Sequential and Simultaneous Decision Making for Optimizing Health Care Resource Flexibilities

Adelina Gnanlet†
Department of Management, Mihaylo College of Business and Economics, California State University at Fullerton, Fullerton, CA 92834, e-mail: agnanlet@fullerton.edu

Wendell G. Gilland
Department of Operations, Technology, and Innovation, Kenan-Flagler Business School, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, e-mail: Wendell_Gilland@unc.edu

ABSTRACT

Health care administrators commonly employ two types of resource flexibilities (demand upgrades and staffing flexibility) to efficiently coordinate two critical internal resources, nursing staff and beds, and an external resource (contract nurses) to satisfy stochastic patient demand. Under demand upgrades, when beds are unavailable for patients in a less acute unit, patients are upgraded to a more acute unit if space is available in that unit. Under staffing flexibility, nurses cross-trained to work in more than one unit are used in addition to dedicated and contract nurses. Resource decisions (beds and staffing) can be made at a single point in time (simultaneous decision making) or at different points in time (sequential decision making). In this article, we address the following questions: for each flexibility configuration, under sequential and simultaneous decision making, what is the optimal resource level required to meet stochastic demand at minimum cost? Is one type of flexibility (e.g., demand upgrades) better than the other type of flexibility (e.g., staffing flexibility)? We use two-stage stochastic programming to find optimal resource levels for two nonhomogeneous hospital units that face stochastic demand following a continuous, general distribution. We conduct a full-factorial numerical experiment and find that the benefit of using staffing flexibility on average is greater than the benefit of using demand upgrades. However, the two types of flexibilities have a positive interaction effect and they complement each other. The type of flexibility and decision timing has an independent effect on system performance (capacity and staffing costs). The benefits of cross-training can be largely realized even if beds and staffing levels have been determined prior to the establishment of a cross-training initiative.


INTRODUCTION

Hospital administrators coordinate internal resources—bed space and nursing staff—to provide the highest quality of care to patients and to reduce adverse
outcomes such as high mortality rates and longer length of stay. Aiken, Clarke, Sloane, Sochalski, and Silber (2002), Needleman, Buerhaus, Mattke, Stewart, and Zelevinsky (2002), and Hassmiller and Cozine (2006) show the positive impact of nurse staffing in providing high-quality care to patients. Hospital administrators have to plan for nurses and bed space so that both nursing staff and beds are available at the right time and right quantity to treat patients.

In a typical health care supply chain, when efficient utilization of internal resources at the hospital is insufficient to meet patient needs (customers), hospitals turn to third-party agencies for additional resources. Shortages of regular staff are supplemented by hiring contract nurses from an external agency for a shift or two until demand stabilizes. Contract nurses are more expensive than regular nurses. The American Hospital Directory indicates that contract nurses cost as a percentage of total operating cost increased steadily from 1.4% to 3.8% over a 5-year period (Shoemaker & Howell, 2004). Recently, the Service Annual Survey by the U.S. Census Bureau (U.S. Census Bureau, 2008) found that 21.5% of the personnel cost in 2006 was incurred by temporary staff and contracted employees. This leads to the question of how to coordinate key internal resources (nursing staff and bed space) and external resource (contract nurses) so that a large number of patients are served at minimum cost?

Hospitals usually hire contract nurses on ad hoc basis inducing higher unplanned costs for the hospital. In order to reduce labor cost by contract staff, hospitals cross-train (Lyons, 1992; Siferd & Benton, 1992) their full-time nursing staff, enabling them to float between units within a specialization. These units within a specialization have different acuity levels, but are similar enough to cross-train nurses. Cross-training (floating/flexibility) of nurses helps to meet heavy demand by using nursing hours from another unit with lean demand. Inman, Blumenfeld, and Ko (2005) show that cross-training not only helps to meet variable demand, but can also reduce staffing costs for the hospital as well as improve morale and job satisfaction for the nurses. Many hospitals have reaped financial benefits from successful implementation of cross-training programs (Altimier & Sanders, 1999; Snyder & Nethersole-Chong, 1999). Cross-training nurses is generally referred to as “staffing flexibility” in the operations literature and “floating” in health care literature.

When bed space, another internal resource at the hospital, is inadequate to meet patient demand, patients are upgraded to a higher acuity/more sophisticated unit where bed spaces are available. The more sophisticated unit where the patient is upgraded usually has the necessary equipment to more than meet the needs of the patient moved in from the simple unit, and therefore is more costly to operate. For example, a gynecological surgery patient may, due to lack of beds in the recommended unit, be accommodated in the vascular and urological surgery unit. The vascular and urological surgery unit is more complex and hence has sufficient resources to treat a gynecology patient. This type of upgrades is called “demand upgrade,” “downward substitution,” or “capacity flexibility” in the operations literature. In the case where bed space is unavailable in any unit, the patient is directed to a nearby hospital.

Therefore, when internal resources are insufficient to meet patient needs, hospital administrators decide to float nurses (staffing flexibility), hire nurses from an external agency (contract nurses), upgrades patients (demand upgrades) to a
more sophisticated unit, or turn away patients. The decision in using flexibility of internal resources or using external resources depends on the total cost to the hospital and demand variability. In this article, we examine the following questions: how much bed space and nursing staff is required under different types of flexibility (e.g., staffing flexibility, demand upgrades, or both) while coordinating internal hospital resources, external agency resource, and patients? How does the type of flexibility affect the optimal capacity (bed space) and staffing (nursing staff) decisions? How much cross-training should be conducted under each type of flexibility?

Hospitals plan for bed space and staffing within different time frames. Usually bed space, hereafter called the capacity of a unit, is determined when the unit is constructed. Staffing decisions, on the other hand, are typically made once a year. It is quite difficult to anticipate future needs of the hospital, including the type of flexibility to be used, so hospitals often make myopic decisions regarding both capacity and staffing. This article was motivated by staffing and capacity decisions at University of North Carolina (UNC) Hospitals. UNC Hospitals is a public academic medical center operated by the state of North Carolina. It has 726 licensed beds and includes a children’s hospital, memorial hospital, neurosciences hospital, women’s hospital, and a newly constructed cancer hospital. UNC Hospitals very recently decided to hire a pool of nurses who could be floated across multiple units. Though capacity and staffing decisions were made early in the development process (when units were constructed), staff flexibility decisions had to be made at a much later date. This raises the question: is the timing of capacity and staffing decisions important? In this article, we study the extent to which the timing of capacity and staffing decisions affects system performance. In our context, system performance refers to the total capacity and staffing costs of the two units we analyze. When capacity decisions are made in the first period and staffing decisions are made in the second period, we call this sequential decision making. This can also be considered as decentralized decision making, where capacity and staffing decisions are made by different departments without any information coordination. In order to test the deviation of sequential decision making from optimal performance, we also develop a simultaneous decision-making case, where all capacity and staffing decisions are made at the same point in time. This can be interpreted as centralized decision making, where staffing and capacity decisions are made by a single department.

Wright, Bretthauer, and Cote (2006) determine the effect of using mandatory nurse-to-patient ratios while minimizing nursing costs and minimizing undesirable schedules for nurses. The nonlinear programming formulation includes a planning level, where the total number of nurses are determined and a scheduling level, where detailed shift schedules are decided. In this article, we determine aggregate nursing levels under varying types of flexibility, and determine capacity (bed spaces) as well. Detailed shift scheduling for nurses is beyond the scope of this article, but we determine staffing and capacity allocations after the realization of demand.

We find that centralized decision making (i.e., deciding for capacity, total staff, and/or float nurses simultaneously) yields better performance compared to making capacity and staffing decisions separately, as expected. If managers are unable to make decisions for all resources (capacity, total staff, float nurses) at the
same time, they should at least integrate capacity and total staff decisions and plan for float nurses later when needed. This strategy achieves most of the benefit for simultaneous decision making. Deciding on internal resources (capacity, total staff, and/or float nurses) at the same time (centralized decision making) becomes vital under staffing flexibility, especially when external resources are expensive (i.e., contract nurse cost is high). We also find that hospitals obtain greater benefit when planning for both types of flexibility (staffing flexibility and demand upgrades) at the same time rather than using one type of flexibility now and adding another level of flexibility later. There is a positive interaction between the two types of flexibility. When demand in the more sophisticated unit is higher than demand in the simple unit, staffing flexibility provides greater benefits to hospital managers than demand upgrades. When demand in the more sophisticated unit is less than the simple unit, however, use of demand upgrades will yield greater benefit. Staffing flexibility is a better option when the cost of external resources (contract nurses) for the more sophisticated unit is high, while demand upgrades is a better option when the cost of external resources for the simple unit is high.

**LITERATURE REVIEW**

Since the 1970s, operations management researchers (Warner & Prawda, 1972; Abernathy, Baloff, Hershey, & Wandel, 1973; Warner, 1976; Siferd & Benton, 1992; Bard & Purnomo, 2005; Easton & Goodale, 2005; Wright et al., 2006) have formulated mathematical models for nurse scheduling considering various aspects such as absenteeism, nurse–patient ratios, and flexible schedules. These articles depict the operational decisions of day-to-day scheduling rather than aggregated staffing decisions. Most recently, Wright et al. (2006) studied the impact of mandatory nurse-to-patient ratio policy on nursing costs and scheduling preference. In our article, we determine long-term (aggregated) staffing decisions of hospitals when they use different types of nurses (regular nurses, float nurses, contract nurses) to meet stochastic patient demand. Nursing literature has emphasized the importance of adequate nurse staffing on patient care. Czaplinski and Donna (1998), Blegen, Goode, and Reed (1998), Hall (2003), and Heinz (2004) show the positive impact of nurse staffing on patient outcomes such as length of stay, mortality rates, and infection rates. Robertson, Dowd, and Mahmud (1997), McCue, Mark, and Harless (2003), Mark (2004), and Woech-Maldonado, Meret-Hanke, Neff, and Mor (2005) analyze the effect of nurse staffing on quality of patient care and financial performance. Shullanberger (2000) presents an integrative review of nurse staffing literature to evaluate different nurse staffing models.

Due to the rise in personnel costs and shortage of full-time nurses, health care administrators implement cross-training programs (Lyons, 1992; Bergman, 1994; Wheaton, 1996) and float nurses when needed. The most significant benefit of cross-training is cost efficiency for hospitals (Altimier & Sanders, 1999; Snyder & Nethersole-Chong, 1999; Gilbert & Counsell, 2000) but other frequently cited benefits are marketability and job satisfaction for nurses. Robertson et al. (1997), Li and King (1999), and Inman et al. (2005) formulate staffing models considering the effects of cross-training nurses. Trivedi and Warners (1976) formulates a branch and bound algorithm for allocation of core staff (dedicated nurses) and float staff
(flexible nurses) with the objective of minimizing workload among all nurses. In our article, we compare the benefits of using float nurses over dedicated nurses and determine the optimal staffing and timing decisions that minimize cost but meet stochastic patient demand.

Smith-Daniels, Schweikhart, and Smith-Daniels (1988) present an exhaustive review on capacity management in health care services and also present future research topics. Lovejoy and Li (2002) and Cochran and Bharti (2006) analyze the impact of bed capacity on patient services and strategies to increase hospital beds. Bretthauer and Cote (1998) decide for equipment, facility, and workforce capacity using queueing models when patients flow from one department to another. Li and Benton (2003) used structural equation modeling to determine the factors that affect capacity decisions (which includes facility, workforce, equipment) in hospitals and thereby determine impact of capacity decisions on cost and quality performance. Jack and Powers (2004) present a framework that classifies different types of flexibility strategies used in health care when demand varies. In our article, we determine optimal number of beds required in two nonhomogeneous hospital units and also evaluate the impact of patient upgrades from one unit to another in order to effectively utilize beds. Our main contribution toward the literature is in evaluating the use of two types of flexibility in tandem and determining the best staffing, capacity, and timing decisions for two nonhomogeneous hospital units facing stochastic demand.

PROBLEM DEFINITION

In this section, we will first discuss the four types of flexibility configurations then explain the timelines for capacity and staffing decisions under each flexibility configuration. We consider two nonhomogeneous hospital units, one being a complex unit with high patient acuity and the other being a simple unit with lower patient acuity. Patients in the simple unit, in some configurations, can be upgraded to the complex unit, but not vice versa. For notation purposes, we assume unit 1 to be the complex unit and unit 2 to be the simple unit. These two units can be staffed with three types of nursing staff, depending on availability. The in-house regular nurses are assigned to their home unit (hereafter referred to as dedicated nurses) while the in-house cross-trained nurses (hereafter referred to as float nurses) are assigned to either of the two units based on demand. Dedicated and float nurses are full-time employees of the hospital and are paid wages even when demand is low. The dedicated nurses and float nurses are called total staff in the notation and formulation. If patient demand is still not met, then contract nurses from an outside agency (hereafter referred to as contract nurses) are hired (at a higher cost than either unit). The two nonhomogeneous units have different capacities, measured in terms of number of bed spaces. Nursing staff (number of dedicated and float nurses) and capacity (number of bed spaces) for each unit are the decision variables in our model. We assume that one patient needs only one bed space for his/her treatment and that one nurse treats one patient. When estimating demand distribution in nursing hour requirements, the differences in patient acuity between two units are also considered.
Fig. 1: Four configurations.

This section describes the model using two types of flexibility under four configurations. Figure 1 shows the network representation of the four configurations.

**Configuration 1 (base case no flexibility)**
When bed space and nursing staff are available, the patient is admitted and treatment proceeds. Nursing staff, in this configuration, consists of only dedicated nurses; no nurses have been cross-trained. If bed space is available when a patient arrives but nursing staff is not available, contract nurses are hired at a cost of $s_i$ for unit $i$. If a bed space is not available, the patient is directed to another hospital and the system incurs a penalty cost of $p_i$ for unit $i$. There are no upgrades from a lower unit to a higher unit.

**Configuration 2 (demand upgrades)**
When both capacity and nursing staff are available, the patient for unit $i$ is admitted to unit $i$. Nursing staff consists of dedicated nurses and contract nurses. Here again, there are no float nurses. Unlike configuration 1, in this configuration, when capacity is not available in unit 2, that patient is admitted to unit 1 provided unused capacity is available in unit 1. Such upgrades to unit 1 are allowed only when demand in unit 1 is first met. Because unit 1 is a complex/sophisticated unit, the equipment is sufficient to treat patients from unit 2 but the dedicated nurses in unit 1 are not trained to handle patients from unit 2 and so contract nurses are hired to
treat patients upgraded to unit 1. Patients for unit 2 are turned away at a cost of $p_2$ if capacity is not available in unit 2 and if upgrade to unit 1 is not possible. If enough capacity in unit 1 is not available, patients for unit 1 are turned away at a cost of $p_1$. If capacity is available in unit $i$ but dedicated staff is not available in unit $i$, then contract nurses are hired for unit $i$.

**Configuration 3 (staffing flexibility)**

Unlike configuration 2, in this configuration we do not have demand upgrades. If both capacity and nursing staff are available, patient is admitted for treatment at the appropriate unit. If capacity is not available, patients are turned away, incurring a penalty cost of $p_i$. In configuration 3, we use three types of nursing staff (dedicated, float, and contract nurses). Dedicated nurses are trained to work only in their home unit. Cost of wages for them is $h_i$ for unit $i$. Float nurses are cross-trained to work in both unit 1 and unit 2. Cost of wages for float nurses is $t$. We assume float nurses are equally productive in both units. The third type of nursing staff are the contract nurses who are hired at a cost of $s_i$ for unit $i$. In each unit, dedicated nurses are assigned first. If demand exceeds the number of dedicated nurses, float nurses are used. Float nurses are first assigned to unit 1 because unit 1 is the complex unit and hiring contract nurses for unit 1 is more expensive. Any remaining float nurses are assigned to unit 2. If dedicated and float nurses are still not able to meet demand, contract nurses are hired as needed.

**Configuration 4 (demand upgrades and staffing flexibility)**

This configuration is highly flexible. Both types of flexibility, demand upgrades and staffing flexibility, are used in this model. If capacity is not available in both units, patients are turned away, incurring a penalty cost of $p_i$. When capacity is not enough in unit 2 patients are upgraded to unit 1, provided there are beds not being used by unit 1 patients. Here again, similar to configuration 3, three types of nursing staff are used. Dedicated nurses are assigned first to meet the demand in unit $i$, followed by assigning float nurses to unit 1. Any excess float nurses are assigned to unit 2, as needed. Finally, contract nurses are hired to meet the remaining demand in each unit.

**Timeline for Decision Making**

This section motivates and explains the framework for the timing of decision making. In all cases, the actual allocation of nursing staff and beds are made after demand is realized. The timing issues we discuss here relate to decisions regarding capacity (number of beds) and nurse staffing levels (both dedicated and float). As indicated in Figure 2, we consider two types of sequential decision making and simultaneous decision making.

UNC Hospital, for example, had decided on capacity and total staffing levels initially when the units were created. Now, in order to reduce labor costs, they are planning to implement cross-training programs. The cross-training program chooses some dedicated nurses and trains them to float to other units. This motivated us to study the impact of decision timing on system performance. We consider two cases of sequential decision making. Sequential decision making
Figure 2: Timeline for configurations.

```
<table>
<thead>
<tr>
<th>Decision Making Case</th>
<th>Timeline</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case 1</td>
<td>Decide on optimal capacities and optimal staffing levels</td>
</tr>
<tr>
<td></td>
<td>Decide on optimal staffing flexibility</td>
</tr>
<tr>
<td></td>
<td>Demand realization</td>
</tr>
<tr>
<td></td>
<td>Allocation of staff based on available capacity and realized demand</td>
</tr>
<tr>
<td></td>
<td>Incurred cost</td>
</tr>
<tr>
<td>Case 2</td>
<td>Decide on optimal capacities</td>
</tr>
<tr>
<td></td>
<td>Decide on optimal staffing level and staffing flexibility</td>
</tr>
<tr>
<td></td>
<td>Demand realization</td>
</tr>
<tr>
<td></td>
<td>Allocation of staff based on available capacity and realized demand</td>
</tr>
<tr>
<td></td>
<td>Incurred cost</td>
</tr>
<tr>
<td>Simultaneous</td>
<td>Decide on optimal capacities, optimal staffing level, and staffing flexibility</td>
</tr>
<tr>
<td></td>
<td>Demand realization</td>
</tr>
<tr>
<td></td>
<td>Allocation of staff based on available capacity and realized demand</td>
</tr>
<tr>
<td></td>
<td>Incurred cost</td>
</tr>
</tbody>
</table>
```

case 1 represents the UNC system, so capacity and total staffing decisions are made initially. Given capacity and total staff, we later determine the number of float nurses to train in order to minimize total expected cost. This type of decision making is applicable only in configuration 3 and configuration 4, where staffing flexibility is modeled. At times, capacity and staffing decisions are decentralized due to organizational structure. Sequential decision making case 2 represents this scenario. In sequential decision making case 2, the capacity decisions are made originally and all staffing decisions (including dedicated and float nurses) are made later. This type of decision making is applicable for all four configurations. With simultaneous decision making, all capacity and staffing (both dedicated and float nurses) decisions are made at the same time.

**MODEL FORMULATION**

This section discusses model formulation and solution for each flexibility configuration under sequential and simultaneous decision making. The two hospital units, each of which has a pool of nursing resources and fixed capacity (bed spaces), face stochastic demand. In practice, supply and demand for each unit is measured as the number of full-time-equivalent (FTE) nurses required each day. The FTEs can also be represented as nursing hour requirements for each unit. In our model, all allocations and demand are measured on a continuous scale, consistent with nursing hours as the unit of measure. Capacity is also approximated to continuous scale for simplicity. Demand for each unit \(i\) is stochastic and follows a general, continuous distribution with the cumulative distribution function \(\Phi_i\). Our model implicitly assumes a 1:1 nurse-to-patient ratio for staffing. Incorporating alternative staffing assumptions into our model could easily be achieved by scaling
demand for nursing staff. The realization of demand is represented as $d_i$ for unit $i$. The following list summarizes the notation in our models.

Cost parameters:

- $p_i$: penalty for not satisfying patient demand per hour in unit $i$
- $f_i$: operating cost per bed space in unit $i$ per hour
- $h_i$: wages for a dedicated nurse for unit $i$ per hour
- $t$: wages for a float nurse per hour
- $s_i$: contract wages for hiring contract nurses in unit $i$ per hour.

Capacity decisions:

- $k_i$: capacity (number of bed spaces) available in unit $i$.

Staffing decisions:

- $z_i$: number of total staff available in unit $i$
- $n_i$: number of dedicated nurses available in unit $i$
- $e_i$: number of float (cross-trained) nurses available in unit $i$
- $e$: number of float (cross-trained) nurses available for both units.

Capacity allocations:

- $a_i$: number of bed spaces allocated to treat patients in unit $i$
- $a_u$: number of bed spaces in unit 1 allocated to treat patients from unit 2.

Staffing allocations:

- $x_i$: number of dedicated nurses allocated to unit $i$
- $y_i$: number of float (cross-trained) nurses allocated to unit $i$.

Demand parameters:

- $d_i$: number of patients to be treated in unit $i$—realization of stochastic demand $\Phi_i$: stochastic demand for unit $i$.

Under all configurations, total staff in unit $i$ is the sum of dedicated nurses and float nurses $\sum_i z_i = \sum_i n_i + e$. Configurations 1 and 2 do not use any float nurses, so $z_i = n_i$. To compare sequential case 1 and case 2 decision making, we take $\sum_i e_i = e$. The following list enumerates the assumptions that have to hold between cost parameters so that trivial solutions are eliminated from the model. (i) $h_2 < h_1 < t < s_2 < s_1$; (ii) $s_2 < h_1 + h_2$; (iii) $f_2 < f_1 < p_2 < p_1$; (iv) $p_i > s_i + f_i$, $i = 1, 2$; (v) $p_2 > s_1 + f_1$. The first assumption prioritizes allocation of nursing staff. Contract nurses are used only when dedicated and float nurses are insufficient to meet demand. The second assumption is required to avert the possibility of hiring dedicated nurses for both unit 1 and unit 2 instead of using a contract nurse. The third assumption prioritizes allocations for capacity. Because unit 1 is complex, the per unit capacity and staffing costs are expensive compared to the equivalent cost.
in unit 2. The fourth assumption prevents the model from losing excess demand. The fifth assumption prevents the scenario where it is better off to lose patients in unit 2 than upgrade them to unit 1.

All models are formulated as two-stage stochastic programming, with the second stage being the actual assignment of patients and nurses to floors, after demand has been realized. Demand in both units follow general continuous distributions. All second-stage formulations (capacity and staffing allocations after demand is realized) minimize the sum of contract nurse cost and/or penalty cost with resource constraints. Second-stage decisions (capacity and staffing allocations) for all configurations are convex in their objective functions, so decisions are determined using first-order conditions. The expected value of the second-stage objective function is then substituted into the first-stage objective function. We prove convexity of the first-stage objective function, and determine first-order conditions. Solving first-order conditions, we get optimal values for decision variables of interest. In order to present more discussion, we present detailed formulations, analysis, and results in our Technical Report (available at http://www.business.fullerton.edu/management/agnanlet).

Configuration 1: No Flexibility

Configuration 1 is the base case configuration without any type of flexibility. Given the lack of flexibility in configuration 1, the second-stage allocations are straightforward in all cases.

Sequential decision making

Sequential decision making in configuration 1 has two periods, as seen in Figure 3.

**Period 1:** After capacity \( (k_i) \) is decided in period 1 by minimizing the sum of fixed capacity cost and the expected penalty cost for turning away patients, staffing decisions are made in period 2, both using a newsvendor approach. The nurse wages and expected contract nurse cost is minimized to determine optimal staffing decisions \( (z_i) \) in period 2.

\[
\begin{align*}
\text{Min} & \quad f_i \cdot k_i + \mathbb{E}[\Phi_i (p_i \cdot (d_i - \min(k_i, d_i)))], \\
\text{subject to} & \quad k_i \geq 0, \quad i = 1, 2.
\end{align*}
\]

**Period 2:** Period 2 determines staffing decisions \( (z_i) \) given capacity decisions \( (k_i) \) from period 1:

\[
\begin{align*}
\text{Min} & \quad h_i \cdot z_i + \mathbb{E}[s_i \cdot (\min(k_i, d_i) - \min(z_i, d_i)) + p_i (d_i - \min(k_i, d_i))], \\
\text{subject to} & \quad k_i \geq z_i \geq 0, \quad i = 1, 2.
\end{align*}
\]

Because the formulation closely represents the newsvendor problem, first-order conditions are used to derive closed form expressions for \( k_i \) and \( z_i \), which are given in Equation (1):
**Simultaneous decision making**

Simultaneous decision making has only one period as shown in Figure 3, where both capacity ($k_i$) and total staff ($z_i$) are decided. Capacity and staffing allocations are made after demand is realized. At the end of the period, cost is incurred.

\[
\text{Min}_{k_i,z_i} \quad f_i \cdot k_i + h_i \cdot z_i + E_{\Phi_i} \left[ s_i \cdot (\min(k_i, d_i) - \min(z_i, d_i)) \right] + p_i \cdot (d_i - \min(k_i, d_i)),
\]

subject to \quad $k_i, z_i \geq 0$, \quad $i = 1, 2$.

It is easily proven that the objective function is convex. Optimal capacity ($k_i^*$) and staffing ($z_i^*$) levels, assuming $k_i^* \geq z_i^*$, are given in Equation (2):

\[
\Phi_i(k_i^*) = \frac{p_i - f_i - s_i}{p_i - s_i}; \quad \Phi_i(z_i^*) = \frac{s_i - h_i}{s_i}.
\]

Optimal capacity and staffing are still determined by a newsvendor-type relationships, but optimal capacity now depends also on the cost of contract nurses. It is easily shown that simultaneous $k_i^*$ is less than sequential $k_i^*$. When Equation (2) yields $z_i^* < k_i^*$, as would happen when the cost of contract nurses ($s_i$) is large in relation to the cost of internal resources ($f_i, h_i$) the optimal solution comes from constraining $z_i = k_i$ in the objective function and solving for $k_i$. The optimal solution is given in Equation (3):

\[
\Phi_i(k_i^*) = \Phi_i(z_i^*) = \frac{p_i - f_i - h_i}{p_i}.
\]
Configuration 2: Demand Upgrades

Configuration 2 allows for one type of flexibility, demand upgrades. The formulations for sequential and simultaneous decision making while planning for demand upgrades are given subsequently.

Sequential decision making

Sequential decision making in configuration 2 has two periods. In the first period, capacity decisions are made assuming demand upgrades are possible after demand is realized. Capacity is decided by minimizing the sum of fixed capacity cost and expected penalty cost under stochastic demand allowing for upgrades. In the second period, given capacity from period 1, total staff \( z_i \) is determined by minimizing the sum of total wage and expected contract nurse cost. The timeline for configuration 2 is depicted in Figure 3.

**Period 1:**

\[
\begin{align*}
\text{Min} & \quad \sum_i (f_i \cdot k_i) + E[\Phi_1 (p_1 (d_1 - \min(d_1, k_1)) \\
& \quad + p_2 (d_2 - \min(d_2, k_2 + (k_1 - d_1)^+)))] \\
\text{subject to} & \quad k_i \geq 0, \quad i = 1, 2.
\end{align*}
\]

Because the objective function is convex in \( k_i \), we obtain optimal capacity \( (k_i^*) \) for period 1 by solving first-order conditions simultaneously.

**Period 2:** In period 2, given capacity \( (k_i^*) \), we determine optimal staffing \( (z_i^*) \).

\[
\begin{align*}
\text{Min} & \quad \sum_i (h_i \cdot z_i) + E[\Phi_2 (s_1 (a_1^* + a_u^* - x_1^*) + s_2 (a_2^* - x_2^*) + p_1 (d_1 - a_1^*) \\
& \quad + p_2 (d_2 - a_2^* - a_u^*))],
\end{align*}
\]

subject to \( z_i \geq 0, \quad i = 1, 2, \)

where, \( a_1^* = \min(k_i, d_i); \ a_u^* = \min(k_1 - \min(k_1, d_1), \ d_2 - \min(k_2, d_2)); \ x_1^* = \min(z_i, \min(k_i, d_i)) \) are optimal capacity and staffing allocation variables in period 2 after the demand is realized. Beds that are allocated to patients in their respective units are represented by optimal capacity allocation variables \( a_1^* \) and \( a_2^* \). Any excess patients in unit 2 are assigned to unit 1 if beds are available in unit 1. Optimal capacity allocation variable \( a_u^* \) represents beds in unit 1 that are used for patients of unit 2. Optimal allocation of staff to beds \( a_i^* \) is represented by \( x_i^* \). Any nursing shortfall is met by contract nurses. Contract nurses in unit 1 also treat patients upgraded from unit 2 to unit 1 \( (a_u^*) \). Because the objective function is convex in \( z_i \), we obtain optimal staffing \( (z_i^*) \) for period 2 by solving first-order conditions simultaneously (refer to Technical Report).

Simultaneous decision making

In simultaneous decision making, capacity \( (k_i) \) and staffing decisions \( (z_i) \) are made at the same time. After demand is realized, capacity and staffing allocations are
made assuming that demand upgrades are possible. At the end of the period, capacity cost, wage, expected penalty, and contract nurse costs are incurred. The timeline for simultaneous decision making is depicted in Figure 3.

\[
\begin{align*}
\min_{k_i, z_i} & \quad \sum_i (f_i \cdot k_i + h_i \cdot z_i) + \mathbb{E}_q \left[ s_1 (a_1^* + a_u^* - x_1^*) + s_2 (a_2^* - x_2^*) \right] \\
& \quad + p_1(d_1 - a_1^*) + p_2(d_2 - a_2^* - a_u^*)
\end{align*}
\]

subject to \( k_i, z_i \geq 0, \ i = 1, 2, \)

where, \( a_i^* = \min(k_i, d_i) \); \( a_u^* = \min(k_1 - \min(k_1, d_1), d_2 - \min(k_2, d_2)) \); \( x_i^* = \min(z_i, \min(k_i, d_i)) \) are optimal capacity and staffing allocation variables, with \( a_u^* \) representing the beds in unit 1 that are used for patients of unit 2. Beds are assigned to patients in their respective units. Any excess patients in unit 2 are assigned to unit 1 if beds are available in unit 1. Nurses are allocated to beds in their units that have appropriate patients; any nursing shortfall is met by using contract nurses. Contract nurses in unit 1 also treat upgraded patients from unit 2. The first-stage objective function is the expected value of capacity and staffing allocations substituted from the second-stage optimization. Because staffing and bed allocations are made simultaneously, we may incur contract nurse cost for using contract nurses in units 1 and 2, and we also may incur penalty cost for turning away patients from unit 1 and unit 2 after upgrades. Analysis of the hessian shows that the first-stage objective function is partially convex and first-order conditions can only be used to determine optimal values of staffing \( (z_i^*) \). For each capacity level in unit 1 and unit 2, we found optimal staffing levels using first-order conditions (refer to Technical Report). A search of costs associated with each capacity level generated optimal capacity decisions for simultaneous decision making.

Configuration 3: Staffing Flexibility

Configuration 3 allows for staffing flexibility but no demand upgrades. Formulations for two types of sequential decision making and a simultaneous decision making for configuration 3 is given subsequently.

Sequential decision making—case 1

Sequential decision making case 1 has two periods. In period 1, capacity \( (k_i) \) and staffing decisions \( (z_i) \) (only dedicated nurses) are made without the knowledge that flexibility will be allowed in period 2. In period 2, given the capacity and total staff available, the optimal number of float nurses is determined. The timeline for sequential decision making case 1 for configuration 3 is shown in Figure 4.

Period 1: In the first period of configuration 3, we optimize for \( k_i \) and \( z_i \). This formulation is similar to optimizing \( k_i \) and \( z_i \) simultaneously in configuration 1, so we use \( k_i^* \) and \( z_i^* \) from configuration 1—simultaneous decision making.
Minimize \( \sum_i (f_i \cdot k_i + h_i \cdot z_i) + E \Phi_i \sum_i [(s_i \cdot (\min(k_i, d_i) - \min(z_i, k_i, d_i)) + p_i \cdot (d_i - \min(k_i, d_i))] \)

subject to \( k_i, z_i \geq 0, \ i = 1, 2. \)

The results of this period 1 optimization are the same as for simultaneous decision making in configuration 1.

**Period 2:** In period 2, the optimal number of float nurses \((e_i)\) is determined under the assumption that preliminary staffing and capacity decisions have already been made. We determine optimal float resource \(e_i^*\), given \(k_i^*\) and \(z_i^*\) from period 1.

Minimize \( \sum_i ((t - h_i) \cdot e_i) + E \Phi_i \left[ \sum_i s_i \cdot (a_i^* - x_i^* - y_i^*) + p_i \cdot (d_i - a_i^*) \right] \)

subject to \( 0 \leq e_i \leq z_i, \ i = 1, 2, \)

where \( a_i^* = \min(k_i, d_i); x_i^* = \min(z_i, d_i); y_1^* = \min(\min(k_1, d_1) - \min(z_1 - e_1, d_1), e_1 + e_2); y_2^* = \min(\min(k_2, d_2) - \min(z_2 - e_2, d_2), e_1 + e_2 - y_1^*) \) are the optimal allocation of capacity, dedicated nurses, and float nurses to unit \( i \) after demand is realized. Dedicated nurses, float nurses, and beds are assigned to each floor based on patient demand. Analysis of second-order conditions proves that the first-stage objective function is convex and first-order conditions (refer to Technical Report) can be used to determine optimal values of float nurses \((e_i^*)\).

**Sequential decision making—case 2**

Sequential decision making case 2 has two periods. In the first period, capacity \((k_i)\) is determined. In the second period, given the capacity from the first period, optimal dedicated \((n_i)\) and float nurses \((e)\) are determined. The timeline for sequential decision making case 2 for configuration 3 is shown in Figure 4.

**Period 1:** In period 1, we only determine capacity for each unit by minimizing the objective function given subsequently. This formulation is similar to optimizing capacity \((k_i)\) for configuration 1 (no flexibility) under sequential decision making.

Minimize \( \sum_i (f_i \cdot k_i) + E \Phi_i p_i \cdot (d_i - \min(k_i, d_i)) \)

subject to \( k_i \geq 0, \ i = 1, 2. \)

First-order conditions prove that these capacities follow a simple newsvendor relationship and optimal capacity is given in Equation (1).

**Period 2:**

Minimize \( \sum_i (h_i \cdot n_i) + t \cdot e + E \Phi_i \left[ \sum_i s_i \cdot (a_i^* - x_i^* - y_i^*) + p_i \cdot (d_i - a_i^*) \right] \)

subject to \( n_i + e \leq k_i; \ n_i, e \geq 0, \ i = 1, 2, \)
Figure 4: Timeline for configurations 3 and 4.

Sequential Decision Making - Case 1

<table>
<thead>
<tr>
<th>Period 1: Resource levels</th>
<th>Period 2 - Stage 1: Cross-training</th>
<th>Period 2 - Stage 2: Allocate capacity and staff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decide on $k_1, k_2, z_1, z_2$</td>
<td>Given $k_1, k_2, z_1, z_2$;</td>
<td>Demand is realized $d_1, d_2$;</td>
</tr>
</tbody>
</table>

Sequential Decision Making - Case 2

<table>
<thead>
<tr>
<th>Period 1: Capacity</th>
<th>Period 2 - Stage 1: Staffing</th>
<th>Period 2 - Stage 2: Allocate capacity and staff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decide on $k_1, k_2$</td>
<td>Given $k_1, k_2$;</td>
<td>Demand is realized $d_1, d_2$;</td>
</tr>
<tr>
<td>Decide on $n_1, n_2, e$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Simultaneous Decision Making

<table>
<thead>
<tr>
<th>Stage 1: Capacity and staff</th>
<th>Stage 2: Allocate capacity and staff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decide on $k_1, k_2, n_1, n_2, e$</td>
<td>Demand is realized $d_1, d_2$;</td>
</tr>
</tbody>
</table>

where $a_i^* = \min(k_i, d_i); x_i^* = \min(n_i, d_i); y_i^* = \min(\min(d_1, k_1) - \min(z_1, d_1), e); y_2^* = \min(\min(d_2, k_2) - \min(z_2, d_2), e - y_1)$ are the optimal capacity, dedicated nurse, and float nurse allocations after demand realization. Second-order conditions prove that the objective function is convex in the decision variables. First-order conditions are necessary and sufficient to find the optimum dedicated nurses ($n_i^*$) and float nurses ($e^*$), given optimal capacity ($k_i^*$) from period 1 (refer to Technical Report).

**Simultaneous decision making**

In simultaneous decision making, all capacity and staffing decisions are made in the same period. Capacity ($k_i$), dedicated nurses ($n_i$), and float nurses ($e$) are decided before demand realization, while capacity and staffing allocations are made after demand realization. The timeline for simultaneous decision making for configuration 3 is shown in Figure 4.

\[
\min_{k_i, n_i, e} \sum_i (f_i \cdot k_i + h_i \cdot n_i) + t \cdot e
\]

\[
+ E_{\Phi_i} \left[ \sum_i \left( s_i \cdot (a_i^* - x_i^* - y_i^*) + p_i \cdot (d_i - a_i^*) \right) \right],
\]

subject to $n_i + e \leq k_i; k_i, n_i, e \geq 0, \ i = 1, 2,$
where $a_i^* = \min(k_i, d_i); x_i^* = \min(n_i, d_i); y_1^* = \min(\min(d_1, k_1) - \min(z_1, d_1), e); y_2^* = \min(\min(d_2, k_2) - \min(z_2, d_2), e - y_1^*); z_i^* = \min(\min(d_i, k_i) - \min(z_i, d_i), e);$ are optimal capacity, dedicated nurse, and float nurse allocations obtained after demand realization. For simultaneous decision making, the optimal stage-2 allocations can be summarized as follows: (i) demand for each floor is accommodated until all beds on that floor are filled; (ii) dedicated nurses are used, as needed, to care for patients on their unit—unused dedicated nurses are idle; (iii) float nurses are first used in unit 1 to cover the shortfall between patients and dedicated staff; and (iv) any remaining float nurses are used in unit 2 to cover the shortfall between patients and dedicated staff. Second-order conditions prove that the objective function is convex in its decision variables $k_i, n_i,$ and $e.$ Solving for $k_i$, using first-order conditions, we get Equation (4). Optimal $n_i$ and $e$ are obtained by solving first-order conditions (refer to Technical Report).

$$
\Phi_i(k_i^*) = \frac{p_i - f_i - s_i}{p_i - s_i}.
$$

When a solution obtained by solving first-order conditions indicates that $n_i^* + e^* > k_i^*$, the optimal solution comes from constraining $n_i + e = k_i$ in the objective function and solving for $k_i$ and $n_i.$ This scenario will occur when the cost of contract nurses is large in relation to the cost of internal resources. Optimal capacity given in Equation (4) is equal to the optimal capacity for no flexibility case under simultaneous decision making (refer to Equation [2]). The first-order conditions for $n_i$ and $e$ (refer to Technical Report) show that the staffing decisions including staffing flexibility decisions are independent of capacity decisions even when these decisions are made simultaneously. Additional results comparing optimal decisions across different configurations are given in the section titled “Analytical Results and Discussion.”

**Configuration 4: Demand Upgrades and Staffing Flexibility**

Configuration 4 has the highest level of flexibility among our models. It uses both demand upgrades and staffing flexibility. The timeline for decision making for configuration 4 is exactly the same as in configuration 3 except that capacity decisions include the possibility for demand upgrades. The three types of decision making for configuration 4 are shown in Figure 4.

**Sequential decision making—case 1**

In sequential decision making case 1, there are two periods. In the first period, both optimal capacity and total staff are determined, assuming demand upgrades. In the second period, given capacity and total staff from period 1, the optimal number of float nurses required to meet demand at minimum cost is determined. Period 1 sequential decision making under configuration 4 is the same model as configuration 2 simultaneous decision making. Therefore, the optimal capacity ($k_i^*$) and staffing ($z_i^*$) for this case are the same optimal values obtained from simultaneous decision making in configuration 2. Staffing flexibility is determined in period 2. Given the values for $k_i$ and $z_i$, we determine the optimal number of float nurses ($e_i$) in period 2. The objective function is convex, so first-order
conditions are necessary and sufficient (refer to Technical Report) to obtain the optimal number of float nurses \((e^*_i)\).

**Sequential decision making—case 2**

Sequential decision making case 2 has two periods. In the first period, capacity is determined assuming demand upgrades in period 2, while in the second period, dedicated and float staff are determined. Optimal capacity \((k^*_i)\) under period 1 is the same as optimal capacity in configuration 2 under sequential decision making. In the second period, given \(k^*_i\) from period 1, we determine the optimal number of dedicated staff, \(n_i\) and float staff, \(e\). Convexity of the period 2 objective function makes first-order conditions necessary and sufficient to obtain the optimal number of dedicated nurses \((n^*_i)\) and float nurses \((e^*)\).

**Simultaneous decision making**

Under simultaneous decision making, all capacity and staffing decisions are made simultaneously assuming demand upgrades and staffing flexibility.

\[
\begin{align*}
\text{Min} \quad & \sum_i (f_i \cdot k_i + h_i \cdot n_i) + t \cdot e + E \cdot [s_i \cdot (a^*_i + a^*_u - x^*_1 - y^*_1)] \\
& + s_2 \cdot (a^*_1 - x^*_2 - y^*_2) + p_1 \cdot (d_1 - a^*_1) + p_2 \cdot (d_2 - a^*_2 - a^*_u)]
\end{align*}
\]

subject to \(n_i + e \leq k_i; k_i, n_i, e \geq 0, i = 1, 2,\)

where \(a^*_i = \min(k_i, d_i); a^*_u = \min(k_1 - \min(k_1, d_1), d_2 - \min(k_2, d_2)); x^*_1 = \min(n_i, d_i); y^*_1 = \min(\min(d_1, k_1) - \min(z_1, d_1), e); y^*_2 = \min(\min(d_2, k_2) - \min(z_2, d_2), e - y^*_1)\) are optimal second-stage capacity, dedicated nurse, and float nurse allocations. For simultaneous decision making, stage-2 allocations can be summarized as follows: (i) patients are admitted to their desired unit as long as beds are available; any excess beds in unit 1 are used to accommodate overflow patients from unit 2 as needed; (ii) dedicated nurses are used to care for traditional patients on their unit, any excess dedicated nurses are idle; (iii) float nurses are used in unit 1 to care for any upgraded patients from floor 2 and/or traditional patients who were not assigned a dedicated nurse; (iv) any remaining float nurses are used in unit 2 to care for patients not assigned a dedicated unit 2 nurse; and (v) contract nurses are assigned to meet excess patient demand. The first-stage objective function is solved by substituting the second-stage allocations in the first-stage objective. Analysis of second-order conditions reveals that we can conclude analytically that the staffing levels \((n_1, n_2, e)\) are unique for a given set of capacity decisions \((k_1, k_2)\). For each capacity level, first-order conditions for \(n_1, n_2,\) and \(e\) are used to determine optimal dedicated and float nurses for simultaneous decision making. A search of costs associated with each capacity level generated optimal capacity decisions for simultaneous decision making.

**Analytical Results and Discussion**

Analytical results obtained by comparing different configurations and timelines of decision making is presented in this section while design of a multi-factorial
numerical experiment and results and discussion are presented in the section titled “Numerical Experiments.”

Optimal capacity and staffing levels
We prove that the first-stage objective function of sequential decision making for all configurations and the simultaneous case for no flexibility and staffing flexibility are convex in their capacity and staffing decision variables. We used first-order conditions, deriving closed form expressions where possible to determine optimal capacity and staffing levels for these cases. Under simultaneous decision making in demand upgrades and full flexibility, our numerical analysis suggests that the objective function is convex in both capacity and staffing decisions, but analytically we could only prove that given a capacity level, the objective function is convex in staffing decisions. Hospital administrators can easily determine optimal values for capacity, total staff, and/or float nurse once relevant costs are determined.

Simultaneous versus sequential decision making
Simultaneous decision making is the least-constrained model. Therefore, simultaneous decision making yields lower costs than sequential decision making (for percentage cost improvement of simultaneous decision making over sequential decision making, refer to the section titled “No flexibility versus staffing flexibility versus demand upgrades”). For the no flexibility case, optimal staffing decisions ($z_i$) are the same for simultaneous and sequential decision making, but optimal capacity ($k_i$) in units 1 and 2 under sequential decision making is higher than optimal capacity in units 1 and 2 under simultaneous decision making. Therefore, suboptimality of sequential decision making is because of suboptimal capacity decisions rather than staffing decisions. The demand upgrade configuration also shows that total staffing under simultaneous decision making and sequential decision making are equal. Therefore, the higher cost for the most constrained sequential decision making case is caused by suboptimal capacity decisions rather than staffing decisions. Fewer beds under simultaneous decision making leads to lower cost as compared to sequential decision making. Under staffing flexibility, sequential case 2 is the most constrained system, with capacity, staffing, and staffing flexibility decisions made separately. From closed-form solutions we find that the optimal number of beds in units 1 and 2 are the same for both sequential case 1 and simultaneous decision making, while the optimal number of beds under sequential case 2 is higher. Staffing decisions are the same for sequential case 2 and simultaneous decision making indicating that the higher cost incurred by sequential case 2 is due to suboptimal capacity decisions made in period 1 rather than staffing decisions.

From this discussion we infer that, for any type of flexibility, making staffing and capacity decisions separately (sequential case 2) allows both units to have more capacity than making staffing and capacity decisions simultaneously (sequential case 1 and simultaneous). Making early capacity decision is the main cause of suboptimality for sequential decision making case 2 compared to simultaneous decision making. This helps hospital managers understand that making capacity decisions separate from total staff and/or float nurses (decentralized decision
making), will allow hospitals to end up with more capacity than needed to meet patient requirements at minimum cost.

We also find that for staffing flexibility configuration, cost of contract nurses has significant impact on capacity and staffing decisions. Sequential decision making case 2 has \( \frac{f_2 - s_i}{p_i - s_i} \) more beds than simultaneous decision making, indicating that the difference in capacity \( k_i \) will be higher when contract nurse cost \( s_i \) is high, penalty cost \( p_i \) is low, or operating cost \( f_i \) is high. Deciding for internal resources (capacity, total staff, and/or float nurse) simultaneously (centralized decision making) becomes vital under staffing flexibility, especially when external resources are expensive (i.e., contract nurse cost is high).

Optimal staffing decisions for no flexibility and staffing flexibility are the same under sequential and simultaneous decision making and independent of capacity decisions. This indicates that there is no interaction between staffing and capacity decision for no flexibility and staffing flexibility configurations but integrating capacity and staffing decisions yield higher benefits.

**Comparison between configurations**

Due to the complexity of first-order conditions when using a general demand distribution, we analytically proved the following results for the case when patient demand is distributed uniformly between 0 and \( b_i \) for unit \( i \).

**Capacity decisions:** Optimal capacity decisions for unit 1 and unit 2 are the same for no flexibility and staffing flexibility configurations. We find that capacity in unit 1 \( k_1 \) is higher in demand upgrade configuration compared to staffing flexibility configuration when \( b_2 \cdot p_2 \geq 1 \) (easily proved using closed-form equations) under sequential decision making. The capacity in unit 1 under sequential decision making for the no flexibility and staffing flexibility configurations is higher than under simultaneous decision making in no flexibility and staffing flexibility cases, but lower than in sequential decision making in the demand upgrade configuration.

**Staffing decisions:** Under sequential and simultaneous decision making, optimal staffing in unit 1 for demand upgrades is higher than optimal staffing for the no-flexibility configuration. Optimal staffing for unit 2 is the same for no-flexibility and demand upgrades configurations and numerically we find the optimal number of beds in unit 2 for demand upgrades is less than the optimal number of beds in unit 2 for the no-flexibility configuration.

**NUMERICAL EXPERIMENTS**

In the previous section, we formulated and analyzed 10 models with different types of flexibility and decision-making timelines. We proved convexity for most models and identified first-order conditions for optimal decisions under a general demand distribution. In this section, we choose parameter values for costs, conduct numerical analysis, and discuss additional results obtained through numerical experimentation. The purpose of numerical experimentation is two-fold: to compare
Table 1: Values of cost parameters.

<table>
<thead>
<tr>
<th>Cost Parameter Levels</th>
<th>s₁ ($/hour)</th>
<th>s₂ ($/hour)</th>
<th>f₁ ($/hour)</th>
<th>f₂ ($/hour)</th>
<th>p₁ ($/hour)</th>
<th>p₂ ($/hour)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>30</td>
<td>25</td>
<td>9</td>
<td>5</td>
<td>51</td>
<td>45</td>
</tr>
<tr>
<td>High</td>
<td>35</td>
<td>28</td>
<td>12</td>
<td>7</td>
<td>60</td>
<td>50</td>
</tr>
</tbody>
</table>

all models under a common structure and derive insights that are hard to prove analytically.

Parameter Values
To fully understand the effect of flexibility, decision timing, and parameter values on resource decisions and performance of a multi-floor hospital unit, we developed a full-factorial scenario analysis for optimal policies. We consider two levels, a low-cost level and a high-cost level for each cost parameter as shown in Table 1, to evaluate the effect of varying costs between two units in a hospital.

Staffing cost parameters
UNC Hospital is an academic medical center with 726 licensed beds. We systematically interviewed nurses and directors in the nurse staffing and planning departments at UNC Hospital to get estimates for wages of dedicated and float nurses, as well as fixed and operating costs of bed capacity. Registered nurses (RNs), who are represented as dedicated nurses in our model, receive an hourly wage between $20 and $24 (excluding benefits) at UNC Hospital. We chose RNs to represent dedicated and float nurses because Licensed Practical Nurses (LPNs) and Nursing Assistants (NAs) are not frequently floated. Contract nurses typically receive between $28/hour and $37/hour (excluding benefits) depending on type of unit. In order to capture the cost difference between dissimilar hospital units, we assumed higher wages, higher contract nurse cost, and higher fixed cost for the sophisticated unit (unit 1 in our model) and lower costs for the simple unit (unit 2 in our model). Hourly wages for dedicated nurses in unit 1 and unit 2 are used as a base cost parameter and hence we selected only one set of cost values for $h₁ and $h₂. For our analysis, wages for dedicated nurses are $22/hour for unit 1 and $20/hour for unit 2. Float nurses at UNC Hospital receive additional training with a preceptor before floating to a new department and sometimes receive higher wages than dedicated nurses. In our model, we assumed wages paid for nurses cross-trained to work in unit 1 as $24/hour and unit 2 as $22/hour. Contract nurse cost ($s_i$) is given in Table 1.

Capacity cost parameters
UNC recently built a 50-bed cancer hospital and expanded capacity on the adult floor. This enabled us to get an estimate of fixed operating cost of a single bed space. The initial investment cost, which includes construction of the building and buying/leasing of equipment, is a fixed cost and is approximately $1,000,000 per
Table 2: Values for demand distribution.

<table>
<thead>
<tr>
<th>Demand Parameters</th>
<th>Distribution 1</th>
<th>Distribution 2</th>
<th>Distribution 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unit 1</td>
<td>[0,20]</td>
<td>[0,20]</td>
<td>[0,25]</td>
</tr>
<tr>
<td>Unit 2</td>
<td>[0,20]</td>
<td>[0,25]</td>
<td>[0,20]</td>
</tr>
</tbody>
</table>

bed. The daily operating cost is a variable cost for maintaining the bed. The daily operating charges include any specialty equipment that is leased (e.g., pumps, cylinders) and the cost of supplies (e.g., tubes, hoses). In our model, we amortized the fixed cost (investment cost) into hourly costs and added hourly operating costs to determine values for \( f_i \). Depending on the type of floor, operating costs per bed varies, so we considered a high cost and a low cost for \( f_i \) as shown in Table 1.

Demand distribution

Demand is assumed to follow a uniform distribution in both unit 1 and unit 2. As shown in Table 2, three demand scenarios are used for demand distribution in the numerical analysis. In each case, the lower bound of the uniform distribution is set to zero. (A lower bound that is greater than zero would add a fixed cost to each configuration in our model.) Distribution 1 models two units facing the same demand, distribution 2 assumes higher demand in unit 2, and distribution 3 captures higher demand in unit 1. A full-factorial experiment for six parameters with two cost levels (low and high) yields 64 different scenarios for each of the three demand distributions, resulting in a total of 192 scenarios for each of the 10 models.

Numerical Insights and Discussion

This section presents the summary of results and explains the insights derived from numerical analysis.

Simultaneous versus sequential decision making

This section compares simultaneous and sequential decision making for the four configurations. The percentage cost improvement (cost savings) for the timing alternatives under the 192 scenarios is shown in Table 3.

Table 3 shows maximum, average, and minimum percentage cost improvement while implementing different timelines of decision making across the four configurations. All 192 scenarios show the same cost improvement relationship for the different decision making timelines. The positive value for minimum percentage cost improvement indicates that cost improvement is in the same direction for all 192 scenarios. In sequential case 1, though suboptimal compared to simultaneous decision making, most of the benefit is obtained by the integrated staffing and capacity optimization as observed by marginal but significant percentage cost improvement of sequential case 1 over simultaneous decision making in Table 3. This implies that the benefits of cross-training can be largely realized even if capacity and staffing levels have been determined prior to the establishment of a
**Table 3:** Sequential versus simultaneous decision making: Maximum/average/minimum percentage cost improvement.

<table>
<thead>
<tr>
<th>Percentage Cost Improvement</th>
<th>Configurations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Simultaneous versus</td>
<td>–</td>
</tr>
<tr>
<td>sequential case 1</td>
<td></td>
</tr>
<tr>
<td>Simultaneous versus</td>
<td>6.0%/2.88%/</td>
</tr>
<tr>
<td>sequential case 2</td>
<td>1.15%</td>
</tr>
<tr>
<td>Sequential case 1 versus</td>
<td>–</td>
</tr>
<tr>
<td>sequential case 2</td>
<td></td>
</tr>
</tbody>
</table>

**Table 4:** Sequential case 2 versus simultaneous decision making: Full flexibility (cost savings in dollars per hour and dollars per year).

<table>
<thead>
<tr>
<th>Cost Savings ($)</th>
<th>Distribution 1</th>
<th>Distribution 2</th>
<th>Distribution 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum cost savings per hour</td>
<td>$50.06</td>
<td>$41.73</td>
<td>$77.94</td>
</tr>
<tr>
<td>Average cost savings per hour</td>
<td>$24.83</td>
<td>$24.53</td>
<td>$32.59</td>
</tr>
<tr>
<td>Minimum cost savings per hour</td>
<td>$7.99</td>
<td>$8.27</td>
<td>$9.73</td>
</tr>
<tr>
<td>Maximum cost savings per year</td>
<td>$438,490.60</td>
<td>$365,519.80</td>
<td>$682,728.10</td>
</tr>
<tr>
<td>Average cost savings per year</td>
<td>$217,455.90</td>
<td>$214,955.40</td>
<td>$285,498.70</td>
</tr>
<tr>
<td>Minimum cost savings per year</td>
<td>$69,966.12</td>
<td>$72,427.68</td>
<td>$85,243.56</td>
</tr>
</tbody>
</table>

*cross-training initiative.* If hospital managers are unable to decide for all resources (capacity, total staff, and/or float nurses) at the same time, they should at least decide for capacity and total staff at the same time and decide for float nurses when needed. This way most of the benefit due to integration will be achieved. The percentage cost improvement in integrating capacity and staffing decisions (simultaneous vs. sequential case 2) is around 2.84% (2.88%, 2.87%, 2.86%, 2.75% are the average values from Table 3), irrespective of the level of flexibility, which indicates that there is limited interaction effect between the type of flexibility and the timing of decision making.

In order to better understand the importance of the percentage cost improvement values, Table 4 presents yearly cost savings in dollar amount when making simultaneous decision making instead of sequential decision making case 2 under full flexibility. The highest benefit that simultaneous decision making yields over sequential case 2 is the scenario where capacity cost in unit 2 ($f_2$) and penalty cost in unit 2 ($p_2$) is low and the rest of the cost parameters are high, irrespective of demand distribution. When flexibility is extended to more than two units, cost savings can be expected to be even greater (analysis of more than two units is beyond the scope of this article). The minimum benefit is seen when the capacity cost in unit 1 ($f_1$) and the penalty cost for both units ($p_1, p_2$) are high and the rest of the cost parameters are low. Under this set of parameters, hospitals choose to have more dedicated capacity and staff in order to avoid turning away of patients, and therefore take little advantage of flexibility.
No flexibility versus staffing flexibility versus demand upgrades

The maximum, average, and minimum percentage cost improvement under sequential decision making cases 1 and 2 when shifting from one configuration to another is shown in Tables 5 and 6, respectively. Table 7 shows the benefit of added flexibility under simultaneous decision making. A positive minimum percentage cost improvement indicates that cost improvement for all 192 scenarios is in the same direction.

The average percentage cost improvement under both sequential case 2 decision making (Table 6) and simultaneous decision making (Table 7) show that the benefit of using staffing flexibility over no flexibility (base case) is greater than using demand upgrades over no flexibility (base case). Under sequential case 2 decision making, the average percentage cost improvement staffing flexibility over base case is 1.94% while average percentage cost improvement for demand upgrades over the base case is 1.39%. Under simultaneous decision making the average percentage cost improvement for staffing flexibility over the base case is 1.83%, as compared to a 1.38% increase from demand upgrades over the base case. 67.7% of scenarios show the benefit of staffing flexibility is greater than the benefit of demand upgrades. The intuition behind staffing flexibility being more valuable

Table 5: Sequential decision making case 1: Maximum/average/minimum percentage cost improvement.

<table>
<thead>
<tr>
<th>From/To</th>
<th>Configuration 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Configuration 3</td>
<td>3.39%/1.56%/0.42%</td>
</tr>
</tbody>
</table>

Table 6: Sequential decision making case 2: Maximum/average/minimum percentage cost improvement.

<table>
<thead>
<tr>
<th>From/To</th>
<th>Configuration 1</th>
<th>Configuration 2</th>
<th>Configuration 3</th>
<th>Configuration 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.47%/1.39%/1.87%</td>
<td>2.95%/1.94%/1.13%</td>
<td>5.72%/3.78%/2.14%</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2.68%/0.55%/-2.36%</td>
<td>4.37%/2.37%/1.27%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>3.67%/1.94%/0.62%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 7: Simultaneous decision making: Maximum/average/minimum percentage cost improvement.

<table>
<thead>
<tr>
<th>From/To</th>
<th>Configuration 1</th>
<th>Configuration 2</th>
<th>Configuration 3</th>
<th>Configuration 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.50%/1.38%/0.09%</td>
<td>3.00%/1.83%/1.15%</td>
<td>5.54%/3.42%/1.92%</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2.20%/0.45%/-2.65%</td>
<td>4.15%/2.06%/0.47%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>3.14%/1.61%/0.65%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
than demand upgrades is as follows: once capacity is allocated to meet demand, staffing needs are met by any means, even if it leads to using contract nurses and incurring higher staffing costs, which reduces the benefit of demand upgrades. Comparing the two types of flexibility, staffing flexibility provides greater benefit in the majority of instances compared to demand upgrades.

A system that has both flexibilities (configuration 4) yields a greater benefit than the sum of the benefits from using staffing flexibility and demand upgrades separately. Under sequential decision making (case 2), the additive benefit of using staffing flexibility and demand upgrade is 3.33%, while the benefit of using staffing flexibility and demand upgrade in a single system (configuration 4) is 3.78%. Similarly, under simultaneous decision making, the additive benefit is 3.21% while the benefit of using both flexibilities (configuration 4) is 3.42%. Staffing flexibility and demand upgrades have slight positive interaction. Thus, the two types of flexibility complement each other. The benefit to a hospital is greater when planning for both types of flexibility at the same time rather than using one type of flexibility now and adding another level of flexibility later.

In order to better understand the importance of percentage cost improvement values, Tables 8 and 9 show yearly cost savings if the hospital used full flexibility instead of the base case under sequential and simultaneous decision making, respectively. It also shows cost savings under three different demand distribution parameters. The maximum cost savings if full flexibility is used instead of the base case is $425,429 under sequential decision making. This is achieved for the scenario where contract nurse cost of both units ($s_1$ and $s_2$) and capacity cost in unit 2 ($f_2$) is high and rest of the cost parameters are low. The minimum cost savings

<table>
<thead>
<tr>
<th>Table 8: Configuration 1 to configuration 4: Sequential decision making (cost savings in dollars per hour and dollars per year).</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cost Savings ($)</td>
</tr>
<tr>
<td>Maximum cost savings per hour</td>
</tr>
<tr>
<td>Average cost savings per hour</td>
</tr>
<tr>
<td>Minimum cost savings per hour</td>
</tr>
<tr>
<td>Maximum cost savings per year</td>
</tr>
<tr>
<td>Average cost savings per year</td>
</tr>
<tr>
<td>Minimum cost savings per year</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 9: Configuration 1 to configuration 4: Simultaneous decision making (cost savings in dollars per hour and dollars per year).</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cost Savings ($)</td>
</tr>
<tr>
<td>Maximum cost savings per hour</td>
</tr>
<tr>
<td>Average cost savings per hour</td>
</tr>
<tr>
<td>Minimum cost savings per hour</td>
</tr>
<tr>
<td>Maximum cost savings per year</td>
</tr>
<tr>
<td>Average cost savings per year</td>
</tr>
<tr>
<td>Minimum cost savings per year</td>
</tr>
</tbody>
</table>
under sequential decision making occur under the opposite scenario, where $s_1$, $s_2$, and $f_2$ are low and rest of the cost parameters are high. The maximum cost savings under simultaneous decision making is $404,870 for the scenario where $s_1$, $f_2$, and $p_1$ are high and rest of the cost parameters are low. When contract nurse cost in unit 1 is high, staffing flexibility helps alleviate the staffing costs and when operating cost in unit 2 is high, demand upgrades helps to alleviate cost of turning away patients. The minimum cost savings is obtained for the reverse scenario where $s_2$, $f_1$, and $p_2$ are high and rest of the cost parameters are low.

Comparing staffing and capacity decisions among the four configurations for all 192 scenarios, the following observations can be made.

(i) Optimal capacity in unit 1 under simultaneous decision making is less than optimal capacity in unit 1 under sequential decision making, irrespective of type of flexibility (this result is proved analytically in the section titled “Analytical Results and Discussion” for configurations 1 and 3, and demonstrated numerically for all 192 scenarios in configurations 2 and 4).

(ii) Optimal capacity in unit 2 is lower in demand upgrade configuration than in no flexibility or staffing flexibility configurations. The possibility of upgrading patients from unit 2 to unit 1 lowers capacity in unit 2 (observed for all 192 scenarios).

(iii) Total available staff ($n_1 + n_2 + e$) in staffing flexibility configuration is greater than total available staff ($z_1 + z_2$) in no flexibility or demand upgrade configuration (observed for all 192 scenarios).

(iv) Both dedicated staff in unit 1 ($n_1$) and float staff ($e$) are higher under simultaneous decision making compared to sequential case 1 for staffing flexibility configuration (observed for all 192 scenarios). Although the number of staff hired is higher in simultaneous than sequential case 1 (number of beds is the same for both types of decision making, which is proved analytically), the overall cost is less for simultaneous decision making compared to sequential case 1 decision making. Therefore, the higher cost for sequential case 1 is attributed to suboptimal staffing decisions in periods 1 and 2 rather than capacity decisions, while higher cost of sequential case 2 is attributed to suboptimal capacity decisions rather than staffing decisions as proved in the section titled “Analytical Results and Discussion.”

**Implications of demand**

For each of the three demand distributions (when demand in unit 1 and unit 2 are equal, demand in unit 1 is high, demand in unit 2 is high), we compare the benefits of using staffing flexibility over no flexibility (base case) for all scenarios (64 in total) generated by all cost parameters. The benefit of staffing flexibility is highest when demand in unit 1 is more than demand in unit 2. Higher demand in unit 1 leads to higher utilization of float nurses because float nurses have priority allocation to unit 1. When demand in unit 2 is more than demand in unit 1, the benefit of staffing flexibility is least, but still significant. Numerical analysis for demand upgrade over no flexibility showed that benefit of demand upgrades is
highest when demand in unit 2 is more than demand in unit 1 (for all 64 scenarios). When there is higher demand in unit 2 than in unit 1, upgrading patients from unit 2 to unit 1 is helpful. When demand in unit 1 is more than demand in unit 2, the benefit of upgrade is least, but significant.

**Effect of cost parameters on capacity and staffing decisions**

In this section, we analyze the trends exhibited by capacity and staffing variables when cost parameters vary under four configurations in simultaneous and sequential decision making. The benefit of using staffing flexibility is higher when the cost of contract nurses in unit 1 \( (s_1) \) is high, irrespective of values assumed for other cost parameters. This is because unit 1, being a complex unit, gets priority over unit 2 to utilize float nurses. Under demand upgrades, the benefit of demand upgrades is highest when the contract nurse cost in unit 2 \( (s_2) \) is high, irrespective of values taken by other cost parameters.

Under simultaneous decision making and demand upgrades, as penalty cost in unit 1 increases, the optimal capacity and dedicated staff initially increases in unit 1 as seen in Figure 5. After a threshold value of penalty cost, capacity in unit 1 keeps increasing while dedicated staff in unit 1 decreases. This phenomenon is observed because at high penalty cost, capacity in unit 1 is quite high and capacity in unit 2 is relatively low leading to greater possibility of upgrades to unit 1. More patient upgrades imply higher use of contract nurses who can treat upgraded patients in unit 1. The benefit of using contract nurses in unit 1 for both unit 1 patients and upgraded patients is greater than the benefit of increasing (permanent) dedicated staff in unit 1 for unit 1 patients and hiring contract nurses for upgraded patients. Therefore, dedicated staff in unit 1 decreases at high penalty cost.

Under staffing flexibility, the effect of cost parameters on capacity and staffing decisions under simultaneous decision making is summarized in Table 10. The table indicates that the level of staffing flexibility changes with the capacity in unit 1. Flexibility \( (e^\ast) \) is more sensitive to a change in capacity of unit 1 than capacity of unit 2 because of priority assignment of float nurses to unit 1.

**Figure 5:** Configuration 2 simultaneous decision making: Effect of increase in penalty cost in unit 1.
Table 10: Effect of cost parameters on the decision variables in configuration 3 under simultaneous decision making.

<table>
<thead>
<tr>
<th>Parameter Increase</th>
<th>$k_1^*$</th>
<th>$k_2^*$</th>
<th>$n_1^*$</th>
<th>$n_2^*$</th>
<th>$e^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_1$</td>
<td>↑</td>
<td>↔</td>
<td>↑</td>
<td>↓</td>
<td>↑</td>
</tr>
<tr>
<td>$p_2$</td>
<td>↔</td>
<td>↑</td>
<td>↔</td>
<td>↔</td>
<td>↔</td>
</tr>
<tr>
<td>$f_1$</td>
<td>↓</td>
<td>↔</td>
<td>↓</td>
<td>↑</td>
<td>↓</td>
</tr>
<tr>
<td>$f_2$</td>
<td>↔</td>
<td>↓</td>
<td>↔</td>
<td>↔</td>
<td>↔</td>
</tr>
<tr>
<td>$s_1$</td>
<td>↓</td>
<td>↔</td>
<td>↑</td>
<td>↓</td>
<td>↑</td>
</tr>
<tr>
<td>$s_2$</td>
<td>↓</td>
<td>↓</td>
<td>↑</td>
<td>↑</td>
<td>↓</td>
</tr>
</tbody>
</table>

Implementation Issues

This subsection discusses several issues faced by managers who use insights from our model to help guide their decision making. The demand in our model is expressed in terms of full-time equivalent nurses required within a particular skill category. Our model should be applied to one skill category (e.g., RNs) at a time. Most hospitals float both RNs and NAs, and also use per diem RNs and NAs, so our staffing model provides insight for both RN staffing and NA staffing. The percentage of LPNs in most hospitals is generally small enough that they are not cross-trained or hired as contract nurses, but if hospitals do float LPNs our staffing model could also guide this decision.

Many hospitals use FTEs per shift as a base unit to determine daily staffing needs. Past staffing data can be used to fit a demand distribution and then used in our model. Managers should be careful to use budgeted staffing data, and not the actual staffing, because the unit may be under or over staffed and only budgeted staffing reflects the actual daily demand. Also, other staffing systems might use nursing hours instead of FTEs and so careful understanding of the measurement units for staffing is required. As nursing managers move closer toward demand realization, they possess additional information about the demand distributions and the extent of demand uncertainty. Our model assumes that demand is independently and identically distributed throughout the planning horizon.

Bed allocation in many hospitals is based on the channel through which patients arrive. Patients are allocated beds under the following channel priority system at UNC Hospital: (i) emergency room (ER) patients who need immediate assistance after triage, (ii) clinic visits, (iii) transfer patients from other hospitals, and (iv) direct admissions. When beds are not available, some patients are allowed to wait, while other patients may remain at or be diverted to another hospital. For demand upgrade flexibility in our model, we considered two extremes of satisfying patient demand. First, when beds are not available, patients are upgraded to similar units within the same specialization typically in lower acuity units. Second, when beds are not available in both units, patients are diverted to another hospital. Depicting these two scenarios, we assume independent demand in our model. However, when beds are unavailable even after considering upgrades, hospitals sometimes place patients in the holdout area or ER for up to 24 hours. Patients in the holdout area or ER waiting for beds to become available cause additional
cost to be incurred by the hospital. These patients usually are admitted within 24 hours. Demand between successive days is slightly correlated due to the fact that patients in the holdout or ER are eventually admitted. We do not consider this scenario in our model. Because our model uses flexible nurses and demand upgrades between similar units in the same specialization, our model has a better applicability to lower acuity units than intensive care unit (ICU) or step down units. When applying our model to ICU or step down units, managers should be careful to check if nurses are flexible enough to float between these units.

Within a hospital, there are typically different levels of flexible nurses, including float nurses who are least flexible (floats within a specialization), flex team nurses, travel nurses, and contract nurses who are highly flexible (floats between specialization if required). In our model, we consider two extreme types of flexible nurses used to meet short-term staffing needs in hospitals: in-house nurses who are trained to float between similar units under a specialization and contract nurses who are hired at short notice for a shift or two to any unit in any specialization. Our model determines the optimal number of float nurses to train, considering the fact that short-term variability in staffing needs is met by float nurses and contract nurses within a specialization. Nurses floating between specialization are not considered. While implementing our model, managers should understand that we assume the total staff pool is constant over a period of time. We do not account for long-term absenteeism, such as maternity leave or unfilled positions. The change in nurse supply arising from long-term absenteeism is usually filled by travel nurses or flex team nurses.

CONCLUSIONS

In this article, we consider two types of flexibility used to coordinate two important resources (beds and nursing staff) and satisfy stochastic demand at minimum cost. We analyze four flexibility configurations (no flexibility, staffing flexibility, demand upgrades, and full flexibility) under simultaneous decision making and sequential decision making. We prove convexity of the objective cost function for all models, and for each flexibility configuration/decision making timeline we determine optimal capacity and staffing decisions. Hospital managers can easily determine optimal staffing and capacity decisions after estimating relevant costs.

Centralized decision making (i.e., deciding on capacity, total staff, and/or float nurses simultaneously) yields greater benefit than decentralized decision making (i.e., making capacity and staffing decisions separately), as expected. The higher cost of sequential case 1 decision making compared to simultaneous decision making is due to suboptimal staffing decisions (rather than capacity decisions), while the higher cost of sequential case 2 compared to simultaneous decision making is because of suboptimal capacity decisions (rather than staffing decisions) for any type of flexibility. When hospital managers plan for internal resources (capacity and total staff) separately from float nurses, the hospital ends up with more staff than needed to meet patient demand at minimum cost. When capacity decisions are made separately from total staff and/or float nurses (decentralized), the hospital ends up with more capacity than needed to meet patient requirements at minimum cost.
Simultaneous decision making of capacity, staffing, and flexibility offers only a small improvement over a system where flexibility decisions are made later. Thus, the benefits of cross-training can be largely realized even if capacity and staffing levels have been determined prior to the establishment of a cross-training initiative. If managers are unable to decide for all resources (capacity, total staff, and/or float nurses) at the same time, they should at least decide for capacity and total staff at the same time, and then decide for float nurses when needed. This way most of the benefit due to integration will be achieved.

The two types of flexibility, demand upgrades and staffing flexibility, have a positive interaction effect between them. Hospital managers can reap the greatest benefit when they plan for both types of flexibility at the same time, rather than using one type of flexibility now and adding another level of flexibility later. The benefit of using full flexibility instead of no flexibility is consistent across all cost and demand parameter values, while the benefit of using simultaneous decision making over sequential decision making is approximately the same across different flexibility scenarios.

On average, the benefit of using staffing flexibility is greater than the benefit of using demand upgrades. The timing of decisions and type of flexibility have limited interaction; their effect on system performance (measured in terms of cost savings) is largely independent.

Hospital managers should be aware that staffing flexibility yields greater benefit when demand in the more sophisticated unit is higher than demand in the simple unit, while demand upgrades yield greater benefit when demand in the more sophisticated unit is less than the simple unit. When the cost of external resources (contract nurse) for the more sophisticated unit is high, staffing flexibility yields greater benefit than demand upgrades. When the cost of external resources for the simple unit is high, demand upgrades yield greater benefit than staffing flexibility.

In this article, we explore the benefits and trade-offs in employing different types of flexibility while coordinating bed spaces and nursing staff. The results shown in this article will help managers not only understand the advantages of resource flexibility, but also understand the impact of system costs on flexibility decisions and benefits. It will also help hospital administrators recognize the importance of including staffing and capacity flexibility in the planning phases, rather than only using it as an ad hoc mechanism to meet daily demand. [Received: February 2007. Accepted: March 2009.]

REFERENCES


**Adelina Gnanlet** is an assistant professor at Mihaylo College of Business and Economics at California State University, Fullerton. She received her PhD in
operations management from the Kenan-Flagler Business School at University of North Carolina at Chapel Hill. She earned an MBA in operations and systems from IIT Chennai, India, and holds a BE in Industrial Engineering from Guindy Engineering College, Anna University, India. Her research interests include impact of productivity, use of resource flexibility, and measurement of productivity in health care systems, use of meta-heuristics in forecasting, and impact of foreign VCs on health care firms.

**Wendell G. Gilland** (PhD, Stanford University) is an associate professor at the Kenan-Flagler Business School at University of North Carolina. His research focuses on analyzing the ability of operational flexibility, such as substitutable products, flexible resources, and priority pricing, to meet diverse customer needs. His research has been published in *Operations Research, Naval Research Logistics, Production and Operations Management*, and the *Journal of Operations Management*. 
This document is a scanned copy of a printed document. No warranty is given about the accuracy of the copy. Users should refer to the original published version of the material.