Health Care Manage Sci (2006) 9: 391–404
DOI 10.1007/s10729-006-0002-4

A network flow approach to optimizing hospital bed capacity decisions

Elif Akcali · Murray J Côté · Chin Lin

Received: July 2005 /Accepted: June 2006

Abstract The delivery of cost-effective and quality hospital-based health care remains an important and ongoing challenge for the American health care industry. Despite numerous advances in medical procedures and technologies, a growing array of outpatient health care options, limits on inpatient reimbursements, and almost two decades of hospital contraction and consolidation, annual inpatient admissions in the United States are currently at levels not seen since the early 1980s. This combination of increased demand and diminished resources makes planning for hospital bed capacity a difficult problem for health care decision makers. We examine this problem by developing a network flow model that incorporates facility performance and budget constraints to determine optimal hospital bed capacity over a finite planning horizon. Under modest assumptions, we demonstrate that for realistic sized capacity planning problems, our network formulation is not computationally intensive, and allows us to obtain optimal bed capacity plans quickly.

Keywords Bed planning · Capacity planning · Hospitals · Mixed-integer models · Network flow models · Planning horizon

1 Introduction

Capacity planning is central to the pursuit of balancing the quality of health care delivered with the cost of providing that care. Such planning involves predicting the quantity and particular attributes of resources required to deliver health care service at specified levels of cost and quality. In general, successful health care capacity planning must address a variety of issues, including the duration of the planning horizon (i.e., operational, tactical, and strategic), the level of care provided (i.e., primary, secondary, and tertiary), the type of care (i.e., inpatient and/or outpatient), the amount, capability, cost, and types of available or desired resources (i.e., doctors, nurses, technicians, medical and clinical support staff, facilities including buildings, rooms, beds, parking spaces, medical diagnostic and monitoring equipment, or any other element that constitutes an "input" to the delivery of health care) as well as the customer service metrics or performance measures expected for the facility (e.g., patient length of stay, likelihood of full capacity where all inpatient beds or examining rooms are occupied, utilization of providers and facilities, and financial performance such as having expenses within or below budget).

While capacity planning has challenged health care decision makers and researchers for decades [1–3],

E. Akcali · C. Lin
Department of Industrial and Systems Engineering,
University of Florida,
Gainesville, FL 32611-6595, USA

E. Akcali
e-mail: akcali@ise.ufl.edu

C. Lin
e-mail: cindylin@ufl.edu

M. J. Côté (✉)
Division of Health Care Policy and Research
Department of Medicine, University of Colorado at Denver
and Health Sciences Center,
13611 East Colfax Avenue, Suite 100,
Aurora, CO 80045-5701, USA
e-mail: murray.cote@uchsc.edu

there is a renewed sense of urgency to address this problem. In addition to the perennial struggle between the continually increasing costs of highly specialized and scarce inputs (i.e., skilled and flexible staff, advanced clinical and medical technology and equipment, physical space and supplies) and declining government and private reimbursements [4, 5], the demand for inpatient care has been growing substantially. According to the American Hospital Association (AHA), while average length of stay (ALOS) remained unchanged at 5.7 days, all community hospital volume statistics increased from 2002 to 2003: inpatient admissions by 0.9% to 34.8 million, total hospital-based outpatient visits by 1.2% to 563.2 million, emergency department visits by 1.0% to 111.1 million, adjusted average daily census (i.e., average number of inpatients and outpatients receiving care per day) by 0.9% to 894,000, and average inpatient occupancy rate increased by 1.9% to 66.8% [6]. However, the number of hospitals of all types decreased by 30 to 5,764, there were 32 fewer community hospitals, and 8,000 fewer community hospital beds in 2003 [6].

In this paper, we focus on aggregate hospital bed capacity planning decisions. We develop a model to simultaneously determine the timing and magnitude of changes in bed capacity that minimizes capacity cost (including the cost of changing capacity as well as the cost of operating capacity) while maintaining a desired level of facility performance (e.g., limiting a patient's expected delay before being admitted to a bed and keeping expenses within budget) over a finite planning horizon. We divide the planning horizon into discrete time periods of equal length, and assume that the system achieves steady state in each of these intervals. This allows us to use queuing methodology to analyze system performance, but this typically leads to nonlinear equations in our formulation. As hospital bed capacity must be integer valued, our planning model is a large-scale nonlinear integer optimization model that minimizes total cost while achieving a targeted level of system performance. We show that some practical considerations lead to simplifications in the model, which leads to a network flow formulation for the problem that can be solved in polynomial time.

A variety of problems that arise in the context of transportation, finance, manufacturing, and service systems can be modeled as network flow models [7]. A network is a collection of (capacitated or uncapacitated) nodes and (directed/undirected and capacitated/uncapacitated) arcs, where the arcs link one node to another and carry flow from one node to another. Well-known network flow models are the shortest path, maximum flow, and minimum total cost flow formulations, for which efficient solution algorithms exist [7]. In our work, we show that the capacity planning model we consider can be transformed into a shortest path model, where the objective is to find the path from the source node to the sink node with the shortest length (i.e., the minimum cost bed capacity plan from the current period to the final period of a given planning horizon).

The remainder of this paper is organized as follows. Section 2 provides a brief overview of the history and current research in hospital bed planning. In Section 3, we describe the system and give three models for planning hospital bed capacity. In Section 4, using data from a medium-sized medical center, we provide a computational study to illustrate how the model formulations can be used and how changes in problem parameters can affect our ability to obtain an optimal solution. Section 5 offers several practical extensions of our model. Last, we give concluding remarks and discuss future research directions in Section 6.

2 Hospital bed planning

During the 1990s, many hospitals in the United States reported having too many beds and were exploring strategies to reduce space [8–13]. Less than a decade later, in part due to renewed growth in demand for inpatient services [6, 14], most hospitals are currently facing considerable space and resource restrictions forcing them to contemplate expensive renovations and/or new construction projects to increase bed capacity [14–16]. However, whether hospitals, in fact, need the additional capacity appears to be unresolved [8, 11]. On one hand, increased inpatient admissions coupled with fewer hospitals and fewer hospital beds would support the argument in favor of capacity increases [6, 14]. Conversely, level or decreasing average length of stays and a corresponding decrease in the average inpatient occupancy rate may imply that existing capacity is sufficient [8]. Regardless, determining the optimal number and organization of hospital beds continues to be a challenge.

The ability to anticipate bed demand and match it with the appropriate bed supply is critical to effective bed planning. Health care decision makers know that both will be influenced by a number of factors. Factors internal to the decision makers include containing the costs associated with operating, contracting, and expanding current bed capacity, reducing bed assignment waiting, maintaining quality of care when patients are placed in inappropriate units (e.g., an intensive care patient may have to be placed in a cardiac unit), eliminating emergency department bottlenecks (i.e., keeping patients in the emergency department after initial treatment due to unavailability of beds in the appropriate care unit), and reducing the probability of diverting patients to other hospitals due to lack of bed capacity [11, 12]. Externally, factors facing decision-makers include atypical changes in community health (e.g., severe flu strains), annual holidays (e.g., Thanksgiving), and the availability, size, and composition of appropriate medical personnel.

Historically, starting with the Hill–Burton Act of 1946, bed capacity planning has tended to be based on target occupancy levels (TOLs) that are assumed to reflect capacity levels that achieve an appropriate balance of cost, patient delays, and resource utilization. TOLs are derived using analytic models of typical hospitals in different categories and are based on acceptable patient delays for different services. However, Green and Nguyen [12], used queuing models to investigate the relationship between occupancy levels and delay, and concluded that using TOLs as the primary determinant of bed capacity is inadequate and may lead to excessive delays for beds. In particular, a TOL does not necessarily correspond to a desired service level, and there is a need to quantify the desired service level and measure its cost implications accurately.

Ryan [17] provides a capacity expansion model with exponential demand and continuous time intervals and continuous facility sizes. In the context of health care planning, however, it is more realistic to model capacity expansion as the product of limited, discrete choices as routine planning sessions (e.g., bimonthly or quarterly) where capacity increases or decreases occur in some fixed bed amount such as a 20-bed unit. Bretthauer and Côté [18] model a general health care delivery system as a network of queuing stations and incorporate the queuing network into an optimization framework to determine the optimal capacity levels subject to a specified level of system performance (e.g., average total time spent at the facility). They use an algorithm combining branch-and-bound with outer approximation cutting plane method to solve the nonlinear optimization problem with discrete variables, but a disadvantage of this algorithm is that in the worst case the algorithm could require complete enumeration of all integer solutions, leading to very large solution times.

3 Problem formulation

In the bed capacity planning problem, we start with a planning horizon of length T indexed by t=1, 2, ..., T. Let λ_t denote the aggregate patient arrival rate in period t, 1/μ be the ALOS per patient, and the service rate per bed per day is given by μ or 1/ALOS and the service rate per bed over period t is μ_t. In practice, there are alternative patient streams (including admissions from the emergency department, admissions from referrals, and elective admissions) for each of which the typical length of stay may be different. As the objective of our work is to provide an aggregate planning tool for bed capacity management, we assume that the arrival rates for different patient streams can be combined and a representative value for the average length of stay per patient (regardless type of services required by the patient) can be determined. Note that while ALOS has been relatively stable over time [6], the actual λ_t for a given facility will not be known until the demand presents itself. Therefore, for the purposes of capacity planning, λ_t can be forecasted by a seasonally adjusted trend-line, for example [19]. Let α_t denote the maximum allowable expected delay for a patient before the patient is admitted to a bed in period t. We note that the number of beds in the system in a given period can be limited due to other resource limitations including as the physical size of the facility and/or the amount and type of personnel available. Let c_0 be the initial bed capacity in the hospital. Last, there is a budget limit on the amount of monetary resources that can be allocated to purchasing additional bed capacity denoted by γ_t.

We have three types of decision variables. Let x_t be number of beds in period t. Let x_t^+ be the amount of increase in bed capacity at the beginning of period t, and $x_t^-$the amount of decrease in bed capacity at the beginning of period t. Let $f(x_t, \lambda_t, \mu_t)$ denote the expected patient waiting cost as a function of number of beds x_t, patient arrival rate λ_t, and average service rate μ_t in period t. Similarly, let $g(x_{t-1}, x_t)$ denote the cost of changing bed capacity from x_{t-1} to x_t (i.e., the cost of increasing or decreasing the existing bed capacity) in period t. Let $h(x_t)$ denote the cost of operating x_t beds in period t. Finally, the expected delay for a patient in period t is a function of number of beds x_t, patient arrival rate λ_t, and service rate per bed μ$_t$, denoted by $w(x_t, \lambda_t, \mu_t)$. We can formulate the bed capacity planning (BCP) problem as a nonlinear integer programming formulation as follows:

$$\min \sum_{t=1}^{T} f(x_t, \lambda_t, \mu_t) + \sum_{t=1}^{T} g(x_{t-1}, x_t) + \sum_{t=1}^{T} h(x_t) \qquad (1)$$

subject to

$$w(x_t, \lambda_t, \mu_t) \le \alpha_1 \qquad \forall_t \qquad (2)$$

$$x_0 = c_0 \qquad (3)$$

$$x_{t-1} + x_t^+ - x_t^- = x_t \qquad \forall_t \qquad (4)$$

$$g(x_{t-1}, x_t) \le \gamma_t \qquad \forall_t \qquad (5)$$

$$x_t, x_t^+, x_t^- \text{ are discrete variables} \qquad \forall_t \qquad (6)$$

The objective function (1) minimizes the total cost of patient waiting, changing the bed capacity, and operating the existing bed capacity. Constraint (2) imposes a maximum allowable limit on the expected patient waiting. For example, in order to quantify the expected delay for a patient to be admitted to a bed, we assume that the hospital can be represented as a *GI*/*G*/*s* queueing system and use the expected waiting time approximation provided by Bitran and Tirupati [20, 21] to calculate a patient's expected wait for a hospital bed. Constraint (3) sets the initial bed capacity while constraint (4) is a flow balance equation stating that the number of beds available in a period is equal to the number of beds available in the previous period plus the increase in bed capacity minus the decrease in bed capacity. Constraint (5) is the budget constraint that limits the amount of the funds allocated to changing capacity. Last, constraint (6) ensures that the number of beds available and changes in bed capacity are integer valued.

3.1 Restricted bed capacity planning problem

It should be readily apparent that the number of integer variables associated with the BCP problem could be quite large as there is no restriction on how many beds can be added or removed from service. For example, community hospitals may have 500 or more beds [6]. In practice, bed capacity is increased or decreased in batches, and is typically changed in integer multiples of a base value, say, in multiples of 10 or 25 corresponding to the size of a unit. As a result, there are only a limited number of choices for changing capacity in each period. Therefore, constraints that capture the change in capacity can be replaced by a set of discrete alternative constraints, requiring that only one alternative is chosen in the solution for each period. Then, the original non-linear integer programming problem becomes a nonlinear binary (i.e., zero–one) integer programming problem, which we refer to as the restricted bed capacity planning (RBCP) problem.

In the RBCP problem, we are given a base value of *B* in multiples of which the bed capacity can be increased or decreased and we let *n* be the number of possible distinct levels of capacity increase or decrease. That is, given bed capacity *c* in period *t*, the bed capacity in period *t*+1 can be one of $(c-nB)^+$, $(c-(n-1)B)^+, \ldots, (c-B)^+, c, c+B, \ldots, c+(n-1)B$, $c+nB$, where $(x)^+ = \max\{0, x\}$. We assume that all acquired new additional capacity is available and becomes effective capacity in the same period. Let $z_{it}^+ = 1$ if the available bed capacity is increased by *iB* at the beginning of period *t* for *i*=1, 2, ..., *n*; and 0 otherwise. Similarly, let $z_{it}^- = 1$ if the bed capacity is decreased by *iB* at the beginning of period *t* for *i*=1, 2, ..., *n*; and 0 otherwise. We can now formulate the RBCP problem as a nonlinear zero–one integer programming problem as follows:

$$\text{Min}\sum_{t=1}^{T} f(x_t, \lambda_t, \mu_t) + \sum_{t=1}^{T} g(x_{t-1}, x_t) + \sum_{t=1}^{T} h(x_t) \qquad (7)$$

subject to

$$w(x_t, \lambda_t, \mu_t) \le \alpha_t \qquad \forall_t \qquad (8)$$

$$x_0 = c_0 \qquad (9)$$

$$x_{t-1} + \sum_{i=1}^{n} iBz_{it}^+ - \sum_{i=1}^{n} iBz_{it}^- = x_t \qquad \forall_t \qquad (10)$$

$$\sum_{i=1}^{n} z_{it}^+ + \sum_{i=1}^{n} z_{it}^- \le 1 \qquad \forall_t \qquad (11)$$

$$g(x_{t-1}, x_t) \le \gamma_t \qquad \forall_t \qquad (12)$$

$$x_t \ge 0 \qquad \forall_t \qquad (13)$$

$$z_{it}^+, z_{it}^- \in \{0, 1\} \qquad \forall_t \qquad (14)$$

As in the BCP problem, objective function (7) minimizes the total cost of patient delay, changing the bed capacity, and operating the existing bed capacity, constraint (8) imposes a maximum allowable limit on the expected patient delay, constraint (9) sets the initial bed capacity, and constraint (10) is a flow balance equation. Constraint (11) ensures that only one choice for changing the capacity is allowed in each period. Constraint (12) imposes the budget constraint on the amount of money allocated to changing bed capacity. Constraints (13) and (14) ensure the non-negativity of the bed capacity level and capacity level selection decision variables, respectively.

An attractive feature of the RBCP problem is that a network representation can be developed. Consider a T-partite graph with T layers each representing a time period t=1, 2, ..., T in the planning horizon. Let (t,c) denote the system when there are c beds in period t. Let $C(c)$ be the set of reachable capacity levels in the next period if the capacity in the current period is c, and we have $C(c) = \{(c - nB)^+, (c - (n-1)B)^+, \ldots, (c - B)^+, c, c + B, \ldots, c + (n-1)B, c + nB\}$. Let S_t be the set of all capacity levels reachable in period t from all capacity levels in period $t-1$. Let d_s be a superficial source node connected only to node (0,x_0) with zero arc length. Node (0,x_0) represents the beginning state where there are x_0 beds in the hospital at time t=0. Let (0,x_0) be connected to all nodes $(1,x')$ for $x' \in C(x_0)$. If $w(x', \lambda_1, \mu_1) \leq \alpha_1$ (i.e., the patient waiting time constraint is not violated) and $g(x_0, x') \leq \gamma_1$ (i.e., the budget constraint is not violated), then the length of these arcs are given by $g(x_0, x') + f(x', \lambda_1, \mu_1) + h(x')$ (i.e., the total cost of changing the bed capacity from x_0 to x', expected patient waiting cost with x' beds in the system, and cost of operating x' beds). However, if either constraint is violated, then the length of the corresponding arc is set to M, where M is a very large number. Similarly, let each node (t,x) for $x \in S_t$ and t=1, 2, ..., $T-1$ be connected to $(t+1, x')$ for $x' \in C(x)$ with length $g(x, x') + f\left(x', \lambda_{t+1}, \mu_{t+1}\right) + h(x')$ if $w(x', \lambda_{t+1}, \mu_{t+1}) \leq \alpha_{t+1}$ and $g(x_t, x') \leq \gamma_{t+1}$, and M otherwise. Last, let each node (T,x) for $x \in S_t$ be connected to a superficial sink node d_t with an arc of zero length.

Figure 1 provides an example of the network representation for the RBCP problem where c_0=300, B=25, n=1, and T=4. In this figure, a path from the superficial source node to the superficial sink node represents a plan for the bed capacity over the planning horizon. The shortest path without containing any arc with cost M yields the capacity plan with total minimum cost that obeys the patient waiting time and budget constraints over the planning horizon. If no such path can be found (i.e., the shortest path contains at least one arc with cost M), then the problem is infeasible and no capacity plan that obeys the waiting time and budget constraints over the planning horizon can be found.

Recalling that there are n distinct levels to increase or decrease capacity, the general network flow representation for the RBCP problem has $2nt + 1$ nodes in layer t for t=1, 2, ..., T. Therefore, there are a total of $nT(T+1) + T + 2$ nodes (including the superficial sink and source nodes) and the shortest path for the RBCP problem can be found in $O\left(n^2 T^4\right)$ time using Dijsktra's algorithm [7].

3.2 Restricted bed capacity planning problem with shuttering

In the RBCP problem, we assume that the cost of increasing or decreasing bed capacity is uniform. In practice, however, decreasing bed capacity can be achieved by shuttering existing bed capacity. That is, a hospital unit is closed and the personnel may be reassigned to other units in the hospital or laid off, thereby, reducing the effective bed capacity. On the other hand, increases in bed capacity can be accomplished two ways. If the existing capacity is larger than the effective capacity implying that shuttered capacity is available, then restoring a shuttered unit into operation by reallocating personnel to this unit can increase bed capacity. However, if the existing capacity is equal to the effective capacity implying that no shuttered capacity is available, then bed capacity can only be increased through a capital investment to open a new unit and purchase new beds. We can incorporate this practical concern into our formulation easily by changing the definition of the objective function by keeping track of the effective and existing bed capacity in the hospital. We now distinguish between two types of capacity changes, where $g(x_0|x', \overline{x}|\overline{x}')$ is the cost of changing effective capacity from x_0 to x' via shuttering and the existing bed capacity from $\overline{x}$ to $\overline{x}'$ via acquiring additional capacity where $\overline{x}' \geq \max\{x', \overline{x}\}$. As before, we assume that all acquired new additional capacity becomes effective capacity in the same period. The formulation is still a nonlinear zero–one integer programming problem, which we refer to as the restricted bed capacity planning with shuttering (RBCPwS) problem.

As with the RBCP problem, a network representation can be developed for the RBCPwS problem. Consider a T-partite graph with T layers each representing a time period t=1, 2, ..., T in the planning horizon. Let $(t, c|\overline{c})$ denote an effective capacity of c and an existing capacity of $\overline{c}$ in time period t, and $c \leq \overline{c}$. Let $C_1(c|\overline{c})$ denote the set of reachable capacity levels via shuttering only, $C_2(c|\overline{c})$ the set of reachable capacity levels by acquiring new additional capacity and $C(c|\overline{c}) = C_1(c|\overline{c}) \cup C_2(c|\overline{c})$ the set of all reachable capacity levels in the next period if the effective and existing bed capacity in the current period are c and $\overline{c}$, respectively. If we have $c + nB \leq \overline{c}$, then we

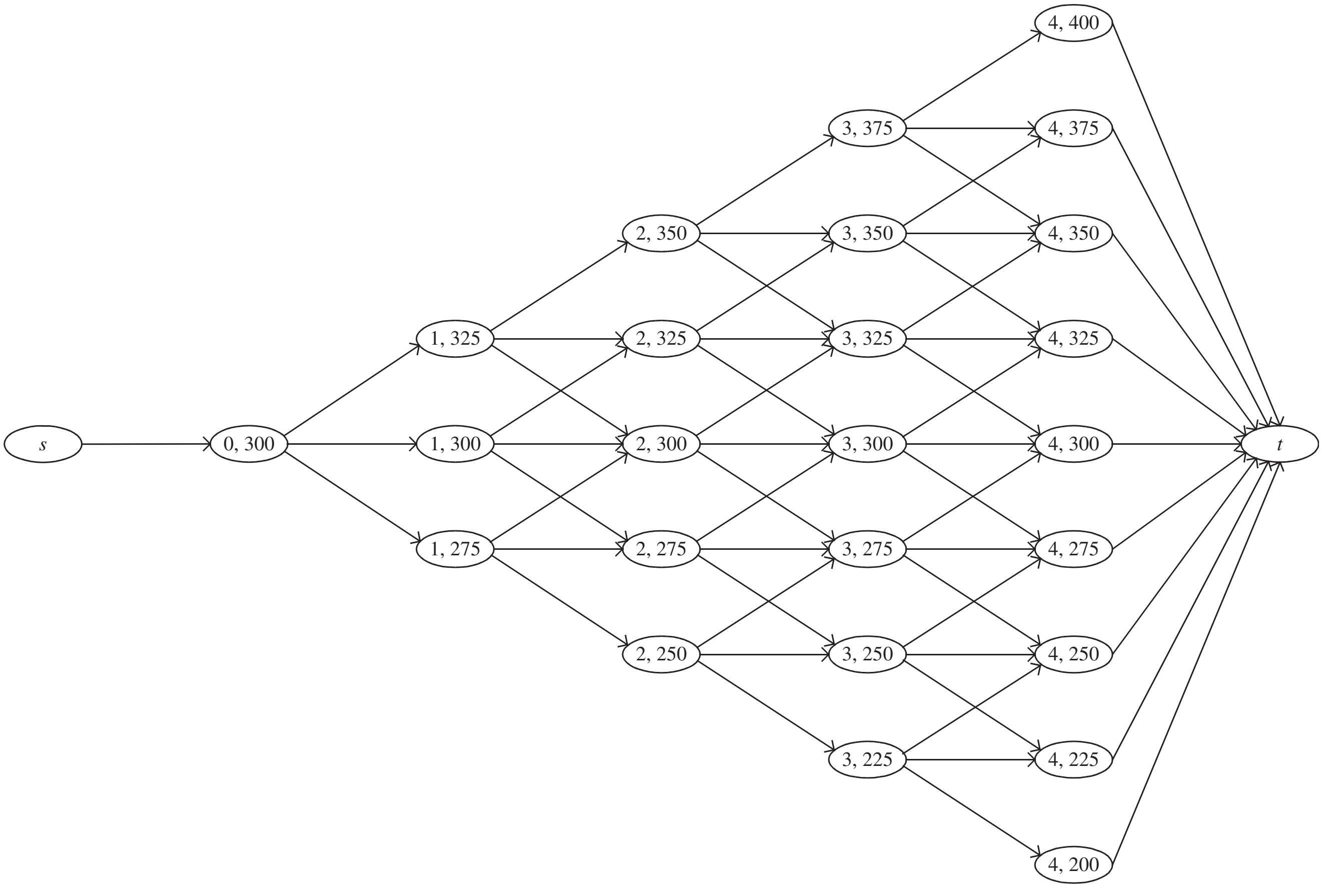


Fig. 1 Network flow representation for RBCP with c_0=300, B=25, n=1, and T=4

have $C_1(c|\overline{c}) = \{((c-nB)^+|\overline{c}),\ldots,((c-B)^+|\overline{c}),(c|\overline{c}),(c+B|\overline{c}),\ldots,(c+nB|\overline{c})\}$ and $C_2(c|\overline{c}) = \{\emptyset\}$. Also if we have $c \le \overline{c} \le c+nB$, we have $C_1(c|\overline{c}) = \{((c-nB)^+|\overline{c}),\ldots,((c-B)^+|\overline{c}),(c|\overline{c}),(c+B|\overline{c}),\ldots,(\overline{c}|\overline{c})\}$ and $C_2(c|\overline{c}) = \{(\overline{c}+B|\overline{c}+B),\ldots,(c+nB|c+nB)\}$. Again, let d_s be a superficial source node connected only to node $(0,x_0|\overline{x})$ with zero arc length. Node $(0,x_0|\overline{x})$ represents the beginning state where there are $\overline{x}$ beds in the system and x_0 in operating condition at t=0. Let $(0,x_0|\overline{x})$ be connected to all nodes $(1,x'|\overline{x})$ in $C(x_0|\overline{x})$. Provided both the patient waiting time constraint and the budget constraint are not violated, then the length of these arcs are given by $g(x_0|\overline{x},x'|\overline{x}')+f(x',\lambda_1,\mu_1)+h(x')$ (i.e., the total cost of changing the effective bed capacity from x_0 to x' via shuttering and the existing bed capacity from $\overline{x}$ to $\overline{x}'$ via new bed acquisition, expected patient waiting cost with x' beds in the system, and cost of operating x' beds). If either of these constraints is violated, then the length of the corresponding arc is set to M. Similarly, each node $(t,x|\overline{x})$ for $(x|\overline{x}) \in S_t$ and t=1, 2, ..., $T-1$ be connected to all nodes $(t+1,x'|\overline{x}')$ in $C(x|\overline{x})$ with length $g(x|\overline{x},x'|\overline{x}')+f(x',\lambda_{t+1},\mu_{t+1})+h(x')$ if $w(x',\lambda_{t+1},\mu_{t+1}) \le \alpha_{t+1}$ and $g(x|\overline{x}.x'|\overline{x}) \le \gamma_1$, and M otherwise. Finally, let each node $(T,x|\overline{x})$ for $(x|\overline{x}) \in S_t$ be connected to a superficial sink node d_t with length zero.

Figure 2 provides an example of the network representation for the RBCPwS problem where c_0=275, $\overline{c}_0$=300, B=25, n=1, and T=4. For ease of exposition, the thin arcs represent opening, maintaining, or shuttering of existing capacity, whereas the thick arcs represent the acquisition of new capacity. In this network, a path from the superficial source node to the superficial sink node represents a plan for the bed capacity throughout the planning horizon. As before, the shortest path without containing any arc with cost M in the network yields the capacity plan with total minimum cost that obeys the patient waiting time and budget constraints throughout the planning horizon by allowing capacity changes via shuttering and/or acquiring additional capacity. If no such path can be found (i.e., the shortest path contains at least one arc with cost M), the problem is infeasible and no capacity plan that obeys the waiting time and budget constraints over the planning horizon can be found.

Recalling the RBCP problem, we have specified the number of arcs and nodes in the network to determine

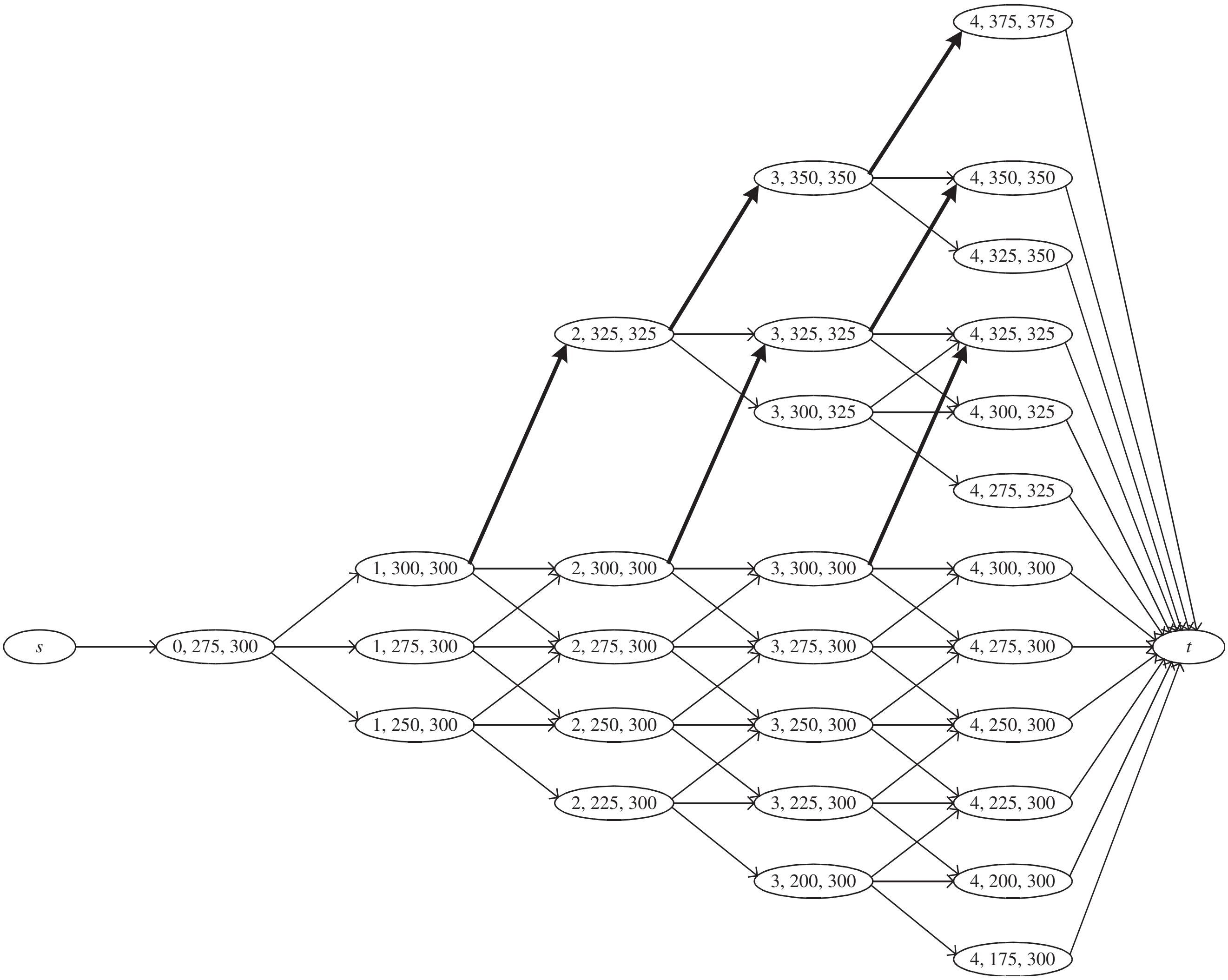


Fig. 2 Network flow representation for RBCPwS with c_0=275, , B=25, n=1, and T=4

the time to obtain the optimal solution. However, for the RBCPwS problem, since existing capacity can be increased further through capital acquisition, the analysis becomes slightly more complicated and dependent on the initial state (i.e., the amount of effective and existing bed capacity). If no additional capacity has to be purchased throughout the planning horizon, then the RBCPwS and RBCP networks are identical and the size of the RBCP network is a lower bound on the size of the RBCPwS network. If additional capacity has to be purchased in a period, at most $(t-1)n^2$ can be added to the network in that period. Hence, if the initial effective capacity is equal to the existing capacity, then there can be at most a total of $nT(T+1)+T+2+T(T+1)(T-1)n^2/6$ nodes (including the superficial sink and source nodes) in the network, and the shortest path can be found again in $O(\mathrm{n}^4T^6)$ time using Dijsktra's algorithm [7].

4 Illustration of the model

In this section, we illustrate the practical applicability and computational behavior of our model through two experiments. In the first experiment, we illustrate how our model can be used to develop bed capacity plans. In the second experiment, we quantify the time (in CPU seconds) needed to obtain optimal solutions. In both experiments, we use the RBCPwS formulation and its associated network model.

4.1 A representative decision-making scenario

To set the stage for the computational experiments that follow, we present a representative decision-making scenario based upon a real-world application of our model to a medium-sized, nongovernment, not-for-profit, general medical and surgical medical center.

Table 1 Parameter settings for the base scenario, S1

Parameter	Value
Length of the planning horizon	T=8 quarters, t=1, 2, ..., 8
Forecasted demand per time period t	$\widehat{\lambda}_t = s_{\text{mod}(t,4)}u(a+bt)$ where s_i is a quarterly seasonal index (i.e., s_1=0.8, s_2=1.0, s_3=1.2, and s_4=1.0), u is a uniformly distributed random number (i.e., u~U[0.8,1.2]), a=6,400, and b=128
Initial existing bed capacity	$c_0 = 350$
Initial effective bed capacity	$\bar{c} = 350$
Number of levels of capacity increase or decrease	n=2
Incremental amount of capacity change	B=10
Cost to operate an effective bed	\$2,000/bed
Cost to shutter an effective bed	\$2,500/bed
Cost to reactivate a shuttered bed	\$2,500/bed
Cost to acquire a new bed (i.e., expand capacity through capital investment)	\$200,000/bed
Coefficient of variation for arrivals	$ca_t = 0.5$
Coefficient of variation for service	$cs_t = 0.5$
Maximum expected delay per patient	$\alpha_t = 1$ h
Cost of waiting	\$300/hour
Service rate	μ_t=15.8 patients per bed

Administration at this facility provided us with information about their facility, capacity planning decision-making processes, and facility-specific data for bed size, bed operating cost, bed acquisition cost, and quarterly patient demand. However, note that at the request of the facility's administration, the data presented here have been modified to protect their identity, but are representative of similar-sized facilities.

This facility would like to determine an optimal bed capacity plan for the next eight quarters, corresponding to its operational, budgetary, and strategic planning periods. Because capacity planning may involve a substantial capital commitment, it is imperative that any capacity expansion plan be carefully developed and justified based upon the facility's current and expected demand. The facility's decision makers would like to minimize the total capacity cost associated with the cost of changing capacity as well as the cost of operating capacity while ensuring that the average time a patient should wait for a bed does not exceed one hour (an internal benchmark for bed assignment). At this facility, both existing and effective bed capacities are 350 beds, capacity change can occur in increments of 10 beds, and there are two levels of capacity increase (i.e., initially, bed capacity can range from 330 to 370 beds, in 10 bed increments). Based on information from the facility's administration, it costs \$2,000/day to operate an effective bed, \$2,500/day to either shutter an effective bed or reactivate a shuttered bed, and \$200,000/bed to expand bed capacity through capital investment. Last, because of seasonal migration (or "snow birds"), demand at the facility can be highly variable throughout the year, and we were provided with data and guidance on values related to patient arrival rates and service times.

4.2 Experiment 1—an application of the model

The intent of this experiment is to illustrate how our network flow model can be used to make bed capacity decisions and generate a T-period capacity plan. Our base scenario was described in Section 4.1, and we refer to it as S1. Table 1 lists the relevant parameter settings for S1, and other experimental scenarios relative to this scenario are given in Table 2.

At the outset, we provide an estimated range of demand for the facility over the planning horizon. Normally, a single seasonally adjusted trendline would be computed to forecast the patient arrival rate based on historic demand data. Instead, to illustrate the extent of variation in demand, Figure 3 displays a set of simulated patient arrival rates over the planning horizon based upon the scenarios given in Table 2. We note that some of the parameter changes directly impact the patient arrival rate, and different patient arrival rates are generated. In S1, S3, S4, S5, and S6, the

Table 2 Scenario descriptions for experiment 1

Scenario	Description	Parameter Change
S0	Level demand	b=0
S1	Base scenario	
S2	Increased rate of demand	b=256
S3	Higher demand variability	ca_t=2.0
S4	Higher service variability	cs_t=2.0
S5	Higher cost of waiting per patient	\$1,200/hour
S6	Smaller maximum expected delay per patient	α_t=0.25 of an hour

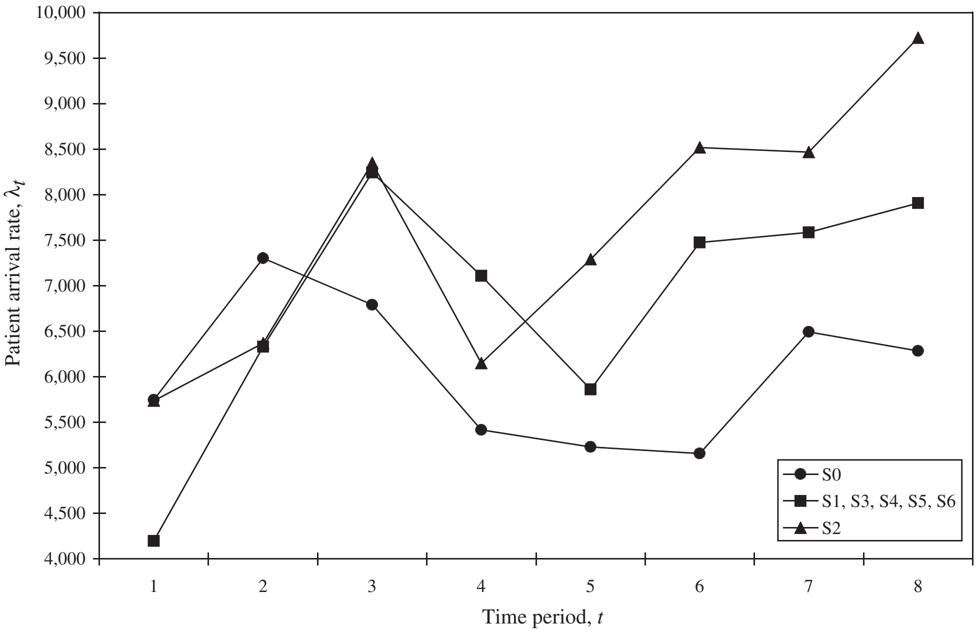


Fig. 3 Patient arrival rates per quarter for experiment 1

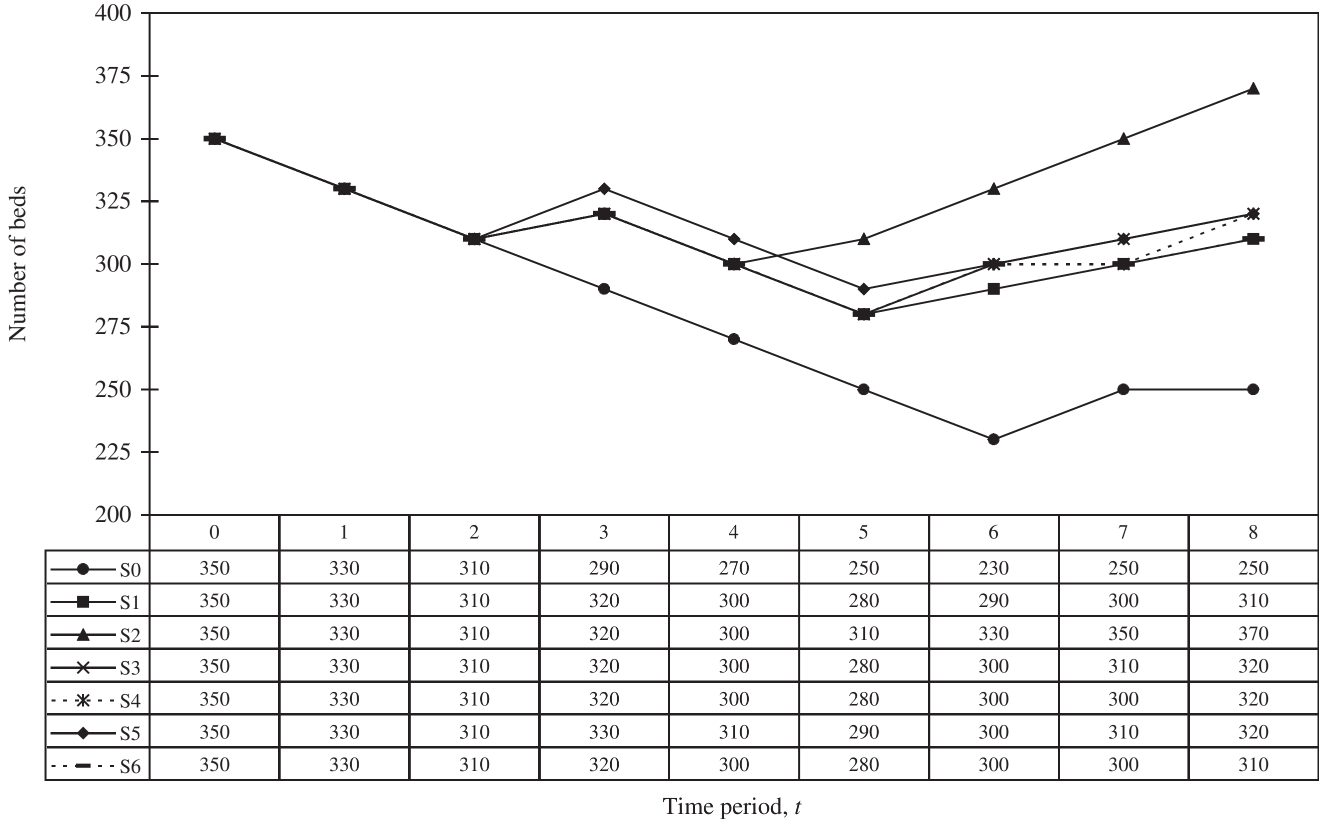
changed parameters do not impact the arrival rate function, so these scenarios have identical arrival rates. (Note that with S3, higher variability in the arrival rate impacts the performance constraint for average waiting time, not the arrival rate function.) For S0 and S2, the patient arrival rate function has no trend and a higher trend compared to S1, respectively. Hence, arrival rates generated for these scenarios are significantly different from each other and S1.

We have implemented our network flow approach using the C++ programming language and solved for the scenarios using a personal computer with 3.0 GHz Pentium IV processor and 512 MB RAM memory. We obtained the optimal solution for each scenario and the results are depicted in Figure 4, where each line represents the optimal capacity plan that corresponds to one of the seven scenarios.

In considering Figure 4, we have the following observations. For S1, we first observe a general reduction in the bed capacity, then a gradual increase near the end of the planning horizon. The initial bed capacity seems to be higher than needed, and as a result, the bed capacity is reduced to reduce total costs over the planning horizon while maintaining the average waiting time constraint. Of course, when the demand increases due to the underlying trend, the bed capacity is increased. When demand is level as in S0, a lower envelope is formed relative to the base case (i.e., the bed capacity for S0 is less than or equal to the base case). Similarly, with an increased rate of demand as in S2, an upper envelope is formed relative to the base case. With increased variation as in S3 and S4, the optimal capacity plans are similar to S1's capacity plan but tend to require higher capacity when the arrival rate is increasing. When the arrival rate increases in periods 6, 7, and 8, because the higher arrival variability and higher service variability affect the average waiting time constraint, more capacity is required to keep from violating this performance constraint. Likewise, with a higher cost of waiting per patient as in S5 or a tighter average waiting time performance constraint as in S6, the optimal capacity plans tend to require capacity slightly higher than the base case. Not surprisingly, the net result of this experiment indicates that optimal bed plans are driven substantially by changes in demand. While health care decision makers may not be able to affect overall demand for their services, if they can reduce variability in arrivals [22] or are willing to tolerate a less stringent performance constraint, less capacity will be required.

4.3 Experiment 2—assessing the impact of problem parameters

As we have discussed earlier, an upper bound on the size of the network (i.e., number of nodes in the network) representing a problem instance of RBCPwS can be characterized in terms of the number of levels for capacity increase or decrease and the number of time periods in the planning horizon. The ratio of the effective bed capacity to existing bed capacity impacts the size of the



	0	1	2	3	4	5	6	7	8
—●— S0	350	330	310	290	270	250	230	250	250
—■— S1	350	330	310	320	300	280	290	300	310
—▲— S2	350	330	310	320	300	310	330	350	370
—×— S3	350	330	310	320	300	280	300	310	320
- - ✱ - - S4	350	330	310	320	300	280	300	300	320
—◆— S5	350	330	310	330	310	290	300	310	320
- - ▬ - - S6	350	330	310	320	300	280	300	300	310

Fig. 4 Optimal capacity plans for experiment 1

network. The size of the network can also be used to quantify the computing time required to obtain the optimal solution. The time required to build the network and find the optimal solution may change as the number of levels increases, the planning horizon length increases, or the ratio of effective to existing bed capacity changes. In order to illustrate the change in computational time, this experiment has two parts: 1) the impact of effective to existing bed capacity and 2) the impact of changes to the number of levels of bed capacity and the length of the planning horizon.

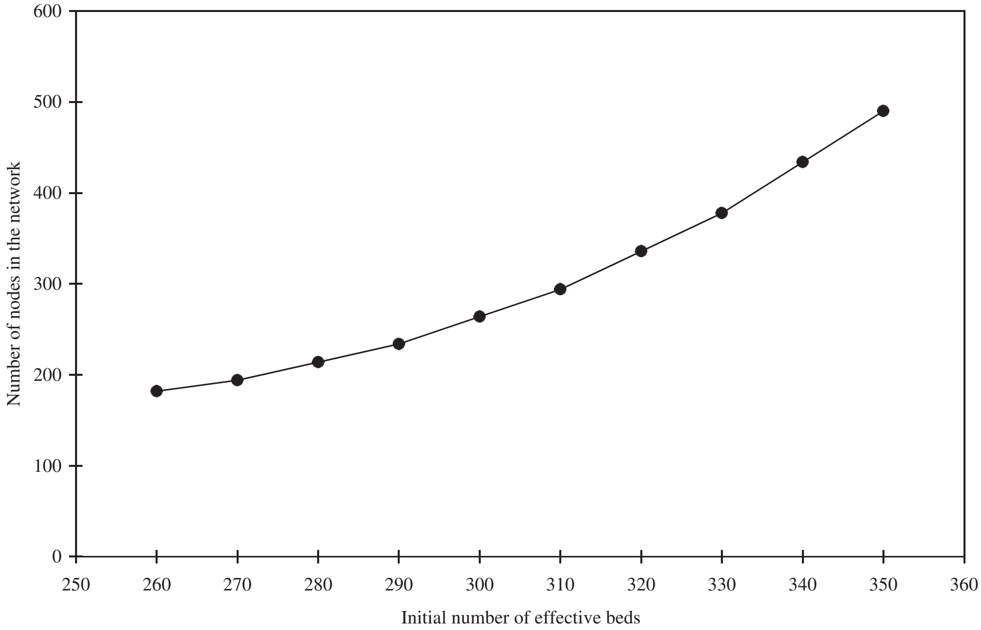
In the first part of this experiment, we fix the number of levels to vary bed capacity and the duration of the planning horizon in addition to some other problem parameters constant and examine the impact of different ratios of existing to effective bed capacity. Using the assumptions for the base case scenario, S1, from the previous experiment, we consider ten different levels of the effective bed capacity in the interval [260, 350]. We generated 30 random test instances for each of these levels and the summary results are provided in Figure 5 and Table 2.

In Figure 5 we depict the number of nodes in the network, and in Table 3 we report the time to build the network and time to obtain the solution for each level of the initial effective bed capacity. The number of nodes increases as the ratio of effective bed capacity to existing bed capacity approaches one, and this behavior is clearly depicted in Figure 5. However, in Table 3, we see that an increase in the size of the network increases the time to build the network only slightly, and its impact on the time to obtain the solution is almost negligible. Therefore, our solution method is robust to changes in the problem size that are induced by the initial effective bed capacity.

For the second part of this experiment, we vary the number of levels to change bed capacity as well as the duration of the planning horizon. We consider four different levels to vary the bed capacity (i.e., n=2, 3, 4, 5) where capacity is increased in increments of B=10, and three different time horizons (i.e., T=8, 12, 16) that correspond to two-, three-, and four-year planning horizons. Therefore, we have 12 settings in total. For each setting, we generated 30 random instances and for each of the instances, we also generated the effective bed capacity as a fraction of the existing bed capacity. The summary results for this experiment are provided in Table 4.

From Table 4, as the number of levels of capacity change and the length of the planning horizon increases,

Fig. 5 Number of nodes in the network as a function of initial effective bed capacity



the number of nodes in the network increases. The increase in the number of nodes impacts the total time required to obtain the optimal solution. However, a closer examination of the results reveals the increase in the number of nodes in the network has a direct impact on the time required to build the network, and has almost no impact on the time to obtain the solution. Only in the setting with the largest test instances (i.e., n=5 and T=16) do we observe an increase in the time to obtain the optimal solution. Even in that case, the maximum solution time is still less than a few seconds. Therefore, our solution method is robust to changes in the problem size induced by the number of levels of capacity change and the duration of the planning horizon.

5 Extensions

In this section, we discuss several extensions to our model. These extensions may arise out of practical considerations associated with how our model addresses facility performance.

5.1 Controlling the magnitude of violation of the performance constraints

In our model, we treat the performance constraints as a hard constraint. That is, if a particular capacity level violates the performance constraint, then a solution with that particular capacity level is not feasible, and is dropped from further consideration. However, the

Table 3 Summary statistics for the RBCPwS problem's solution time (in CPU seconds) as a function of initial effective capacity

Initial Level of Effective Bed Capacity	Time (in CPU seconds) to						Total Time (in CPU seconds)		
	Build the Network			Obtain the Solution					
	Min.	Avg.	Max.	Min.	Avg.	Max.	Min.	Avg.	Max.
260	0.02	0.03	0.16	0.00	0.00	0.00	0.02	0.03	0.16
270	0.03	0.03	0.05	0.00	0.00	0.00	0.03	0.03	0.05
280	0.03	0.03	0.05	0.00	0.00	0.00	0.03	0.03	0.05
290	0.03	0.03	0.05	0.00	0.00	0.02	0.03	0.04	0.05
300	0.03	0.04	0.05	0.00	0.00	0.00	0.03	0.04	0.05
310	0.03	0.04	0.05	0.00	0.00	0.00	0.03	0.04	0.05
320	0.03	0.04	0.05	0.00	0.00	0.02	0.03	0.04	0.05
330	0.03	0.05	0.06	0.00	0.00	0.00	0.03	0.05	0.06
340	0.05	0.07	0.08	0.00	0.00	0.00	0.05	0.07	0.08
350	0.08	0.09	0.09	0.00	0.00	0.02	0.08	0.09	0.09

Table 4 Summary statistics for the RBCPwS problem's solution as a function of capacity levels and the length of the planning horizon

(n, T)	Number of		Time (in CPU seconds) to		Total time (in CPU seconds)
	Nodes in the Network Avg. (Min., Max.)	Arcs in the Network Avg. (Min., Max.)	Build the Network Avg. (Min., Max.)	Obtain the Solution Avg. (Min., Max.)	Avg. (Min., Max.)
(2,8)	323.7 (194, 490)	1,213.9 (737, 1,865)	0.04 (0.02, 0.14)	0 (0, 0)	0.04 (0.02, 0.14)
(3,8)	707.1 (451, 982)	3,566.3 (2,226, 5,061)	0.13 (0.06, 0.20)	0 (0, 0.02)	0.13 (0.06, 0.20)
(4,8)	1,257.5 (858, 1,642)	7,989.3 (5,273, 10,665)	0.31 (0.17, 0.44)	0 (0, 0.02)	0.31 (0.17, 0.44)
(5,8)	1,980.0 (1,455, 2,470)	15,199.0 (10,834, 19,349)	0.64 (0.41, 0.89)	0 (0, 0.02)	0.64 (0.41, 0.89)
(2,12)	997.7 (662, 1,338)	4,067.3 (2,661, 5,521)	0.19 (0.09, 0.34)	0 (0, 0.02)	0.19 (0.09, 0.34)
(3,12)	2,296.2 (1,697, 2,858)	12,850.4 (9,328, 16,195)	0.63 (0.38, 0.86)	0 (0, 0.02)	0.63 (0.39, 0.86)
(4,12)	4,165.3 (3,278, 4,950)	29,679.0 (22,957, 35,669)	1.72 (1.16, 2.25)	0 (0, 0.02)	1.72 (1.16, 2.25)
(5,12)	6,607.5 (5,469, 7,614)	57,209.8 (46,688, 66,583)	4.17 (3.02, 5.31)	0 (0, 0.02)	4.17 (3.02, 5.31)
(2,16)	2,468.2 (1,574, 3,282)	10,586.7 (6,637, 14,225)	0.62 (0.31, 0.99)	0 (0, 0.02)	0.62 (0.31, 0.99)
(3,16)	5,659.9 (4,110, 6,954)	33,550.5 (23,969, 41,609)	2.25 (1.36, 3.36)	0 (0, 0.02)	2.25 (1.36, 3.36)
(4,16)	10,211.9 (8,022, 11,986)	77,310.6 (59,925, 91,473)	6.83 (4.83, 8.84)	0 (0, 0.02)	6.83 (4.83, 8.84)
(5,16)	16,128.0 (13,278, 18,378)	148,628.7 (120,987, 170,537)	57.57 (13.13, 138.22)	0.49 (0, 1.25)	58.06 (13.13, 139.28)

performance constraint can be modeled as a soft constraint where we can deliberately allow the violation of the performance constraint while incurring a penalty cost to be added to the objective function. We can justify this constraint by noting that lags typically exist between capacity levels so there might be periods of time where the facility is operating above its typical utilization and the capacity expansion cannot occur quickly enough to allow the organization to react to the change in demand.

To illustrate how our model can be reformulated with the soft constraint, let v_t be the amount of the violation, s_t be the amount of slack in the performance constraint, and $\pi(v_t)$ be the penalty cost incurred for violating the performance constraint in period t. Then, considering the RBCP problem, objective function (7) would be replaced with:

$$\text{Min}\sum_{t=1}^{T} f(x_t,\lambda_t,\mu_t) + \sum_{t=1}^{T} g(x_{t-1},x_t) + \sum_{t=1}^{T} h(x_t) + \sum_{i=1}^{T}\pi(v_t) \tag{15}$$

Similarly, constraints (8) and (13) would be replaced with:

$$w(x_t,\lambda_t,\mu_t) - v_t + s_t = \alpha_t \qquad \forall t \tag{16}$$

$$\begin{aligned} v_t &= \max\{w(x_t,\lambda_t,\mu_t) - \alpha_t, 0\} \\ s_t &= \max\{\alpha_t - w(x_t,\lambda_t,\mu_t), 0\} \qquad \forall t \end{aligned} \tag{17}$$

$$x_t, v_t, s_t \geq 0 \qquad \forall t \tag{18}$$

It is easy to observe that this modified version of the RBCP problem can still be formulated and solved as a network flow problem. The variables v_t and s_t are calculated in constraint (17) once $w(x_t,\lambda_t,\mu_t)$ is known. The only modification of the network is to include the cost associated with violating the performance constraint.

5.2 Inclusion of multiple performance constraints

When evaluating a hospital, we recognize that the average waiting time to be assigned a bed or having expenses within budget are not the only metrics to assess facility performance. Indeed, it may be necessary to include measures for facility utilization, likelihood of patient diversion, and the like. Regardless, we note that more performance constraints can easily be added to the formulations for the BCP, RBCP, and

RBCPwS problems. An increase in the number of performance constraints does not increase the time to obtain the solution significantly, as there is only a need to take these additional constraints into account in setting up the network and assigning a large arc cost in case any of the constraints are violated. Therefore, our modeling approach is robust and additional constraints can be considered without increasing the complexity of the formulation significantly.

6 Concluding remarks and future research directions

We have presented a network flow approach to optimize bed capacity planning decisions for hospitals. Our model incorporates the reasonable concerns associated with determining hospital bed size, such as a finite planning horizon, an upper bound on the average waiting time before a patient is admitted to a hospital bed, and a budget constraint that limits the amount of money that can be allocated to changing bed capacity. Further, our model accommodates capacity change through shuttering, as well as expansion of bed capacity through new capital investment. Our series of computational experiments illustrated both the ease of implementation of our model and the sensitivity of the computational time required to obtain the optimal solution to several problem parameters. We have also discussed extensions of our model in the form of soft performance constraints and multiple performance constraints.

Our model is based on a generic view of a hospital where we have assumed that the demand (i.e., patient arrivals) and service (i.e., beds) components are homogeneous. From an aggregate planning perspective, such uniformity may be acceptable. However, in order to apply this research to operational decision support for health care delivery, there are additional avenues of research worth pursuing. First, if cost depends on all previous stages, for example, the cost of maintaining the beds depends not only on the number of beds but also the duration of the beds are in the system, then the number of vertices in the network will be exponential with respect to T and the optimal solution to the network will not be solved with a polynomial time algorithm. Consequently, alternative model formulations and solution techniques to determine the optimal bed plan would be necessary. Second, recognizing that hospital beds are not identical, facility capacity could be separated to distinguish the various specialties, with specialty-specific demand rates, lengths of stays, and costs. In determining the average waiting time associated with being assigned a bed, we have used closed-form approximations to calculate this statistic. Therefore, we are implicitly assuming that this general distribution accounts for different types of patients that require different types of hospital-based health care. This may not necessarily be the case, and should be investigated further. Third, our work can be expanded to include multiple types of patients (e.g., electives, admissions coming through the emergency department, and referrals from physicians). Also, in estimating the cost of patient waiting, we assume that this cost is identical regardless of patient type. Clearly, for example, there should be different waiting costs associated placing an emergency department admission in an appropriate unit versus an inappropriate unit. As such, representations of patient waiting cost need to be developed in the presence of congested, heterogeneous resources. Fourth, the current form of our model does not account for the potential time delay that may exist between the decision to expand capacity and actually starting to use the new capacity. Our current model formulations would have to be amended to include the length of delay relative to a capacity expansion (e.g., if a capacity expansion requires k time periods and we need to use the capacity in period t, the decision to expand should occur on or before period $t-k$) and reconciliation of multiple capacity expansions over the planning horizon (e.g., if a capacity expansion decision is made in period t, can the facility make another decision in subsequent periods until $t+k$ when the earlier decision comes into effect). However, because these capacity expansion considerations would destroy the underlying polynomially bounded network structure of the current model, other solution methodologies would have to be developed. Last, as evidenced by the current nurse shortage [23] and the ongoing debate regarding nurse-to-patient ratios [24], the ability to use physical capacity hinges upon the availability of suitable medical personnel. A natural extension of our model would be to incorporate workforce planning to simultaneously determine the quantity and composition of the health care resources to construct a comprehensive capacity plan.

Acknowledgments Funding for this research was provided by the Division of Design, Manufacture, and Industrial Innovation of the National Science Foundation (Award #0323664).

References

1. Pierskalla WP, Brailer D (1994) Applications of operations research in health care delivery. In: Pollock S, Barnett A, Rothkopf M (eds) Beyond the profit motive: public sector

applications and methodology. Handbooks in OR&MS, vol 6. North-Holland, New York

2. Pierskalla WP, Wilson D (1989) Review of operations research improvements in patient care delivery systems. Working paper, University of Pennsylvania, Philadelphia
3. Smith-Daniels VL, Schweikhart SB, Smith-Daniels DE (1988) Capacity management in health care services: review and future research directions. Decis Sci 19:899–918
4. Pierskalla WP (2001) Health care delivery. Presented at the National Science Foundation Workshop on Engineering the Service Sector, Atlanta
5. Sainfort F (2001) Where is OR/MS in the present crises in health care delivery. Presented at the Institute for Operations Research and the Management Sciences 2001 Annual Meeting, Miami
6. American Hospital Association (2005) Hospital Statistics™ 2005 edn. Health Forum, Chicago
7. Ahuja RK, Magnanti TL, Orlin JB (1993) Network flows: theory, algorithms, and applications. Prentice Hall, Englewood Cliffs
8. Bazzoli GJ, Brewster LR, Liu G, Kuo S (2003) Does U.S. hospital capacity need to be expanded? Health Aff 22: 40–54
9. Berwick DM (1996) We can cut costs and improve care at the same time. Med Econ 180:185–187
10. Cerne F, Montague J (1994) Capacity crisis. Hosp Health Netw 68:30–40
11. Green LV (2002/2003) How many hospital beds? Inquiry 39:400–412
12. Green LV, Nguyen V (2001) Strategies for cutting hospital beds: the impact on patient service. Health Serv Res 36: 421–442
13. Hayward C (1998) What are we going to do with all our excess capacity? Health Care Strateg Manage 16:20–23
14. Coile Jr RC (2002) Futurescan 2002: a forecast of healthcare trends 2002–2006. Health Administration, Chicago
15. Bellandi D (1999) Running at capacity. Mod Healthc 110: 112–113
16. Johnson AM (1997) Capacity planning for the future. J Health Care Finance 24:72–75
17. Ryan SM (2004) Capacity expansion for random exponential demand growth with lead times. Manage Sci 50:740–748
18. Bretthauer KM, Côté MJ (1998) A model for planning resource requirements in health care organizations. Decis Sci 29:243–270
19. Côté MJ, Tucker SL (2001) Four methodologies to improve healthcare demand forecasting. Healthc Financ Manage 55: 54–58
20. Bitran GR, Tirupati D (1989) Tradeoff curves, targeting and balancing in manufacturing queueing networks. Oper Res 37:547–564
21. Bitran GR, Tirupati D (1989) Capacity planning in manufacturing networks with discrete options. Ann Oper Res 17: 119–135
22. Litvak E, Buerhaus PI, Davidoff F et al (2005) Managing unnecessary variability in patient demand to reduce nursing stress and improve patient safety. Joint Comm J Quality Saf 31:330–338
23. Aiken LA, Clarke SP, Sloane DM, et al (2001) Nurses' reports on hospital care in five countries. Health Aff 20: 43–53
24. Aiken LA, Clarke SP, Sloane DM et al (2002) Hospital nurse staffing and patient mortality, nurse burnout, and job dissatisfaction. JAMA 288:1987–1993