/(\nu-1)}$$

the approximation $\widetilde{V}_n(S)$ achieves its lower bound from Proposition 2 and the greedy solution Y given by $y_{i0j_i} := x_{i0}$, $y_{i0j} = 0$ for $j \neq j_i$, and $y_{ijk} = 0$ is optimal for $\widetilde{V}_n(S)$.

In other words, if all the floors have enough excess capacity for a given state S (condition (23)), then it is optimal to assign all patients currently waiting to their ideal floors. The factor $(1-(1-\alpha)^{1/(\nu|F|)})^{-1}$ in the definition of θ_j grows fast as α approaches zero. To improve the result, note that in the DSS the variable $\sum_{i\in Q_j} \hat{g}_{it}$ has Poisson distributions with parameter $\sum_{i\in Q_j} \lambda_{it}$. When the Poisson parameter is large (like in the WCMH case), the central limit theorem can be used to approximate the Poisson distribution with a Normal distribution with mean and variance equal to the Poisson parameter. Hence, we can obtain a similar result to Proposition 5 by using

$$\theta_j = \max_{t\in M}\left\{\sum_{i\in Q_j} \hat{g}_{it} + z_{\beta/2}\left(\sum_{i\in Q_j} \hat{g}_{it}\right)^{1/2}\right\}, \tag{24}$$

where $\beta = 1-(1-\alpha)^{1/(\nu|F|)}$. For example, for the case $n = 2$, $1-\alpha = 0.99$, and using a sample size of $\nu = 500$ we have $z_{\beta/2} \approx 4$. Hence, if the residual capacity of floor j is greater than the corresponding θ_j for all $j \in F$, the greedy solution is optimal for $\widetilde{V}_2(S)$ with probability at least 0.99.

Note that the numbers θ_j do not depend on the state S. For a given floor, we use the term "primary" patients to refer to those patients for whom the floor is ideal, while all other patients for whom the floor is feasible are labeled "secondary." Hospital regulations establish that a number of beds on each floor should be reserved for primary patients (so-called "crash beds").

Therefore, the decision space can be reduced to only those decisions in $\mathcal{Y}(S)$ that obey the above rules. More precisely, we define a decision rule as a function f assigning one and only one decision $f(R) := Y \in \mathcal{Y}(S)$ for a given rule-parameter R, where $R := [r_{jl}]$ is an $|F| \times 3$ matrix such that $r_{j1} \leqslant r_{j2} \leqslant r_{j3}$ for all $j \in F$. Ideally, we would choose the r_{j3} values to be the θ_j values from Proposition 5 or from (24), but in practice those values are too large or imprecise. Hence, we let the r_{j3} vary as part of a decision rule. The process of assigning a decision Y for a given matrix R is done in three consecutive steps: proactive transfer; primary assignment; and secondary assignment.

In proactive transfer, we transfer secondary patients from a floor j in which there are fewer than r_{j2} beds available to another floor in which there are more than r_{j3} beds available. When there is more than one transfer floor option, the floor with the lowest transferring cost is chosen. The objective is to increase the capacity in each floor for potential or currently waiting primary patients. In primary assignment, patients currently waiting are assigned to their corresponding ideal floors as long as floor capacities permit. If there are floor conflicts among primary patients, those with the highest assignment reward have priority. Finally, in secondary assignment, all remaining waiting patients are assigned to floors with at least r_{j1} beds of remaining capacity, where we again resolve floor-conflict assignments according to the highest assignment reward. The details of the assignment algorithms for each of the three steps are given in the online supplement.

When $r_{j2} = 0$, no proactive transfers occur out of floor j. Further, when $r_{j1} = 0$, no bed reservation occurs for primary patients on floor j. The current decision rule used at WCMH corresponds to $R = 0$, that is, it includes neither proactive transfers nor bed reservations.

In practice, we only consider a finite set $\mathcal{R}(S)$ of rule-parameter matrices and the following approximation:

$$\widehat{V}_2(S) := \min_{R\in\mathcal{R}(S)}\left\{C(f(R)) + \frac{1}{\nu}\sum_{j=1}^{\nu} V_1(\hat{S}_j)\right\}.$$

It is not difficult to show that Proposition 3 still holds if we replace $\widetilde{V}$ by $\widehat{V}$.

Accordingly, we obtain the following revised real-time algorithm.

ALGORITHM 1. Given state S and sample size ν, execute the following sequence:

```
min ← ∞,
Generate a random sample (G_1, D_1), ..., (G_ν, D_ν),
for all R ∈ R(S) do
    s ← 0,
    for j = 1, ..., ν do
        Update S using f(R) and (G_j, D_j) to yield Ŝ_j,
        s ← s + V_1(Ŝ_j),
    end for,
    if C(f(R)) + s/ν < min then
        min ← C(f(R)) + s/ν,
        R̂ ← R,
    end if,
end for,
return R̂.
```

4.4. Approximation Algorithm Performance

To compare the performance of the solutions obtained from our approximation algorithm to the optimal solutions, we designed a simulation in a simplified hospital setting.

Because obtaining optimal solutions for problem instances of a realistic size is very time-consuming, we were restricted to exploring relatively small problem instances. In this case, we consider a hospital defined by the following parameters: $F = \{1, 2, 3\}$; $Q = \{1, 2, 3\}$; $F_1 = \{1, 2, 3\}$; $F_2 = \{1, 2, 3\}$; $F_3 = \{3\}$; $c_j = 3$, $j \in F$; $b_{ijk} = 200$, $i \in Q$, $k \in F_i$. We set the assignment rewards to

$$[a_{ij}] = \begin{bmatrix} 1{,}000 & 900 & 900 \\ 900 & 1{,}000 & 900 \\ — & — & 1{,}000 \end{bmatrix}.$$

We assumed arrival rates of three, two, and one patients per day for the Poisson processes associated with each of the three categories of patients, respectively. Hence, because a 24-hour day has 96 15-minute intervals, we use $\lambda_{1t} = 3/96$, $\lambda_{2t} = 2/96$, and $\lambda_{3t} = 1/96$ for all t. Finally, we assumed that the length-of-stay for all patient categories has a Poisson distribution with average of one day per patient. Therefore, $p_{ijt} = 1 - \exp(-1/96)$ for all i, j, t.

We randomly generated 30 initial states and 46 decision rule matrices R with various degrees of proactivity levels. The details of these decision rules are provided in the online supplement. For each initial state S, we computed optimal values $V_2(S), V_3(S)$, and $V_4(S)$ and corresponding optimal decision rules by using dynamic programming. We also computed $\widehat{V}_2(S), \widehat{V}_3(S)$, and $\widehat{V}_4(S)$ and corresponding decision rules using our heuristic. Table 1 shows the average relative error between $V_n(S)$ and $\widehat{V}_n(S)$, as well as the percentage of times that the rule chosen by the heuristic coincided with the optimal decision rule for $n = 2, 3, 4$. In addition, we compared the performance of the prior rule used by the hospital to the optimal solution for each of the states in our sample. We found that the approximation method significantly outperforms the prior WCMH rule as the number of stages increases. Note that for two stages, the prior WCMH rule was always optimal because it involved only one decision period, so a myopic approach is optimal by definition. It is also interesting to note that the heuristic performs better as n increases. This pattern is due to the fact that for relatively low values of n, the predictive ability of the heuristic is very robust. We expect that at some value of n that trend will not persist. However, because the exponential growth of the state space makes computing optimal solutions prohibitive for $n \geqslant 5$, it is not possible to determine that cutoff point. For example, solving for the optimal policy took 30 minutes for the four-stage problem but it took 74 hours to analyze a single decision rule for the five-stage problem. That extrapolates into 370 days of simulation time to evaluate the five-stage problem. By contrast, our heuristic is able to determine a recommended action for problems of this size in less than five seconds and for realistic problems in less than two minutes.

Table 1. Comparing performance.

	Relative error (%)		Worst-case error (%)		Optimal (%)	
Number of stages (n)	Heuristic	Prior rule	Heuristic	Prior rule	Heuristic	Prior rule
2	5.5	0.0	34.9	0.0	53.3	100.0
3	1.1	10.0	13.6	26.9	86.7	53.3
4	0.6	14.2	10.4	40.1	90.0	40.0

5. Computational Experiments

We conducted several computational experiments to fine-tune some of the parameters eventually used in the implementation of the DSS at WCMH.

5.1. Tractable Decision Rule Space Determination

Because the number of potential decision rules is very large, we conducted more computational experiments to identify a tractable number of decision rules to be used in the WCMH DSS. The challenge was to identify a small set of decision rules that yield good performance results. A total of 46 decision rule matrices R were considered, and each simulation depicted a hypothetical one-month period for WCMH ($=96 \times 30 = 2{,}880$ time periods). Based on historical data on arrival rates for patients of each category, we simulated a total of 455 patient arrivals. We computed 14 simulation runs, all starting from the same initial state. Details of the simulation design are provided in the online supplement.

5.1.1. Simulation Results. For each simulation run, we used the rolling-horizon Algorithm 1, computing $\widehat{V}_{2,880}(S)$ in each iteration, and updating states according to the decision rule $\widehat{R}$ returned by the algorithm and the simulated arrivals and departures of patients.

The following variables were recorded:

1. number of transfers;
2. number of "last-minute" transfers (reactive transfers);
3. number of patients assigned to their ideal floor;
4. number of patients assigned to an alternate floor;
5. number of times each decision rule was invoked;
6. number of periods waiting before assignment.

The objective of this analysis was to successively add decision rules and identify a "point of diminishing returns" beyond which the addition of decision rules resulted in little or no improvement in results.

Table 2 shows that the initial addition of decision rules makes a dramatic difference in reducing the number of reactive transfers. However, beyond four decision rules, the incremental improvements diminish. In other words, a relatively small set of decision rules can achieve performance similar to that of a larger set. By studying the most frequently used decision rules, we learned that as long as a varied set was used, we obtain good results.

Figure 1 shows the average delay before making a bed assignment for different levels of capacity utilization.

Table 2. Simulation results.

Simulation run	Decision rules	Total transfers	Reactive transfers	Assigned to ideal	Assigned to an alternate
1	1	73	73	402	53
2	2	80	51	382	73
3	4	33	12	363	92
4	6	33	12	363	92
5	10	34	12	364	91
6	14	32	11	367	88
7	18	31	11	367	88
8	22	30	11	367	88
9	26	30	10	367	88
10	30	31	10	369	86
11	34	29	9	369	86
12	38	29	10	370	85
13	42	28	9	371	84
14	46	28	8	371	84

In Figure 1, we compare simulation runs 1, 2, 3, and 14 with a "passive" strategy in which no assigned patient is transferred between floors to make room for newly arrived patients. Simulation runs 4–13 were omitted because the similarity in performance resulted in a very cluttered graph. Figure 1 shows that the reactive decision rule used by WCMH results in significantly better performance than the passive decision rule. However, the large number of last-minute transfers, most of which occur at higher levels of capacity utilization, result in prolonged waiting times during critical "crunch" periods. This is illustrated in Table 2, which shows that all 73 of the transfers were reactive transfers. By contrast, the inclusion of additional decision rules in runs 2–14 shows an important pattern in reduced delays prior to making a bed assignment and in the number of reactive transfers. In both cases, there is a dramatic improvement in performance that quickly levels off. The last significant performance improvement is found at Run 3, where four decision rules are included. For Run 3, the time delay before a bed assignment can be made remains stable and low, even when capacity utilization is high. In addition, the number of reactive transfers is dramatically less than the number incurred using the hospital's prior rule (12 compared to 73). Run 14, not surprisingly, yields the best results in terms of time to floor assignment and number of reactive transfers. However, the incremental improvement is small and, more significantly, the amount of time needed to analyze all 46 decision rules is prohibitively long.

Figure 1. Waiting time as a function of occupancy.

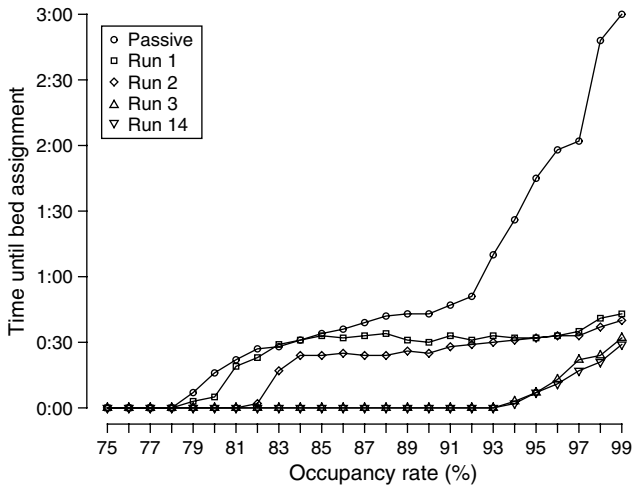


Table 3. Patient categories at WCMH.

Category	Description	Avg. length of stay	a_i ($)
1	Short stay general medicine	2.51	7,464
2	Long stay general medicine	4.86	11,616
3	Short stay general surgery	2.54	11,075
4	Long stay general surgery	8.42	27,653
5	Pediatrics	2.24	4,396
6	Short stay telemetry	1.94	8,268
7	Long stay telemetry	4.71	14,095
8	Clean general surgery	3.81	24,381
9	Short stay critical care	4.47	12,716
10	Long stay critical care	16.29	61,323
11	Surgical gynecology	3.05	11,894
12	Obstetrics/special gynecology	2.42	4,639

6. DSS at WCMH

Based on the simulation conducted in §5, the DSS was encoded with the decision rules depicted in simulation Run 3 (see the online supplement). Table 3 shows the patient categories that were identified at WCMH based on floor assignment and length of stay for the different patient populations the hospital serves. Table 3 also shows the expected reimbursement for each category, again based on historical averages.

Prior to implementing the DSS, hospital administrators were asked to participate in an exercise to determine the appropriate values of the a_{ij} and b_{ijk}. The values for the a_{ij} were initially set to the expected reimbursement of a patient of category i depicted in Table 3, and then scaled down to reflect the desirability of assigning a patient of that category to the different floors. For example, $a_{1,1} = 7{,}464$ because Floor 1 is the floor that specializes in cardiology and general medicine. Floor 2 specializes in general surgery and pediatrics (although the staff are still capable of providing health care service for a general medical patient) and was therefore considered an "alternate" floor with $a_{1,2} = 7{,}164$.

Determining the values of the b_{ijk} was also done interactively. Because determining the values of the b_{ijk} does not require iterating through multiple time periods, it was possible to represent one instance of (10)–(11) in a Microsoft Excel spreadsheet which was then solved using the standard Excel Solver engine. The initial values of all b_{ijk} were set to 100. Then, a baseline hypothetical state was created that was similar to an "average day" in terms of census, staffing, and patient mix at WCMH. Nursing administrators

were then asked to evaluate the patient assignment/transfer decisions that were obtained by solving the patient allocation problem in response to different events. The values of the b_{ijk} were modified, based on feedback from nursing administrators, until decisions consistent with hospital strategy were obtained. The process of determining the appropriate values for the b_{ijk} took approximately one hour and involved evaluating 100 different events and associated responses. Microsoft Excel proved to be a valuable tool because the graphical layout improved the ability of the nursing administrators to understand and critique the events and associated responses. A sensitivity analysis regarding the impact of cost parameters on DSS recommendations is presented in the online supplement.

The DSS was programmed using version 1.4.2 of the Java 2 Platform, Standard Edition (J2SE) and all optimization problems were solved using lp_solve v5.1. The application was run on a desktop PC with a 3 GHz Pentium 4 processor and 512 MB of RAM.

The 18-day trial period began January 17, 2005. The DSS included the decision rule set identified in the simulation described in §5. During the trial period, a total of 292 patients were admitted, an average of 16.22 admissions per day, which is roughly 10% higher than the 14.79 admissions per day average obtained from historical data. This increase in the number of admissions per day was associated with a greater than average number of medical patients, and is consistent with expected seasonal fluctuations.

In total, 83% of the actions recommended by the DSS (269 out of 324) were followed. The DSS allowed the user to enter the reason why a given recommendation was not taken. This information was provided for 47 of the 55 recommendations that were not taken. A total of 20 recommendations were not followed because, based on some preliminary information on patients currently being evaluated in the ED, the bed manager was concerned that following the recommendation would complicate the ability of the hospital to take an admission that was "imminent" but not "official" in the sense that no admission orders had been written. Concerns related to future staffing were the cause of rejecting 11 recommendations and represented cases where the bed manager was hesitant to bring a floor to full capacity, when the ability to staff the floor properly 16–24 hours in the future was in question. The remaining 16 instances were related to: gender compatibility (4); "staff too busy" (6); and a range of clinical concerns (6).

We measured the impact on the number of last-minute transfers using the DSS, and also the actual time needed to transfer each patient admitted from the ED to the floor. To determine the impact on last-minute transfers, a real-time simulation was conducted using the hospital's previous, reactive assignment decision rule, and the status updates obtained during the trial period. This enabled us to compare the number of last-minute transfers and preemptive transfers occurring during the trial period to the number that would have resulted during the trial period had the hospital used its prior decision rule. Table 4 shows the results (note that under the prior decision rule, all in-house transfers were last-minute transfers).

Table 4. Impact on patient transfers.

Type of transfer	DSS	Prior rule
Preemptive	21	0
Last minute	7	69
Total	28	69

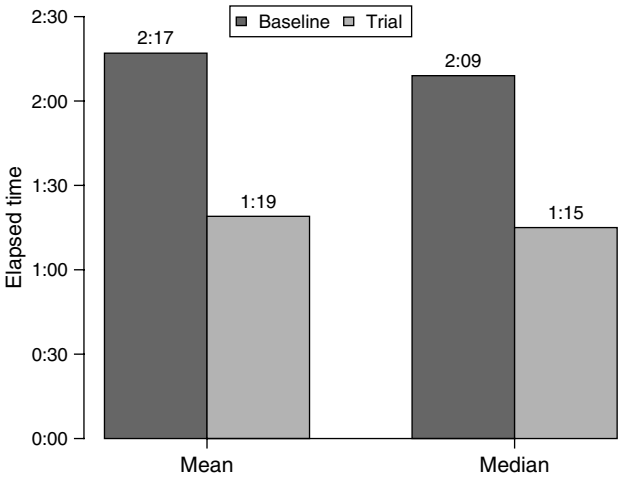
Overall, the total number of in-house transfers was less using the DSS (28 versus 69 using the prior rule), and the reduction in the number of last-minute transfers was even more pronounced (7 versus 69 using the prior rule). The significant reduction in the number of last-minute transfers came at a cost. The number of patients assigned to alternate floors by the DSS was higher than under the hospital's prior assignment rule.

In terms of the number of patients assigned to alternate floors during the trial period, a total of 105 of this type of assignments were made using the DSS, as opposed to only 56 using the prior rule. These results together with the results from Table 4 show that ensuring that beds were available when needed was achieved primarily through the assignment of patients to alternate floors as opposed to transfers. This is a significant positive finding because in-house transfers, even if done preemptively, are time-consuming and disrupt the continuity of care for the patient that is transferred.

Baseline data was obtained from the hospital to determine whether the reduction in last-minute transfers was associated with the expected reduction in the average waiting time before transferring patients from the ED to the floors. The hospital provided data on waiting time in the ED collected at different times over the last half of 2004. The baseline data only contained observations made when beds were available in the hospital. The important implication is that these patients were not being delayed because no beds were available in the hospital (which does happen), but because the available beds were not on the floors to which these patients could be assigned. This baseline data showed very little monthly fluctuation. The Anderson-Darling test (Stephens 1974) was performed on the baseline transfer time data provided by the hospital. The results, with test statistic $A^2 = 0.31$, did not refute the null hypothesis that the data come from a lognormal distribution. The observed decrease in transfer time during the trial period, depicted in Figure 2, was significant at $\alpha = 0.001$.

The impact of the number of available beds on waiting times was also studied. The reason for this analysis is that presumably when each floor has several beds available, achieving short waits should not be difficult. However, as the number of available beds decreases, following decision rules that ensure that remaining beds are allocated in a manner that reflects expected arrival and discharge

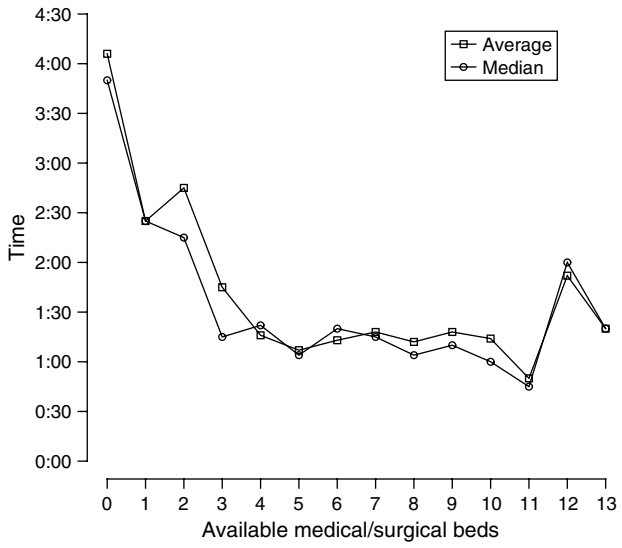
Figure 2. Impact on transfer time.



patterns, should result in shorter waiting times than following decision rules that do not. The results depicted in Figure 2 show that aggressive assignment and reallocation strategies, represented by some of the rules in the decision rule set, enable the hospital to achieve short waiting times even when the number of available beds is low. In addition, inspection of the daily census from each day in the data set that provided the baseline waiting time of two hours and 17 minutes shows that the total number of available beds was at least seven.

Figure 3 plots the average and median waiting times based on the number of available beds at the time admission orders were written. The trend line shows that, on average, the waiting time is slightly over one hour. When the number of available beds decreases to less than four, the waiting times begin to increase. This suggests that the approach developed here is effective, even when remaining capacity is low.

Figure 3. Wait time and bed availability.



6.1. Impact on the Hospital

The reduction in waiting time, almost one hour per patient on average, has significant quality of care and financial repercussions. The impact on quality of care is obvious. Admitted patients are moved more quickly to a floor, and the process of carrying out the admission orders begins in a more timely fashion. For patients waiting to be seen by a physician in the ED, the fact that admitted patients are vacating more quickly means they will have shorter waits. For WCMH, which averages 280 ED admissions per month, a one-hour reduction for each of these admissions translates into a gain of approximately 3,360 additional bed-hours in the ED per year. Each bed in the ED averages $258/hour in charges, so if these additional hours are used to see additional patients, the expected benefit to WCMH is $866,880 per year. During the trial implementation, the additional bed-hours in the ED were used to see additional patients, and according to James Papadakos, Chief Financial Officer at WCMH, the anticipated increase in the number of patients treated and in revenues were observed.

Assigning patients based on the recommendations provided by the DSS also improved capacity utilization. Although not as dramatic as the reduction in waiting time, the resulting improvement has significant quality of care and financial implications. During the 18-day study, by following the recommendations provided by the DSS, WCMH was able to accept four patients totaling 11 bed-days that otherwise would have been diverted to a different hospital. This means the hospital was able to provide services to more patients in the community as a result of diverting fewer ambulances. This is an important accomplishment because diverting ambulances results in a delay in the provision of emergency care which is clearly undesirable. From a financial standpoint, based on WCMH data, the hospital expects to generate $3,043 in charges per bed-day. An increase of 11 bed-days corresponds to an expected increase of $33,473 in charges and translates into an annualized increase of $678,758 in charges. Because WCMH only collects, on average, 39% of the amount billed, the total expected increase in charges of $1,545,638 corresponds to an expected increase of $602,798 in actual revenue. This reflects a 1% increase in revenue from operations, which is not trivial in the not-for-profit health care setting where operating margins of 1%–2% are considered a success. Furthermore, this increase in revenue is achieved using existing physical and staff resources, so that very little additional cost is incurred. As a result, a large portion of the $602,798 in additional revenue is retained.

Finally, WCMH, like all hospitals, routinely tracks patient safety and staffing effectiveness indicators. One of these indicators, the amount of time needed to transfer a patient from the ED to the floor, has been the primary focus of this paper. The other indicators include fall rates, medication errors, hospital-acquired infection rates, restraint usage, length-of-stay, and patient satisfaction surveys that are sent to patients postdischarge. A very important finding

of a key metric for WCMH was that there was no statistically significant change in any of these other indicators. This offers further confirmation that quality of care for admitted patients is not being compromised by taking the actions recommended by the DSS. The DSS implementation was considered a success by WCMH and the hospital has opted to continue using the system as a tool to aid in patient assignment decisions. The hospital has also decided to create an "operations manager" position, an individual who will work with the DSS and work to identify other opportunities to improve patient flow and hospital efficiency. WCMH has continued to use the DSS to aid in patient allocation decisions. Furthermore, the effectiveness of optimality-based resource utilization tools has prompted WCMH to initiate work with us on tools to improve the utilization of their cardiac diagnostic facilities.

7. Electronic Companion

An electronic companion to this paper is available as part of the online version that can be found at http://or.journal.informs.org/.

Acknowledgments

The authors gratefully acknowledge Allison Breault, Vice President for Patient Care Services at Windham Hospital, for her time, support, and invaluable contributions to this research. In addition, the authors thank the anonymous associate editor and referees for their helpful, constructive comments that greatly improved the quality of this manuscript.

References

Albareda, M., E. Fernandez. 2000. The stochastic generalized assignment problem with Bernoulli demands. *TOP* **8**(2) 165–190.

Bayley, M. D., J. S. Schwartz, F. S. Shofer. 2005. The financial burden of emergency department congestion and hospital crowding for chest pain patients awaiting admission. *Ann. Emergency Medicine* **45**(2) 110–117.

Brandeau, M. L., F. Sainfort, W. P. Pierskalla, eds. 2004. *Operations Research and Health Care: A Handbook of Methods and Applications*. International Series in Operations Research and Management Science. Springer, New York.

Caissie, L. 2004. Fax patient reports to save time. *Patient Flow Weekly* **1**(3). http://www.hcpro.com.

Cox, T. F., J. F. Birchall, H. Wong. 1985. Optimising the queuing system for an ear, nose and throat outpatient clinic. *J. Appl. Statist.* **12**(1) 113–126.

Derlet, R. W., J. R. Richards. 2000. Overcrowding in the nation's emergency departments: Complex causes and disturbing effects. *Ann. Emergency Medicine* **35**(1) 63–68.

Dodge, G. 2001. California's emergency services and trauma care system. Testimony to Assembly Committee on Health by Emergency Nurses Association, California State Council, Des Plaines, IL, February 13.

Eckstein, M., L. S. Chan. 2004. The effect of emergency department crowding on paramedic ambulance availability. *Ann. Emergency Medicine* **43**(1) 100–105.

Flagle, C. D. 2002. Some origins of operations research in healthcare. *Oper. Res.* **50**(1) 52–60.

General Accounting Office (GAO). 2003. Hospital emergency departments: Crowded conditions vary among hospitals and communities. United States General Accounting Office, GAO-03–460, March. http://www.gao.gov.

Gerchak, Y., D. Gupta, M. Henig. 1996. Reservation planning for elective surgery under uncertain demand for emergency surgery. *Management Sci.* **42**(3) 321–334.

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

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

Green, L., S. Savin, B. Wang. 2006. Managing patient service in a diagnostic medical facility. *Oper. Res.* **54**(1) 11–25.

Jackson, R. R. P., J. D. Welch, J. Fry. 1964. Appointment systems in hospitals and general practice. *Oper. Res. Quart.* **15**(3) 219–232.

Kellerman, A. 2001. Emergency care in California: No emergency? *Health Affairs, Web Exclusives Supplement*, March. http://content.healthaffairs.org/webexclusives.

Krein, S., M. Casey. 1998. Research on managed care organizations in rural communities. *J. Rural Health* **14**(3) 180–199.

Litvak, E., M. C. Long, A. Cooper, M. L. McManus. 2001. Emergency room diversion: Causes and solutions. *Academic Emergency Medicine* **8**(11) 1108–1110.

Mine, H., M. Fukushima, K. Ishikawa, I. Sawa. 1983. An algorithm for the assignment problem with side constraints. *Memoirs of the Faculty of Engineering*, XLV, Part 4. Kyoto University, Kyoto, Japan.

Parker, P. 2004. Move care to a higher level with emergency systems. *Nursing Management* **35**(9) 82–86.

Schull, M. J., M. Vermeulen, G. Slaughter. 2004. Emergency department crowding and thrombolysis delays in acute myocardial infarction. *Ann. Emergency Medicine* **44**(6) 577–585.

Schull, M. J., K. Lazier, M. Vermeulen, S. Mawhinney, L. J. Morrison. 2003. Emergency department contributors to ambulance diversion: A quantitative analysis. *Ann. Emergency Medicine* **41**(4) 467–476.

Solberg, L., B. Asplin, R. Weinick, D. Magid. 2003. Emergency department crowding: Consensus development of potential measures. *Ann. Emergency Medicine* **42**(6) 824–834.

Stephens, M. A. 1974. EDF statistics for goodness of fit and some comparisons. *J. Amer. Statist. Assoc.* **69**(347) 730–737.

Tucker, J. B., J. E. Barone, J. Cecere, R. G. Blabey, R. Chan-Kook. 1999. Using queueing theory to determine operating room staffing needs. *J. Trauma, Injury, Infection, and Critical Care* **46**(1) 71–79.

Wilson, M. E. 2001. Overcrowding and other crises: How can we survive? *J. Emergency Nursing* **27**(3) 225–227.