Manufacturing & Service Operations Management

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

Managing Hospital Inpatient Bed Capacity Through Partitioning Care into Focused Wings

Thomas J. Best, Burhaneddin Sandıkçı, Donald D. Eisenstein, David O. Meltzer

MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 17, No. 2, Spring 2015, pp. 157–176
ISSN 1523-4614 (print) | ISSN 1526-5498 (online)

http://dx.doi.org/10.1287/msom.2014.0516

Managing Hospital Inpatient Bed Capacity Through Partitioning Care into Focused Wings

Thomas J. Best, Burhaneddin Sandıkçı, Donald D. Eisenstein
University of Chicago Booth School of Business, Chicago, Illinois 60637
{thomas.j.best@chicagobooth.edu, burhan@chicagobooth.edu, don.eisenstein@chicagobooth.edu}

David O. Meltzer
University of Chicago Medical Center, Chicago, Illinois 60637, dmeltzer@medicine.bsd.uchicago.edu

We consider the partitioning of care types into wings from the perspective of a hospital administrator who wishes to optimize the use of a fixed number of beds that provide services for heterogeneous care types. The hospital administrator decides on the number of wings to form, the number of beds to allocate to each wing, and the set of care types to assign to each wing to maximize the total utility to the hospital. The administrator faces an inherent trade-off between forming large wings to pool demand and bed capacity, and forming specialized wings to focus on narrow ranges of care types. Specialized wings not only provide advantages from focused care but also allow the protection of beds for high-utility care types. We provide an optimization model for the wing formation decision and address the advantages of focus endogenously in our model. Using data from a large urban teaching hospital in the United States along with a national database, we report on a number of managerial insights. In particular, as the overall demand increases across all care types, wings are formed to reserve more beds for higher-utility types, which leads to higher overall hospital utility but also some disparity across types, such as increased hospital access for some and decreased access for others. Furthermore, overall bed occupancy decreases as the hospital is split into wings. However, if sufficient focus is attained, shorter lengths-of-stay associated with focused care may increase overall patient throughput. We also observe that when patients are willing to wait longer for admission, the hospital tends to form more wings. This implies that hospitals that garner longer waits can form more specialized wings and thereby benefit from focused care, whereas hospitals that cannot will tend to form fewer, if any, wings, choosing to pool demand and bed capacity.

Keywords: hospital bed capacity management; care partitioning; focus
History: Received: October 2, 2013; accepted: October 25, 2014. Published online in *Articles in Advance* February 23, 2015.

1. Introduction

Inpatient care continues to be the single most significant source of revenue for hospitals in the United States (American Hospital Association 2011), but the availability of inpatient beds is under strain. Average bed occupancy, the ratio of the number of occupied beds to available bed capacity, has increased in the last decade, reversing a 30-year trend (National Center for Health Statistics 2012). Capacity is especially strained in large cities, with particular challenges arising for teaching hospitals and hospitals near recent closures (Bazzoli et al. 2003). The strain is expected to worsen with the increasing insurance coverage of the Affordable Care Act and the aging of a large demographic (Litvak and Bisognano 2011).

Adding new beds to address the strain may not always be feasible for reasons including financial, legislative, staffing, and/or physical constraints. It is commonly accepted that "the cost of building and equipping the space for an average hospital bed exceeds $1 million" (Hall 2012, p. 180). Several survey results (e.g., see Bazzoli et al. 2003) indicate that hospitals, in general, refrain from capital intensive investments to deal with capacity problems, resorting instead to better management of existing capacity. As discussed in Green (2012), managing existing capacity is a complex issue, since hospital administrators must allocate beds across many distinct care types, at times leading to ample beds for some and insufficient beds for others.

When a patient is admitted into a hospital bed, he is typically admitted into a specialized unit or wing with some expertise for his care. Forming specialized wings provides some advantages to the hospital. For example, a cardiology wing may be given a fixed number of beds that allows the hospital to better coordinate the demand for and supply of cardiologists, cardiac surgeons and nurses, as well as better manage the inventory of specialized equipment and supplies, and to further improve and standardize processes within the wing. It is not uncommon for a wing to combine a number of care types, pooling a shared allocation of beds, with critical considerations being the ability for the same nursing unit to be able to adequately care for all patients in the wing. The more a wing specializes, the more a hospital can take advantages of focused care; quality, costs, and lengths-of-stay may all improve. Moreover, forming specialized wings allows the hospital more control over its overall bed capacity by reserving beds for some care types while restricting beds for others. On the other hand, forming specialized wings works against the pooling of demand and capacity and thus leaves the hospital more susceptible to stochastic fluctuations.

The purpose of this paper is to model, at a strategic level, the trade-off between specialization and pooling of hospital bed capacity in an optimization framework to help a hospital decide how best to form wings. The *wing formation decision* has three components: (1) How many wings to form? (2) How many beds to allocate to each wing? (3) How to partition the care types among the wings? Each patient provides some utility to the hospital depending on his care type. This utility is chosen by the hospital to represent the contribution of the patient to its mission. At the hospital's discretion, it can be a measure of financial or societal benefit. Our model seeks to find the wing formation that maximizes the total expected daily utility of the hospital.

The effects of focus are modeled endogenously, so that forming wings that care for a narrow and cohesive set of patients will benefit from increased per-patient utility (because of improved quality and/or reduced costs) and/or shorter lengths-of-stay (because of more efficient handling of processes). As a countervailing effect, we incorporate a queueing system to model the advantages of pooling care types into the same wing. Patients arrive according to a stochastic process and occupy a bed for a random amount of time that depends on both the care type of the patient and the structure of his assigned wing. Furthermore, patients have stochastic tolerance times while waiting to be admitted into a bed. For example, on average, patients in urgent need of care will abandon quickly if care is not provided, while patients waiting for elective or scheduled procedures may wait longer. The waiting of patients may occur in a physical queue (e.g., the emergency room) or in a virtual queue when patients wait elsewhere for admittance for nonemergency procedures.

The wing formation decision is a critical strategic decision for a hospital. Consider the case at the University of Chicago Medical Center (UCMC) that served as a primary motivation for our model. As is typical of urban academic medical centers, UCMC strives to be the preferred destination for the most complex of patient cases that will further its mission in preeminent care, clinical research, and medical education. But UCMC also has an important commitment to its local community, of which about 30% lives below the poverty level and 48% is on public programs such as Medicaid and Medicare. With seven other hospitals closing in the area since 1986, UCMC (and, in particular, its emergency department) has had to serve as the first line of medical care for many in the community. Overall, the demand for care exceeds the bed capacity at UCMC; thus, UCMC must necessarily make difficult decisions about which patients should be given priority in admission.

In a typical year, UCMC admits about 14,000 adults to its 300 medical-surgical beds. Just under half of these admits are made through its emergency department, and these patients are most frequently assigned to the general medicine (GEN) service. Prior to 2006, wing definitions were lax: at times, the GEN service occupied nearly one-third of the beds throughout the hospital, causing considerable strain on the bed capacity available for the other 17 services. This resulted in frequent rescheduling and sometimes cancellation of impending surgeries because of a lack of inpatient beds for postsurgery care, while at other times, preventing physicians from admitting patients from their own patient panels. The lax wings also put strain on the quality of care, since nurses on a given floor attended to a wider range of care types, resulting in frequent calls to physicians for help on unfamiliar treatments.

The overflow of GEN patients also meant that UCMC struggled to meet its financial obligations. A significant fraction of GEN patients are reimbursed by public programs. As a result, just 35% of the inpatient admissions to UCMC in 2005 were funded with private insurance, while other local academic medical centers, such as Northwestern Memorial Hospital (55%), McGaw Hospital of Loyola University (50%), and Rush University Medical Center (45%), were considerably higher. In fact, UCMC has estimated that a percentage point increase in private pay boosts its annual profits by more than $8 million.

Forming wings is a primary tool UCMC now uses to help plan its flow of patients and its finances. Because they are a capacity constrained hospital, UCMC must make choices about which patients can be admitted for treatment. Forming strict wings is a way to make admit decisions in a purposeful and consistent manner, rather than leaving these decisions to chance or political maneuvering. Under supervision of the Illinois Department of Public Health, UCMC formed four strict wings in July 2006. A wing with 129 beds was reserved for the surgical care types, another with 72 beds was assigned to cancer treatment, and a third with 30 beds was dedicated to cardiology. GEN was placed into the fourth wing with 69 beds, which was reduced to 35 by May 2009, as UCMC established an arrangement with a nearby hospital to transfer GEN patients with mutual consent, where UCMC physicians helped administer their care. Karlin (2013, p. 531) describes the implementation of these wings: "deciding who would be allowed in a particular institutional space, and then convincing others that allocating people on the basis of newly created biofinancial categories is reasonable, was the UCMC's first step in trying to manage what it regarded as a financial crisis." As Karlin documents in detail, the UCMC wings are not without controversy, and therefore it is a goal of this paper to offer a useful model and insights to better understand the implications of strict wings.

The wing formations beginning in 2006 at UCMC are one type of flexibility a hospital might use. Care types within a wing flexibly share beds, but the boundary between wings is strict, not allowing care types to overflow from their assigned wing into another. It is conceivable to improve operational efficiency by providing some additional flexibility through allowing a general care type to admit to multiple wings. Starting in early 2012, UCMC began allowing GEN patients to overflow from their primary wing into a limited number of beds in a select secondary wing, but such overflows were used very infrequently because of issues such as quality of care as well as other administrative and operational concerns. Overall, to benefit from additional flexibility at the expense of some loss of focus, such overflow policies for a "general" care type might be used by a hospital, because nursing teams in any wing are typically cross-trained to care for such general patients (see, for example, Shi et al. 2013). In §8, we describe a heuristic approach to account for such overflows within our model of strict wings and why in general such overflow flexibility is difficult to handle in a tractable way.

In practice, the strategic decision of forming wings is a static one, at least for many months, since it is impractical for a hospital to change the allocations frequently because of administrative, equipment, and staffing issues. For example, UCMC did not change its four-wing structure since it was first established except for the allocation of beds between the wings, which is revised, on average, about twice a year.

Our work is, to the best of our knowledge, the first to model the full wing formation decision in an optimization framework and furthermore to endogenize the advantages of focus within the optimization. We apply our model to realistic size instances using data from UCMC and national databases, and we uncover a number of insights. In particular, we find that as the overall demand increases across all care types, wings are formed to allocate an increasing number of beds to higher-utility care types. This behavior leads to disparities among care types, as hospital access increases for some and decreases for others. Furthermore, in general, overall hospital bed occupancy decreases with the number of wings formed, since partitioned capacity faces stochastic demand. However, when sufficient focus is attained, shorter lengths-of-stay associated with focused care may enable more patients to receive care (i.e., the overall patient throughput can increase). We also observe that when patients are willing to wait longer for admission, forming more wings becomes preferable. This implies that hospitals that are able to garner longer waits, perhaps because of the focus of the hospital on elective services, or a lack of alternatives for care that are convenient or of competitive quality, will more readily be able to form specialized wings and thereby reap the benefits of focus. On the other hand, hospitals that are not able to garner longer waits will form fewer, if any, wings, relying instead on the advantages of pooling beds and demand.

The rest of this paper is organized as follows. We provide a review of the related literature in §2. In §3, we present a general framework for our optimization model, which presents computational challenges when solving realistic size instances. To overcome these challenges, we take a bounding approach and develop a dynamic programming (DP)-based heuristic in §4 to obtain a feasible solution. In §5, we illustrate an application of our model, followed by numerical results and insights in §6. In §7, we prove upper bounds for our problem and demonstrate that our heuristic solutions perform fairly well (average optimality gap = 1%) on realistic instances. We provide proofs along with other technical details in an online supplement (available as supplemental material at http://dx.doi.org/10.1287/msom.2014.0516).

2. Related Work

Revenue Management. The wing formation of a hospital can be viewed as a revenue management (RM) strategy. A hospital can use wings to reserve beds for high-utility care types, capturing more of their demand, and also restrict access for low-utility types, causing more of their demand to go elsewhere. Such a management approach is commonly referred to as quantity-based RM (Talluri and van Ryzin 2004). Chapman and Carmel (1992) recommend the use of quantity-based RM in hospitals. They suggest reserving capacity for lucrative elective surgeries and restricting capacity for other noncritical procedures. This is in lieu of pricing-based RM, which they label as less feasible or influential since third parties pay for most inpatient care at fixed rates.

Ayvaz and Huh (2010) study the *dynamic* allocation of a fixed capacity of a single healthcare resource to two patient types, one of which, say urgent patients, leaves the hospital immediately if not served upon arrival and the other, say nonurgent patients, waits indefinitely until service. Each served patient generates some revenue, but there are costs associated with waiting and leaving. Ignoring the variation in the lengths-of-stay of patients, their model maximizes the total net revenue by deciding on the fraction of the pooled resource to reserve for the urgent class using a function of the number of backlogged nonurgent patients at each time point. They characterize the structure of the optimal policy but recognize the difficulty of finding and implementing the optimal policy in practice, so they suggest a simple heuristic policy. In contrast to our work, the frequent reformation of the wings, which is at the centerpiece of the work of Ayvaz and Huh (2010), violates the premises of the wing formation problem studied in this paper. In a similar vein, Stanciu et al. (2010) study protection levels for hospital operating rooms when *nested* solutions are allowed. Nested solutions, where higher-utility types can access the capacity reserved for lower-utility types but not vice versa, are discouraged by the UCMC system that motivated our work. In addition, to the best of our knowledge, the rich RM literature has not yet addressed the effects of focus, which we consider in our framework.

Healthcare Operations. The wing formation problem has been studied in the literature under various restrictions. Gorunescu et al. (2002) and Kokangul (2008), for example, restrict the general problem to optimizing the bed capacity of a single wing, which is taken as a particular unit in a hospital—or as the entire hospital without considering the heterogeneity across different services—and investigate measures of lost demand, bed occupancy, and/or waiting times. Relaxing the single wing assumption, several researchers evaluate a prespecified set of wing formations, so they avoid the combinatorial optimization aspect of the general problem. Dumas (1984) and Harper and Shahani (2002) use simulation, whereas Green and Nguyen (2001) use queueing formulae to evaluate and compare a few wing formations in terms of their effects on bed occupancy and patient waiting times. Another thread in this line of research combines the previous two threads by considering the bed allocation decisions for a fixed number of wings and a given assignment of care types to these wings. Kao and Tung (1981) provide one of the early models to allocate beds across a number of wings to minimize total demand overflow. Vassilacopoulos (1985) and Huang (1998) use simulation modeling to allocate beds across wings subject to thresholds set for admission delay for emergency patients, for bed occupancies, and for queue lengths, and report on applications of their models in real hospitals. Ma and Demeulemeester (2013) iterate between deterministic optimization and stochastic simulation models. The deterministic models decide on the patient mix and volume that maximizes the financial contribution to the hospital with a given bed capacity, and subsequently they allocate the bed capacity across patient types. Given the output of the deterministic models, the stochastic model evaluates various operational metrics for a few operating policies. They illustrate their framework for a hypothetical hospital with two wings and nine care types. A significant limitation of this work is the fact that it ignores emergency patients by assuming that such patients are treated in a separate wing than elective patients (i.e., the emergency department) with a dedicated capacity, which is not justifiable for many academic medical centers including UCMC, because a significant fraction (more than 40% for UCMC) of such patients ultimately require inpatient care. Our work distinguishes itself from this line of research by considering the *whole* wing formation problem, and also by incorporating the effects of focused care into this problem, which has not yet been dealt with.

Wing formation is a strategic level decision. Unlike the examples in the RM literature, practical users of this approach would not alter wing definitions in frequent intervals nor are they allowed nested solutions, because—among many other negative side effects—such alterations go against the benefits of focused wings. Similar to the cited RM work, many researchers have studied healthcare capacity allocation decisions under a more tactical setting that allow such changes. We refer the reader to the excellent surveys of Hall (2012) and Hulshof et al. (2012) for detailed discussions of this vast literature on hospital operations.

Pooling and Flexibility. Our work also relates to the rich literature on pooling/splitting queues (see, for example, Whitt 1999 and Van Dijk and Van der Sluis 2008, as well as the references therein). The main insight from this literature is that for homogeneous customers, pooling capacities is superior in terms of average customer waiting time and the total capacity requirements, but the superiority is no longer certain when customers are heterogeneous or different performance metrics are used. Most similar to our work, Whitt (1999) considers the partitioning of multiple customer classes into service groups, while simultaneously deciding on the number of servers for each group. He decides on the partitions to minimize the total number of servers needed to achieve targets on expected waiting time of each class, whereas we consider the total number of servers (i.e., beds) to be given and decide on the partitions to maximize total utility derived from the set of patients seen in the hospital. In contrast to our work, he assumes that the customer service times are unaffected by the partitioning, whereas we make the service times dependent on the level of focus. We are unaware of any queueing papers that adjust the customer service times and/or the utilities derived from the customers as a function of the formed partition, which we achieve through explicitly modeling focus.

Partitioning of customer types into classes is also known as the indexing problem in the RM literature. The number of feasible partitions for many realistic instances can be so large that it becomes quickly intractable to select the optimal partition, so a common heuristic in RM is to consider only contiguous partitions from a fixed sequence of types (Talluri and van Ryzin 2004). From a solution perspective, our approach (detailed in §4.1) and that of Whitt (1999) also follow this common heuristic.

Our work also relates to the literature on process flexibility, or cross-training of workers. In the context of our hospital problem, such flexibility means cross-training some wings so that patients of a particular care type(s) can be treated in those wings. Unfortunately, the wing formation problem becomes significantly more difficult when this type of flexibility is allowed. The chain flexibility that was first analyzed by Jordan and Graves (1995) is a simple yet powerful type of limited flexibility that achieves the operational efficiency that is almost equal to that of full flexibility. However, such flexibility is shown to be outperformed by other designs when stochasticity and heterogeneity are present. Wallace and Whitt (2005), for example, consider a queueing system in which each server is trained to handle a pair of customer types with possibly different priorities over either type. They use simulation and many-server asymptotics to approximate the best total server count and the best assignment of servers to customer types. However, they also caution that their solutions might be suboptimal if the types have substantially different service time distributions. Because of the difficulties associated with optimizing under such flexibility, we leave the study of optimal wing formation when such flexibility is allowed for future research.

Focused Care. There is a growing literature on quantifying the effects of focus in healthcare. Green (2012) discusses the general trade-offs of having dedicated or pooled hospital beds, calling for more work on identifying when dedicated beds are more efficient. Vanberkel et al. (2012) provide a survey of the operations literature on focus and pooling. They also investigate the trade-offs in pooling and focus on various metrics using a single-server queue with two customer types. Muller (2010) surveys the medical literature on the advantages and motivations of “service line specialization” (analogous to focused care). Although there is no agreement on the overall impact of focus for a hospital, there is a general consensus on its effect on individual metrics. Green and Nguyen (2001) and Green (2004), for example, note that focused care can shorten lengths-of-stay, lower mortality rates, or reduce readmission rates. Such potential benefits are also confirmed by recent empirical papers; see, e.g., Hyer et al. (2009) and KC and Terwiesch (2011) for shorter lengths-of-stay, and Clark and Huckman (2012) for lower mortality rates. While most of the recent research concerning focused care concentrates on using empirical models to quantify its effect on individual metrics, there is a general lack of incorporating these effects in analytical models that address questions related to hospital design. Our work makes an important initial attempt toward this goal.

System Load. When faced with resource shortages, healthcare providers are often forced to make rationing decisions. Such decisions oftentimes manifest themselves as the shortening of the lengths-of-stay in the hospital. This is the predominant finding in the literature (see, for example, Hellmann et al. 1962 and Strauss et al. 1986) and is also supported by our conversations with UCMC, although the relationship can be nuanced. Berk and Moinzadeh (1998) study the impact of early discharge decisions on the quality of care in a single unit of a hospital. They find that inclusion of an early discharge option improves system accessibility significantly and does not jeopardize care equity among patients. More strikingly, they report that introduction of an early discharge option has more pronounced effect on increasing the capacity of the unit than adding beds to the unit with no early discharges. KC and Terwiesch (2009) also find that health workers accelerate their service rate (hence, decrease the length-of-stay) as load increases, but long periods of increased workload may decrease the service rate and also cause a reduction in the overall quality of care. In a separate study, KC and Terwiesch (2012) report that lengths-of-stay in an intensive care unit decrease as its occupancy increases, but that shorter lengths-of-stay are associated with a higher probability of readmission. Given this variety of findings, we incorporate the influence of system load into our model by adjusting the lengths-of-stay, which we describe in §5.1.

3. The Model

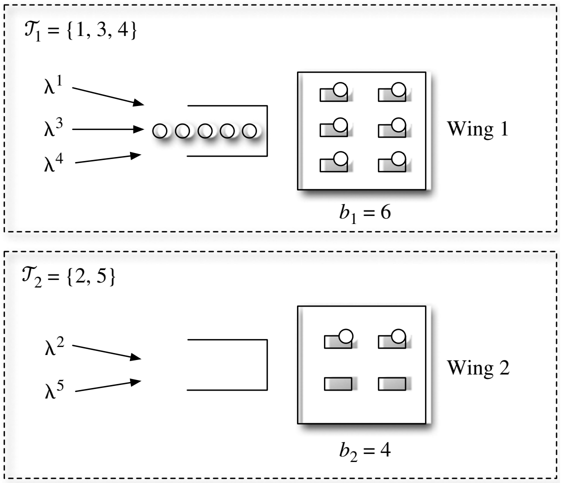
In this section, we set up the general framework for our optimization model. We consider the wing formation decision from the perspective of a hospital administrator whose goal is to most effectively use her limited bed capacity to help achieve the mission of the hospital. She achieves her goal by controlling the number of wings to form within the hospital, the number of beds to allocate to each wing, and the partitioning of care types to each wing. The total number of beds available for allocation and the set of care types that can be seen in the hospital are fixed and known. Each care type corresponds to a category, identifiable upon admission, corresponding to some clinical and/or surgical group in the hospital (e.g., cardiology, orthopedics). Each care type can be assigned to only one wing, so that an arriving patient of a particular type can only be admitted if a bed is available in his assigned wing. If, upon arrival, a patient faces no available beds in his assigned wing, then he may wait for some time for an opening before seeking service elsewhere. Each patient, if admitted, occupies a bed for a random duration and provides a utility to the hospital, which corresponds to a general measure of contribution to the hospital's mission. There are also effects of focus derived from forming a specialized wing, for example, in the form of decreased lengths-of-stay or increased utility. The administrator seeks to maximize the total expected patient-derived utility of the hospital.

We formalize the above discussion with the following notation. Let B denote the total number of beds, and $\mathcal{C}=\{1,2,\dots,T\}$ the index set of care types. Admission requests for care type $i\in\mathcal{C}$ arrive stochastically with daily rate λ^i. The decisions are to determine (i) the number of wings w to form; (ii) the number of beds $b_j\ge 0$ to allocate to wing j, for $j=1,\dots,w$; and (iii) the nonempty set of care types $\mathcal{T}_j$ to assign to wing j, for $j=1,\dots,w$, such that the assignment forms a partition of the set $\mathcal{C}$, which is illustrated in Figure 1.

Let $\mathbf{b}=\{b_1,b_2,\dots,b_w\}$ and $\boldsymbol{\mathcal{T}}=\{\mathcal{T}_1,\mathcal{T}_2,\dots,\mathcal{T}_w\}$, so that a wing formation can be described by the vector of decision variables $(w,\mathbf{b},\boldsymbol{\mathcal{T}})$, with feasible region Ψ, where

$$\Psi=\left\{(w,\mathbf{b},\boldsymbol{\mathcal{T}})\colon \sum_{j=1}^{w} b_j\le B,\ \boldsymbol{\mathcal{T}} \text{ is a partition of } \mathcal{C},\right.$$
$$\left. w\ge 1 \text{ integer, and } b_j\ge 0 \text{ integer for all } j\right\}.$$

Figure 1 **An Illustration of Wing Formation with $B=10$, $\mathcal{C}=\{1,2,\dots,5\}$, and $w=2$**



Let $u^i(\psi)$ denote the expected utility gained from a patient of type $i\in\mathcal{C}$ under a wing formation $\psi\in\Psi$. By specifying the dependency on ψ, we emphasize that utility is a function of the level of focus, which is determined by how the wings are formed. In §5.1, we specify a fairly robust form for this function. We allow the utility to be any measure of contribution to the hospital's mission. In §6.2, we use a utility measure called case-mix index that fits the mission of UCMC and similar hospitals, and we also discuss how our model can be used to minimize a function of patient waiting times.

Let $p^i(\psi)$ denote the steady-state probability that an arriving patient of type $i\in\mathcal{C}$ abandons the system before gaining access to a bed in his assigned wing under the wing formation $\psi\in\Psi$. Henceforth, unless otherwise noted, for simplicity we suppress the dependency of $p^i(\psi)$ and $u^i(\psi)$ on ψ. We write the expected daily utility from wing j as the sum of the daily utilities from each care type admitted into the wing:

$$U_j=\sum_{i\in\mathcal{T}_j}\lambda^i\left(1-p^i\right)u^i \quad \text{for } j=1,\dots,w. \tag{1}$$

We can now write the optimization problem faced by the hospital administrator as

$$(P)\quad Z=\max_{(w,\mathbf{b},\boldsymbol{\mathcal{T}})}\left\{\sum_{j=1}^{w}U_j\colon (w,\mathbf{b},\boldsymbol{\mathcal{T}})\in\Psi\right\}.$$

4. Solving the Model

This section develops a tractable solution methodology to solve realistic instances of problem (P). The full problem is, in general, intractable, because of the combinatorial nature of the wing formation decision and the intricacies involved in computing the abandonment probability p^i for $i\in\mathcal{C}$. To handle these issues, in §4.1 we introduce a restriction to problem (P) which leads in §4.2 to an efficient DP-based solution. We select a queueing system in §4.3 that is both tractable and robust. In §4.4, we show how to easily optimize (P) when only the bed allocation vector $\mathbf{b}$ is to be determined (i.e., when the number of wings w and the assignment of care types $\boldsymbol{\mathcal{T}}$ are fixed), which is particularly useful when a small number of beds are added to or removed from usage in the hospital.

4.1. A Restrictive Sequence

The problem (P) can be written as a large-scale generalized set partitioning and capacity allocation problem having nonconvex functionals. However, realistic instances of such a formulation are intractable. For example, the number of possible care type partitions for an instance similar to UCMC in size with $T = 18$ is over 6.8×10^{11}. Therefore, we propose an approximation that relies on restricting the feasible region Ψ but also produces high-quality solutions, which we demonstrate in §7.

We reduce the feasible region Ψ by imposing the restriction that wings are formed by making "cuts" in a fixed sequence $\mathcal{S}$ of the care types. Under such a restriction, if two care types are in the same wing, then all care types between them in the sequence $\mathcal{S}$ must also be in the wing. For example, suppose $T = 3$ and we impose a sequence $\mathcal{S} = (1, 3, 2)$. Feasible solutions under this restriction require any wing that includes care types 1 and 2 to also include type 3. The overall effect of such a restriction is enormous computational savings, which increases exponentially fast with the number of care types T. In the case of UCMC, the number of possible care type partitions drops down to a manageable 1.3×10^5. Note that this restriction is a commonly used heuristic for solving the indexing problem in RM (Talluri and van Ryzin 2004).

Given a sequence $\mathcal{S}$, the restricted wing formation problem, which provides a lower bound for Z, is

$$(P_{\mathcal{S}}) \quad Z_{\mathcal{S}} = \max_{(w,\mathbf{b},\boldsymbol{\mathcal{T}})} \left\{ \sum_{j=1}^{w} U_j \colon (w, \mathbf{b}, \boldsymbol{\mathcal{T}}) \in \{\Psi \text{ and } (\mathcal{T}_j \text{ are cuts in the sequence } \mathcal{S})\} \right\}.$$

In general, one should try as many sequences as reasonably possible. We explore some approaches to forming a sequence $\mathcal{S}$ in §5.2, and we illustrate the quality of these approximate solutions in §7.

4.2. A Dynamic Programming Solution for $(P_{\mathcal{S}})$

The restricted problem $(P_{\mathcal{S}})$ can be solved efficiently through a DP reformulation. Define the state space of the DP as $\Xi = \Xi' \cup \{(0,0,0)\}$, where $(0,0,0)$ is an auxiliary state introduced for convenience and $\Xi' = \{(j,k,l)\colon j = 1, \ldots, T;\ k = 0, \ldots, B;\ l = j, \ldots T\}$, where j, k, and l, respectively, index wings, beds, and care types. For state $\xi = (j,k,l) \in \Xi$, the set of actions is $\mathcal{A}(\xi) = \{(b,t)\colon b = 0, \ldots, B-k;\ t = 1, \ldots T-l\}$, where b indicates the number of beds to allocate to the next wing and t indicates the number of types to assign to the next wing. More specifically, while in state $\xi = (j,k,l)$, taking action $a = (b,t) \in \mathcal{A}(\xi)$ transitions the system to a new state $\xi' = (j+1, k+b, l+t)$, which is interpreted as forming a new wing, labeled $j+1$, with a capacity of b beds for the care types that are indexed as $\{l+1, l+2, \ldots, l+t\}$ in the sequence $\mathcal{S}$. The expected reward associated with this state-action pair is $r(\xi, a) = U_{j+1}$, which is computed using Equation (1).

The reward-to-go function $V(\xi)$ for $\xi = (j,k,l) \in \Xi'$ denotes the maximum expected total utility that can be obtained by solving the wing formation subproblem for the remaining $B-k$ beds and the remaining $T-l$ care types in the sequence $\mathcal{S}$. So, the maximum expected total utility for the full problem is given by $V(0,0,0)$ and can be computed by recursively solving the following optimality equations

$$V(\xi) = \begin{cases} 0 & \text{if } \xi = (\cdot,\cdot,T), (\cdot,B,\cdot), \\ \max\limits_{a \in \mathcal{A}(\xi)} \{r(\xi,a) + V(\xi')\} & \text{otherwise}, \end{cases} \qquad \forall\, \xi \in \Xi. \tag{2}$$

The number of states for this DP is $O(T^2 B)$, each having $O(TB)$ actions associated with it. Hence, the total number of actions to choose from is $O(T^3 B^2)$. The optimality Equations (2) can be solved by backward induction. As noted earlier, an action $a \in \mathcal{A}(\xi)$ for state $\xi \in \Xi$ indicates a new wing. For computational speedup, the total utility associated with each action (wing) only needs to be computed once and saved to memory. The DP can then be solved in linear time in the number of actions.

Our approach makes it easy to add various types of constraints to our model without any added computational burden, by simply removing some actions from the DP. For example, the hospital may want to prevent wings from pooling particular care types together, or restrict the total number of care types allowed in a wing. The actions of the DP can be further restricted to ensure some minimum/maximum numbers of beds per wing and/or per care type, and make such bounds depend on the care types allocated to the wing. Constraints can also be added, for example, that mandate certain limits on abandonment probabilities or levels of slack capacity within select wings.

Our approach is similar to that of Graves and Redfield (1988), who consider an analogous problem of assigning manufacturing tasks to stations, whereas we assign care types to wings. To overcome the difficulties with the evaluation of each possible assignment, they exploit a natural precedence relation among the tasks, similar to our idea of fixing a sequence of care types, to obtain a tractable DP solution.

4.3. The Queueing System: Calculating U_j

The DP detailed in §4.2 requires computing the overall wing utility U_j for every possible action. Our solution method is therefore open to any tractable approach for computing the U_j. At the core of determining U_j is computing p^i, the fraction of arrivals that abandons the wing without being seen, for all care types i that are assigned to wing j (see Equation (1)). One reasonable approach for estimating p^i is to consider an abandonment queueing model, in which patients queue for some random duration before abandoning the system and seeking care elsewhere. We can think of some patients as queueing in a common location such as an emergency department, and others as waiting from home or another care facility. Some patients will have little ability or desire to wait for care when the need is urgent or there are other competitive options for care; and thus the patient will abandon quickly. In other cases, a patient may be able and willing to wait longer when perhaps admission is not time critical (such as an elective procedure) or the hospital in question is perceived to provide high-quality care; and thus the patient may abandon slowly.

Abandonment queueing models, in general, are notoriously difficult, in that closed-form exact expressions for their performance measures are rare (Ward 2012). The most robust yet tractable version is the $M/M/b_j+M$ queueing model. To apply this model in our setting, the set of care types $\mathcal{T}_j$ (i.e., those types that are assigned to wing j) form a single class, seeking care according to a stationary Poisson process with rate $\Lambda_j = \sum_{i\in\mathcal{T}_j} \lambda^i$. Patients are immediately admitted to any of the available beds in the wing. All patients in a wing are assumed to have a common exponential length-of-stay distribution. We denote the mean length-of-stay of a patient of type $i \in \mathcal{C}$ with $s^i(\psi)$ to indicate that lengths-of-stay depend on the level of focus of the wing, and henceforth suppress its dependency on ψ for convenience. Patients assigned to the same wing wait for an open bed in a first-come-first-served queue for an exponentially distributed amount of time with mean q_j. So the parameter q_j can be interpreted as the average willingness-to-wait of the patients assigned to the wing. As a result, all care types in this single-class $M/M/b_j+M$ model have identical steady-state abandonment probabilities, which we denote by p_j for wing j. The abandonment probability p_j for wing j is given by Zeltyn (2004)

$$p_j = \frac{1+(d_j/b_j-1)X_j}{\mathcal{E}_j+(d_j/b_j)X_j}, \tag{3}$$

where $d_j = \sum_{i\in\mathcal{T}_j} \lambda^i s^i$ indicates the aggregate (single-class) average bed demand for wing j,

$$X_j = 1+\sum_{n=1}^{\infty}\prod_{k=1}^{n}\frac{\Lambda_j q_j}{\Lambda_j q_j/(d_j/b_j)+k}, \quad \text{and}$$

$$\mathcal{E}_j = \frac{\sum_{k=0}^{b_j-1} d_j^k/k!}{d_j^{b_j-1}/(b_j-1)!}.$$

For our first set of numerical experiments in §6.2, we implement the single-class $M/M/b_j+M$ model because of its tractability and replace p^i in Equation (1) with p_j of Equation (3). In our implementation, we compute X_j by truncating the infinite summation, but we also show in Online Appendix C that our estimate of X_j can be made arbitrarily close. In §6.1, we describe evidence that supports our exponential length-of-stay assumption when forming wings. Furthermore, we explore an extension in §6.3, in which we allow two patient classes (say, an urgent and a nonurgent class) in each wing so that the arrival rate, length-of-stay, utility, willingness-to-wait parameters are all class dependent, and one class has priority over the other.

Further extensions of the $M/M/b_j+M$ model are possible but rely on approximations. See, for example, Zeltyn and Mandelbaum (2005) for the $M/M/b+GI$ queue, Iravani and Balcıoğlu (2008) for the $M/GI/b+GI$ queue, and Whitt (2006) for the $G/GI/b+GI$ queue. Hypothetically, one can also use simulation to include more features of the real setting when evaluating Equation (1). But for realistic instances, simulation-based optimization quickly becomes impractical. For UCMC, with $T=18$ and $B=300$, there are over 50,000 wings to simulate for each fixed sequence $\mathcal{S}$. From our experiments, we estimate it would take about six months using simulation to evaluate all the wings for a given sequence. In practice, one would also want to try many different sequences, which renders simulation even less attractive.

4.4. Reallocating Beds Across a Fixed Set of Wings

After a hospital has implemented a wing formation $(w, \mathbf{b}, \boldsymbol{\mathcal{T}})$, it may occasionally require updating as parameter estimates (such as patient arrival rates, lengths-of-stay, or utilities) might change over time, or beds may be added or removed. For example, as noted in §1, UCMC made no changes to the number of wings nor the assignment of types to the wings since it first started this practice; only the allocation of beds were changed. This is probably natural, since organizational issues are mostly tied to how the care types are defined and partitioned into wings. For example, each wing at UCMC is a managerial entity with an appointed patient flow director.

The problem becomes significantly easier when the number of wings w and the assignment of care types $\boldsymbol{\mathcal{T}}$ are fixed, in which case the only decision is the allocation $\mathbf{b}$ of the beds to the wings. If the effects of focused care are also not considered, then one can use the references cited in §2 to address this restricted problem. Considering the effects of focus, we propose an integer linear program (IP) to optimize the allocation of $\mathbf{b}$. Let $c_{jk} := U_j(k) - U_j(k-1)$ be the marginal utility gained by allocating the kth bed to wing j, for $k = 1, \ldots, B$, $j = 1, \ldots, w$. The marginals c_{jk} are nonlinear in k, because of the nonlinear queueing dynamics. However, calculating the $B \times w$ marginals takes only a few seconds for our queueing model in §4.3.

Given the marginals c_{jk}, an optimal bed (re)allocation can be determined by the following IP: $\max\{\sum_{j,k} c_{jk}x_{jk}\colon\ \sum_{j,k} x_{jk} \le B, x_{jk} \ge x_{j,k+1}$, and $x_{jk} \in \{0,1\}$ for $k = 1, \ldots, B$, $j = 1, \ldots, w\}$, where the decision variable $x_{jk} = 1$ if the kth bed is allocated to wing j, and 0 otherwise. Solving this IP with any commercial solver takes only a few seconds for realistic-sized problems.

5. Applying the Model

This section illustrates an application of our model for a hospital like UCMC. We present empirical data in §5.1 to motivate specific representations for the utilities $u^i(\psi)$ and lengths-of-stay $s^i(\psi)$ for care type $i \in \mathcal{C}$. We discuss how to select a sequence $\mathcal{S}$ in §5.2, and propose a particularly useful one for our instances.

5.1. Functional Forms for $s^i(\psi)$ and $u^i(\psi)$

We discussed in §2 that the impact of focus and demand load on operational metrics, such as reduced lengths-of-stay and improved quality, have been widely recognized in the literature. Despite these empirical observations, there is a lack of analytical models that capture such effects.

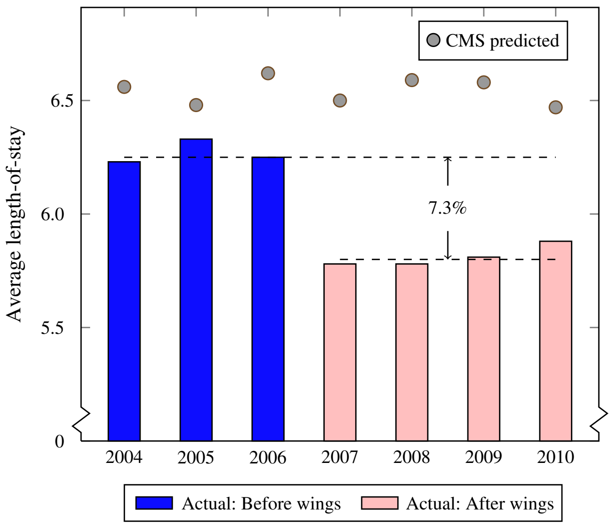
To model such effects endogenously within our optimization framework, we propose functional forms for both $s^i(\psi)$ and $u^i(\psi)$ for care type $i \in \mathcal{C}$ under wing formation $\psi \in \Psi$. We choose stylized yet nontrivial functional forms. Let m^i be the nominal mean length-of-stay of a patient of type i, without any consideration of focus. Similarly, let v^i be the nominal utility gained from a patient of type i. We also introduce a *length-of-stay scaling factor* $\alpha^i(\psi) \le 1$ to adjust the nominal length-of-stay m^i, and a *utility scaling factor* $\gamma^i(\psi)$ to adjust the nominal utility v^i. In particular, we let

$$\begin{aligned} s^i(\psi) &= (1 - \alpha^i(\psi)) \cdot m^i \text{ and} \\ u^i(\psi) &= (1 + \gamma^i(\psi)) \cdot v^i. \end{aligned} \tag{4}$$

The scaling factors α^i and γ^i can be quite general. For example, the benefits arising from forming a wing with a narrow and cohesive set of care types can be captured by assigning positive values for α^i and γ^i, resulting in reduced average length-of-stay and increased average per-patient utility. On the other hand, if there are inefficiencies or disutilities from forming a wing with particular care types, such effects can be captured by negative values for α^i and γ^i.

We require $\alpha^i \le 1$ to avoid a negative length-of-stay s^i, but, in practice, typical values of α^i would max out at around 10% because it is difficult to achieve a larger reduction in lengths-of-stay without violating medical care standards (see, for example, Hyer et al. 2009, Capkun et al. 2012). Figure 2 displays the average length-of-stay of all inpatient cases at UCMC before and after strict wings are formed. The lengths-of-stay decreased, on average, by 7.3% after wings are formed (t-test: $p < 0.001$). While the patient mix as well as the medical standards might have changed during this period, these changes do not explain the observed decrease. To control for these changes, we perform another test comparing the actual lengths-of-stay for the mix of patients seen at UCMC to their predicted lengths-of-stay. We obtain the predicted length-of-stay of a patient from the national average length-of-stay for the year he was seen at UCMC and for his diagnosis related grouping (DRG), which is published in the federal database of the Centers for Medicare and Medicaid Services (CMS). For each patient, we take the difference between his actual and predicted length-of-stay and test for any change in the average difference before and after wings are formed. We find strong evidence that the average difference before wings is significantly smaller than the average difference after wings (t-test: $p < 0.001$).

Figure 2 **(Color online) Change in Average Lengths-of-Stay of All Adult Patients Discharged from UCMC**



To help explain the decrease in lengths-of-stay after wings were formed, we interviewed the patient flow director for a wing at UCMC. He first explained that it was his job to admit or deny access to his wing. When the wing is full (which is often the case), he is forced to refuse or delay incoming requests. Since *he* controls all new admits to the wing, he knows that whenever he discharges a patient, he is able to fill the bed with one of *his* care types—a motivation to turn over beds that simply did not exist before strict wings were formed when empty beds might be filled by any care type. When the wing becomes increasingly full, he walks the hall looking for empty beds. Additionally, he leads daily meetings with other doctors, nurses, and case managers assigned to his wing to review every case. His care providers are focused on a narrow set of care types, and thus they can more easily coordinate care during the stay of a patient and arrange for discharge. As the wing becomes full, he reported that he works with staff to cut the non-value-adding time a patient spends in the hospital (e.g., the time spent waiting for paperwork or for arranging transport and care after discharge from the hospital). The flow director was clear that patients are not over-expedited to avoid any increase in medical errors or readmission rates. We therefore conclude that the formation of the wings at UCMC created an environment at the hospital where the managers, physicians, nurses, etc. felt an increased sense of ownership over the beds in their wing, which motivated a more efficient use of their beds (particularly, in heavily loaded wings), that did not exist before the formation of the wings.

We model α^i using a generalized logistic function. Given a wing formation $\psi = (w, \mathbf{b}, \boldsymbol{\mathcal{T}})$, we set

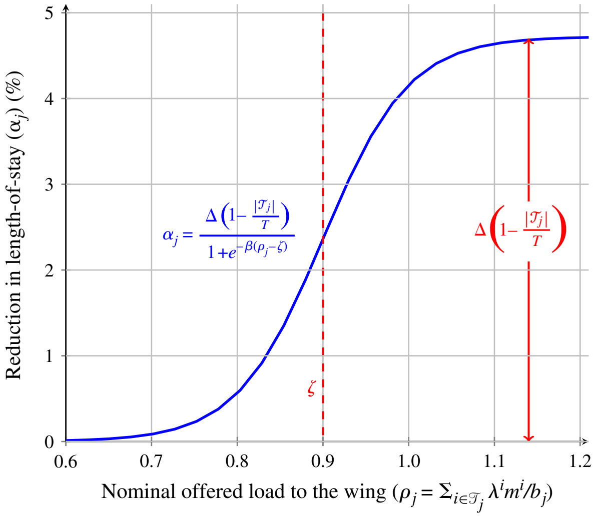
$$\text{length-of-stay scaling factors:}\quad \alpha^i \leftarrow \alpha_j = \frac{\Delta(1-\frac{|\mathcal{T}_j|}{T})}{1+e^{-\beta(\rho_j-\zeta)}} \quad \text{for } i \in \mathcal{T}_j,\ j=1,\ldots,w. \tag{5}$$

Although our framework allows for a unique scaling factor for each care type i, for illustrative purposes we choose to set α^i the same for all care types sharing the same wing j. The parameters Δ, β, ζ, and T of Equation (5) are exogenous, while ρ_j and $\mathcal{T}_j$ are determined by ψ. For wing j, $\rho_j = \sum_{i \in \mathcal{T}_j} \lambda^i m^i / b_j$ is the nominal offered load to the wing, and $|\mathcal{T}_j|$ is the number of care types assigned to the wing.

Figure 3 illustrates a plot of α_j for a particular set of parameter values. The level of focus of the wing is represented by the height of the curve. If a wing is formed with only a few care types, and thus more focus, then the term $\Delta(1 - |\mathcal{T}_j|/T)$ is larger, increasing the height of the curve, hence the potential reduction in the lengths-of-stay in the wing. At the other extreme, if the hospital pools all care types into a single wing, the curve becomes flat at 0% and there is no adjustment to the nominal lengths-of-stay.

The impact of load is determined by the x axis. When ρ_j is small, there is negligible adjustment to the nominal lengths-of-stay for the wing. But as ρ_j increases, the adjustment increases until it asymptotically reaches $\Delta(1-|\mathcal{T}_j|/T)$ representing the maximum possible reduction. The S-shape of the curve captures the concerns described by the UCMC flow director. That is, there is minimal incentive to reduce lengths-of-stay when a wing has plenty of slack capacity; however, this incentive increases with load; however, a wing under heavy load can only reduce lengths-of-stay by so much without jeopardizing the quality of care.

Figure 3 **(Color online) An Illustration of the Length-of-Stay Scaling Factor α_j When $\Delta = 0.05$, $\beta = 20$, $\zeta = 0.9$, $|\mathcal{T}_j| = 1$, and $T = 18$**



In our numerical runs, we check the sensitivity of our results on the parameter Δ, which controls the asymptote of the curve. The higher the value of Δ, the more potential reductions there are to length-of-stay because of increased importance of focus. The parameter $\beta > 0$ controls how steeply the curve slopes, and $\zeta > 0$ specifies the horizontal coordinate of the point of steepest slope. When solving a particular instance, Δ, β, ζ, and T are fixed; hence, as our DP considers forming a particular wing j, we compute a new α_j as a function of $|\mathcal{T}_j|$ and ρ_j.

We choose a simple functional form for γ^i. Given a wing formation $\psi = (w, \mathbf{b}, \boldsymbol{\mathcal{T}})$, we set

$$\text{utility scaling factors:}\quad \gamma^i \leftarrow \gamma_j = \eta\left(1-\frac{|\mathcal{T}_j|}{T}\right) \quad \text{for } i \in \mathcal{T}_j,\ j=1,\ldots,w. \tag{6}$$

This expression is identical to the numerator of Equation (5), except that Δ is replaced with another exogenous parameter η. In our numerical runs, we check the sensitivity of our results on the parameter η, and when solving a particular instance, we compute a new γ_j during our optimization as $\mathcal{T}_j$ changes in a wing.

In general, our model is robust—one can change the parameters Δ, β, ζ, and η to fine-tune the curves for the particular characteristics under consideration or use different functional forms altogether.

5.2. Selecting a Sequence $\mathcal{S}$

Our DP from §4.2 can take any sequence $\mathcal{S}$. We use the following sequence for our numerical study in §6:

$$\bar{\mathcal{S}}\text{: the sequence formed when care types are sorted by nominal daily utility } v^i/m^i. \quad (7)$$

When restricted to cuts in $\bar{\mathcal{S}}$, solutions will tend to form wings with care types of similar daily utility. This sequence is intuitively effective, since it allows some wings to function as protectors of beds for care types of higher utility and other wings to function as limiting beds for types of lower utility. Furthermore, if a wing is formed with care types of similar daily utility, then the hospital administrator is mostly indifferent to the resulting competition for beds within the wing.

In our numerical experiments, we find that $\bar{\mathcal{S}}$ yields high-quality solutions as discussed in §7; $\bar{\mathcal{S}}$ is also supported by Proposition 1, which states with some stipulations that if two care types have the same nominal daily utility, then they should be placed together in a wing. Its proof is given in Online Appendix A.

Proposition 1. *Consider an instance with* $\mathcal{C} = \{1, 2\}$ *under the queueing system of* §4.3. *If* $v^1/m^1 = v^2/m^2$, $q \geq \max\{m^1, m^2\}$, *and there are no focus effects, then* $U^{1,2} \geq U^1 + U^2$, *where* U^i *is the expected utility from a wing for type* $i \in \mathcal{C}$ *with* $b^i \geq 0$ *beds, and* $U^{1,2}$ *is the expected utility from a pooled wing with* $b^1 + b^2$ *beds.*

There are certainly additional considerations besides utility when forming a sequence $\mathcal{S}$, and in practice one would try many candidate sequences. For example, a hospital can leverage similarities in the diseases and the care delivery processes of particular types by placing them adjacent in the sequence; that is, trading off some similarities in utility for strong cohesion of care. For example, the UCMC types gynecology (GYN) and gynecology/oncology (GOC) treat relatively similar diseases, and nurses with training in GOC are typically cross-trained in GYN and vice versa, so it is natural to place them adjacent in the sequence so that the model has an easy option of combining them into the same wing. In contrast, care types that are undesirable to have in the same wing can be separated in the sequence. In §7, we investigate the robustness and importance of selecting a sequence by comparing the utility-based sequence $\bar{\mathcal{S}}$ to a medically based sequence, as well as randomly generated sequences.

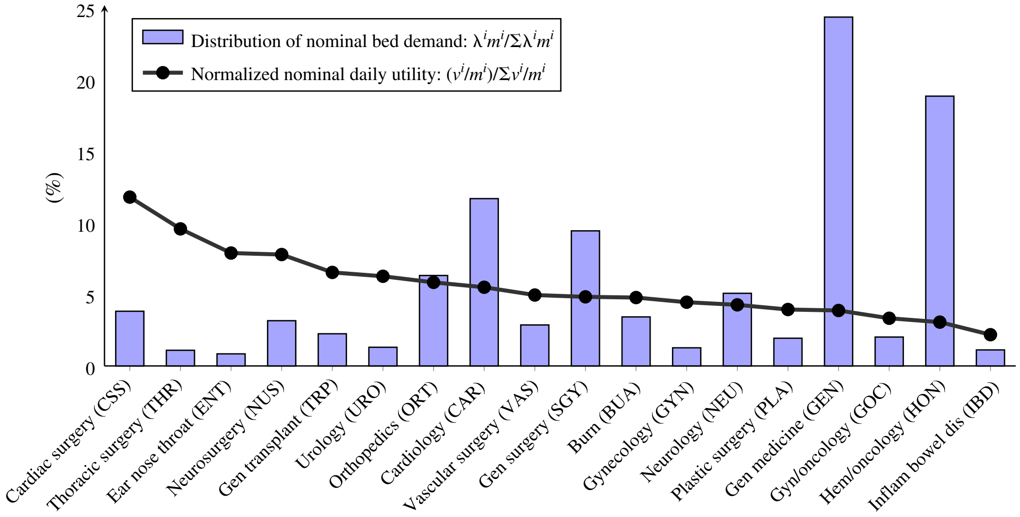
6. Numerical Study

This section summarizes our numerical results and the insights gained about the wing formation decisions. In particular, §6.1 summarizes our estimation of exogenous model parameters, and §§6.2 and 6.3 present a set of results along with a discussion of their managerial implications.

6.1. Setting Model Parameters

Defining the Set of Care Types ($\mathcal{C}$). Defining each care type broadly, and hence choosing a small number T of care types, forgoes the potential benefits of optimization and focus. On the other hand, defining each care type narrowly, and hence choosing a large T, may cause the solution to be nonsensical, since it may be impossible to correctly identify patients upon arrival into their designated care type. We focus on adult inpatient care only. And we exclude ICU, obstetrics, and psychiatric care, since such services typically have beds exclusively assigned for their use. We take the 18 medical-surgical services (see the x-axis labels in Figure 4) that are predefined at UCMC, which

Figure 4 **(Color online) An Illustration of the Variability in the Demands and Utilities of the 18 Care Types Seen at UCMC**



represent about 99% of the inpatient care at UCMC. Such a choice is natural and suitable for implementation, since patients are currently assigned to one of these services upon arrival at UCMC. Lacking predefined care types, one could use classification tools similar to those discussed in Harper and Shahani (2002).

Estimating the Arrival Rates (λ^i). True arrival rates are impossible to know for a capacity constrained hospital like UCMC because of the censoring that occurs when demand exceeds capacity. Instead we use the Nationwide Inpatient Sample (NIS), part of the Healthcare Cost and Utilization Project (http://www.hcup-us.ahrq.gov, accessed June 27, 2009). NIS is the largest publicly available, all-payer inpatient care database for U.S. hospitals. It contains over eight million discharges per year from over 1,000 hospitals across 40 states. Restricting the database to large urban teaching hospitals only, we estimate the *relative* arrival rate of each care type. Such a data set is less censored than data from a single hospital, because care would more likely be received at some other hospital when a particular hospital is at capacity.

We treat the obtained distribution of arrival rates across care types as fixed. However, in sensitivity analysis, we scale this distribution with a constant factor as we change the nominal offered load ρ to the hospital, where ρ equals the ratio of the overall arrival rate $\sum_i \lambda^i$ to the overall nominal service rate, which is the product of the number of beds B and the reciprocal of the mean length-of-stay $\sum_i \lambda^i m^i / \sum_i \lambda^i$, and so $\rho = \sum_i \lambda^i m^i / B$.

Estimating the Nominal Utilities (v^i). To illustrate our model, we use a commonly applied measure of care complexity, namely, the DRG relative weight. Each patient is classified by one of over 500 DRGs upon discharge from a hospital. Each DRG has an associated *relative weight* published by CMS. The relative weight of a DRG considers the resources consumed for the care provided, and hence it is a measure of the cost of care and relates closely to the complexity of care and profitability for the hospital. Hospitals compute the average relative weight of all admitted patients to obtain their case-mix index (CMI) and use it as one of the primary sources of comparison with competing hospitals. A hospital such as UCMC strives to increase its CMI (Karlin 2013) in order to ensure financial health and sufficient complexity for its research and teaching missions. We set the nominal utility v^i to be the average DRG relative weight, weighted by the relative volume of each DRG designated within care type i across a typical year at UCMC, and therefore our objective function seeks to maximize the total expected relative weights per day of the hospital.

Estimating the Nominal Lengths-of-Stay (m^i). We use data at UCMC for the year prior to the implementation of wings to estimate m^i, the average length-of-stay of the patients seen within care type i. Our estimated m^i values range from 2.8 days for GYN to 9.4 days for cardiac surgery (CSS).

We compute the coefficient of variation (CV) for all possible combinations of care types that can be assigned to a wing. Across all $2^{18} - 1$ combinations, the average CV is 1.09, with a standard deviation of only 0.06, with the first and 99th percentiles of 0.97 and 1.21, respectively. Since the CV for the lengths-of-stay are close to 1, our use of the exponential distribution for the numerical runs in §6.2 corresponds to the suggestion in Green (2006). Our statistical analysis shows that the lengths-of-stay data more closely follow a log-normal distribution for most care types (which follows Shi et al. 2013). However, simulating the wing formations we later identify in this section using log-normal instead of exponential, we find that the change in hospital utility is at most 0.09%, so we conclude that our analytical model using exponential distributions is sufficiently accurate. In §6.3, we do relax the exponential assumption somewhat, by allowing a mixture of exponential distributions.

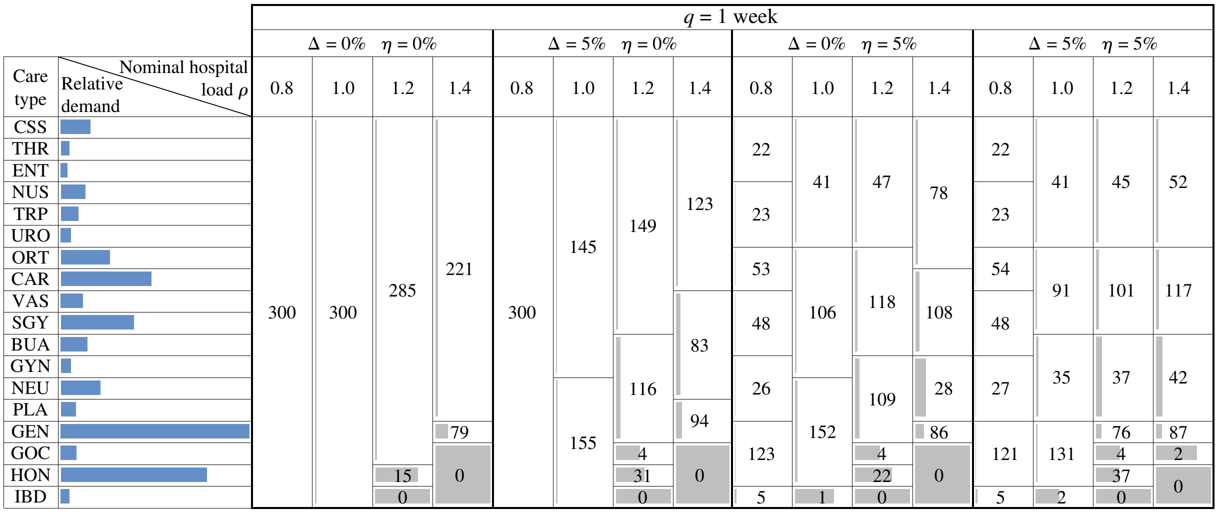
Figure 4 displays the distribution of bed demand across the 18 care types as well as the normalized daily utility of each type. Care types with very similar bed demand may create significantly different daily utility for the hospital—for example, care types THR and IBD have almost identical bed demand, but THR provides more than three times the daily utility. This indicates that it may be advantageous to form wings to restrict beds for low-utility types with large demand, and form other wings to reserve beds for care types of higher utility, especially if overall capacity is strained or there are significant benefits from focused care.

To test this intuition, we now apply our model to the data from Figure 4 under multiple instances of model parameters. For all instances solved, we set the total number of inpatient beds $B = 300$. For the logistic function parameters, extensive numerical tests suggest that the structure of the solutions are less sensitive to reasonable changes in β and ζ than to changes in Δ. Thus, we simplify our exposition, and fix $\beta = 20$ and $\zeta = 0.9$ (as in Figure 3), so that there is a negligible reduction in nominal length-of-stay for relatively low levels of load (i.e., less than 70%), and it gradually increases reaching its peak rate at 90% load and leveling off at about 110% load, indicating no further clinically justifiable reduction in length-of-stay.

6.2. Results

Each of the 16 columns in Figure 5 depicts the solution we obtain for a problem instance using the DP

Figure 5 (Color online) Impact of Offered Load and Focused Care



Percent change in overall hospital utility as compared to the DP-based solution

		Δ = 0% η = 0%				Δ = 5% η = 0%				Δ = 0% η = 5%				Δ = 5% η = 5%			
	ρ	0.8	1.0	1.2	1.4	0.8	1.0	1.2	1.4	0.8	1.0	1.2	1.4	0.8	1.0	1.2	1.4
(A)	One wing	0.0	0.0	−5.0	−9.8	0.0	−0.2	−6.1	−11.5	−3.7	−1.9	−7.0	−12.3	−3.8	−3.1	−9.3	−14.7
(B)	The four UCMC wings	−0.6	−5.0	−8.6	−12.3	−0.5	−3.9	−7.7	−11.8	−1.9	−4.4	−8.2	−12.5	−1.8	−4.3	−8.5	−12.6

Percent change in overall hospital bed occupancy as compared to the DP-based solution

		Δ = 0% η = 0%				Δ = 5% η = 0%				Δ = 0% η = 5%				Δ = 5% η = 5%			
	ρ	0.8	1.0	1.2	1.4	0.8	1.0	1.2	1.4	0.8	1.0	1.2	1.4	0.8	1.0	1.2	1.4
(C)	One wing	0.0	0.0	+1.5	+1.0	0.0	+2.3	+1.8	+1.8	+0.4	+2.2	+2.7	+1.7	+1.3	+4.5	+2.6	+2.3
(D)	The four UCMC wings	−0.6	−4.4	−0.6	+0.6	−1.1	−2.9	−0.9	+1.1	−0.2	−2.3	+0.6	+1.3	+0.2	−0.8	−0.1	+1.6

from §4.2 with the sequence $\bar{\mathcal{S}}$. Three parameters identifying each instance vary along the horizontal. We vary Δ, the parameter controlling the maximum possible reduction in lengths-of-stay from focused care, from 0 (no gain) to 5% (a conservative maximum reduction). Similarly, we vary η, the parameter controlling the maximum possible gain in utility from focused care, from 0 to 5%. Finally, we also vary the parameter ρ, the nominal offered load to the hospital, from 0.80 to 1.40. We can therefore identify each instance with the triplet (Δ, η, ρ). In all of these instances, the average willingness-to-wait for all wings, q, is set to one week.

The care types are sorted from top to bottom along the vertical axis of Figure 5, in decreasing daily utility (i.e., the $\bar{\mathcal{S}}$ sequence). The distribution of nominal bed demand from Figure 4 is repeated beside each care type label. Within each depicted solution, wings are denoted by horizontal dividing lines, with the number of beds allocated to each wing indicated. For example, for the first column, where $(\Delta, \eta, \rho) = (0, 0, 0.8)$, the solution keeps all care types together into one wing with all 300 beds. Intuitively, when bed demand is low and there are no potential benefits from focus, there is no reason to segregate patient types; one should instead take full advantage of pooling capacity.

The proportion of gray shading in each wing corresponds to the expected fraction of patients that abandon the queue for the wing before receiving care. More shading also implies longer waits for those patients willing to wait for admission. Observe that wings with significant shading tend to have care types of lower daily utility. For example, the third column $(0, 0, 1.2)$ indicates that IBD patients are never admitted (the solid shading corresponding to zero beds allocated), and about 20% of HON patients are expected to receive care at the hospital. On the other hand, only a very small fraction of the top 16 care types (the ones that are allocated 285 beds) leave the hospital without being seen.

Row (A) displays, for each instance, the percent change in overall hospital utility when the depicted solution is compared to the *one-wing solution* (i.e., a pooled wing with 300 beds shared by all 18 types), while row (B) displays the same when compared to the four wings that were implemented by UCMC in July 2006 (i.e., one wing with 69 beds for GEN, another with 30 for CAR, a third with 72 for HON, and a final wing with 129 beds for all other types). Rows (C) and (D) make the same comparisons for overall hospital bed occupancy.

We now discuss managerial insights from our model that are supported by these numerical results.

In an overloaded hospital (i.e., $\rho \geq 1$), increased load leads to more restricting of beds for low-utility care types. As hospital demand increases beyond available bed capacity, it becomes increasingly advantageous to restrict beds to lower-utility types while reserving beds for higher-utility types. Compare, for example, columns $(0, 0, 1.0)$, $(0, 0, 1.2)$, and $(0, 0, 1.4)$ in Figure 5. As load increases from 1.0 to 1.2, the DP-based solution moves from a fully pooled design to a formation in which the two types with lowest utility are severely restricted. As the load further increases to 1.4, these last two types along with GOC are restricted further with no beds allocated. A similar phenomenon is also observed by comparing the last three columns under each (Δ, η) combination. As overall demand grows, the hospital increasingly restricts access to lower-utility care types, thereby shifting beds "up" for the use of higher-utility care types. In fact, with enough demand, one would form a focused hospital reserving all of the bed capacity for the highest-utility care type (e.g., a completely focused facility dedicated to CSS), unless constraints are added.

Splitting into wings tends to decrease bed occupancy. Row (C) shows that as wings are formed, the overall bed occupancy decreases. This is because most beds are allocated to high-utility care types with some slack capacity (observe the very little gray shading for high-utility wings).

Proposition 2 formalizes this observation by establishing conditions under which splitting into wings always decreases bed occupancy. Its proof is given in Online Appendix A.

Proposition 2. *Consider an instance of* (P) *under the queueing system of* §4.3, *and the one-wing solution (denoted "O") and a multiwing solution (denoted "M"). If the mean willingness-to-wait in "O" is not smaller than the mean willingness-to-wait in any wing in "M" or the nominal mean length-of-stay of any care type, and the focus-adjusted mean length-of-stay of each care type in "O" is not smaller than that for "M," then the bed occupancy of "O" is higher than the weighted average bed occupancy of "M."*

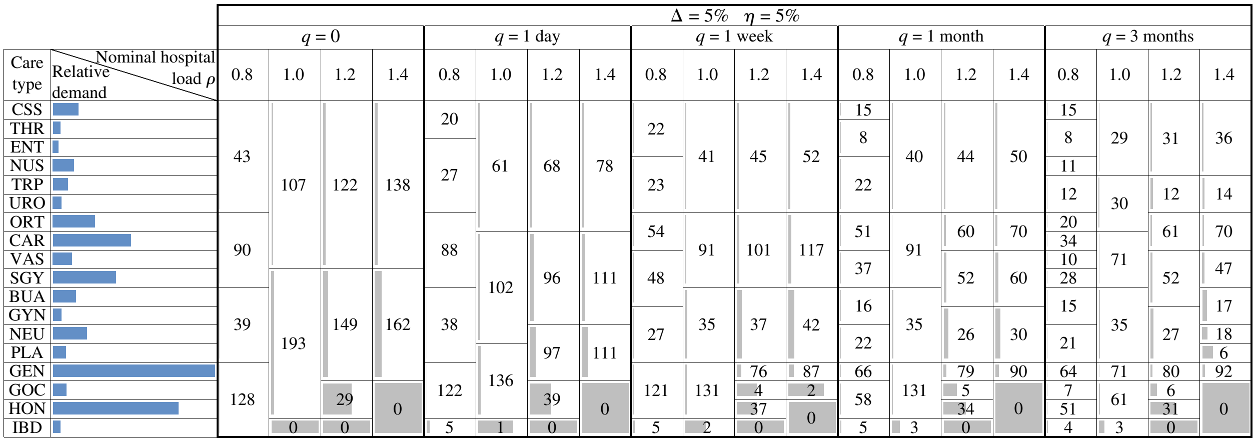
Despite the decrease in bed occupancy, we note that increased wing formation may also lead to serving more patients. This occurs when the benefits of focus are sufficiently large so that lengths-of-stay decrease significantly. From Little's Law, the throughput of the hospital can increase by as much as a factor of $1/(1-\Delta)$. In fact, our model can be easily adapted to maximize patient admission rates (i.e., throughput), by setting $u^i = 1$ for all care types. Setting each $u^i = 1$ also minimizes a linear function of the expected waiting times. To see this, let W_j be the expected waiting time for wing j prior to admission. When $u^i = 1$ for all i under the queueing system of §4.3, using $p_j = W_j/q_j$ (Garnett et al. 2002), the objective function of problem (P) can be rearranged as $\sum_{i\in\mathcal{I}} \lambda^i - \sum_{j=1}^{w} W_j(\Lambda_j/q_j)$. Observe that all terms except W_j in this objective are constants, so the objective function is now equivalent to minimizing (a linear function of) expected waiting times.

Increased benefits from focus lead to more wing formation. As the potential benefits of focus increase, it becomes more advantageous to form wings, while forgoing some of the benefits of pooling. An increase in Δ promotes wing formation due to a potential decrease in lengths-of-stay, and thus as ρ also gets large the hospital administrator has more incentive to form focused wings to decrease the bed demand through reducing lengths-of-stay and therefore admitting more patients, resulting in higher overall utility. This phenomenon is most clearly depicted in Figure 5 by comparing each instance in the first four columns with its counterpart in the next four columns, so that only the parameter Δ is changed from 0% to 5% in each comparison.

When $\eta > 0$, a focused wing has the potential of increasing the utility per patient served. Consider the four columns under $(0, 5\%, \cdot)$. When the load is small, $\rho = 0.8$, the system has plenty of slack capacity; thus forming wings to increase utility, and appropriately allocating capacity to each wing so that almost no patients are turned away, is more beneficial than advantages of pooling, and so considerable wing formation occurs. But as ρ increases, the advantages of pooling become a factor to offset the utility advantages of focused wing formation, and so the number of wings formed when $\rho \geq 1.0$ is more moderate compared to $\rho = 0.8$, but it again increases with ρ. When $\Delta = 5\%$ and $\eta = 5\%$ (the last four columns), the solutions mostly mimic those under $(0, 5\%, \cdot)$ and so considerable wing formation occurs because focused wings gain advantages in both lengths-of-stay and patient utility. The objective of maximizing utility is also clearly reflected on these results—they demonstrate the more pronounced impact of the utility benefits of focused care on the structure of the wing formation in comparison to the length-of-stay benefits of focused care.

Increased hospital utility coincides with unequal access. As shown in row (A) of Figure 5, wing formation can lead to substantial increases in overall hospital utility. However, such gains in utility are accompanied by disparity in access, as indicated by the gray shading in each wing. Consider the $(0, 0, 1.2)$ instance that reserves 285 beds for the 16 care types with the highest utility, 15 beds for HON, and no beds for IBD. Forming the wings as indicated by our model increases the overall utility by about 5% compared to the one-wing formation. However, only about 20% of HON patients on average and no IBD patients are admitted under the DP-based solution, while 97% of the higher-utility patients gain access. The one-wing formation can be viewed as more equitable in that each care type receives equal access.

Figure 6 (Color online) Impact of Offered Load and Willingness-to-Wait for Service



Percent change in overall hospital utility as compared to the DP-based solution

		q = 0				q = 1 day				q = 1 week				q = 1 month				q = 3 months			
		0.8	1.0	1.2	1.4	0.8	1.0	1.2	1.4	0.8	1.0	1.2	1.4	0.8	1.0	1.2	1.4	0.8	1.0	1.2	1.4
(A)	One wing	−3.0	−2.3	−6.8	−11.5	−3.4	−2.6	−7.5	−12.8	−3.8	−3.1	−9.3	−14.7	−4.1	−3.6	−10.6	−15.9	−4.3	−3.9	−11.5	−16.7
(B)	The four UCMC wings	−2.3	−5.4	−7.9	−11.2	−1.9	−4.6	−7.5	−11.3	−1.8	−4.3	−8.5	−12.6	−1.8	−4.5	−9.8	−13.7	−1.9	−4.8	−10.6	−14.3

Percent change in overall hospital bed occupancy as compared to the DP-based solution

		q = 0				q = 1 day				q = 1 week				q = 1 month				q = 3 months			
		0.8	1.0	1.2	1.4	0.8	1.0	1.2	1.4	0.8	1.0	1.2	1.4	0.8	1.0	1.2	1.4	0.8	1.0	1.2	1.4
(C)	One wing	+1.4	+3.5	+5.0	+4.4	+1.6	+4.6	+4.1	+4.1	+1.3	+4.5	+2.6	+2.3	+1.3	+3.9	+1.6	+1.2	+1.4	+4.2	+1.0	+0.8
(D)	The four UCMC wings	−1.0	−3.3	+0.1	+1.4	−0.1	−1.5	+0.5	+2.5	+0.2	−0.8	−0.1	−1.6	+0.4	−1.2	−0.9	+0.9	+0.6	−0.9	−1.3	+0.7

The four wings at UCMC. Rows (B) and (D), respectively, compare the utility and occupancy performances of the four wings implemented at UCMC to the DP-based solution. Since the UCMC is capacity constrained and enjoys benefits of focus, the last two columns are the most relevant. We observe that UCMC falls significantly short in utility of the DP-based solution, whereas it surpasses in occupancy. This trade-off between utility and occupancy is certainly affected by some UCMC constraints that are not considered in the DP. For example, contrary to the DP model, UCMC has allocated some number of beds to every care type regardless of their utility. In addition, UCMC allocated more beds for HON than the DP-based solution, in spite of its relatively low daily utility (see Figure 4). This discrepancy may be caused by the fact that the utility measure used in the DP model ignores insurance reimbursements, while in reality HON patients at UCMC have a significantly larger fraction of high-paying private insurance than many other care types. Nevertheless, the actual constraints and the objectives used by UCMC to form their wings were not formalized, and so it is not possible to precisely mimic their decision process within the DP.

Increased willingness-to-wait leads to greater wing formation. As patients become more willing (or able) to wait for a bed, the randomness in the arrival stream becomes less important, as patients buffer in queue. Therefore, the advantages of pooling care types together is mitigated, and thus we expect more wings to be formed. To test for this intuition, in Figure 6, we vary q, the average time patients are willing to wait for a bed, from 0 (not willing to wait) to three months, as well as the offered load ρ from 0.80 to 1.4. In each of these instances we set the parameters for focused care at $\Delta = \eta = 5\%$. These results provide evidence in support of our intuition. This phenomenon is most clearly displayed when $q = 3$ months (the last four columns). Considerable wing formation occurs in this case, since longer average willingness-to-wait diminishes the advantages of pooling. Therefore, because patients are willing to wait longer, and $\Delta > 0$ and $\eta > 0$, the hospital takes full advantage of forming focused wings.

The willingness-to-wait of a patient may also be interpreted as a model for the type of hospital. For example, a less renowned hospital may not have the ability to induce a large q. Our model predicts that such a hospital may tend to form fewer wings, becoming a facility for more general care. In contrast, a more renowned hospital may invoke a larger q and thus can form more wings and take advantage of focused care. These results also have implications for urgency of care. That is, a care type corresponding to more elective care, or care for which some notice or plan can be accommodated, can be handled by more focused wings, whereas emergency care is more economically handled by pooling capacity.

6.3. Extension to Two Patient Classes

In practice, some care types may exhibit within-type heterogeneity that merits an additional level of classification. For example, consider two patients in need of an angioplasty, one for a planned procedure and the other after a sudden cardiac arrest. Both patients would be treated by the cardiology service, but when they both request admission, the hospital is likely to give higher priority to the cardiac arrest, who is less able to wait for care. Furthermore, they may have different lengths-of-stay and may yield different utilities.

To model such heterogeneity, we extend the single-class $M/M/b_j + M$ queueing system to include two classes. In particular, each care type is now composed of two classes, say an urgent class and a nonurgent class. Each class has its own arrival rate, length-of-stay, utility, and willingness-to-wait. We assume urgent requests are not able to wait long, so we set the average willingness-to-wait of the urgent class to be no larger than that of the nonurgent class. We also assume urgent requests have non-preemptive priority: as a bed opens up in a wing, any awaiting urgent patients for that wing are admitted before nonurgent ones.

When applying the revised model, we define the urgent class to be patients admitted through the emergency department and adjust the arrival rate of each class accordingly. Our data indicate that some care types have more of their demand arrive as urgent (e.g., 71% of GEN), while others have less (e.g., 31% of HON). Overall, our data indicate that about 44% of all admits are urgent.

We use the same solution methodology as before, but the only additional challenge is the computation of the abandonment probability for each class. When computing these probabilities for a wing, we first employ a standard t-test with 99% confidence level to check if the two classes in the wing differ in their average lengths-of-stay. If they do not differ, then we use the exact expressions provided in Sarhangian and Balcıoğlu (2013) to calculate the abandonment probabilities. If classes also differ in their lengths-of-stay, then approximations are necessary. We proceed by calculating the abandonment probabilities by solving the balance equations of a birth-death process with a sufficiently large but *finite* queue capacity that limits the queue blocking probability to less than 10^{-6}.

To isolate the impact of prioritizing urgent patients, we re-solve the instances in Figure 6 using the two-class model assuming identical average willingness-to-wait for both classes. When compared to the solutions in Figure 6, we find only minor differences in the wings formed, implying that, if urgent and nonurgent patients have similar willingness-to-wait, our single-class model would be a good substitute for the more sophisticated two-class model for the purpose of forming wings. However, the two models may deviate significantly for other metrics. For example, we observe that the overall bed occupancy has decreased with the two-class model, which can be explained by the preemptive priority: the priority given to the urgent class extends the delay for the nonurgent class patients compared to their delay in the single-class setting, and therefore a higher fraction of patients leave the hospital without being seen.

Next, we vary the willingness-to-wait for both classes in such a way that the overall average willingness-to-wait for the entire system matches that for the single-class model. That is, when the urgent class becomes less willing to wait, on average, the nonurgent class becomes more willing to wait. We have observed in our single-class results that as patients are less willing to wait pooling becomes the preferred choice, but if they are more willing to wait it is better to form more focused wings. In this new set of experiments, we find that there is a slight tendency toward pooling as the average willingness-to-wait of the urgent class decreases, while that of the nonurgent class is simultaneously increasing. The priority access of the urgent class plays the main role in this pattern. A slight decrease in the willingness-to-wait of the urgent class is more influential than the same order increase in the willingness-to-wait of the nonurgent class, because the urgent class is given priority access to the hospital, and therefore the pooling forces win.

7. Quality of Solutions and Robustness of Sequence Selection

We are able to evaluate the quality of the DP-based solutions obtained via the methodology of §§4.1 and 4.2 using the utility-based sequence $\bar{S}$. By exploiting the special structure of $\bar{S}$, we construct an upper bound for the optimal objective value Z under the queueing system of §4.3. We provide the technical details, including the proofs of the results in this section, in Online Appendix B. In summary, we transform an instance Θ of problem (P) into a new instance $\hat{\Theta}$, by splitting the original care types into subtypes, so that all subtypes have the same arrival rate and the same length-of-stay. Solving $\hat{\Theta}$ results in an upper bound $Z(\hat{\Theta})$ for the optimal objective value $Z(\Theta)$ of the original instance Θ, because every candidate solution in the feasible region of Θ has an equivalent candidate solution in the feasible region of $\hat{\Theta}$.

Theorem 1. $Z(\hat{\Theta}) \geq Z(\Theta)$.

An upper bound on $Z(\Theta)$ is only useful if it can be computed efficiently. At first glance, $Z(\hat{\Theta})$ is seemingly even more difficult to compute than $Z(\Theta)$, because our transformation has increased the number of care types by splitting original types into subtypes. To efficiently compute an upper bound, we construct a modified problem $(\hat{P})$ that augments the set of decision variables of (P). The technical details of this construction are provided in §B.2 of Online Appendix B.

Theorems 2 and 3 state that an instance of problem $(\hat{P})$ based on the data of $\hat{\Theta}$ can be solved to optimality by applying the DP from §4.2 along with the utility-based sequence $\bar{\mathcal{S}}$ of subtypes, and its optimal value $\hat{Z}(\hat{\Theta})$ is an upper bound for $Z(\Theta)$

Theorem 2. *An instance of problem $(\hat{P})$ can be solved to optimality using the DP of §4.2 by making cuts in the utility-sorted sequence $\bar{\mathcal{S}}$.*

Theorem 3. $\hat{Z}(\hat{\Theta}) \geq Z(\Theta)$.

Table 1 summarizes the results of using the upper bound $\hat{Z}(\hat{\Theta})$ to evaluate the quality of the solutions in §6.2. For all scenarios in this table, we set $T = 18$, $B = 300$, $\beta = 20$, $\zeta = 0.9$, and take v^i and m^i for all $i \in \mathcal{C}$ from the data displayed in Figure 4. In many of our computational tests, we find negligible optimality gap: its average is 1.01%, but in some cases (particularly, when η, Δ, and ρ are high) it can be as high as 2.91%. These findings suggest that our heuristic would perform well in many practical situations.

Whereas we were able to derive bounds to show the quality of our DP-based solutions because of the special structure of $\bar{\mathcal{S}}$, we now demonstrate that other sequences tend to also work very well—that the key to good wing formations is much more dependent on the DP optimization, rather than the selection of a sequence. For this purpose, we generate 1,000 random sequences, and re-solve each instance (i.e., column) in Figures 5 and 6 for each one of these sequences. For a given instance, we compute $\phi_i = (Z_i - Z_{\bar{\mathcal{S}}})/Z_{\bar{\mathcal{S}}}$, for $i = 1, 2, \ldots, 1{,}000$, to quantify the relative performance of the random sequence i over the utility-based sequence $\bar{\mathcal{S}}$, where $Z_{\bar{\mathcal{S}}}$ and Z_i, respectively, denote the objective value associated with the sequence $\bar{\mathcal{S}}$ and that associated with the random sequence $i = 1, 2, \ldots, 1{,}000$.

The results show that $\bar{\mathcal{S}}$ outperforms the random sequences for a significant majority (82.2%) of the cases but only by a narrow margin. The values of ϕ_i across all instances range from -1.85% to $+0.36\%$ with an average of -0.23%, which indicate that there are some random sequences that produce slightly better objective values than $\bar{\mathcal{S}}$, but, on average, $\bar{\mathcal{S}}$ is a better sequence in maximizing utility. Nevertheless, as illustrated by the average value, the differences are so small that we conclude that the underlying value of the wing solutions obtained is primarily attributable to the DP optimization as opposed to a choice in sequence.

We further scrutinize $\bar{\mathcal{S}}$ by comparing it to a medically based sequence. We identify a medically based sequence by obtaining a set of cohesion scores for each pair of care types from UCMC, which we use as edge weights in a complete graph formed from the 18 UCMC care types. Solving for a longest Hamiltonian path in this graph results in the medically based sequence: BUA-PLA-VAS-THR-CSS-CAR-GEN-NEU-NUS-ENT-HON-IBD-SGY-URO-TRP-ORT-GOC-GYN. This sequence has some important differences from $\bar{\mathcal{S}}$. For example, the cohesive GOC and GYN are now adjacent, whereas they are separated by three other types in $\bar{\mathcal{S}}$. Yet, across all problem instances in Figures 5 and 6, the average loss in utility by switching to this medically based sequence is only 0.17%, following our previous insights made with random sequences. Furthermore, we emphasize that the main insights highlighted in §6.2 continue to hold using the medically based sequence.

In light of these results, we argue that a hospital administrator in practice will be able to try different sequences, and many, if not all of them, will be expected to perform well. So the administrator is easily able to deviate from using $\bar{\mathcal{S}}$ and instead uses a more preferred and perhaps easier to implement medically based sequence (or any others that are deemed appropriate).

Table 1 **Optimality Gap: Percentage Difference Between Lower and Upper Bounds**

Willingness-to-wait parameter (q)	Utility scaling parameter (η) (%)	Length-of-stay scaling parameter (Δ) (%)	Nominal hospital load (ρ)			
			0.8	1.0	1.2	1.4
0	0	0	0.00	0.00	0.01	0.40
		5	0.00	0.06	0.92	1.26
	5	0	0.28	0.19	1.77	1.92
		5	0.28	0.36	2.39	2.83
1 week	0	0	0.00	0.89	0.88	0.86
		5	0.00	1.25	1.99	2.07
	5	0	0.64	1.60	2.72	2.34
		5	0.61	1.81	2.77	2.91
3 months	0	0	0.00	0.46	0.41	0.38
		5	0.00	0.49	1.13	1.06
	5	0	0.25	0.92	1.60	1.33
		5	0.23	1.21	1.57	1.49

8. Concluding Remarks

The effective use of hospital bed capacity is becoming increasingly critical as financial concerns worsen and demand grows. Forming wings (i.e., the partitioning of beds and care types into specialized wings) is an important strategy for a hospital to manage its available beds. We are the first to consider wing formation decisions within an optimization framework and, furthermore, to consider the critical tradeoff between pooling capacity and focusing care. We capture effects from pooling by embedding sophisticated yet tractable queueing models within our optimization framework. We capture effects from focus in two distinct ways, each motivated by the literature and our collaboration with UCMC; focused care may affect the utility gained from each patient and/or a patient's length-of-stay.

We apply our framework to data from UCMC and a national database and report a number of insights: (i) As the overall demand for care increases, the hospital will tend to form wings to reserve beds for higher-utility patients and restrict beds for lower-utility patients; (ii) as the hospital splits into wings to improve its utility, it suffers from reduced bed occupancy and increased disparity in access to care; (iii) as patients are willing (or able) to wait longer for care, the benefits of pooling become mitigated for the hospital, and therefore forming more specialized wings becomes increasingly beneficial; and (iv) when there exists a significant amount of urgent demand that are less able to wait, the benefits of pooling dominate those of specialized wings. An interesting implication of these findings is that hospitals that can garner longer waits (maybe because of their reputation for better care or a lack of alternatives) will be able to form more specialized wings and thereby reap more benefits of focus. On the other hand, hospitals that cannot garner long waits will tend to form few, if any, wings, instead choosing to pool demand and capacity.

Our framework is capable of handling important nuances of a realistic environment. It can be used to make recommendations for an existing hospital as well as for designing a new hospital. It is agnostic to how the hospital's mission or utility is defined; this can indeed be financial in nature, but it can also represent some altruistic measure of care.

The case when the planning horizon consists of seasonal variations in the problem parameter values, such as significant changes in patient arrival patterns, can be handled by revising the state definition and optimality equations of our DP formulation, so that the total utility associated with a particular wing formation is now an accumulation of the utilities across all seasons in the planning horizon, as opposed to an average season.

This work considered the wing formation decisions for an individual hospital. The more difficult question of forming wings when multiple hospitals strategically make such decisions is left for future research. In such a strategic environment, hospitals may both cooperate (as UCMC does to manage the overflow of GEN patients to a community hospital) or compete.

Future research might also investigate the impact of additional flexibility. For example, care types might be assigned a primary wing but have access when needed (perhaps with low priority) to a secondary wing(s). This type of flexibility is intractable under our optimization framework, since it prevents one from pricing each formed wing independently of the others. However, one can heuristically extend our current framework to allow a "general" care type with high demand to have access to secondary wings. This would be most practically applicable to GEN patients if one assumes that there is a sufficient supply of GEN patients to fill slack beds in select wings (as is approximately the case for many urban hospitals). Then one could use our approach; except, when pricing select wings in our DP, make the assumption that some fraction of the average unoccupied beds in the wing can be filled with GEN patients, thereby increasing the overall utility of the patients in these wings. The DP would then tend to form wings with a bit more slack bed capacity since the cost of bed overage has been reduced. This is a heuristic, since in practice such overflow additions to the wing may indeed impact the primary care types of the wing. Handling this type of primary/secondary wing flexibility with a more accurate optimization approach is left for future research.

There is also need for empirical research regarding the effects of focus on lengths-of-stay, utilities, and patients' willingness-to-wait. The functional forms we use for $u^i(\psi)$ and $s^i(\psi)$ are motivated by some empirical evidence and interviews with hospital administrators, but further efforts may estimate different parameter values for a particular setting or suggest completely new functional forms. It is, however, important to note that our optimization framework and solution methods are robust to such refinements.

Our framework assumed that the reductions in lengths-of-stay because of focused care are modest—in our numerical runs, for example, we used at most a 5% reduction. Our current framework does not easily extend to the case when excessive reductions in lengths-of-stay cause adverse medical events that essentially result in an increase in readmission rates to the hospital. Incorporating such repercussions of a change in lengths-of-stay is left for future research.

Supplemental Material

Supplemental material to this paper is available at http://dx.doi.org/10.1287/msom.2014.0516.

Acknowledgments

The authors thank former editor Stephen Graves, an anonymous associate editor, and three anonymous referees for their constructive suggestions that helped improve the paper. The authors also thank the following individuals from the University of Chicago Medical Center for detailed discussions and for providing data: Elmer Abbo (assistant professor of medicine), Jean Blake (former director of Clinical Operations Integration), and Stephen Weber (chief medical officer); and the following individuals at NorthShore University HealthSystem for providing comments on an earlier draft of the paper: Jesse Peterson Hall (president of Highland Park Hospital) and Ken Anderson (chief medical quality officer). The authors alone are responsible for interpreting and reporting any information from these individuals. The authors also thank Barış Ata (professor of operations management) for many helpful comments on earlier drafts. Thomas Best is supported by the Merck Quantitative Sciences Graduate Fellowship in Health Economics. Thomas Best, Burhaneddin Sandıkçı, and Donald Eisenstein gratefully acknowledge the financial support from the University of Chicago Booth School of Business. This research used resources from the University of Chicago Research Computing Center.

References

American Hospital Association (2011) TrendWatch Chartbook: Trends affecting hospitals and health systems. Accessed January 31, 2015, http://www.aha.org/research/reports/tw/chartbook/2011chartbook.shtml.

Ayvaz N, Huh WT (2010) Allocation of hospital capacity to multiple types of patients. *J. Revenue and Pricing Management* 9(5): 386–398.

Bazzoli GJ, Brewster LR, Liu G, Kuo S (2003) Does U.S hospital capacity need to be expanded? *Health Affairs* 22(6):40–54.

Berk E, Moinzadeh K (1998) The impact of discharge decisions on health care quality. *Management Sci.* 44(3):400–415.

Capkun V, Messner M, Rissbacher C (2012) Service specialization and operational performance in hospitals. *Internat. J. Oper. Production Management* 32(4):468–495.

Chapman SN, Carmel JI (1992) Demand/capacity management in health care: An application of yield management. *Health Care Management Rev.* 17(4):45–54.

Clark JR, Huckman RS (2012) Broadening focus: Spillovers, complementarities, and specialization in the hospital industry. *Management Sci.* 58(4):708–722.

Dumas MB (1984) Simulation modeling for hospital bed planning. *Simulation* 43(2):43–69.

Garnett O, Mandelbaum A, Reiman M (2002) Designing a call center with impatient customers. *Manufacturing Service Oper. Management* 4(3):208–227.

Gorunescu F, McClean SI, Millard PH (2002) A queueing model for bed-occupancy management and planning of hospitals. *J. Oper. Res. Soc.* 53(1):19–24.

Graves SC, Redfield CH (1988) Equipment selection and task assignment for multiproduct assembly system design. *Internat. J. Flexible Manufacturing Systems* 1(1):31–50.

Green LV (2004) Capacity planning and management in hospitals. Brandeau ML, Sainfort F, Pierskalla WP, eds. *Operations Research and Health Care: A Handbook of Methods and Applications* (Kluwer Academic Publishers, Boston), 15–41.

Green LV (2006) Queueing analysis in healthcare. Hall RW, ed. *Patient Flow: Reducing Delay in Healthcare Delivery* (Springer, New York), 281–307.

Green LV (2012) The vital role of operations analysis in improving healthcare delivery. *Manufacturing Service Oper. Management* 14(4):488–494.

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

Hall R (2012) Bed assignment and bed management. Hall R, ed. *Handbook of Healthcare System Scheduling* (Springer, New York), 177–200.

Harper PR, Shahani AK (2002) Modelling for the planning and management of bed capacities in hospitals. *J. Oper. Res. Soc.* 53(1):11–18.

Hellmann LM, Kohl SG, Palmer J (1962) Early hospital discharge in obstetrics. *Lancet* 279(7223):227–232.

Huang X-M (1998) Decision making support in reshaping hospital medical services. *Health Care Management Sci.* 1(2):165–173.

Hulshof PJH, Kortbeek N, Boucherie RJ, Hans EW, Bakker PJM (2012) Taxonomic classification of planning decisions in health care: A structured review of the state of the art in OR/MS. *Health Systems* 1(2):129–175.

Hyer NL, Wemmerlöv U, Morris JA (2009) Performance analysis of a focused hospital unit: The case of an integrated trauma center. *J. Oper. Management* 27(3):203–219.

Iravani F, Balcıoǧlu B (2008) Approximations for the $M/GI/N+GI$ type call center. *Queueing Systems* 58(2):137–153.

Jordan WC, Graves SC (1995) Principles on the benefits of manufacturing process flexibility. *Management Sci.* 41(4):577–594.

Kao EPC, Tung GG (1981) Bed allocation in a public health care delivery system. *Management Sci.* 27(5):507–520.

Karlin J (2013) Loss and gain in translation: Financial epidemiology on the south side of Chicago. *Public Culture* 25(371): 523–550.

KC DS, Terwiesch C (2009) Impact of workload on service time and patient safety: An econometric analysis of hospital operations. *Management Sci.* 55(9):1486–1498.

KC DS, Terwiesch C (2011) The effects of focus on performance: Evidence from California hospitals. *Management Sci.* 57(11): 1897–1912.

KC DS, Terwiesch C (2012) An econometric analysis of patient flows in the cardiac intensive care unit. *Manufacturing Service Oper. Management* 14(1):50–65.

Kokangul A (2008) A combination of deterministic and stochastic approaches to optimize bed capacity in a hospital unit. *Comput. Methods and Programs in Biomedicine* 90(1):56–65.

Litvak E, Bisognano M (2011) More patients, less payment: Increasing hospital efficiency in the aftermath of health reform. *Health Affairs* 30(1):76–80.

Ma G, Demeulemeester E (2013) A multilevel integrative approach to hospital case mix and capacity planning. *Comput. Oper. Res.* 40(9):2198–2207.

Muller N (2010) Do general, community hospitals compete by specializing in high volume, high revenue-generating service lines? Ph.D. thesis, University of Virginia, Richmond.

National Center for Health Statistics (2012) Health, United States, 2011: With special feature on socioeconomic status and health. National Center for Health Statistics, Hyattsville, MD. http://www.cdc.gov/nchs/data/hus/hus11.pdf.

Sarhangian V, Balcıoǧlu B (2013) Waiting time analysis of multiclass queues with impatient customers. *Probab. Engrg. Inform. Sci.* 27(3):333–352.

Shi P, Dai J, Ding D, Ang J, Chou M, Xin J, Sim J (2013) Patient flow from emergency department to inpatient wards: Empirical observations from a Singaporean hospital, Working paper, Georgia Institute of Technology, Atlanta.

Stanciu A, Vargas L, May J (2010) A revenue management approach for managing operating room capacity. *Proc. 2010 Winter Simulation Conf. (WSC)* (IEEE, New York), 2444–2454.

Strauss MJ, LoGerfo JP, Yeltatzie JA, Temkin N, Hudson LD (1986) Rationing of intensive care unit services: An everyday occurrence. *J. Amer. Medical Assoc.* 255(9):1143–1146.

Talluri KT, van Ryzin GJ (2004) *The Theory and Practice of Revenue Management* (Springer, New York).

Vanberkel PT, Boucherie RJ, Hans EW, Hurink J, Litvak N (2012) Efficiency evaluation for pooling resources in health care. *OR Spectrum* 34(2):371–390.

Van Dijk NM, Van der Sluis E (2008) To pool or not to pool in call centers. *Production Oper. Management* 17(3):296–305.

Vassilacopoulos G (1985) A simulation model for bed allocation to hospital inpatient departments. *Simulation* 45(5):233–241.

Wallace RB, Whitt W (2005) A staffing algorithm for call centers with skill-based routing. *Manufacturing Service Oper. Management* 7(4):276–294.

Ward AR (2012) Asymptotic analysis of queueing systems with reneging: A survey of results for FIFO, single class models. *Surveys Oper. Res. Management Sci.* 17(1):1–14.

Whitt W (1999) Partitioning customers into service groups. *Management Sci.* 45(11):1579–1592.

Whitt W (2006) Fluid models for many-server queues with abandonments. *Oper. Res.* 54(1):37–54.

Zeltyn S (2004) Call centers with impatient customers: Exact analysis and many-server asymptotics of the $M/M/n+G$ queue. Ph.D. thesis, Technion–Israel Institute of Technology, Haifa.

Zeltyn S, Mandelbaum A (2005) Call centers with impatient customers: Many-server asymptotics of the $M/M/n+G$ queue. *Queueing Systems* 51(3–4):361–402.