Manufacturing & Service Operations Management

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

Multiresource Allocation Scheduling in Dynamic Environments

Woonghee Tim Huh, Nan Liu, Van-Anh Truong,

M&SOM MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 15, No. 2, Spring 2013, pp. 280–291
ISSN 1523-4614 (print) | ISSN 1526-5498 (online)

informs®
http://dx.doi.org/10.1287/msom.1120.0415

Multiresource Allocation Scheduling in Dynamic Environments

Woonghee Tim Huh
Sauder School of Business, University of British Columbia, Vancouver, British Columbia V6T 1Z2, Canada,
tim.huh@sauder.ubc.ca

Nan Liu
Department of Health Policy and Management, Mailman School of Public Health, Columbia University,
New York, New York 10032, nl2320@columbia.edu

Van-Anh Truong
Department of Industrial Engineering and Operations Research, Columbia University,
New York, New York 10027, vatruong@ieor.columbia.edu

Motivated by service capacity-management problems in healthcare contexts, we consider a multiresource allocation problem with two classes of jobs (elective and emergency) in a dynamic and nonstationary environment. Emergency jobs need to be served immediately, whereas elective jobs can wait. Distributional information about demand and resource availability is continually updated, and we allow jobs to renege. We prove that our formulation is convex, and the optimal amount of capacity reserved for emergency jobs in each period decreases with the number of elective jobs waiting for service. However, the optimal policy is difficult to compute exactly. We develop the idea of a *limit policy* starting at a particular time, and use this policy to obtain upper and lower bounds on the decisions of an optimal policy in each period, and also to develop several computationally efficient policies. We show in computational experiments that our best policy performs within 1.8% of an optimal policy on average.

Key words: multiresource allocation; Markov decision process; healthcare operations management
History: Received: December 2, 2011; accepted: September 14, 2012. Published online in *Articles in Advance* January 28, 2013.

1. Introduction

We consider a multiresource allocation problem with two classes of jobs (elective and emergency) in a dynamic and nonstationary environment. Emergency jobs need to be performed immediately, whereas elective jobs can wait. This paper is primarily motivated by service capacity-management problems in healthcare contexts, where a limited amount of capacity must be allocated among distinct patient demand streams. Examples include walk-in and scheduled patients in a primary-care facility, and emergency and nonemergency patients for testing (such as magnetic resonance imaging) or a surgical procedure.

In managing such systems, the manager can choose how many elective jobs (patients) to allocate to each day, and thus how much capacity remaining in the day can be reserved for emergency jobs (patients). We refer to these decisions as *allocation scheduling* decisions. (We use the terms *patients* and *jobs* interchangeably.) The main goal of allocation scheduling is to fulfill demand for elective patients in a timely manner, and to leave sufficient slack capacity to meet emergency demand. In making the above trade-off in allocation scheduling, the decision maker must anticipate the demand for emergency and elective jobs, as well as the pattern of resource availability over time. Allocation scheduling is further complicated by the fact that any job may require *multiple* resources (e.g., surgeons, nurses, and an operating room and equipment), and a lack of any necessary resource could result in cancellation or postponement.

Compounding the complexity is the fact that scheduling decisions are often made in environments where information about demand and resource availability is highly dynamic, nonstationary, and correlated. For example, in surgical scheduling, several factors account for nonstationarity and correlation:

1. *Staffing patterns*. Salaried staff accounts for most of the surgical-suite cost (Dexter et al. 1999), and staffing scheduling is subject to time-of-week and time-of-year fluctuations.
2. *Medical equipment*. The availability of such devices that can reduce surgical time (for example, see Kuttenkuler 2004) affects the consumption rate of other resources such as operating rooms.
3. *Patient scheduling pattern and demand growth*. Surgical demand is nonstationary and subject to periodicity and trend, as evidenced by Moore et al. (2008).
4. *Cyclic treatment*. For certain surgical subspecialities (for example, chemotherapy and colorectal liver

metastases), demand is correlated over time because a request for a procedure typically results in subsequent requests.

In this paper, we consider an allocation scheduling problem in such a dynamic environment, where demand and capacity constraints may be random, nonstationary, and time correlated. Requests for elective patients arrive in each period, and a decision must be made to fulfill a number of these requests in the period and wait-list the rest. This decision must satisfy capacity constraints for the period with respect to *multiple* types of resources. There is a per-patient per-period cost for wait-listing, and wait-listed patients may renege. After the scheduling decision has been made for the period, emergency demand arises. Emergency demand that exceeds available capacity must be satisfied using surge capacity at a cost. The decision maker must determine a scheduling policy to minimize the total discounted cost over a finite horizon.

Although the standard tools of Markov decision processes (MDPs) can be used to derive the structure of the optimal policy, MDPs cannot be used as a computational tool in this setting because the computation explodes in general with the length of the horizon. We analyze the optimal policy and derive efficient approximations as well as upper and lower bounds on the optimal decisions, based on which we propose an efficient scheduling policy.

Our work is closely related to those dealing with the allocation of medical service capacity among distinct demand streams. This topic has attracted growing attention in the operations management literature (Gupta 2007). In general, three types of decision problems have been considered: (1) who to serve next, (2) when to schedule the arriving patient, and (3) how much capacity to reserve for a particular class of patients.

For the first problem (who to serve next), Green et al. (2006) analyzed the problem of scheduling patients for a diagnostic facility shared by outpatients, inpatients, and emergency patients. They assumed only one patient will arrive or will be served in a single period. In the second problem (when to schedule), referred to as advanced scheduling, patients are scheduled into future dates upon their arrival. Patrick et al. (2008) presented a method for dynamically scheduling multipriority patients to a diagnostic facility, and Liu et al. (2010) developed dynamic policies for a primary-care clinic taking into account patients' cancellation and no-show behavior. Advanced scheduling is used in contexts where it is important to fix appointment dates soon after they are requested.

The third problem (how much capacity to reserve) is the subject of our paper. Gerchak et al. (1996) considered the problem of reserving surgical capacity for emergency cases when the same operating rooms are also used for elective cases, and characterized the structure of the optimal scheduling policy. Ayvaz and Huh (2010) extended the work of Gerchak et al. (1996) by considering independent but nonstationary arrivals and capacity realizations in each period. They also considered the possibility of allowing same-day service for elective cases, the option of rejecting elective cases, and multiple classes of elective cases. Both sets of authors use MDP tools for analysis and computation, and their methodology cannot be readily adapted to evolving information about demands and capacities.

Our contributions in this work can be summarized as follows. We formulate the allocation scheduling problem in fully dynamic environments; our model is the first to exploit evolving and possibly correlated information about the distribution of demand and capacity, to the best our knowledge. Our model explicitly captures "resource uncertainty" (Cardoen et al. 2010) involving not just a single resource, but multiple resources. We prove that, similar to the problem with simpler independent and identically distributed (i.i.d.) demand, the optimal amount of capacity reserved for emergency patients in each period decreases with the number of patients waiting for elective patients, but the optimal policy is difficult to compute exactly because of state-space explosion. To circumvent this difficulty, we develop a methodology based on the idea of a *limit policy* starting at a particular time. Limit policies are ancillary policies that we use as a means to define our proposed policies, where limit policies are used to approximate the value function in the Bellman equation. Computational results show that our proposed policies perform well.

The remainder of this paper is organized as follows. In §2, we introduce our MDP formulation for the multiresource allocation scheduling problem. We analyze this dynamic model in §3. In §4, we introduce methods to calculate upper and lower bounds on the optimal decisions and derive the approximate scheduling policies. In §5, we prove that the optimal decision in each period is bounded by those of the approximate polices. In §6, we present the results of our numerical experiments. We provide our concluding remarks in §7.

2. Model

In this section, we provide the mathematical description of the multiresource allocation scheduling problem and introduce some of the notations used throughout this paper.

We consider a finite planning horizon of T periods, numbered $t = 1, \ldots, T$. Demands for elective and emergency patients over the periods are random variables denoted by d_t and e_t, respectively, $t = 1, \ldots, T$. We use $\mathbf{d}_t$ to denote the vector consisting of d_t and e_t. The number of elective surgeries scheduled in period t is q_t, $t = 1, \ldots, T$. Any remaining capacity is used to satisfy emergency surgeries. We give special notation to two important sums. We use Q_s to denote the cumulative number of elective surgeries scheduled by time s, or $\sum_{t=1}^{s} q_t$, and D_s to denote the cumulative number of requests from elective patients by time s, or $\sum_{t=1}^{s} d_t$.

We assume that each patient uses n resources. The available quantities of these resources are specified by a nonnegative vector $\mathbf{u}_t$ in each period t. The number q_t of elective surgeries scheduled requires an amount $\mathbf{A}_{t1} q_t$ of the resources, whereas the number e_t of emergency surgeries that arise requires an amount $\mathbf{A}_{t2} e_t$ of the resources. The vectors $\mathbf{A}_{t1}$ and $\mathbf{A}_{t2}$ are column vectors in $\mathbb{R}^n_+$. We call the $n \times 2$ matrix $\mathbf{A}_t := [\mathbf{A}_{t1} \mid \mathbf{A}_{t2}]$ formed by these vectors the *utilization matrix* for period t. We require that the scheduled number q_t of elective surgeries must not exceed the available capacity at t, i.e., $\mathbf{A}_{t1} q_t \le \mathbf{u}_t$.

The events in each period occur in the following sequence.

(i) At the beginning of each period t, there are $w_{t-1} \ge 0$ elective patient requests on the waiting list. The number of elective surgery requests for the period, namely, d_t, is observed and added to the waiting list. The capacity vector $\mathbf{u}_t$ and the utilization matrix $\mathbf{A}_t$ are then observed.

(ii) The manager decides the number q_t of elective patient requests to fulfill in the period, reserving enough spare capacity for emergency requests that may arrive later in the period. There is a per-unit penalty b for each elective patient request on the waiting list that is not fulfilled in the period. After the penalty has been charged, the waiting list may be reduced by a random fraction $\xi_t \in [0, 1]$ due to patient reneging. Each loss of a patient causes a loss of revenue c. We call the total cost due to waiting and reneging patients in each period t the *time-t waiting cost*.

(iii) After the value of q_t has been determined, the number of emergency patient requests for the period, namely, e_t, is observed and fulfilled with the remaining capacity for the period and additional *surge capacity* as needed. The surge capacity used of resource j is charged at a unit penalty rate of p_j, $j = 1, 2, \ldots, n$, and we let $\mathbf{p}$ be a column vector consisting of p_j's. We call the total penalty cost due to use of surge capacity in period t as the *time-t overtime cost*.

Our model assumes a time-dependent (rather than wait time-dependent) reneging rate. Though it is difficult to deal with wait time-dependent abandonment rates in general, our model can handle the special case where patients renege if their wait time exceeds an exponentially distributed "tolerance" threshold. Indeed, the queuing literature often makes such an assumption for technical tractability (Ward and Glynn 2003).

If elective patients must be notified of their appointment at least L periods in advance, then we can introduce a *scheduling lead time* of L periods. In every period t, a decision is made to assign q_t patients who are among the w_t on the waiting list at t to receive elective surgery in period $t + L$. For simplicity, we assume that $L = 0$ in the rest of the paper. However, all of our results extend naturally when L is a positive integer.

One unique feature of our model is that all random quantities introduced above, such as $\mathbf{d}_t$, $\mathbf{u}_t$, and $\mathbf{A}_t$, are allowed to be correlated with each other and correlated over time. Because of the correlation structure, our model can evolve in a way that is dependent on past history. We also note that our model cannot be reduced to a single-resource case by identifying the bottleneck resource. The reason is that the bottleneck resource is policy dependent. A strength of our model is that it can strategically match capacity with demand.

We assume that, for each period t, we have what we call an *information set* that is denoted by $\mathcal{F}_t$. The information set $\mathcal{F}_t$ contains all of the information that is available just before the allocation decision is made in period t (i.e., at the end of step (i) in period t), including all past demands and capacities. In particular, because d_t, $\mathbf{A}_t$ and $\mathbf{u}_t$ are observed at the beginning of period t, these quantities are known deterministically given $\mathcal{F}_t$, whereas e_t and ξ_t are observed after the allocation decision in period t, and so they do not belong to $\mathcal{F}_t$, but to $\mathcal{F}_{t+1}$. The information set $\mathcal{F}_t$ is unaffected by any decision, and is therefore common to all policies. Note that $\mathcal{F}_t$ determines the distribution of demands, costs, and capacities for all current and future periods $\{t, t+1, \ldots, T\}$. The random variables e_t and ξ_t are distributed to a joint distribution that is conditional upon $\mathcal{F}_t$, but for ease of notation, we may not represent their dependency on $\mathcal{F}_t$ when there is no ambiguity.

There is a discount factor of α. The goal of the problem is to find a feasible scheduling policy (i.e., one that respects the capacity constraints) that minimizes the total expected discounted cost over the planning horizon. We consider only policies that are *nonanticipatory*, i.e., at time s, the information that a feasible policy can use consists only of $\mathcal{F}_s$ and the current waiting list. We use superscripts P and OPT to refer to a given policy P and an optimal policy, respectively. Given a policy P, the state of waiting list in the system evolves following the equation:

$$w_t^P = (w_{t-1}^P + d_t - q_t^P)(1 - \xi_t). \tag{1}$$

Note that in our model, we assume that all variables are continuous variables.

3. Structure of the Optimal Policy

In this section, we first formulate the problem as an MDP, which provides a framework for finding an allocation decision that provides the optimal trade-off between overplanning and underplanning for emergency patients.

3.1. MDP Formulation

The MDP problem that we formulate has the objective of minimizing the total discounted cost over the finite horizon of length T. The decision to make is the number of elective surgeries scheduled in period t, q_t, and it takes place at step (ii) in §2. Let B_t represent the number of elective surgeries waiting to be scheduled at period t including those arriving at period t, i.e.,

$$B_t = w_{t-1} + d_t. \tag{2}$$

The state at period t (i.e., the information based on which the scheduling decision is made), then, is denoted by (B_t, F_t). Notice that F_t contains all the information on past demand and capacities. The decision q_t depends on the state and should belong to the set

$$\mathscr{R}(B_t, F_t) \equiv \{q\colon \mathbf{A}_{t1} q \le \mathbf{u}_t,\ 0 \le q \le B_t\}, \tag{3}$$

which represents a feasible set of capacity allocation in period t. Let $\mathbf{p}^\tau$ represent the transpose of $\mathbf{p}$. The expected cost incurred in period t, given the decision q_t, can be written as

$$\begin{aligned} L(q_t, B_t, F_t) = {}& (b + c\mathbf{E}[\xi_t \mid F_t])(B_t - q_t) \\ & + \mathbf{p}^\tau \mathbf{E}(\mathbf{A}_{t2} e_t + \mathbf{A}_{t1} q_t - \mathbf{u}_t)^+. \end{aligned} \tag{4}$$

Consider a finite planning horizon of T periods, where $\alpha \in [0, 1]$ is the discounting factor. Let $V_t(B_t, F_t)$ denote the optimal waiting and overtime costs incurred from period t to T when the state at the end of step (i) in period t is (B_t, F_t). Notice that in the next period $t+1$, the number of outstanding elective patients is updated by

$$B_{t+1} = (1 - \xi_t)(B_t - q_t) + d_{t+1},$$

which follows from (1) and (2). Thus, the Bellman equation can be formulated as follows:

$$V_t(B_t, F_t) = \min_{q_t \in \mathscr{R}(B_t, F_t)} G_t(q_t, B_t, F_t), \tag{5}$$

where

$$\begin{aligned} & G_t(q_t, B_t, F_t) \\ & \quad = L(q_t, B_t, F_t) + \alpha \mathbf{E}[V_{t+1}(B_{t+1}, F_{t+1}) \mid F_t] \\ & \quad = L(q_t, B_t, F_t) \\ & \qquad + \alpha \mathbf{E}\{V_{t+1}((1 - \xi_t)(B_t - q_t) + d_{t+1}, F_{t+1}) \mid F_t\}, \end{aligned} \tag{6}$$

where the terminal function is given by $V_{T+1}(B_{T+1}, F_{T+1}) = vB_{T+1}$ for some salvage value $v \ge 0$.

The MDP formulation presented in this section is not easy to solve in general because the information state F_t can grow large as the period index t increases and the decision of q_t concerns the availability of multiple resources. We first focus our attention to the single-period cost function in §3.2, and then we study certain structural properties of this MDP in §3.3. We present the proofs in the online appendix (available at http://dx.doi.org/10.1287/msom.1120.0415).

3.2. Properties of the Single-Period Cost Function

For a fixed B_t, the single-period cost function L as defined in (4) is convex in q_t. Therefore, the myopic problem of minimizing L as a function of q_t is not difficult. Following the definition of submodularity of Topkis (1998), we call a function g *submodular* if $g(x_1, y_1) + g(x_2, y_2) \le g(x_1, y_2) + g(x_2, y_1)$ for all $x_1 > x_2$, $y_1 > y_2$. We can show the following results.

Lemma 1. *For fixed F_t, the single-period cost function L in (4) is jointly convex and submodular in B_t and q_t. Furthermore, L is jointly convex and submodular in (B_t, z_t), where $z_t = B_t - q_t$.*

Below we consider a special case when only a single resource constraint exists. In this case, the utilization matrix $\mathbf{A}_t$ becomes a one-by-two matrix, and the resource availability $\mathbf{u}_t$ reduces to a scalar. The single-period cost function can be written as

$$\begin{aligned} L(q_t, B_t, F_t) &= (b + c\mathbf{E}[\xi_t \mid F_t])(B_t - q_t) + p\mathbf{A}_{t2}\mathbf{E}[r_t], \\ & \qquad \text{where } r_t = \left[e_t - \frac{\mathbf{u}_t - \mathbf{A}_{t1} q_t}{\mathbf{A}_{t2}} \right]^+. \end{aligned} \tag{7}$$

Note that r_t represents the number of emergency patients that could not be accommodated in period t by the remaining available capacity after satisfying q_t number of elective surgeries. The myopic problem of finding the optimal q_t for L becomes a variant of the newsvendor problem, where the uncertain demand is given by e_t, and the stocking quantity is $(\mathbf{u}_t - \mathbf{A}_{t1} q_t)/\mathbf{A}_{t2}$, a linear transformation of q_t. Denote the optimal values for q_t and r_t in this myopic problem by q_t^m and r_t^m, respectively.

Lemma 2. *Consider the case of a single resource. Let r_t^{nv} be the* $\max\{0, 1 - (b + c\mathbf{E}[\xi_t \mid F_t])/(p\mathbf{A}_{t1})\}$ *quantile of $e_t \mid F_t$. Then, the value of q_t^m minimizing (7) is given by $(\mathbf{u}_t - \mathbf{A}_{t2} r_t^m)/\mathbf{A}_{t1}$, where r_t^m is the point in the interval* $[\max\{0, (\mathbf{u}_t - \mathbf{A}_{t1} B_t)/\mathbf{A}_{t2}\}, \mathbf{u}_t/\mathbf{A}_{t2}]$ *that is the closest to r_t^{nv}, i.e.,*

$$r_t^m = \begin{cases} \max\{0, (\mathbf{u}_t - \mathbf{A}_{t1} B_t)/\mathbf{A}_{t2}\} & \\ \qquad \textit{if } r_t^{nv} < \max\{0, (\mathbf{u}_t - \mathbf{A}_{t1} B_t)/\mathbf{A}_{t2}\}, & \\ r_t^{nv} \quad \textit{if } \max\{0, (\mathbf{u}_t - \mathbf{A}_{t1} B_t)/\mathbf{A}_{t2}\} \le r_t^{nv} \le \mathbf{u}_t/\mathbf{A}_{t2}, & \\ \mathbf{u}_t/\mathbf{A}_{t2} \quad \textit{if } r_t^{nv} > \mathbf{u}_t/\mathbf{A}_{t2}. & \end{cases}$$

3.3. Analysis of the Dynamic Model

In this section, we identify some structural properties for the optimal policies. Recall the Bellman equations defined in (5) and (6) and that the terminal cost function is $V_{T+1}(B_{T+1}, F_{T+1}) = vB_{T+1}$. The following lemma presents some useful properties of functions $V_t(B_t, F_t)$ and $G_t(q_t, B_t, F_t)$.

Lemma 3. *For fixed F_t, $G_t(\cdot)$ and $V_t(\cdot)$ have the following properties*:

(a) *$G_t(q_t, B_t, F_t)$ is jointly convex and submodular in q_t and B_t,*

(b) *$G_t(q_t, B_t, F_t)$ is increasing in B_t, and*

(c) *$V_t(B_t, F_t)$ is convex and increasing in B_t.*

(d) *Furthermore, G_t is jointly convex and submodular in (B_t, z_t), where $z_t = B_t - q_t$.*

(Note that in this paper, we use the terms "increasing" and "decreasing" to mean "nondecreasing" and "nonincreasing," respectively, unless specified otherwise.)

Let $q_t^{\max}(B_t, F_t)$ represent the largest optimal choice for q_t minimizing $G_t(q_t, B_t, F_t)$, given B_t and F_t. Let $q_t^{\min}(B_t, F_t)$ be the smallest optimal choice for q_t. Because the feasible region for q_t, i.e., $\mathscr{R}(B_t, F_t)$, is increasing in B_t, we have the following theorem as a direct result of Lemma 3.

Theorem 1. *For fixed F_t, both $q_t^{\max}(B_t, F_t)$ and $q_t^{\min}(B_t, F_t)$ are increasing in B_t.*

We can further show that, for fixed F_t, the increment of $q_t^{\min}(B_t, F_t)$ associated with that of B_t is bounded above by the increment of B_t itself; that is, for a one unit increase in B_t, the increment of $q_t^{\min}(B_t, F_t)$ is at most one unit. Similar results also hold for $q_t^{\max}(B_t, F_t)$.

Theorem 2. *For fixed F_t, $q_t^{\min}(B_t + \Delta, F_t) \le q_t^{\min}(B_t, F_t) + \Delta$ and $q_t^{\max}(B_t + \Delta, F_t) \le q_t^{\max}(B_t, F_t) + \Delta$ for any $\Delta > 0$.*

These structural results for the optimal scheduling policies shown above are typically the best ones that can be obtained in such models; see Gerchak et al. (1996) and Ayvaz and Huh (2010).

Next, we consider the impact of resource availability on the optimal policy. Intuitively, if more capacity is made available, then the additional capacity will be distributed between elective cases and emergency cases; that is, the optimal values of q_t and r_t will increase in $\mathbf{u}_t$, but the amount of increase in q_t and the amount of increase in r_t will both be bounded above by some function of how much $\mathbf{u}_t$ increases. Such results have been shown to be true by Ayvaz and Huh (2010) under a setting where the capacities realized in each period are independent of each other. However, in our model, such intuitive monotonicity results do *not* necessarily hold because of correlation between demand and capacity. Consider the following hypothetical case. If a larger capacity realization in this period is strongly correlated with a smaller emergency demand in the next period, then in the current period the manager may want to allocate less capacity for elective demand and reserve more capacity for emergency demand, because she knows that in the next period there is less need to reserve capacity for emergency cases and hence more capacity can be used for elective cases.

However, if we can regulate the dependence structure between demand and capacity in a way such that capacity realization does not influence the demand process, we can still show certain monotonicity results. Under the assumptions of Theorem 3, the information set F_t only needs to contain the demand history, i.e., $F_t = \{d_1, e_1, d_2, e_2, \ldots, d_t\}$, because only the demand process may be correlated over time. Let $q_t^{\max}(\mathbf{u}_t)$ and $q_t^{\min}(\mathbf{u}_t)$ represent the maximum and minimum optimal values, respectively, for q_t given B_t, $\mathbf{A}_t$, $\mathbf{u}_t$, and F_t. We can then show the following results on how $q_t^{\max}(\mathbf{u}_t)$ and $q_t^{\min}(\mathbf{u}_t)$ change in $\mathbf{u}_t$ with all other arguments fixed. Let E_j be an n-by-1 vector where all of its entries are 0, except the jth entry is 1. Let $\mathbf{A}_{t1j}$ represent the jth entry of $\mathbf{A}_{t1}$.

Theorem 3. *Suppose*

(1) *$\{\mathbf{d}_t, t = 1, 2, \ldots, T\}$ is independent of $\{\mathbf{u}_t, t = 1, 2, \ldots, T\}$ and $\{\mathbf{A}_t, t = 1, 2, \ldots, T\}$;*

(2) *$\{\mathbf{u}_t, t = 1, 2, \ldots, T\}$ is a sequence of independent random vectors; and*

(3) *$\{\mathbf{A}_t, t = 1, 2, \ldots, T\}$ is a sequence of independent matrices.*

Then, $q_t^{\max}(\mathbf{u}_t)$ and $q_t^{\min}(\mathbf{u}_t)$ increase with $\mathbf{u}_t$ componentwise for fixed B_t, $\mathbf{A}_t$, and F_t; i.e., $q_t^{\max}(\mathbf{u}_t + \Delta E_j) \ge q_t^{\max}(\mathbf{u}_t)$ and $q_t^{\min}(\mathbf{u}_t + \Delta E_j) \ge q_t^{\min}(\mathbf{u}_t)$ for any scalar $\Delta > 0$ and any fixed $j = 1, 2, \ldots, n$. Furthermore, $q_t^{\max}(\mathbf{u}_t + \Delta E_j) \le q_t^{\max}(\mathbf{u}_t) + \Delta/\mathbf{A}_{t1j}$ and $q_t^{\min}(\mathbf{u}_t + \Delta E_j) \le q_t^{\min}(\mathbf{u}_t) + \Delta/\mathbf{A}_{t1j}$.

Note that condition (1) says that the demand process is independent of the utilization and capacity processes; conditions (2) and (3) imply that the utilization matrix is independent across periods and so is the capacity. Even under these conditions, the demand can still be correlated over time, and the utilization matrix and capacity can also be correlated in any given period.

4. Development of Approximate Scheduling Algorithms

In the previous section, we have derived some structural properties for the optimal scheduling policies. Although these results provide useful insights, they do not address the "curse of dimensionality" in the computation of the optimal policy. The computation is especially problematic because the system state in our models is very large, containing all historical information on demands and capacities. To address this issue, we develop several efficient policies. Our policies are based on the idea of replacing the value function that is commonly used in the computation of optimal allocation quantities with approximations. As we shall show, these approximations capture the long-term impact of a decision in terms of the inevitable and incremental effects on future costs.

4.1. Incremental Cost and Benefit of a Decision

In this section, we will describe a way to account for the long-term impact of a decision, either in terms of the incremental cost that it introduces or in terms of the incremental benefits that it brings, compared to the decisions that have been made before. This new cost accounting scheme is crucial in the development of our approximate scheduling policies. Our approach in this section is to describe the time-t waiting cost for each period t as a sum of contributions from all decisions made in periods $s = 1, \ldots, t$. (Recall that the time-t waiting cost consists of both the waiting cost for those on the waiting list and the penalties associated with reneging in period t.)

Because we consider a capacitated system, the decision in each period impacts the set of possible states that the system can reach in each future period. More specifically, the wait list in period t is gradually determined by the decisions in *each* period $\{1, 2, \ldots, t-1\}$ as follows. Suppose we fix a policy P. For any policy P, we can define two sets of affiliated policies:

- *Lower limit policies.* For any periods $s \in \{1, \ldots, T\}$, we denote by P_s a policy that mimics policy P in periods $\{1, \ldots, s\}$ and then accommodates as many as elective cases as possible in periods $\{s+1, \ldots, T\}$. Thus, for $t \in \{1, \ldots, T\}$,

$$q_t^{P_s} = \begin{cases} q_t^P & \text{if } t \le s, \\ \min\{w_{t-1}^{P_s} + d_t, c_t\} & \text{if } t > s, \end{cases}$$

where c_t is the maximum number of elective patients that can be served in period t, i.e.,

$$c_t = \max\{q \ge 0 \mid \mathbf{A}_{t1} q \le \mathbf{u}_t\}. \qquad (8)$$

We call P_s the *lower limit policy defined at time s*.

- *Upper limit policies.* For any $s \in \{1, \ldots, T\}$, we define the *upper limit policy defined at time s*, denoted by P^s, to be a policy that mimics P in periods $\{1, \ldots, s\}$ and then schedules no more elective patient in periods $\{s+1, \ldots, T\}$.

Intuitively, the *lower* limit policy leads to *shorter* waiting lists, while the *upper* limit policy results in *longer* waiting lists. This intuition is formalized below in Proposition 1. It is helpful to introduce some new notations here. Let $\bar{w}_t^P$ be the size of the waiting list under policy P at the end of period t and right before reneging occurs, i.e.,

$$\bar{w}_t^P = w_{t-1}^P + d_t - q_t^P.$$

(Compare this expression with the definition of w_t^P in (1).) In short, we call $\bar{w}_t^P$ the *prereneging* waiting list at t. This variable will be useful in cost accounting. The following proposition follows directly from the definitions of the lower and upper limit policies, P_s and P^s, and induction.

Proposition 1. *For every policy P and any pair of s and t satisfying $1 \le s \le t \le T$, we have*

$$\bar{w}_t^{P_s} \le \bar{w}_t^P \le \bar{w}_t^{P^s}. \qquad (9)$$

Furthermore, $\bar{w}_t^{P_s}$ is increasing in s, and $\bar{w}_t^{P^s}$ is decreasing in s.

We remark that the inequalities in (9) are tight in the sense that by defining P appropriately, we can achieve either $\bar{w}_t^P = \bar{w}_t^{P_s}$ or $\bar{w}_t^P = \bar{w}_t^{P^s}$. Thus, the interval defined by $\bar{w}_t^{P_s}$ and $\bar{w}_t^{P^s}$ represents the set of values that a policy P could have obtained for $\bar{w}_t^P$. Let

$$\mathcal{R}_s^P(t) = \{\bar{w} \in \mathbb{R}_+ \mid \bar{w}_t^{P_s} \le \bar{w} \le \bar{w}_t^{P^s}\}, \qquad (10)$$

where $\mathbb{R}_+$ represents the set of all nonnegative real numbers, and we call $\mathcal{R}_s^P(t)$ the *feasible region for $\bar{w}_t^P$ as seen at the end of period s*. The following result is a corollary of Proposition 1.

Proposition 2. *For every policy P, $\{\bar{w}_t^P\} = \mathcal{R}_t^P(t) \subset \mathcal{R}_{t-1}^P(t) \subset \cdots \subset R_1^P(t)$.*

We think of each region $\mathcal{R}_s^P(t)$ as containing all possible values for $\bar{w}_t^P$, given that decisions in the interval $[1, s]$ have been finalized. As more decisions are made, and as these sets become more restricted, the achievable range of time-t waiting cost under policy P potentially changes. In particular, we have $\mathcal{R}_t^P(t) = \{\bar{w}_t^P\}$ after period-t decisions have been made. We discuss below how to quantify the cost and benefit brought by these successive restrictions.

For any real number u and v, satisfying $u \le v$, we define the *time-t minimum waiting cost* of the interval $[u, v]$ as the value

$$L([u, v], t) = \min_{\bar{w} \in [u, v]} \alpha^t (b + c\xi_t)\bar{w} = \alpha^t (b + c\xi_t) u;$$

that is, it is the least possible value for the waiting cost at t, given that $\bar{w}$ must be in $[u, v]$. It is easy to see that for each t, the minimum waiting cost is monotone in the set $[u, v]$; i.e., if $[u', v'] \subseteq [u, v]$, then $L([u', v'], t) \ge L([u, v], t)$. This monotonicity property allows us to quantify reductions to the feasible set for $\bar{w}_t^P$ in terms of increases in the resulting minimum waiting costs as follows. Let $L_s^P(t)$ denote the increment in the time-t minimum waiting cost due to additional restrictions imposed by the set $\mathscr{R}_s^P(t)$ on $\mathscr{R}_{s-1}^P(t)$ (i.e., imposed by the decision made in period t). More precisely, at the beginning of period s, $\bar{w}_t^P$ is confined to the set $\mathscr{R}_{s-1}^P(t)$. At the end of period s, the set of possibilities for $\bar{w}_t^P$ is reduced by $\mathscr{R}_s^P(t)$. Accordingly, we define

$$L_s^P(t) = L(\mathscr{R}_s^P(t), t) - L(\mathscr{R}_{s-1}^P(t), t), \tag{11}$$

where we define $\mathscr{R}_0^P(t) = \mathscr{R}_0(t) = \{\bar{w} \in \mathbb{R}_+ \mid 0 \le \bar{w} \le \sum_{s=1}^t d_s\}$. Notice that $L_s^P(t)$ is a random quantity because the sets $\mathscr{R}_s^P(t)$ and $\mathscr{R}_{s-1}^P(t)$ depend on future realizations of demand and capacity, as well as patient reneging. The cost $L_s^P(t)$ is an *incremental cost* because it captures the cost of a restriction induced by an additional decision. By the monotonicity of the minimum waiting cost and Proposition 2, $L_s^P(t)$ is always nonnegative.

Similarly, we define the *time-t maximum waiting cost* of $[u, v]$, where $u \le v$, as the value

$$U([u,v],t) = \max_{\bar{w}\in[u,v]} \alpha^t(b + c\xi_t)\bar{w} = \alpha^t(b + c\xi_t)v.$$

Let $U_s^P(t)$ denote the decrease in the time-t maximum waiting cost due to additional restriction imposed by the set $\mathscr{R}_s^P(t)$ on $\mathscr{R}_{s-1}^P(t)$, i.e.,

$$U_s^P(t) = U(\mathscr{R}_{s-1}^P(t), t) - U(\mathscr{R}_s^P(t), t). \tag{12}$$

We refer to $U_s^P(t)$ as an *incremental benefit* because it captures the benefit induced by an additional decision. As before, $U_s^P(t)$ is a random quantity, and it can be shown that $U_s^P(t)$ is always nonnegative.

Using the incremental cost defined in (11), we are able to show that every contribution to the waiting list in period t can be attributed to some decision made in previous periods $\{1, \ldots, t\}$. Analogously, the length of the time-t waiting list would have been $\sum_{s=1}^t d_s$ if no elective patient had been scheduled in each period up to t, and any reduction of the time-t waiting list (from $\sum_{s=1}^t d_s$) can be attributed to some decision between 1 and t.

Theorem 4. *The waiting cost incurred in period t by policy P, $(b + c\xi_t)\bar{w}_t^P$, satisfies*

$$\alpha^t(b + c\xi_t)\bar{w}_t^P = \sum_{s=1}^t L_s^P(t) \quad \textit{and}$$

$$\alpha^t(b + c\xi_t)\bar{w}_t^P = \alpha^t(b + c\xi_t)\sum_{s=1}^t d_s - \sum_{s=1}^t U_s^P(t),$$

where α is the discount factor per period.

Theorem 4 provides an alternate way of expressing the cost. Define

$$L_s^P = \sum_{t=s}^T L_s^P(t), \tag{13}$$

and call it the *aggregate incremental cost at s*. This captures the total effect, in terms of cost, of the time-s decision q_s^P as it possibly increases the lower bound on the size of the prereneging waiting list for all future periods t. Similarly, define the *aggregate incremental benefit at s* as

$$U_s^P = \sum_{t=s}^T U_s^P(t). \tag{14}$$

Then, the waiting cost during the horizon can be written as follows:

$$\begin{aligned}\sum_{t=1}^T \alpha^t(b + c\xi_t)\bar{w}_t^P &= \sum_{t=1}^T\sum_{s=1}^t L_s^P(t)\\ &= \sum_{s=1}^T\sum_{t=s}^T L_s^P(t) = \sum_{s=1}^T L_s^P,\end{aligned} \tag{15}$$

or equivalently,

$$\begin{aligned}\sum_{t=1}^T \alpha^t(b + c\xi_t)\bar{w}_t^P &= \sum_{t=1}^T \alpha^t(b + c\xi_t)\sum_{s=1}^t d_s - \sum_{t=1}^T\sum_{s=1}^t U_s^P(t)\\ &= \sum_{t=1}^T \alpha^t(b + c\xi_t)\sum_{s=1}^t d_s - \sum_{s=1}^t U_s^P.\end{aligned}$$

Note that so far in this section, we have considered the waiting cost under policy P, and we remind the reader that the other cost is the overtime cost, which we denote by, for each period s,

$$O_s^P = \alpha^s \mathbf{p}^\tau(\mathbf{A}_{s2}e_s + \mathbf{A}_{s1}q_s^P - \mathbf{u}_s)^+. \tag{16}$$

As can be seen from this expression, the overtime cost can be computed easily from realized values such as emergency demand e_s and the number of elective surgeries q_s^P, and there is no need for approximating this cost.

4.2. Two Approximate Policies Based on Incremental Costs and Benefits

In §4.1, we have analyzed the time-t waiting cost in each period t and shown that it can be expressed in terms of either the incremental costs or the incremental benefits of decisions in periods $\{1, \ldots, t\}$. The incremental costs and benefits can be computed for each sample of demands and capacities under a given policy. In this section, we will show that these incremental costs and benefits can be calculated forward in time, as random quantities that depend on future demands, capacities, and reneging behavior. Based on this, we will introduce two policies for our surgical scheduling problem, called the *lower minimization policy* (LM) and the *upper minimization policy* (UM), which are related to the lower limit policy and the upper limit policy, respectively.

We fix the policy to be P and the current period to be s. We also fix w_{s-1}^P, the length of the waiting list carried over from period $s-1$ to s, and d_s, the number of elective patient requests for period s. Let q_s^P be the number of elective patients that are scheduled in period s. Once q_s^P is decided, the set $\mathscr{R}_s^P(t)$ defined in (10), for any $t \geq s$, is not affected by any future decision; thus, the minimum and maximum waiting costs $L(\mathscr{R}_s^P(t), t)$ and $U(\mathscr{R}_s^P(t), t)$, as well as incremental quantities $L_s^P(t)$ and $U_s^P(t)$, are not affected by future decisions either. All of these quantities depend only on exogenously defined random elements (demands, capacities, utilization matrices, and reneging fractions) in the future.

In the following proposition, we establish some key properties (such as monotonicity and convexity) of the incremental cost and benefit as functions of q_s^P. These properties will become useful in ensuring that the scheduling policies that we will propose can be easily computed. These results are shown for a single sample path of information for exogenous random elements, but it should be noted that these properties also hold in the expected sense.

Proposition 3. *Fix the policy P and the size of the waiting list at the beginning of period s, w_{s-1}^P, for some period s. For any period $t \geq s$, the following statements hold for any sample path of random variables between periods s and t, i.e., $\{\mathbf{d}_s, \ldots, \mathbf{d}_t\}$, $\{\mathbf{A}_s, \ldots, \mathbf{A}_t\}$ and $\{\mathbf{u}_s, \ldots, \mathbf{u}_t\}$:*

(a) $L_s^P(t)$ is nonnegative and decreasing in q_s^P, and equals 0 when $q_s^P = \min\{c_s, w_s^P + d_s\}$. Furthermore, $L_s^P(t)$ is convex and continuous as a function of q_s^P.

(b) $U_s^P(t)$ is nonnegative and increasing in q_s^P, and equals 0 when $q_s^P = 0$. In fact, $U_s^P(t)$ is a linear function of q_s^P.

If we interpret the time-t incremental costs $L_s^P(t)$ as the additional cost imposed on period t by the decision in period s, we can compute the aggregate impact of the decision in period s by summing this quantity over all possible values of t—recall how we defined the aggregate incremental cost in (13). Similarly we have defined the aggregate incremental benefit in (14). It is straightforward to verify that the aggregate incremental costs and benefits inherit all the properties of the constituting terms. We have the following corollary to Proposition 3.

Corollary 1. *The statement of Proposition 3 continues to hold when $L_s^P(t)$ is replaced with L_s^P and $U_s^P(t)$ is replaced with U_s^P.*

As in §4.1, the majority of this section has been devoted to the waiting cost, and we remind the reader that the overtime cost is given in (16). Now, we are ready to introduce two scheduling policies for our surgical scheduling problem.

- *Lower minimization policy.* Under this policy, we choose the number of elective surgeries, q_s^{LM}, such that $E[L_s^P + O_s^P \mid F_s]$ is minimized over the feasible set $[0, \min(c_s, w_{s-1}^P + d_s)]$, where P refers to LM. We can think of $L_s^P + O_s^P$ as a proxy for the cost that we can associate with the decision in period s; it is approximate because the impact on the waiting costs is approximated with the aggregate incremental cost at s, L_s^P (which is based on the lower limit policy). Conditioned on F_s, both the size of the waiting list at the end of period $s-1$, w_{s-1}^P, and the new elective patient requests in period s, d_s, are known deterministically. In fact, this search can be performed efficiently because L_s^P is convex in q_s^P (by Corollary 1), and O_s^P is also convex in q_s^P (by the definition of O_s^P).
- *Upper minimization policy.* This policy specifies the choice of $q_s^{\mathrm{UM}} \in [0, \min(c_s, d_s + w_{s-1}^P)]$ such that $E[O_s^P - U_s^P \mid F_s]$ is minimized, where P denotes UM. This expression, $E[O_s^P - U_s^P \mid F_s]$, is another proxy for the cost we associate for period s, where we approximate the savings on waiting costs with the aggregate incremental benefit, U_s^P. As before, we can show that this expression is convex in q_s^P.

In each period, our policies minimize a single-dimensional convex function, which can be performed efficiently if we are able to evaluate the function at a given point efficiently. To evaluate the function at any given point, we need to evaluate the expected cost of running a limit policy, which can be computed using a Monte Carlo simulation. The effort required to compute the cost of a limit policy on a sample path of simulation is at most proportional to the length of the horizon, and thus the computational effort scales very nicely (i.e., polynomially) with the size of the problem.

Compared with a *myopic* policy that determines the scheduling decisions in each period by solving the newsvendor problem EC.1 in the online appendix (see Lemma 2), the two policies introduced above take into account the impact of current decisions on *future* (waiting) costs. Hence, they tend to be less "myopic," and it is hoped that they would perform better than the newsvendor-based myopic policy.

4.3. Comparison to Existing Approaches

The problem of how to use evolving demand forecasts to devise effective supply chain management policies in these settings has been the subject of a considerable body of research. We refer the reader to Iida and Zipkin (2006) and Dong and Lee (2003) for more comprehensive discussions. Many works focus on characterizing the structure of the optimal policy for the single-item, periodic-review, stochastic inventory problem, and in many models, including models with Markov-modulated demands, correlated demand, and forecast evolution (Iida and Zipkin 2006, Gallego and Özer 2001, Zipkin 2000), the optimal policy can be shown to be state dependent. In general, the state space of these problems grows exponentially with the number of periods, and as a result, many authors have developed computationally efficient heuristics to compute policies, e.g., Dong and Lee (2003), Lu et al. (2006), and Iida and Zipkin (2006).

One class of heuristics, which we call *look-ahead optimization* (LA), characterizes the portion that can be computed immediately (without considering future decisions) of the long-term cost of a decision and chooses a policy that minimizes this cost in each period. Chan (1999) and Chan and Muckstadt (1999) were the first to consider this approach, and they studied uncapacitated and capacitated multi-item inventory models with linear costs. They defined a penalty function, which we will call *marginal holding cost*, which accounts for the holding cost incurred over the rest of the horizon due to the inventory ordered in this period. Their policy orders in each period to minimize the expected period's backorder cost plus the marginal holding cost. Levi et al. (2005) showed that in single-item inventory problems, the decisions of LA are lower bounds on the decisions of an optimal policy. Truong (2012) proved that LA is an approximation algorithm with a worst-case performance ratio of 2. Levi et al. (2008) defined another look-ahead cost called *marginal backlogging cost* for the capacitated single-item inventory problems. They proved that the policy that minimizes the marginal backlogging cost plus the period's holding cost in each period has inventory positions that are upper bounds on the inventory positions of an optimal policy.

The algorithms we develop in this paper are inspired by the look-ahead optimization approach above. We also attempt to capture the long-term impact of a scheduling decision and make decisions to optimize this impact. However, there also important differences between our approach and those previously undertaken.

First, previous approaches characterize the *minimum impact* of a decision on two separate categories of future costs, namely, holding and backorder costs, to derive the marginal holding and marginal backorder cost, respectively. The minimization of each of these marginal costs plus the myopic (period's) cost leads to a separate (upper or lower) bound. In our problem setting, we characterize the long-term impact of a decision on a *single* category of cost, namely, the cost having elective patients wait. This cost is analogous to the holding cost in inventory problems. We view the impact of this cost from diametric perspectives. We analyze *both the minimum impact and the maximum impact* of a decision on future waiting costs. As a result, we quantify *both an incremental cost and an incremental benefit* to a decision. We minimize the myopic cost plus the incremental cost to obtain new lower bounds and policies, and we minimize the myopic cost minus the incremental benefit to obtain new upper bounds and a different class of policies. As we will show in our computational experiments, the policies obtained by maximizing the incremental benefits are superior to those obtained by minimizing the incremental costs.

Second, in inventory problems, the problem dynamics are simpler. The marginal holding and backorder cost can be stated in terms of the current state and exogenous future random variables, using closed-form expressions. In our problem, the incremental costs and benefits depend more intricately on the sequence of exogenous random variables that are realized in the future, including resource availability and resource constraints, reneging, and emergency arrivals. Thus, there is no easy way to express these costs and benefits. We must define them implicitly, in terms of the costs incurred by certain *limit policies*. The decisions of a limit policy are predefined, so that the expected cost of the entire policy can be calculated efficiently on every sample path. Because of this implicit definition, the analysis of these policies and the proof of the bounds are considerably more technical. The advantage of working with limit policies, however, is that they are a very rich class of policies, and they provide a general method to generate potential bounds. We believe that this method of generating bounds will find application in many other settings.

5. Bounds on the Optimal Decision

The two policies introduced in §4, LM and UM, are based on the incremental cost and benefit, which are estimated by considering the limit policies. These estimated costs provide lower and upper bounds on the waiting cost. Now, we will show that the *decisions* of UM and LM provide lower and upper bounds on the optimal decision in each period. These bounds provide additional motivation and support for another policy, which schedules the number of elective surgeries in each period to be the average of those suggested by the LM and UM algorithms. We will compare the performance of these policies to the performance of an optimal policy in the next section.

Under LM, the aggregate incremental cost function that we use to determine the number of elective surgeries q_s^{LM} in each period s is based on the logic that all future capacity would be allocated for elective cases. Intuitively, this would overestimate the future capacity employed in elective surgeries, and thus underestimate the penalty cost associated with keeping an elective patient request on the waiting list. Therefore, we expect the resulting waiting list under LM to be larger than that under an optimal scheduling policy. Similarly, we expect the waiting list under UM to be smaller compared with that of an optimal policy. The following results formalize our intuition above.

Theorem 5. *For each $s \in \{1, \dots, T\}$, both $\bar{w}_s^{\mathrm{LM}} \ge \bar{w}_s^{\mathrm{OPT}}$ and $w_s^{\mathrm{LM}} \ge w_s^{\mathrm{OPT}}$ hold almost surely.*

Theorem 6. *For period $s \in \{1, \dots, T\}$, both $\bar{w}_s^{\mathrm{UM}} \le \bar{w}_s^{\mathrm{OPT}}$ and $w_s^{\mathrm{UM}} \le w_s^{\mathrm{OPT}}$ hold almost surely.*

Finally, we summarize below the main result of this section, that LM and UM yield lower and upper bounds on the optimal policy. This result follows from the proofs of Theorems 5 and 6.

Corollary 2. *For each period $s \in \{1, \dots, T\}$, suppose we fix the waiting list size w_{s-1} and information set F_s. Then, the decisions produced by LM and UM satisfy $q_s^{\mathrm{LM}} \le q_s^{\mathrm{OPT}} \le q_s^{\mathrm{UM}}$ almost surely, where q_s^{OPT} is the decision of an optimal policy.*

6. Computational Experience

Although we have noted that the optimal policy for our scheduling problem can be difficult to compute, a number of policies emerge from our discussion so far. We introduced two policies, the lower minimizing and upper minimizing policies, in §4. We now study how they perform. Our experimental setting is not meant to simulate real-life instances of the allocation scheduling problem. Rather, we aim to generate a sufficiently rich set of small problem samples, where the optimal policy can be evaluated with reasonable computational effort, so that we can compare the performance of our algorithms with that of an optimal policy or a naive myopic policy.

We will use the following variation of our basic model to make our experiments more comparable to those of Gerchak et al. (1996). In the experiments, we consider the case where only a single resource constraint exists. Instead of having *cases* as the unit for emergency demand e_t, $t = 1, \dots, T$, we use *units of surge capacity* as the unit to specify e_t. Similarly, we use p to capture the cost of using each unit of surge capacity to satisfy emergency demand. We redefine the column $\mathbf{A}_{2t}$, $t = 1, \dots, T$, so that $\mathbf{A}_{2t} e_t$ captures the amount of the normal capacity at t consumed by a quantity of emergency demand that is equivalent to e_t units of surge capacity. This variation is slightly more flexible in that it allows the model to capture variability in the capacity usage of different emergency cases. It can be shown that all of our results hold under this variation.

We consider horizon lengths $T = 5, 14$ periods. The waiting penalty for postponement of an elective patient by one day is $b = 1, 2, 10$. The penalty for using one unit of surge capacity is $p = 24$. For simplicity, we assume that the discount factor is 1 and there is no reneging. We set average demand and capacity values to be similar to those chosen by Gerchak et al. (1996). The capacity is 960 minutes per day, corresponding to two eight-hour shifts. The demands are randomly generated for each problem instance, as we shall next describe.

We specify the demand structure in each experimental instance as a binary probability tree. Let $U[a, b]$ denote a random variable that is uniformly distributed on $[a, b]$. For each problem instance, we generate the branch probabilities of the binary tree in each period from $U[0, 1]$. We generate the newly arriving demand for elective patients in each period at each node in the tree using $U[0, 12]$. Elective surgeries always take 60 minutes per session. We generate the mean number of minutes demanded by emergency patients at each node from $U[320, 480]$. We assume that the actual demand for emergency patients is uniformly distributed between 0 and twice the mean. Once a problem instance (i.e., a binary tree) is generated, the manager is aware of the demand structure (including the average at each node as well as branch transition probabilities) and the current state, but does not know how the state would evolve and how demand would be realized.

We also consider a class of problem instances, which we call *tight* instances. In these instances, the mean demand for emergency patients at each node is generated from a much higher range, namely, from $U[660, 800]$, much closer to the capacity of 960. The rest of the problem data are generated for these instances in the same way as before.

For each combination of the parameters, we generate a number of instances, i.e., binary probability trees, for the demand structure based on the aforementioned distributional assumption (see Table 1). For each instance, we compute the optimal policy OPT using an exhaustive search. To facilitate computing the optimal policy we consider only integer-valued demand and allocation quantities. In addition to UM and LM, we have implemented two other policies. The first policy, a hybrid one denoted by H, is motivated by the results of §5. Because the outcomes of LM and UM form bounds on OPT, our proposed policy H ensures that the waiting list in each period is the average of what it would have been under LM and UM. The second policy we consider is a myopic one, which minimizes the single-period cost given in (4), and we denote this policy by M.

The outcome of our experiments show that UM is the best-performing policy. In 6,600 problem instances, the average performance of UM is within 1.8% of OPT, compared with 35% for the naive myopic policy M. Under a range of problem configurations, UM's performance is never more than 3.6% away than OPT, whereas *M*'s performance can

Table 1 Performance of Approximate Policies

Set up					Mean of performance ratios				Variance of performance ratios			
T	*p*	*b*	Tight?	No. instances	LM	UM	H	M	LM	UM	H	M
5	24	10	No	1,000	1.0815	1.0056	1.0225	1.8099	0.0967	0.0221	0.0345	0.4434
5	24	2	No	1,000	1.0304	1.0016	1.0067	1.4979	0.0380	0.0136	0.0157	0.2569
5	24	1	No	1,000	1.0164	1.0006	1.0014	1.4178	0.0233	0.0050	0.0053	0.2097
5	24	10	Yes	1,000	1.0018	1.0000	1.0007	1.0043	0.0060	0.0003	0.0025	0.0110
5	24	2	Yes	1,000	1.0001	1.0000	1.0000	1.0002	0.0005	0.0000	0.0002	0.0011
5	24	1	Yes	1,000	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	0.0000
14	24	10	No	100	1.2770	1.0360	1.0980	2.3430	0.0811	0.0319	0.0384	0.3032
14	24	2	No	100	1.0968	1.0047	1.0367	1.6151	0.0494	0.0110	0.0223	0.1605
14	24	1	No	100	1.0622	1.0011	1.0221	1.5112	0.0325	0.0039	0.0127	0.1100
14	24	10	Yes	100	1.0150	1.0003	1.0059	1.0310	0.0144	0.0012	0.0065	0.0245
14	24	2	Yes	100	1.0010	1.0000	1.0004	1.0017	0.0021	0.0000	0.0009	0.0030
14	24	1	Yes	100	1.0002	1.0000	1.0001	1.0003	0.0004	0.0000	0.0001	0.0006

be as much as 234% worse than that of OPT. The myopic policy does not perform well because it does not anticipate the long-term impact of a waiting list. Because the cost of postponing an elective patient by one period is small, a wait list is bad only if it is not resolved over many periods. Because the myopic policy sees only one period ahead, it severely overallocates capacity to emergency surgeries. The performance statistics are summarized in Table 1, where the performance ratio is defined as the ratio of the total discounted cost under the scheduling policy in question to that under an optimal policy.

We see that UM consistently outperforms all other policies across a range of experimental settings. It exhibits both smaller average performance ratio with respect to OPT, as well as smaller variance for the performance ratios. By comparison, LM and H are within 4.8% and 2.4% of OPT on average, respectively. Though they are slightly worse than UM, they seem to be much better than M. Recall that to derive UM and LM, we use the future cost of a limit policy as a substitute for the future cost of an optimal policy. The limit policy used to derive UM allocates all available capacity in each period to emergency surgeries, whereas the limit policy used to derive LM allocates all available capacity to elective cases. Because the cost of overplanning for emergency surgeries and making patients wait is typically much smaller than that of underplanning, the cost of the former limit policy is closer to optimal than that of the latter. Hence, as an approximation, UM is expected to perform better than LM. Because UM is consistently better than LM, H is always worse than UM because its performance lies between those of UM and LM.

We also find that LM, UM, and H all perform much better than the myopic policy M, and thus the myopic policy may not be a good policy to use. All of our policies deteriorate in performance with longer horizons and higher waiting costs, although the degradation in performance for UM is relatively small. All of the policies do better under capacity-tight instances because the decision space in these instances is more highly constrained. Hence the decisions undertaken by different policies are closer together and closer to those of OPT.

7. Conclusion

In this paper, we develop a capacity allocation model for allocation scheduling that explicitly considers multiple resources that are needed to serve two classes of patients with different wait-time sensitivities. Our model allows the demands, resource utilization, and capacity availability to be random, nonstationary, and time correlated. We prove similar structural results for the optimal solutions as in the settings with i.i.d. demands. Our primary theoretical contribution is a method to obtain upper and lower bounds on the decisions of an optimal policy in each period. We also develop several computationally efficient policies that are shown to perform very well in our numerical experiments.

Our work is motivated by problems in healthcare service capacity management, in particular problems in surgical scheduling. Many important features in our model, such as dynamic and nonstationary environment and multiresource constraints, are well recognized in the surgical scheduling context (Cardoen et al. 2010, Moore et al. 2008, Dexter et al. 2005). Previous literature in this area usually assumes a simplified setup, such as i.i.d. demand structure and a single generic resource constraint. Our work is able to bring theory closer to practice by considering a much more general setup.

In our model, an elective patient does not receive an appointment at the time of joining the waiting list. This is not an uncommon practice in public health systems (such as Canada, the United Kingdom, and Australia). In the United Kingdom and Australia, waiting lists are kept for elective patients, and they are considered useful in managing surgery schedules (Edwards 1997).

There are several ways to improve our model for practical application. First, it would be useful to consider patient heterogeneity in resource consumption within each patient type, i.e., emergency and elective. Second, we have assumed that patients only consume resources for one period, and it would be interesting to consider a model that allows patients to occupy some resource (e.g., beds) for multiple periods. Third, our model specifies how many elective patients to admit for the current period but not the timing or sequence of service. It would be interesting to develop a sequential decision model that jointly makes these decisions. All extensions above require substantially revised models and analysis, and we leave them for future research.

Electronic Companion

An electronic companion to this paper is available as part of the online version at http://dx.doi.org/10.1287/msom.1120.0415.

Acknowledgments

The authors thank the editor, associate editor, and referees for constructive and positive feedback.

References

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

Cardoen B, Demeulemeester E, Beliën J (2010) Operating room planning and scheduling: A literature review. *Eur. J. Oper. Res.* 201(3):921–932.

Chan EWM (1999) Markov chain models for multi-echelon supply chains. Doctoral dissertation, Cornell University, Ithaca, NY.

Chan E, Muckstadt J (1999) Markov chain models for multi-echelon supply chains. Working paper, Cornell University, Ithaca, NY.

Dexter F, Marcon E, Epstein RH, Ledolter J (2005) Validation of statistical methods to compare cancellation rates on the day of surgery. *Anesthesia Analgesia* 101(2):465–473.

Dexter F, Macario A, Traub RD, Hopwood M, Lubarsky DA (1999) An operating room scheduling strategy to maximize the use of operating room block time: Computer simulation of patient scheduling and survey of patients preferences for surgical waiting time. *Anesthesia Analgesia* 89(1):7–20.

Dong L, Lee HL (2003) Optimal policies and approximations for a serial multiechelon inventory system with time-correlated demand. *Oper. Res.* 51(6):969–980.

Edwards RT (1997) *NHS Waiting Lists: Towards the Elusive Solution*. Office of Health Economics, London.

Gallego G, Özer Ö (2001) Integrating replenishment decisions with advance demand information. *Managment Sci.* 47(10):1344–1360.

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

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

Gupta D (2007) Surgical suites' operations management. *Production Oper. Management* 16(6):689–700.

Iida T, Zipkin PH (2006) Approximate solutions of a dynamic forecast-inventory model. *Manufacturing Service Oper. Management* 8(4):407–425.

Kuttenkuler J (2004) VCU neurosurgeon develops new device for performing deep brain surgery. Accessed November 30, 2012, http://www.news.vcu.edu/news/VCU_neurosurgeon_develops_new_device_for_performing_deep_brain.

Levi R, Pál M, Roundy R, Shmoys DB (2005) Approximation algorithms for stochastic inventory control models. Michael J, Kaibel V, eds. *Integer Programming and Combinatorial Optimization*, Lecture Notes in Computer Science, Vol. 3509 (Springer, Berlin), 306–320.

Levi R, Roundy RO, Shmoys DB, Truong VA (2008) Approximation algorithms for capacitated stochastic inventory control models. *Oper. Res.* 56(5):1184–1199.

Liu N, Ziya S, Kulkarni VG (2010) Dynamic scheduling of outpatient appointments under patient no-shows and cancellations. *Manufacturing Service Oper. Management* 12(2):347–364.

Lu X, Song J-S, Regan A (2006) Inventory planning with forecast updates: Approximate solutions and cost error bounds. *Oper. Res.* 54(6):1079–1097.

Moore IC, Strum DP, Vargas LG, Thomson DJ (2008) Observations on surgical demand time series: Detection and resolution of holiday variance. *Anesthesiology* 109(3):408–416.

Patrick J, Puterman ML, Queyranne M (2008) Dynamic multipriority patient scheduling for a diagnostic resource. *Oper. Res.* 56(6):1507–1525.

Topkis DM (1998) *Supermodularity and Complementarity* (Princeton University Press, Princeton, NJ).

Truong VA (2012) The pediatric vaccine stockpiling problem. *Vaccine* 30(43):6175–6179.

Ward AR, Glynn PW (2003) A diffusion approximation for a Markovian queue with reneging. *Queueing Systems* 43(1):103–128.

Zipkin PH (2000) *Foundations of Inventory Management*, Vol. 2 (McGraw-Hill, Boston).