Managing Patient Service in a Diagnostic Medical Facility

Linda V. Green, Sergei V. Savin, Ben Wang

Columbia Business School

August 2003, revised April 2004

Abstract

Hospital diagnostic facilities, such as magnetic resonance imaging centers, typically provide service to several diverse patient groups: outpatients, who are scheduled in advance; inpatients, whose demands are generated randomly during the day; and emergency patients, who must be served as soon as possible. Our analysis focuses on two inter-related tasks: designing the outpatient appointment schedule, and establishing dynamic priority rules for admitting patients into service.

We formulate the problem of managing patient demand for diagnostic service as a finite horizon dynamic program and identify properties of the optimal policies. Using empirical data from a major urban hospital, we conduct numerical studies to develop insights on the sensitivity of the optimal policies to the various cost and probability parameters and to evaluate the performance of several heuristic rules for appointment acceptance and patient scheduling.

1 Introduction

Medical diagnostic facilities, such as Magnetic Resonance Imaging (MRI) installations, constitute a critical component of a comprehensive health care system. In an increasingly competitive industry where both costs and customer expectations are rising rapidly, hospital managers are under greater pressure to manage such facilities more efficiently and effectively.

In many cases, hospital-based imaging facilities are accessed by a wide range of patients, from both inside and outside of the hospital. These patients can be grouped into three broad categories, each of which has distinct demand characteristics: outpatients, inpatients and emergency patients. Outpatient appointments are typically scheduled days or weeks in advance and sometimes result in cancellations and no-shows. On the other hand, inpatient demands are usually generated the same day as needed, while emergency patients must be served as soon as possible following the physician's request. The financial characteristics of these three classes are also generally quite different. The hospital typically receives a "fee-for-service" for providing outpatient services, while inpatient care is typically reimbursed based on a DRG (diagnostic-related group) code and is irrespective of what resources are actually used. As a result, hospital managers often view imaging procedures for outpatients as a source of additional revenue but as a cost for inpatients. Emergency patients may fall into either category, largely depending on whether or not they are ultimately admitted as inpatients.

Diagnostic imaging equipment is very expensive. For example, a new MRI costs approximately $2 million with a commensurate cost for building and preparing the space it will occupy. In addition, the purchase of MRIs is typically regulated by the states through a certificate of need (CON) process which is used, among other things, to restrict the supply of MRIs (and other expensive technologies) in order to control costs (much of which is

paid by the states through Medicaid and other insurance programs). Therefore, hospital managers have every incentive to keep these machines fully utilized. This is often done by filling most, if not all, examination slots during the day with outpatient appointments. In particular, little or no slack is allocated for unanticipated inpatient demands or emergencies. This scheduling approach sometimes results in postponing an inpatient exam one or more days, potentially delaying the patients's discharge from the hospital and, therefore, increasing hospital costs, which are typically not reimbursed due to the DRG prospective payment system. Outpatients, too, often experience significant delays.

The management of a diagnostic facility consists of two interrelated tasks: establishing an appointment schedule for outpatients, and designing a system of dynamic priority rules for admitting patients into service in real time. Appointment scheduling consists of determining the duration, number, and timing of examination slots for a particular day. This task may be further complicated by outpatient cancellations and "no-shows". Dynamic priority rules provide real-time control of access to the facility by potentially competing patient classes. In particular, before the beginning of each examination slot (service period) there may be waiting patients from more than one class due to the random arrivals of requests generated by inpatients and emergency patients. If there is an emergency service request, it gets the highest priority. In the absence of an emergency patient, a decision must be made as to whether a (scheduled) outpatient or a (non-scheduled) inpatient should be served next, if both are waiting. Serving an inpatient will result in a delay for one or more outpatients, which will adversely affect perceived service quality and could result in lost business. Yet, selecting an outpatient may ultimately result in longer hospital stays for some inpatients.

These capacity management tasks - appointment scheduling and real-time capacity allocation - are interrelated. On the one hand, the selection of a specific appointment schedule affects the likelihood and timing of both inpatient and outpatient delays which

will affect the choice of the real-time allocation policy. On the other hand, the impact of a particular schedule on the overall performance of the service facility likely depends on the selected real-time allocation policy. Thus, both sets of decisions have to be analyzed within a common decision framework, which we present and analyze in this paper.

Specifically, this paper makes the following contributions:

1. We model the operations of a medical diagnostic facility with several patient types as a dynamic stochastic control problem and establish structural properties of an optimal real-time capacity allocation policy under an arbitrary outpatient appointment schedule. In particular, we show that under mild assumptions about the structure of the cost and revenue functions, the expected profits are optimized by monotone "switching curve" policies.

2. We establish additional structural properties of the optimal real-time capacity allocation controls under a "threshold" outpatient appointment policy often used in practice. We identify conditions for which a simple priority scheme is optimal, and prove that in general, the switching curve is a function of only the service period index and the number of waiting patients belonging to the "critical" class (which is likely to be inpatients). Furthermore, we show that the critical number which governs the switching policy is monotonic in the service period index and is independent of the appointment schedule.

3. We develop a linear heuristic as a simpler alternative to the optimal real-time service policy, and, using a series of numerical studies based on the operational data from an actual hospital MRI facility, show that it works well over a broad range of parameter settings. This heuristic reduces to an even simpler heuristic for some parameter values, and we identify situations in which this simpler heuristic performs well.

4. Though we do not obtain an optimal policy for the more difficult problem of appointment scheduling, we numerically explore the performance of some simple heuristics, and show that one of these, combined with an appropriate service heuristic, works well in many circumstances.

The rest of the paper is organized as follows. We start with a review of the related literature in the next section. In sections 3 and 4, we describe the model and establish the structural properties of an optimal real-time capacity management policy for admitting patients into service under an arbitrary appointment schedule. In section 5, we focus on "threshold" appointment schemes, under which the first specified number of examination slots of the day are scheduled with outpatients and the remaining slots are left open for inpatients. We develop a linear approximation heuristic in section 6, and in section 7, we use the data from the operations of a real MRI diagnostic facility as the basis of an experimental design to numerically study the performance of this heuristic and others, and to explore the performance of some commonly used policies for outpatient scheduling. We conclude in section 8 with a discussion of our results and potential directions for future research.

2 Review of Related Literature

Our analysis of the operations of a medical diagnostic facility shares some traits with a number of stochastic control and scheduling applications.

The dynamic allocation of service capacity among several competing customer classes has been studied in a variety of contexts. Those most similar to our problem include hotel management (Liberman and Yechiali (1978), Bitran and Gilbert (1996)), car rentals (Carrol and Grimes (1995) and Geraghty and Johnson (1997)), airline yield management (Belobaba (1989)), telecommunications (Ross and Tsang (1989), Altman *et al.* (2001),

Örmeci *et al.* (2001)), and call center management (see Gans *et al.* (2002) for a comprehensive review). In particular, the last two research streams often model the service capacity allocation as a dynamic priority queueing control problem. In some of these business environments (e.g. hotel management, car rentals, airline yield management, telecommunications), the provision of service cannot be delayed, i.e., if the demand can't be served at the requested time, the customer is lost. In the diagnostic facility setting, service of both inpatients and outpatients may be delayed with appropriate penalties. In call center research, the analyzed service system is often modeled as a delay system, similar to our case. However, the analysis typically focuses on the properties of the stationary state in the infinite horizon setting. On the contrary, in our model, we analyze transient behavior of the diagnostic system in the finite horizon setting corresponding to a day of service.

As in our case, the scheduling literature considers finite horizon models in which the objective is to identify a service sequence which minimizes either expected or total cost (see e.g. Pinedo (2001)). However, there are two important distinctions. First, in scheduling models, it is assumed that all jobs will be processed by the end of the horizon, while in our case, some patients may be lost or rescheduled to another day. Second, in scheduling models it is assumed that a fixed number of jobs are "released" either in the beginning of the horizon or at some, perhaps, random time during the horizon. In our setting, jobs are released only at pre-specified time intervals, but the actual total number of service requests over the horizon is a random variable.

There is a distinct literature dealing with medical appointment scheduling. Most of these papers (e.g. Bailey (1952), Soriano (1964), Fries and Marathe (1980), and Ho and Lau (1992)) address the scheduling of a single patient class and focus on the impact of policies on both patient and physician waiting times. The problem of allocating medical service capacity between distinct demand streams has received only limited coverage in

the research literature. Gerchak, Gupta and Henig (1996) analyzed this problem in the setting of an operating room where the capacity is shared between elective and emergency surgeries. The focus of their work is on the reservation planning policy for elective patients. Though closely related to the outpatient appointment scheduling problem we consider, their model does not have unscheduled inpatient demand.

3 Model Description

On each working day, we consider N identical service slots (periods), some of which may have been reserved through the appointment system. The assumption of identical length service slots reflects common practice with typical lengths ranging from 30 minutes to an hour. In our analysis, we assume that examination times are the same for all patients, irrespective of their type, and equal to the length of the service slot. In practice, the average service time may depend on the type of the exam performed and may vary from patient to patient within the same exam type. (Some facilities allocate more than one service slot for examinations that typically take longer than average, for example, abdomen scans.) We express the schedule of accepted appointments as an N-dimensional binary vector a: $a_i = 1$ if the i-th appointment slot has been filled and $a_i = 0$ otherwise, $i = 1, ..., N$.

We assume that non-scheduled inpatient and emergency demands occur randomly during the day. In particular, the intensity of inpatient and emergency demands is considered to be relatively low, so that it is unlikely that more than one request for each type of service arrives during each service period. We denote by p_e and p_n the arrival probabilities of emergency and inpatient service requests, respectively, during any service period. We also assume that there is a positive probability of a "no-show" for each scheduled outpatient appointment. We denote by p_s the probability that a scheduled outpatient shows up for the appointment. We assume that the arrival streams of the three patient groups are

independent of each other.

Using these assumptions, we can model the dynamics of the diagnostic facility as a discrete time Markov chain, whose trajectory is determined by the selection of the patient type to be admitted into service in the beginning of each service period. The state of the service system consists of the numbers of (non-scheduled) inpatients and (scheduled) outpatients, (n_i, s_i), waiting for service right after the beginning of i-th service period ("after-the-action state"). For the i-th service slot, we consider the following finite state space $S_i = ((n, s)\,|0 \leq n \leq i-1, 0 \leq s \leq i-1)$. This definition reflects our assumption that there can be at most one arrival of each type of patient during any service period - in particular, the number of patients of each type waiting during the $(i+1)$-st service slot can be higher than those waiting during the i-th slot by at most 1. At the end of each service period, $i = 1, ..., N-1$, there are several possible actions available, depending on whether or not an inpatient or emergency patient request has arrived during the period.

We assume that the diagnostic facility collects a revenue of r_s and r_n per examination for each outpatient and inpatient, respectively. Delaying a service request incurs a waiting cost per period of w_s and w_n for outpatients and inpatients, respectively. Finally, there is a penalty function $f(n, s)$ associated with patients not served by the end of the day. In the simplest case, there are penalty costs π_s and π_n for each outpatient and inpatient not served. Based on information obtained from several hospitals, the following assumptions reflect typical relationships concerning these revenue and cost parameters: (1) the revenue brought in by an outpatient dominates the revenue brought in by an inpatient $(r_s > r_n)$; (2) the waiting cost per service period for an outpatient is greater than that for an inpatient $(w_s > w_n)$ (since the inpatient is in the hospital anyway) and (3) in the linear penalty cost case, the end-of-day penalty cost for an inpatient is greater than that for an outpatient $(\pi_s < \pi_n)$. This last relationship is based on several factors. First, the failure to examine an inpatient by the end of the day typically results in an extra day of hospitalization

for that patient, particularly since an MRI is often required to verify that the patient is ready to be discharged. Second, each additional day translates to an additional net cost for the hospital since payment for most insured patients is based on a prospective (DRG) basis which is independent of the actual time or resources required for treatment. The cost of an extra day is considered by most hospitals to be very high because of variable staffing costs and the opportunity cost associated with not having the bed available for other patients. This latter cost is most relevant for large hospitals that operate at high average occupancy levels, such as the one we studied which has an average occupancy level of over 90%. Finally, if one or more outpatients are still waiting to be examined at the end of the normal workday, the scans will generally be performed using overtime. In some of these cases, the patients leave the facility before being examined and either choose to have their scan done elsewhere or reschedule for another day. In either case, the associated net penalty cost is likely to be significantly lower than the cost of an additional inpatient hospital day.

Using the cost and revenue structure and the system dynamics introduced above, we can formulate the profit maximization problem for a diagnostic facility as a finite-horizon dynamic program. For a given appointment schedule a, let $V_i^a(n,s)$ be the optimal total expected profit over the $(N-i)$-period planning horizon, starting in the i-th service period and ending in the N-th, when the state of the system right after the beginning of the i-th service period is (n,s). Using well-known results on Markov dynamic programming, we can formulate the optimality equation satisfied by the optimal cost function:

$$
\begin{aligned}
& V_i^a(n,s) \\
= \; & -sw_s - nw_n + p_e p_n \left[(1 - p_s a_{i+1}) V_{i+1}^a(n+1,s) + p_s a_{i+1} V_{i+1}^a(n+1,s+1) \right] \\
& + p_e (1 - p_n) \left[(1 - p_s a_{i+1}) V_{i+1}^a(n,s) + p_s a_{i+1} V_{i+1}^a(n,s+1) \right] \\
& + p_n (1 - p_e) \left[(1 - p_s a_{i+1}) H_{i+1}^a(n+1,s) + p_s a_{i+1} H_{i+1}^a(n+1,s+1) \right]
\end{aligned}
$$

$$+(1-p_e)(1-p_n)\left[(1-p_s a_{i+1})H^a_{i+1}(n,s)+p_s a_{i+1}H^a_{i+1}(n,s+1)\right],$$

$$(n,s) \in S_i,\ i=1,...,N, \tag{1}$$

where the optimal actions are determined by maximization operators

$$H^a_i(n,s)=\begin{cases}\max[V^a_i(n-1,s)+r_n, V^a_i(n,s-1)+r_s], & \text{if } n\geq 1, s\geq 1,\\ V^a_i(n-1,0)+r_n, & \text{if } n\geq 1, s=0,\\ V^a_i(0,s-1)+r_s, & \text{if } n=0, s\geq 1,\\ V^a_i(0,0), & \text{if } n=s=0.\end{cases} \tag{2}$$

The boundary condition (at the end of each working day) for the recursion (2) is given by

$$H^a_{N+1}(n,s)=V^a_{N+1}(n,s)=f(n,s), \tag{3}$$

where $f(n,s)$ is assumed to be a known function defined on S_{N+1}.

(1) states that when n inpatients and s outpatients are waiting for service during period i, the waiting penalty cost $-sw_s-nw_n$ is incurred and the system can find itself in one of eight possible states (depending on arrivals of new service requests) before the beginning of period $i+1$. For example, if during i-th service period there is an emergency as well as a non-emergency inpatient arrival, but the outpatient scheduled for $(i+1)$-st slot does not arrive (the probability of this is $p_e p_n(1-p_s a_{i+1})$), the system will start the $(i+1)$-st period in state $(n+1,s)$, since the $(i+1)$-st slot will be used for the emergency patient. If, however, there is no emergency arrival during the i-th service period but both a new inpatient as well as the outpatient scheduled for the $(i+1)$-st slot arrive (the probability of this is $(1-p_e)p_n p_s a_{i+1}$), the $(i+1)$-st period will start in either state $(n+1,s)$ or $(n,s+1)$, depending on whether an outpatient or an inpatient is served during the $(i+1)$-st slot. At the heart of the optimization (1) is the choice (2) between serving outpatient customers who arrive via the appointment system and inpatient customers whose service requests arrive "unexpectedly". Clearly, the optimal choice is influenced by the interplay between

the cost parameters of these two customer groups, as well as by how close the time of the decision is to the end of the working day.

On the strategic level, the diagnostic facility can maximize expected profits by choosing a favorable appointment schedule a. We assume that outpatient demand is high relative to capacity and so all service slots reserved for outpatients will actually be scheduled. Using the notation introduced above, we formulate the appointment scheduling problem as follows:

$$V^* = \max_a \left[V_1^a (0,0) \right], \tag{4}$$

where V^* denotes the optimal expected daily profits assuming the system begins empty each day. This last assumption assures stability and simplifies our analysis. In reality, some hospitals may have inpatients waiting for service in the beginning of the day, particularly as a result of a backlog from the previous day. Our development of optimal real-time allocation policies (section 4) can accommodate this case. The optimization model (1), (3), (4) reflects the interaction between the strategic appointment scheduling and the tactical capacity allocation in the daily operations of the medical diagnostic facility. In the next section we investigate the properties of the optimal real-time capacity allocation policy under an arbitrary appointment scheme.

4 Structural Properties of the Optimal Service Policy

Define G_{N+1} as the class of profit functions defined on S_{N+1} such that for every $g \in G_{N+1}$

$$g(n,s+1) - g(n+1,s) \leq g(n+1,s+1) - g(n+2,s), \tag{5}$$

$$g(n+1,s) - g(n,s+1) \leq g(n+1,s+1) - g(n,s+2), \tag{6}$$

$$g(n,s) - g(n,s+1) \leq g(n+1,s) - g(n+1,s+1), \tag{7}$$

$$g(n,s) - g(n+1,s) \leq g(n+1,s) - g(n+2,s), \tag{8}$$

$$g(n,s) - g(n,s+1) \;\leq\; g(n,s+1) - g(n,s+2), \tag{9}$$

where we have assumed that all the states for which g is evaluated belong to S_{N+1}. Submodularity (7) of the profit function implies that when a diagnostic facility is shared between two different patient classes, the profit improvement resulting from the reduction in the number of waiting patients of one type (e.g., outpatients) is an increasing function of the number of the second type patients (e.g., inpatients) waiting for service. (Note that rearrangement of the terms results in a symmetric result.) The concavity of the profit function also implies that a similar statement is valid with respect to increases in the number of the same type of patients waiting for service. Both of these properties are intuitively appealing since they both underscore the increasing importance of wait reduction as the congestion in the system grows. Using approaches developed in Topkis (1978) and Glasserman and Yao (1994), we can establish that (5) and (6) have special implications for the structure of optimal capacity allocation policies:

Proposition 1

For any $f \in G_{N+1}$ and an arbitrary appointment scheme a, in each service period $i = 1, ..., N$ the optimal capacity allocation policy belongs to the class of monotone "switching curve" policies:

a) For any state $(n,s) \in S_i$, there exists a critical index $n_i^a(s)$ such that an outpatient (inpatient) is selected for service if and only if $n < n_i^a(s)$ $(n \geq n_i^a(s))$.

b) Critical indices $n_i^a(s)$ form monotone sequences: $(s_1 \geq s_2) \Rightarrow n_i^a(s_1) \geq n_i^a(s_2)$.

This proposition establishes that for any given service slot and specified number of patients of a given class (e.g. outpatients), it will be optimal to serve that class only if the number of the other class (e.g. inpatients) is below a critical value. However, this critical value increases as the number of outpatients increases. Therefore, the optimal policy assigns service priority so as to balance the congestion due to the two patient classes.

Switching curve capacity allocation policies outlined in Proposition 1 are similar to the admission control policies established by Altman *et al.* (2001) and Örmeci *et al.* (2001) for loss service systems, and the optimal control policies for make-to-stock/make-to-order production systems derived by Ha (1997) and Carr and Duenyas (2000). However, these papers analyzed stationary state, infinite horizon models, while our results pertain to a transient, finite horizon system.

An example of the switching curve capacity allocation, determined by solving the dynamic program, is presented in Figure 1, where the appointment scheme is described as $a_i = 1, i = 1, ..., 15$, $a_i = 0, i = 16, ..., 20$, and the penalty function is given by $f(n, s) = -s^2\pi_s - n^2\pi_n \in G_{N+1}$, for $\pi_s, \pi_n \geq 0$. The switching curve policies introduced in Proposition 1 reflect the real-time allocation of the capacity between the scheduled and non-scheduled components of the demand for diagnostic services. The decision on which type of waiting patient to serve next is clearly influenced by the problem's parameters. In particular, as the relative importance of a particular patient group increases, the values of the critical indices change accordingly:

Proposition 2

a) For any $f \in G_{N+1}$, and in each service period $i = 1, ..., N$, critical indices $n_i^a(s)$ are non-increasing functions of r_n, w_n, p_n and non-decreasing functions of r_s, w_s, p_s.

b) If $f(n, s) = -n\pi_n - s\pi_s$, $n_i^a(s)$ are non-increasing functions of π_n and non-decreasing functions of π_s.

The general properties of the optimal capacity allocation policy presented in Propositions 1 and 2 are valid for arbitrary appointment schemes. In the next section we consider a particular class of appointment schedules we call "threshold". We show that the special structure of threshold schedules allows us to develop a more detailed characterization of the optimal capacity allocation policy.

5 Threshold Appointment Schedules and Linear Penalty Functions

Threshold appointment schedules, in which all service slots before a specified time are used for scheduling outpatients and later slots are left open, are very common in hospital diagnostic facilities. Since tight control of MRI purchases coupled with growing demand typically result in waits of days or weeks for appointments, we make the assumption that all designated outpatient slots are actually filled in advance with outpatients.

We designate the threshold, a^*, to be the last slot for which an appointment is made, i.e., $a^* = \max(i|a_i = 1)$. Therefore, under this scheme we have $a_i = 1$ if $i \leq a^*$ and $a_i = 0$ if $i > a^*$. Using this special structure, and assuming that the end-of-day penalty function is linear, i.e. $f(n,s) = -n\pi_n - s\pi_s$, we can provide a sharper characterization of the optimal tactical capacity allocation policy:

Proposition 3

Let $\pi_n + r_n + w_n \geq \pi_s + r_s + w_s$ *and consider a service period* $i = 1, ..., N$.

Then, $(w_n \geq w_s) \Rightarrow n_i^a(s) = 0, \forall i = 1, ..., N, \forall s = 0, ..., i$.

Alternatively, $w_n < w_s$ *implies that*

a)

$$V_i^a(n, s+1) - V_i^a(n+1, s) \leq V_{i+1}^a(n, s+1) - V_{i+1}^a(n+1, s), \forall (n,s) \in S_{i-1}. \qquad (10)$$

b) there exists an index n_i^* *such that* $n_i^a(s) = n_i^*, \forall i = 1, ..., N, \forall s = 0, ..., i$.

c) the value of n_i^* *is independent of* p_s *or appointment threshold* a^*.

d) n_i^* *is monotone decreasing with the service slot index* i: $n_{i+1}^* \leq n_i^*$.

The statements of Proposition 3 imply that the condition $\pi_n + r_n + w_n \geq \pi_s + r_s + w_s$ insures what we call a "critical" status for the inpatients: the likelihood of serving them

increases as the end of the day approaches and, in particular, inpatients have priority in the last slot of the day. In addition, inpatients acquire an absolute service priority over the outpatients when $w_n \geq w_s$. We note that similar results can be obtained for the case of $\pi_n + r_n + w_n \leq \pi_s + r_s + w_s$, when outpatients become the critical patient class. Proposition 3 also implies that for the threshold appointment approach, linear penalty costs insure that the optimal switching curves defined in Proposition 1 take the simple form of "switching indices": outpatients are served if and only if the number of waiting inpatients is less than the switching index value n_i^*. It is important to note that this policy is independent of both the number of outpatients in the system as well as the chosen threshold policy - a property which is a direct consequence of the linearity of the penalty cost function. Clearly, the results of Proposition 3 can be restated to yield an analogous result if outpatients are the critical class. Therefore, the statement of the Proposition can be generalized as follows. First, for any set of problem parameters one of the patient classes attains the "critical" status. Second, if the waiting cost for the critical class dominates the waiting cost for the non-critical class, it is optimal to employ "critical first" capacity rationing policy: critical patients should be served whenever possible. Finally, if the waiting cost for the critical patients is less than the waiting cost for non-critical ones, the critical class should be served whenever the number of them waiting equals or exceeds the switching index value.

6 Linear Capacity Allocation Heuristic

Under the optimal capacity allocation policy described by Proposition 3**,** capacity management decisions are based on the time of day (slot index i) and the number of critical customers waiting for service at that time. The fairly complex nature of this optimal capacity allocation, especially for a large number of service slots, creates an incentive to identify heuristic capacity allocation policies which may be easier to compute and im-

plement. The structure of the Bellman's equation (1) and the initial condition (3) with linear penalty function suggest a linear approximation for the optimal value function. The following result formally describes such an approximation:

Proposition 4

Consider the finite horizon dynamic program

$$\begin{aligned}
&\widehat{V}_i^a(n,s) \\
= \;& -sw_s - nw_n + p_e p_n\left[(1-p_s a_{i+1})\,\widehat{V}_{i+1}^a(n+1,s) + p_s a_{i+1}\widehat{V}_{i+1}^a(n+1,s+1)\right] \\
& + p_e(1-p_n)\left[(1-p_s a_{i+1})\,\widehat{V}_{i+1}^a(n,s) + p_s a_{i+1}\widehat{V}_{i+1}^a(n,s+1)\right] \\
& + p_n(1-p_e)\left[(1-p_s a_{i+1})\,\widehat{H}_{i+1}^a(n+1,s) + p_s a_{i+1}\widehat{H}_{i+1}^a(n+1,s+1)\right] \\
& + (1-p_e)(1-p_n)\left[(1-p_s a_{i+1})\widehat{H}_{i+1}^a(n,s) + p_s a_{i+1}\widehat{H}_{i+1}^a(n,s+1)\right], i = 1,...,N, \\
n \;\in\; & Z, s \in Z, \qquad (11)
\end{aligned}$$

with

$$\widehat{H}_i^a(n,s) = \max[\widehat{V}_i^a(n-1,s) + r_n, \widehat{V}_i^a(n,s-1) + r_s], n \in Z, s \in Z \qquad (12)$$

and

$$\widehat{V}_{N+1}^a(n,s) = -s\pi_s - n\pi_n, n \in Z, s \in Z. \qquad (13)$$

Then,

$$\widehat{V}_i^a(n,s) = \alpha_i n + \beta_i s + \gamma_i, \qquad (14)$$

where the values of coefficients α_i, β_i and γ_i are given by:

$$\begin{aligned}
\alpha_i &= -\pi_n - (N-i+1)w_n, \\
\beta_i &= -\pi_s - (N-i+1)w_s, \\
\gamma_i &= \gamma_{i+1} - p_n(\pi_n + (N-i)w_n) - p_s a_i(\pi_s + (N-i)w_s) \\
&\quad + (1-p_e)\max(r_n + \pi_n + (N-i)w_n, r_s + \pi_s + (N-i)w_s), \\
i &= 1,...,N, \gamma_{N+1} = 0. \qquad (15)
\end{aligned}$$

Proposition 4 states that the removal of the special structure of the profit maximization operator $H_i^a(n,s)$ at the states where either n or s (or both) are equal to 0, leads to the linearization of the optimal profit function. It is therefore to be expected that the linear approximation (14) to the optimal value function works well in cases when the system is subjected to high patient loads and therefore the probability of no waiting patient of one or both types is low. Due to the linearity of (14) with respect to n and s, the choice between serving an inpatient and an outpatient for a given service slot i does not depend on the state of the system (n,s). In particular, we observe that for (14), the outpatients are served if and only if the service slot index i does not exceed the critical value i_h^* which using (2) can be expressed as follows:

$$i_h^* = \begin{cases} 0, & \text{for } (w_s = w_n, r_n + \pi_n \geq r_s + \pi_s) \\ & \text{and } \left(w_s \neq w_n, N - \frac{r_n+\pi_n-r_s-\pi_s}{w_s-w_n} \leq 0\right), \\ N, & \text{for } (w_s = w_n, r_n + \pi_n < r_s + \pi_s) \\ & \text{and } \left(w_s \neq w_n, \frac{r_n+\pi_n-r_s-\pi_s}{w_s-w_n} \leq 0\right), \\ \text{Floor } \left[N - \frac{r_n+\pi_n-r_s-\pi_s}{w_s-w_n}\right], & \text{for } \left(w_s \neq w_n,\ 0 < N - \frac{r_n+\pi_n-r_s-\pi_s}{w_s-w_n} < N\right). \end{cases} \tag{16}$$

This linear approximation (LA) heuristic policy suggests that outpatients are served in the beginning of the day ($i \leq i_h^*$) and inpatients at the end of the day ($i > i_h^*$). We explore the performance of this and other heuristic policies in the next section.

7 Numerical Study: Evaluating the Performance of Heuristic Policies

We used data from the operations of the MRI facility in a large urban hospital as a basis to explore the impact of the model parameters on the performance of alternative strategic and tactical policies. Most of this data was collected over a three week period

using a combination of observers and handwritten scheduling documents to determine the total requests and actual arrivals for diagnostic service. We estimated the probability of an inpatient demand to be $p_n = 0.4$, and the probability of an emergency arrival to be $p_e = 0.1$. The estimate of $p_s = 0.84$ was calculated from the total number of actual outpatient arrivals relative to appointments. The number of examination slots per day, $N = 20$, reflected a recently expanded operating schedule due to increasing backlogs. Revenue and cost estimates were based largely on conversations with hospital managers. We use $r_s = \$1000$ as the average fee charged by the hospital for an outpatient MRI exam. Estimating the corresponding revenue for inpatients is much more problematic. As explained earlier, inpatient charges are determined by a lump-sum DRG system. Any given DRG category encompasses a considerable variety of possible resource options (e.g. use of an MRI or not), and inpatients requiring an MRI come from a broad range of DRG categories. In addition, though it is possible that the hospital could lose the revenue from a scheduled outpatient if that outpatient decides not to use the hospital's facility after a long delay, it would be highly unusual for an inpatient request for an MRI not to be met before leaving the facility. We decided to use $r_n = \$200$ as an estimate of the inpatient fee, though admittedly this is somewhat speculative. The waiting costs w_s and w_n were estimated as the opportunity costs of waiting for the duration of one service period for each patient type. Since inpatients are in the hospital anyway, we assume that $w_n = \$0$. For outpatients, we estimate their waiting cost as the average hourly wage by taking the ratio of an estimate of the average annual salary ($30,000/year) and dividing by the number of working hours per day (8) times the number of working days in a year (250), or $w_s = \frac{\$30,000}{8\times 250} = \$15$. This is based on the assumption that the typical service period is one hour. The end-of-day penalty costs are perhaps the hardest parameters to estimate. In our numerical studies, unless stated otherwise, we use the linear penalty function $V_{N+1}^a(n,s) = f(n,s) = -s\pi_s - n\pi_n$, where $\pi_n = \$2000$ and $\pi_s = \$100$. Here

we assume that the penalty associated with an inpatient who is not served by the end of the day is the cost associated with an extra, uncompensated day in the hospital, while the penalty associated with an outpatient is some combination of the cost associated with conducting the examination using overtime and the loss of good will if the patient is not examined that day.

Table 1 provides a summary of this set of problem parameters which we call the "base case". In order to compensate for the unreliability of some of the cost data as well as to represent a spectrum of possible actual operating situations across hospitals, we expanded our numerical experimental design to include a fairly broad range of values around this basic set. We kept the parameter setting for outpatient revenue as $r_s = \$1000$ and varied r_n from a low of \$0 to a high of \$800. While it seemed reasonable to continue to assume that $w_n = \$0$, we used values for w_s ranging from \$10 to \$20. Since data obtained from other hospitals indicated that the assumption of $\pi_n = \$2000$ was on the high end, we used this as the uppermost value in our experiments and added the values of \$500 and \$1000. And based on some initial numerical results and our hypothesis that the results would be highly influenced by the ratio of the inpatient to outpatient penalty cost, we decided to let π_s vary from \$100 to \$300. In our experiments on the impact of the probability parameters, we varied the probabilities of an inpatient arrival, p_n, and an outpatient showing up for an appointment, p_s, from 0 to 1, and allowed the probability of an emergency arrival, p_e, to increase to 0.25.

We first explicitly compute the optimal capacity management policies for the base case. We then numerically explore the performance of several heuristic tactical policies, including the linear heuristic developed previously. Next, we propose and explore the performance of several heuristic policies for scheduling outpatients. This is of particular interest since we do not have any analytical results for the optimal appointment schedule. Finally, we present the results of a simulation that tests the robustness of capacity management ap-

proaches in the case when diagnostic service times are random. For each problem set we studied, the optimal tactical capacity allocation policy for a given appointment schedule was computed by solving the dynamic program (1)-(3) with $f(n,s) = -\pi_n n - \pi_s s$. The optimal appointment threshold was determined through a one-dimensional search for a maximizer of (4) over all possible threshold values $a^* = 0, ..., N$.

7.1 The Base Case

For the parameter values corresponding to our base case described by Table 1, Proposition 3 implies that inpatients are the critical class and that the optimal tactical policy takes the form of switching indices. Figure 2a shows this optimal set of tactical switching indices as well as the optimal threshold, $a^* = 15$. Note that the switching indices, $n_i^* = 1$ for $i \geq 15$ indicate that inpatients have non-preemptive priority for these slots even if outpatients are waiting.

Though this result may suggest that the period at which inpatients first get priority is an upper bound for the optimal value of the appointment threshold, this is generally not true. Figure 2b shows a counter-example for which the optimal policy combines scheduling outpatients into all slots with a tactical policy which always gives inpatients priority (an "inpatients first" policy). The problem data set used in this figure coincides with the base case except for the costs associated with outpatients: $w_s = \pi_s = 0$. Structure of the optimal policy for this case is not surprising since the condition $w_s = w_n$ assures the optimality of the inpatients first tactical policy from Proposition 3, while elimination of the end-of-day penalty for outpatients makes it optimal to schedule as many as possible.

7.2 Heuristic Service Policies

Since the optimal service policy described above is complex from a managerial perspective and because it may be very difficult to obtain accurate estimates for all of the required data, we are interested in exploring the performance of two simpler service policies: "critical first" and the linear approximation (LA) heuristic derived in the previous section. The critical first policy always gives service priority to a patient of the critical type (in the sense of Proposition 3). For example, if $r_n + w_n + \pi_n > r_s + w_s + \pi_s$, the critical first policy becomes an inpatients first policy, which is used by some hospitals motivated by the financial (and perhaps, clinical) risk of potentially postponing inpatient exams to another day. On the other hand, if $r_n + w_n + \pi_n < r_s + w_s + \pi_s$, the critical first policy assigns absolute priority to outpatients, a practice followed by hospitals that want to avoid lost outpatient revenue. Note that by Proposition 3, the critical first policy is optimal for certain parameter settings and that the LA policy coincides with the critical first policy when $i_h^* = 0$ or when $i_h^* = N$.

Table 2 shows the relative profit gap ε (in percentage) between the optimal service policy and the critical first service policy, for a range of values for the "softer" parameters, when each is used with the optimal threshold appointment policy (i.e., the one that maximizes (4)). Note that the application of this heuristic leads to a loss of only 2.5% in operational profits using the base case (highlighted) and that, in general, Table 2 indicates that the critical first approach performs quite well for a wide range of parameters. Note also that all cells in Table 2 for which the critical first heuristic is optimal correspond to parameter sets such that: $r_s + w_s + \pi_s > r_n + w_n + \pi_n$ and $w_s > w_n$, confirming the optimality of the inpatients first policy as indicated by Proposition 3. The remaining cells in Table 2 correspond to the case where $r_s + w_s + \pi_s < r_n + w_n + \pi_n$ and $w_s > w_n$, i.e., the case where the inpatient class has critical status, but the critical first policy is not

necessarily optimal. Another important observation is that the maximum performance gap of 7.7% is achieved for one of these cases where the "status level" of outpatients, $r_s + w_s + \pi_s = 1120$, is also quite close to that of inpatients, $r_n + w_n + \pi_n = 1200$. It seems plausible that for such parameter sets the rough-cut critical first approach may need to be replaced by a more refined alternative. Such an alternative is offered by the more flexible LA heuristic which, for the same set of parameters as in Table 2, replicates the performance of the critical first approach in all but three cases (out of 81). In these three cases,the LA heuristic performs significantly better: 0.8% vs 3.0% for $r_n = 200$, $\pi_n = 1000$, $\pi_s = 100$, $w_s = 10$, 0.3% vs. 5.2% for $r_n = 200$, $\pi_n = 1000$, $\pi_s = 100$, $w_s = 15$, and 0.0% vs. 7.7% for $r_n = 200$, $\pi_n = 1000$, $\pi_s = 100$, $w_s = 20$. Note that all 3 parameter sets correspond to cases where inpatient and outpatient "status levels" are close without one class having absolute service priority over the other.

The advantage of the LA policy over the critical first approach for such cases is illustrated in Figure 3, where we vary the values of π_s and π_n, while keeping the rest of problem parameters as in the base case. Figure 3a compares the performance of the two heuristics in three distinct regions of parameter space. In the top region $r_s + w_s + \pi_s > r_n + w_n + \pi_n$ and so from Proposition 3 and (16), the LA policy reduces to the critical (outpatients) first heuristic and is optimal. In the bottom region, which includes the base case and where $r_s + w_s + \pi_s < r_n + w_n + \pi_n - (N-1)(w_s - w_n)$, the LA and critical first heuristics coincide and their performance is suboptimal. Finally, in the middle band, where the status levels of two patient classes are close ($r_n + w_n + \pi_n \geq r_s + w_s + \pi_s \geq r_n + w_n + \pi_n - (N-1)(w_s - w_n)$), the LA heuristic dominates the critical first approach. The relative performance of the two heuristics is shown in Figure 3b, in which π_n is fixed at the base value of 2000 and π_s is varied. Note that as the ratio $(r_s + w_s + \pi_s)/(r_n + w_n + \pi_n)$ approaches 1, the LA heuristic, unlike critical first, approaches optimality in a monotone fashion. Such stable near-optimal performance of the LA policy over a wide range of problem parameters makes

it a preferred choice for a heuristic service policy. As explained in the previous section, the performance of the LA heuristic should be even better for higher values of the probabilities of patient arrivals p_e, p_n and p_s since as the level of congestion in the service system grows, the assumption underlying the linear approximation becomes increasingly accurate.

7.3 Heuristic Appointment Policies

Since unlike the tactical case, we have no analytical results for the optimal appointment policy, it is particularly important to explore the performance of heuristic appointment policies. We first restrict our attention to threshold policies for two reasons: (1) these are the simplest and the most common in practice; and (2) we can identify the optimal threshold policy by simple enumeration and hence quantify the performance of a heuristic threshold policy. At the end of this section, we show that threshold policies are not necessarily optimal.

We consider three heuristic threshold policies. The first one, which we call the "fill all slots" (FAS) policy, is often used in practice by many facilities attempting to maximize outpatient revenues. The FAS policy allocates all appointment slots to outpatients, i.e., $a^*_{\text{FAS}} = N$.

The second heuristic appointment policy, which we call "balanced", attempts to allocate the capacity to match the expected demand for each patient class. Under this policy, the appointment threshold a^*_{B} is selected so that the number of unscheduled service slots will be equal to the expected number of non-scheduled patient arrivals during the day. Accounting for all possibilities, we get

$$a^*_{\text{B}} = \begin{cases} 0, & \text{for } \frac{N(1-p_n-p_e)}{p_s} \leq 0, \\ \frac{N(1-p_n-p_e)}{p_s}, & \text{for } 0 < \frac{N(1-p_n-p_e)}{p_s} \leq N, \\ N, & \text{for } \frac{N(1-p_n-p_e)}{p_s} > N. \end{cases} \tag{17}$$

The third heuristic appointment policy, which we call "newsvendor", attempts to achieve the most profitable allocation of scheduled and non-scheduled examination slots. Under this policy, the appointment threshold a^*_{N} is selected by disregarding the waiting costs w_n and w_s and deriving an approximation for the expected daily profits achieved under the threshold a^*. The derivation details are presented in the Appendix.

The relative performances of the FAS, balanced and newsvendor heuristics are presented in Tables 3a, 3b, and 3c, respectively. For each of these appointment policies we selected the best "matching" tactical policy by solving the dynamic program (1). Note that since the patient demand probabilities remain fixed in Table 3, the "balanced" appointment threshold, a^*_{B}, is identical for all parameter combinations and equal to 11. a^*_{B} is also the lower bound for the optimal threshold in all of the cases we studied. (Of course, $a^*_{\mathrm{FAS}} = N$ is an upper bound on the optimal threshold). Thus, the balanced heuristic allows too few outpatients into service while the FAS heuristic often allows too many.

These results indicate that the FAS policy is surprisingly good over a wide range of parameter settings, and for the base case values, it is only 4.1% off the optimal performance. As expected, Table 3a shows that the FAS performance decreases as the end-of-day penalty cost for each patient type increases and is generally bad when both are high. It's performance also deteriorates as the outpatient waiting cost, w_s, increases. In contrast, Table 3b shows that the performance of the balanced heuristic improves as the end-of-day penalty for outpatients, π_s, increases, and also as waiting costs increase. Table 3b demonstrates that the balanced heuristic generally performs worse than the FAS policy (in 51 out of 81 cases we studied) and, in particular, for the base case parameters, the balanced heuristic is 9.2% off the optimal. So for the majority of cases we studied, deviations away from the optimal threshold in the direction of limiting outpatients are penalized more than similar deviations in the direction of allowing too many outpatients.

Finally, Table 3c indicates that in a majority of problem cases we studied (72 out of 81), the newsvendor heuristic replicates the performance of the FAS heuristic. In the 9 remaining cases ($r_n = 200$, $\pi_n = 1000$, $\pi_s = 200, 300$ and $r_n = 800$, $\pi_n = 1000$, $\pi_s = 300$), the newsvendor heuristic replicates the balanced one, performing worse than FAS in 6 cases and better in 3 cases. So in most cases (75 out of 81), the more sophisticated newsvendor heuristic "picks" the better of the FAS and the balanced heuristic, but overall, does not improve upon FAS performance over a wide parameter range.

We also evaluated the performance of the combination of the LA heuristic proposed in the last subsection with each of the appointment heuristics. Though our qualitative observations regarding the relative performance of the three appointment heuristics in Table 3 remain valid, the use of the LA policy, as expected, results in lower profits: the performance of the FAS and newsvendor appointment policies for our base case declines to 6.6%, while the performance of the balanced heuristic for the base case is now 11.6%. Table 4 ranks the performance of these three capacity management heuristics over the range of problem parameters we studied. The FAS-LA combination provides the best performance in 56 (out of 81) cases studied, while the Newsvendor-LA policy is best in 53 cases. The balanced appointment policy in combination with the LA capacity allocation is best in 30 cases. It's important to note that that though the FAS-LA pairing has the highest percentage of "hits", the balanced heuristic is the best performer in cases when the end-of-day penalties for both patient classes are high.

For our base case, the FAS-LA policy is the best capacity management recommendation among the heuristic policies we consider. Note that for the base case, the LA capacity allocation reduces to the "inpatients first" policy. The strong performance of this somewhat counterintuitive policy is a consequence, among other factors, of the relatively low probability of a patient arrival during any service slot with no scheduled outpatient. However, consistent with our observation above, if the value of π_s in the base case is increased

from 100 to 200, the balanced heuristic outperforms the FAS and newsvendor heuristics. This example underscores the high degree of sensitivity of the performance of a particular appointment scheme to changes in the penalty parameters.

For our base case, the FAS-LA policy is the best of the capacity management heuristics we considered. Note that for this case, the LA policy reduces to the "inpatients first" policy. However, consistent with our observation above, if the value of π_s in the base case is increased from 100 to 200, the balanced heuristic outperforms the FAS and newsvendor heuristics. This example underscores the high degree of sensitivity of the performance of a particular appointment scheme to changes in the penalty parameters. It is also important to note that our evaluation of these heuristics has been based solely on profitability: for the base case the optimal expected profit is \$8752, while the expected profits under the FAS-LA and balanced-LA policies are \$8393 and \$7947, respectively. Yet, in many facilities, it may be important to consider patient service as well, particularly for scheduled patients. In order to understand the impact of these policies on service, we used a simulation (see below) to estimate the expected number of outpatients left unserved at the end of the operating day. Our results indicate that for the base case, the expected number of unserved outpatients using the optimal policy is 2.6, but increases to 6.6 using the FAS-LA heuristic, and goes down to 0.6 under the balanced-LA policy. So while the FAS-LA heuristic does a good job in terms of achieving near-optimal expected profits, its "service" performance is significantly worse than that of the optimal policy. On the other hand, the balanced-LA heuristic exhibits superior service performance and may be a desirable alternative to the optimal policy in more patient-oriented service environments.

Until this point, we have only considered threshold appointment policies. However, there exist appointment schemes that can result in better performance than threshold policies under certain conditions. An example of such an appointment scheme is a policy under which regularly spaced examination slots are left unscheduled in order to provide

slack uniformly over the day for emergencies and unscheduled inpatients. In general, it seems reasonable to leave approximately $(p_e + p_n)\,N$ equally spaced slots open to accommodate uniformly occurring emergency and inpatient demand. For the base case, with $p_e = 0.1$ and $p_n = 0.4$, this results in a "fill alternative slots" (ALT) policy, with every odd appointment slot being reserved for outpatients. Table 5 compares the expected profits of the best threshold appointment scheme with that of the ALT policy (each appointment policy was combined with the best matching tactical policy) for the following set of problem parameters: p_e=0.1, r_s=1000, p_s=0.84, p_n=0.4, π_n=2000, π_s=100. Note that in this data set the largest values of waiting costs for both inpatient and outpatient delays are relatively high. Clearly, this choice of parameters tilts the balance of performance in the direction of the ALT policy which intentionally leaves empty slots to deal with unscheduled demand. We find that while for the base case, the best threshold policy outperforms the ALT policy by about 20.8%, high waiting costs for both outpatients and inpatients make the performance of threshold policies much worse than that of the ALT policy. It is interesting to note that even when the waiting cost for outpatients is extremely high, the threshold policy outperforms the ALT policy whenever $w_n = 0$.

While the threshold approach to appointment management may not always be optimal, we conjecture that its performance is either optimal or very near optimal for the realistic problem parameters in Table 1. Of course, the verification of this conjecture even for a given set of problem parameters for $N = 20$ service slots would require a substantial computational effort resulting from the necessity to check all 2^N possible appointment schemes.

7.4 The Impact of Stochastic Service Durations

A key assumption of our capacity management model (1)-(3) is that examination times are fixed and equal to the length of each service slot. This assumption allows for a sharp characterization of the optimal policies and may be justified in cases when there is little variablility in the duration of diagnostic exams. However, in some facilities, the same diagnostic capacity is used for many different types of exams and/or there is considerable variability associated with each particular exam type.

To understand the impact of service time variability, we constructed a simulation of the actual MRI facility we studied based on actual service times and using the base case parameters. Figure 4 shows the histogram of these service times recorded over a 3 week period along with the best fit to this distribution. Though the mean duration of the MRI exams during this time (48 minutes) is very close to the length of the facility's fixed service slot (45 minutes) allocated by the facility for each exam, the standard deviation of 26 mins is significant. This variability is largely due to the broad mix of MRI "studies" that fall into six major categories ranging from non-contrast brain studies with a mean of 32 minutes to abdomen studies with a mean of 61 minutes. Using the simulation model, we compared the performance of two sets of policies: the optimal policy; and the FAS appointment policy paired with the LA service policy (equivalent here to "inpatients first"), identified above as the best combination of heuristics for this case.

As in the actual MRI facility, the simulation splits the operating day into 20 slots of 45 minutes each. In each slot, the incidence of inpatient and emergency arrivals are modeled as independent Bernoulli random variables with parameters $p_n = 0.4$ and $p_e = 0.1$, respectively; and the actual arrival times of inpatient and emergency patients within each slot are simulated as independent uniform random variables. Outpatients are assumed to arrive right before the beginning of the slot for which they were scheduled. Diagnostic

service times are generated from the best fit distribution to the actual service times - Weibull with the location parameter of 8.20, the scale parameter of 44.15 and the shape parameter of 1.54. At the end of each service i, the numbers of inpatients (n) and outpatients (s) waiting for service are counted and, if there are no emergency patients waiting, the admission decision is made according to the service policy that is being simulated[1] and the respective incremental revenue contribution is computed. At the same time, the waiting costs are determined based on the numbers of each patient class left waiting. At the end of operating day, (when the time equal to 20 service slots has elapsed), penalty costs are assessed using the numbers of patients of each type left unserved.

We ran the simulation for 50,000 days for each policy and obtained the following results: the mean profit under the optimal policy was \$6558, with a mean standard error (MSE) of \$15 (as compared to \$8752 earned in the case of fixed service times, as indicated in Table 5), while the mean profit under the FAS-LA policy was \$6431 with MSE of \$17 (as compared to \$8174 in the case of fixed service times). These results prompt three observations. First, and not surprisingly, the profits generated by both policies are significantly affected by the uncertainty in the durations of diagnostic exams: mean profits generated by each policy are 20%-25% less than those expected under fixed service times for this set of parameters. Also, the relative ranking of two policies is unchanged; the optimal policy remains a better performer. Finally, the relative performance gap between the two policies is significantly reduced: from 6.6% in the case of fixed service times to 1.9% when real service times are used. This performance gap reduction strengthens the argument of using the simple

[1]Note that the policy optimal for our base case model is describing the capacity allocation decisions after the completions of i-th service with $i = 1, ..., 20$. When the service completion times are simulated, it is possible that more than 20 patients can be served during a particular simulation run. In those cases, we have used "critical first" admission policy for patients who are served after the completion of 20-th diagnostic service. This necessary adjustment is not likely to worsen the performance of the policy since, at the end of day, the absolute service priority for inpatients is likely to be optimal.

FAS-LA heuristic as a capacity management policy in the facility we studied as well as in other MRI facilities with similar parameter values.

8 Discussion

As more and more attention is being given to controlling the ever-escalating costs of health care, it becomes increasingly important to identify ways to use health resources more efficiently. Diagnostic imaging facilities are part of the larger category of health care technology which has been identified as being one of the leading engines of increasing health costs. Yet, due to the complexity of dealing with the competing demands for these machines and a lack of understanding of all the costs (as well as the stochastic dynamics) involved, MRI and other imaging facilities are often managed in ways which result in both under-utilization of a very expensive resource and long patient delays.

The work described in this paper is the first attempt to gain insights into the management of an MRI facility. Our analytical and numerical results strongly indicate that in many cases, performance of such facilities can be significantly improved, often by the use of simple heuristic policies. Though our work doesn't reveal any single overall capacity management policy that works well under virtually all circumstances, it does provide substantial guidance on how such policies should be chosen. In particular, our conclusions based on our study of an actual MRI facility generally support the use of the LA heuristic with a fill-all-slots (FAS) appointment policy for a wide range of realistic parameter values. However, this work also highlights the sensitivity of these policies to several parameters, particularly the end-of-day-penalties, which are often, particularly in the case of outpatients, very hard to estimate. Given the less than 2% (or about $32,000 per year based on 260 operating days) profit gap between the optimal and heuristic policies indicated by our simulation of an actual MRI facility, as well as the relative simplicity of the heuristic

rules compared to the switching curve, we conclude that it would be more beneficial for managers to invest time and effort in improving the accuracy of the most critical data, rather than in an information system to implement the dynamic program algorithm.

The model we propose as well as the results we obtain represent a promising step in furthering the understanding of the management of diagnostic medical systems. Of course, more work remains to be done. One potential complication, which we are currently studying, is that the pattern of demand and the examination schedule may change from day to day during the week. Thus, the time horizon for the analysis of capacity management decisions might more appropriately be extended to a week. This extension would also allow for a more detailed and accurate analysis of inpatient demand, which is sometimes pushed over from one day to the next - a characteristic we couldn't incorporate into our single day framework. Other issues of importance include capacity allocation among multiple magnets in the same facility, and strategies to address the issue of outpatient cancellations. These indicate the need for further research in the area of management of medical diagnostic capacity.

References

Altman, E., Jimenez, T., and G. Koole, "On optimal call admission control in a resource-sharing system", *IEEE Trans. on Communications* **49** (2001), 1659–1668.

Bailey, N. T. J., "A Study of Queues and Appointment Systems in Hospital Outpatient Departments, with Special Reference to Waiting Times", *Journal of the Royal Statistical Society*, Series B, **14** (1952), 185-199.

Belobaba, P. P. "Application of a Probabilistic Decision Model to Airline Seat Inventory Control," *Operations Research*, **37** (1989), 183-197.

Bitran, G. R., and S. M. Gilbert, "Managing Hotel Reservations with Uncertain Arrivals", *Operations Research*, **44** (1996), 35-49.

Carr, S., and I. Duenyas, "Optimal Admission Control and Sequencing in a Make-to-Stock/Make-to-Order Production System", *Operations Research*, **48** (2000), 709-720.

Carrioza, E., Conde, E., and M. Munoz-Marquez, "Admission Policies in Loss Queueing Models with Heterogeneous Arrivals", *Management Science*, 44 (1998), 311-320.

Carrol, W., and R. Grimes, "Evolutionary Change in Product Management: Experiences in the Car Rental Industry", *Interfaces*, **25** (1995), 84.

Fries, B. E., and V. P. Marathe, "Determination of Optimal Variable-Sized Multiple-Block Appointment Systems", *Operations Research*, **29** (1981), 324-345.

Foschini, G. J., and B. Gopinath, "Sharing Memory Optimally", *IEEE Trans. on Comm.*, **31** (1983), 352-360.

Gans, N., Koole, G., and A. Mandelbaum, "Telephone Call Centers: Tutorial, Review, and Research Prospects," *Manufacturing & Service Operations Management*, **5** (2003), 79-141.

Geraghty, M., and E. Johnson, "Revenue Management Saves National Car Rental", *Interfaces*, **27** (1997), 107.

Glasserman, P, and D. Yao, "Monotone Structure in Discrete-Event Systems", John Wiley & Sons, 1994, Chapters 4 and 6.

Ha, A., "Optimal Dynamic Scheduling Policy for a Make-to-Stock Production System", *Operations Research*, **45** (1997), 42-53.

Ho, C.-J., and H.-S. Lau, "Minimizing Total Cost in Scheduling Outpatient Appointments", *Management Science*, **38** (1992) 1750-1765.

Liberman, V., and U. Yechiali, "On the Hotel Overbooking Problem: An Inventory System with Stochastic Cancellations", *Management Science*, **24** (1978), 1117-1126.

Örmeci, L., Burnetas, A., and J. van der Wal, "Admission Policies to a Two-class Loss System," *Stochastic Models* **17** (2001), 513–539.

Pinedo, M., "Scheduling: Theory, Algorithms and Systems", 2ed., Prentice Hall (2001).

Ross, K. W., and D. H. K. Tsang, "The Stochastic Knapsack Problem", *IEEE Trans. on Comm.*, **37** (1989), 740-747.

Soriano, A., "Comparison of Two Scheduling Systems", *Operations Research*, **14** (1966), 388-397.

Topkis, D., "Minimizing a Submodular Function on a Lattice", *Operations Research*, **26** (1978), 305-321.

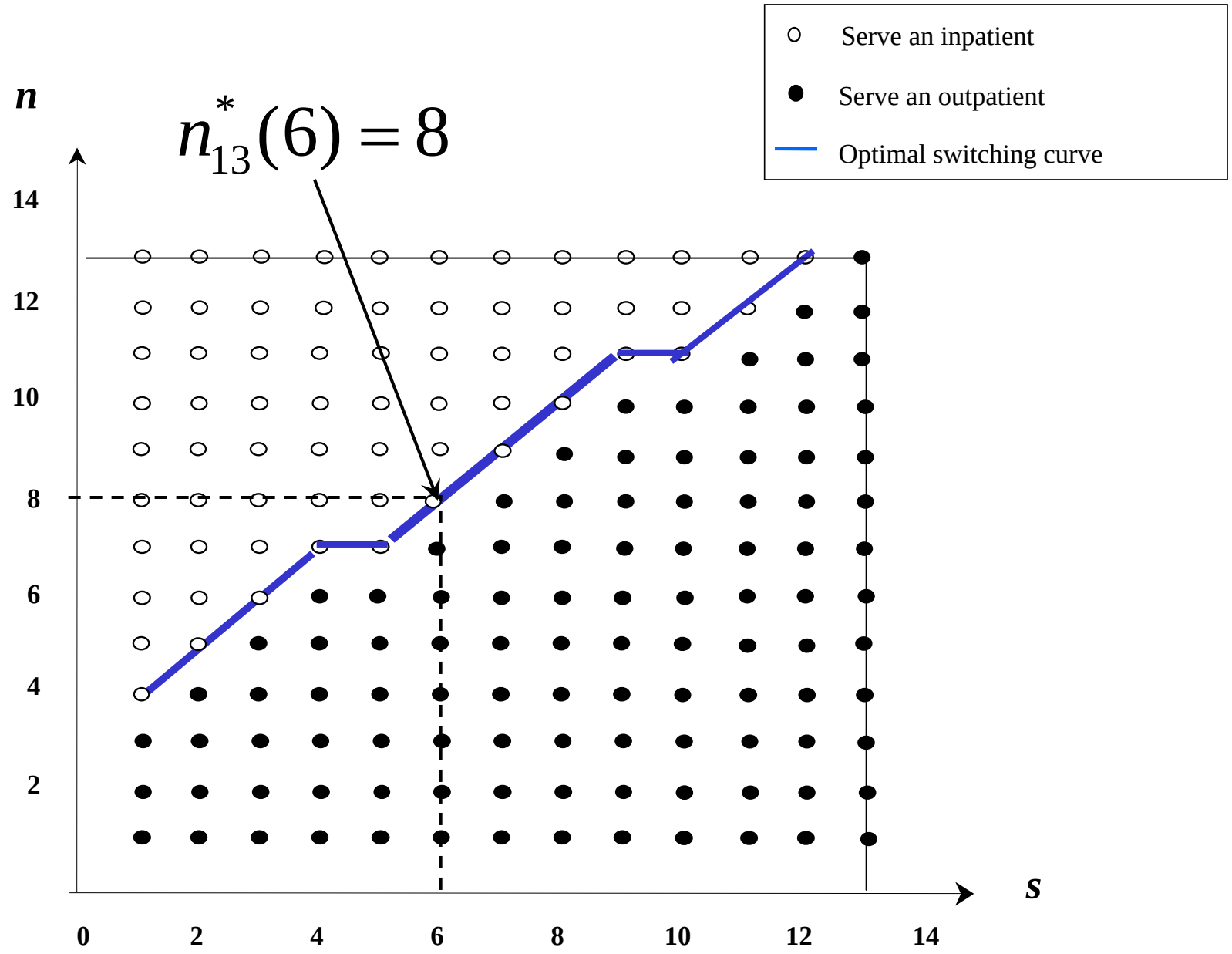

Fig. 1: The switching curve capacity allocation at slot i=13 (p_s=0.84, p_e=0.1, p_n=0.4, r_s=1000, r_n=200, w_s=15, w_n=0, π_s=400, π_n=500, N=20, $f(n,s) = -s^2\pi_s - n^2\pi_n$).

Parameter	p_n	p_e	p_s	r_s	r_n	w_s	w_n	π_s	π_n
Value	0.4	0.1	0.84	\$1000	\$200	\$15	\$0	\$100	\$2000

Table 1. Basic set of problem parameters.

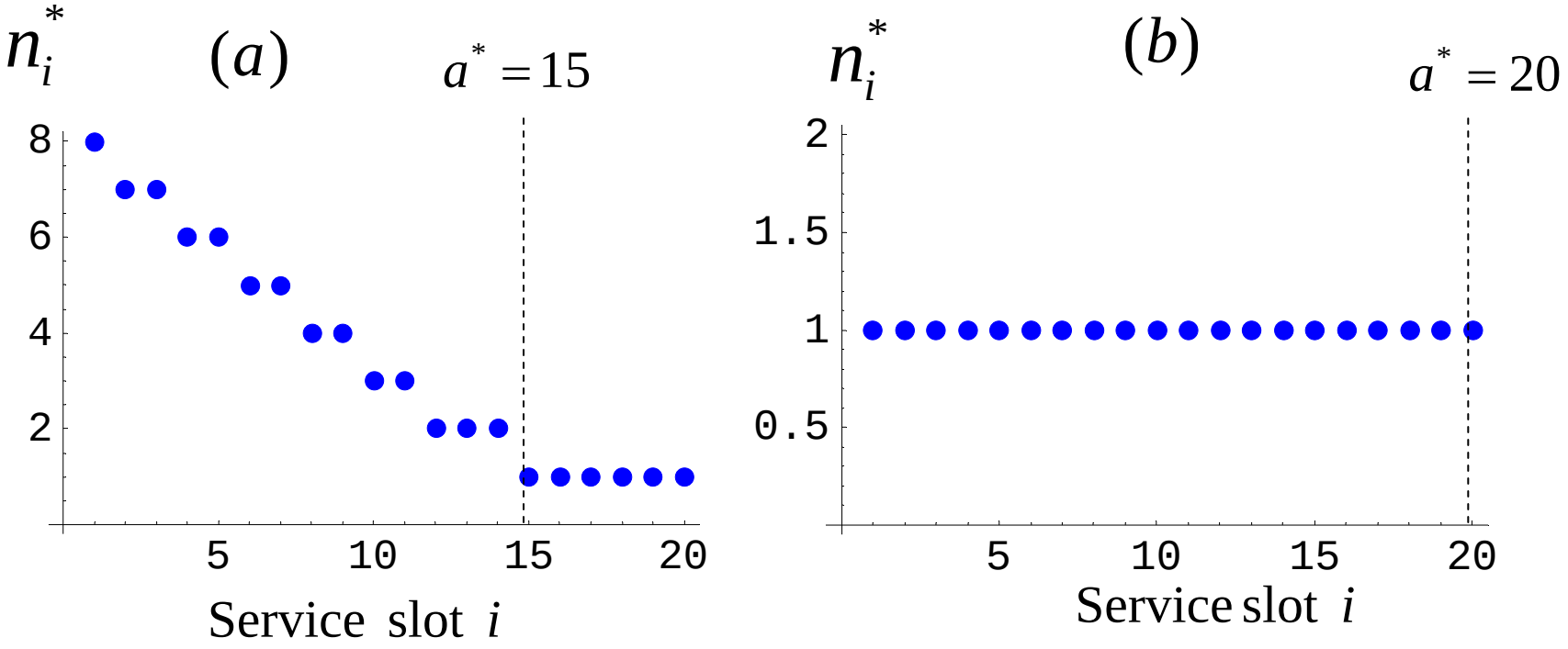

Fig. 2: The optimal switching indices and the optimal appointment threshold for the problem with parameters from Table 1 (a) and for π_s=w_s=0 (b) (the remaining parameters are taken from Table 1).

			w_s=10			w_s=15			w_s=20	
r_n	π_n	π_s=100	π_s=200	π_s=300	π_s=100	π_s=200	π_s=300	π_s=100	π_s=200	π_s=300
0	500	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	1000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	2000	1.8	2.0	2.1	3.2	3.5	3.8	4.9	5.3	5.7
200	500	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	1000	3.0	0.0	0.0	5.2	0.0	0.0	7.7	0.0	0.0
	2000	1.4	1.5	1.6	**2.5**	2.7	2.8	3.8	4.0	4.2
800	500	1.6	2.0	0.0	2.8	3.5	0.0	4.1	5.1	0.0
	1000	1.0	1.1	1.2	1.9	2.0	2.2	2.8	3.0	3.2
	2000	0.8	0.8	0.8	1.4	1.5	1.5	2.1	2.2	2.3

Table 2: Relative profit gaps (in %) between the optimal tactical policy and "critical first" policy (under optimal threshold appointment policy). p_e=0.1, r_s=1000, p_s=0.84, p_n=0.4, w_n=0.

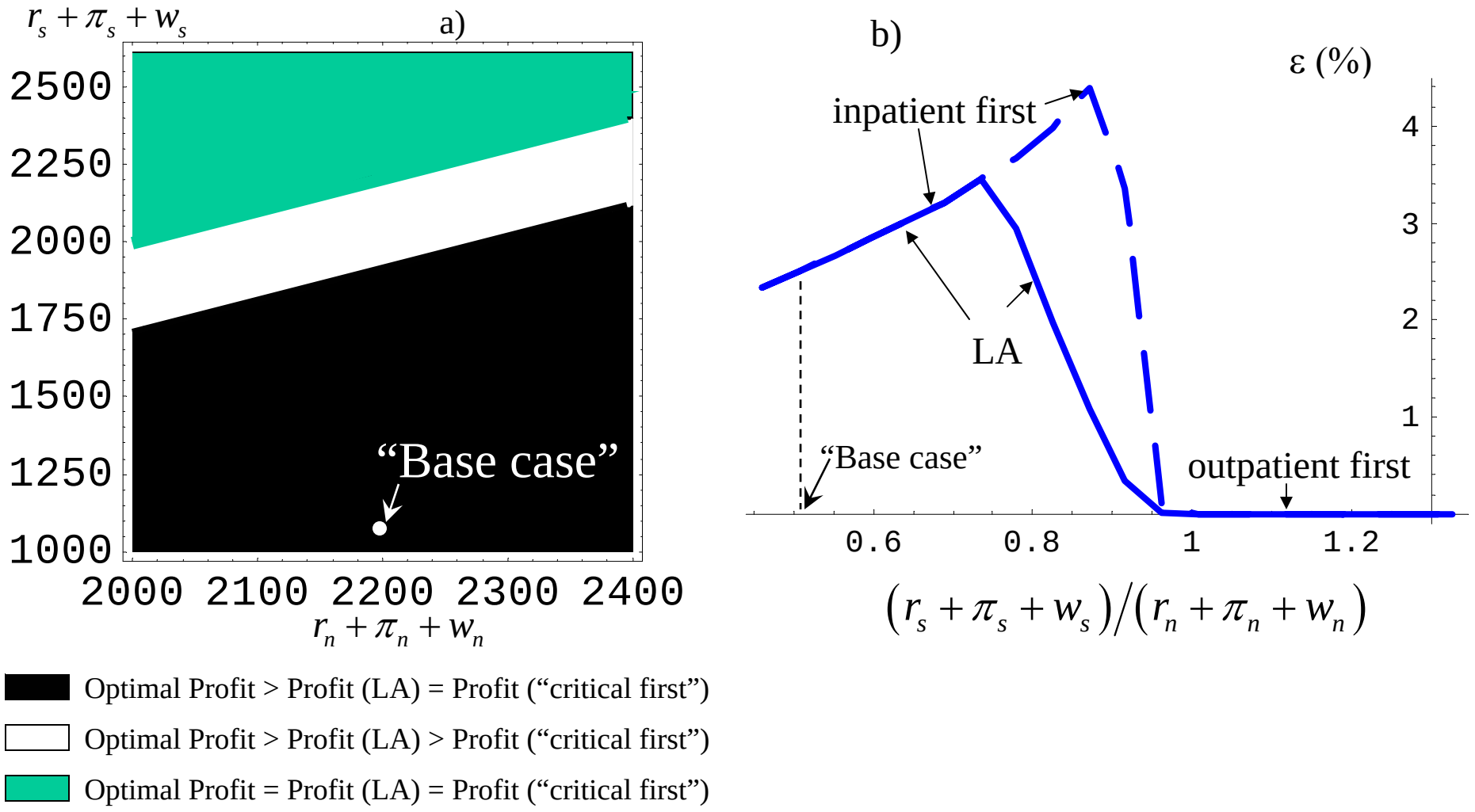

Fig. 3: a) Ranking the LA and "critical first" heuristics (under optimal appointment policy) for different penalty cost values π_n and π_s, b) relative performance (in %) of two heuristics (under optimal appointment policy) as compared to the optimal tactical policy for different values of the outpatient end-of-day penalty cost π_s. The rest of problem parameters are taken from Table 1.

a			w_s=10			w_s=15			w_s=20	
r_n	π_n	π_s=100	π_s=200	π_s=300	π_s=100	π_s=200	π_s=300	π_s=100	π_s=200	π_s=300
0	500	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	1000	0.6	1.3	2.1	0.7	1.4	2.1	0.8	1.5	2.3
	2000	4.0	9.7	16.6	4.9	11.0	18.1	5.9	12.4	19.7
200	500	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.1
	1000	3.1	6.4	7.1	3.8	6.5	7.3	4.4	6.7	7.5
	2000	3.3	8.0	13.6	**4.1**	9.1	14.9	4.9	10.3	16.2
800	500	2.1	5.0	7.3	2.5	5.5	7.5	3.0	6.0	7.6
	1000	2.1	5.1	8.6	2.6	5.7	9.3	3.1	6.4	10.0
	2000	2.2	5.3	8.9	2.7	6.0	9.7	3.2	6.7	10.5

b			w_s=10			w_s=15			w_s=20	
r_n	π_n	π_s=100	π_s=200	π_s=300	π_s=100	π_s=200	π_s=300	π_s=100	π_s=200	π_s=300
0	500	34.2	33.8	33.4	34.0	33.7	33.3	33.9	33.6	33.1
	1000	16.7	16.6	16.5	16.7	16.5	16.4	16.4	16.3	16.2
	2000	12.1	9.7	8.0	11.1	8.8	7.0	9.9	7.9	6.1
200	500	24.6	24.1	23.7	24.4	24.0	23.6	24.3	23.9	23.5
	1000	9.8	9.2	9.2	9.3	9.1	9.1	9.1	9.0	9.0
	2000	10.1	8.0	6.5	**9.2**	7.2	5.7	8.2	6.5	4.9
800	500	6.4	5.2	5.0	5.9	4.9	4.9	5.4	4.9	4.8
	1000	6.5	5.1	4.1	5.8	4.6	3.6	5.2	4.1	3.1
	2000	6.7	5.3	4.3	6.1	4.7	3.7	5.4	4.2	3.1

Tables 3ab: Relative performance (in %) of the "fill all slots" (**a**) and "balanced" (**b**) policies as compared to the optimal appointment policy. p_e=0.1, r_s=1000, p_s=0.84, p_n=0.4, w_n=0. Highlighted cells correspond to the parameters from Table 1.

c			w_s=10			w_s=15			w_s=20	
r_n	π_n	π_s=100	π_s=200	π_s=300	π_s=100	π_s=200	π_s=300	π_s=100	π_s=200	π_s=300
0	500	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	1000	0.6	1.3	2.1	0.7	1.4	2.1	0.8	1.5	2.3
	2000	4.0	9.7	16.6	4.9	11.0	18.1	5.9	12.4	19.7
200	500	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0
	1000	3.1	9.2	9.2	3.8	9.1	9.1	4.3	9.0	9.0
	2000	3.3	8.0	13.6	**4.1**	9.1	14.9	4.9	10.2	16.2
800	500	2.1	4.9	4.9	2.5	5.5	4.9	3.0	6.0	4.8
	1000	2.1	5.1	8.6	2.6	5.7	9.3	3.1	6.4	10.0
	2000	2.2	5.3	8.9	2.7	6.0	9.7	3.2	6.7	10.5

Table 3c: Relative performance (in %) of the "newsvendor" policy as compared to the optimal appointment policy. p_e=0.1, r_s=1000, p_s=0.84, p_n=0.4, w_n=0. Highlighted cells correspond to the parameters from Table 1.

			w_s=10			w_s=15			w_s=20	
r_n	π_n	π_s=100	π_s=200	π_s=300	π_s=100	π_s=200	π_s=300	π_s=100	π_s=200	π_s=300
0	500	F,N	F,N	F,N	F,N	F,N	F,N	F,N	F,N	F,N
	1000	F,N	F,N	F,N	F,N	F,N	F,N	F,N	F,N	F,N
	2000	F,N	F,N,B	B	F,N	B	B	F,N	B	B
200	500	F,N	F,N	F,N	F,N	F,N	F,N	F,N	F,N	F,N
	1000	F,N	F	F	F,N	F	F	F,N	F	F
	2000	F,N	F,N,B	B	**F,N**	B	B	F,N	B	B
800	500	F,N	F,N,B	N,B	F,N	B	N,B	F,N	B	N,B
	1000	F,N	F,N,B	B	F,N	B	B	F,N	B	B
	2000	F,N	F,N,B	B	F,N	B	B	F,N	B	B

Table 4: Best performing threshold heuristic in combination with LA tactical policy: F="Fill All Slots", B="Balanced", N="Newsvendor". Same parameters as in Table 3.

	w_s=15			w_s=100			w_s=300		
	w_n=0	100	300	0	100	300	0	100	300
Best Threshold	8752	8390	8104	7304	4612	4327	6521	3757	1251
"Fill Alt. Slots"	6935	6713	6427	6514	5512	5227	5981	4655	2402

Table 5: Profit values (in $) under "fill alternative slots" appointment policy and the best threshold policy. p_e=0.1, r_s=1000, p_s=0.84, p_n=0.4, r_n=200, π_n=2000, π_s=100. Highlighted cells correspond to parameters from Table 1.

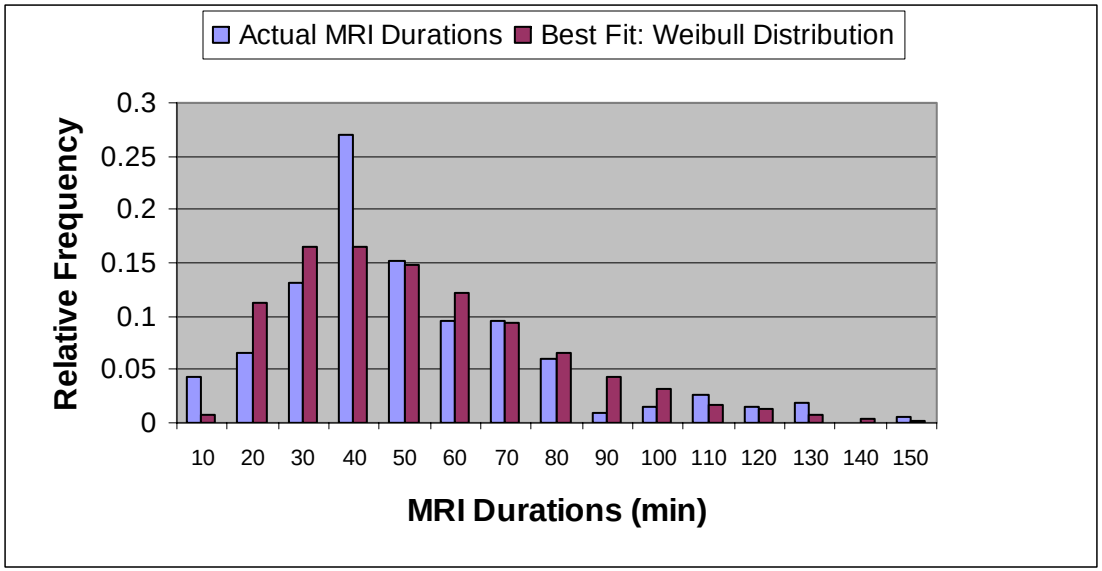

Fig. 4: Histogram of the actual durations of the MRI exams over a 3 week period. The best fit (Weibull with the location parameter of 8.2, the scale parameter of 44.15 and the shape parameter of 1.54) is also shown.