Revenue Management for a Primary-Care Clinic in the Presence of Patient Choice

Diwakar Gupta
Graduate Program in Industrial and Systems Engineering, Department of Mechanical Engineering, University of Minnesota, Minneapolis, Minnesota 55455, guptad@me.umn.edu

Lei Wang
SmartOps Corporation, Pittsburgh, Pennsylvania 15212, lwang@smartops.com

In addition to having uncertain patient arrivals, primary-care clinics also face uncertainty arising from patient choices. Patients have different perceptions of the acuity of their need, different time-of-day preferences, as well as different degrees of loyalty toward their designated primary-care provider (PCP). Advanced access systems are designed to reduce wait and increase satisfaction by allowing patients to choose either a same-day or a scheduled future appointment. However, the clinic must carefully manage patients’ access to physicians’ slots to balance the needs of those who book in advance and those who require a same-day appointment. On the one hand, scheduling too many appointments in advance can lead to capacity shortages when same-day requests arrive. On the other hand, scheduling too few appointments increases patients’ wait time, patient-PCP mismatch, and the possibility of clinic slots going unused.

The capacity management problem facing the clinic is to decide which appointment requests to accept to maximize revenue. We develop a Markov decision process model for the appointment-booking problem in which the patients’ choice behavior is modeled explicitly. When the clinic is served by a single physician, we prove that the optimal policy is a threshold-type policy as long as the choice probabilities satisfy a weak condition. For a multiple-doctor clinic, we partially characterize the structure of the optimal policy. We propose several heuristics and an upper bound. Numerical tests show that the two heuristics based on the partial characterization of the optimal policy are quite accurate. We also study the effect on the clinic’s optimal profit of patients’ loyalty to their PCPs, total clinic load, and load imbalance among physicians.

Subject classifications: health-care operations; patient choice; capacity reservation; revenue management; Markov decision processes.

Area of review: Manufacturing, Service, and Supply Chain Operations.

History: Received August 2005; revisions received May 2006, November 2006; accepted November 2006.

1. Introduction

Many service capacity planning models in the operations research literature account for interarrival and processing time variability. However, a typical primary-care clinic is faced with additional variability stemming from patient choices. For example, some patients will accept any open appointment on the day they call. Others prefer a convenient time and do not mind waiting until a later day. Patients have different degrees of loyalty toward their designated primary-care provider, or PCP. Some book appointments only with their PCP, even when this leads to inconvenient appointment times or extra waiting, whereas others switch easily to alternate providers.

How should a clinic manage access via the appointment-booking process? Responses can range from low cost-low customization at the one end, to high cost-high customization at the other. For example, lying at one end of the spectrum, Minute Clinics (www.minuteclinic.com) reduce uncertainty by catering to only a limited number of routine diagnoses. Appointments are eliminated. Patients can walk in and expect short waiting times. Physicians and nurse practitioners are interchangeable; that is, patients do not have an expectation of consulting the same provider at each visit. At the other end of the spectrum, the Park Nicollet Clinic in Minneapolis accommodates demand and patient choice variability through enhanced services at a premium price (Haeg 2002). Patients enroll with a physician and pay a fixed fee up front. In return, they get fast personalized service from their PCP, who also bills them (or their insurers) for each visit. Providers are thus paid more for their efforts.

A majority of the clinics lie somewhere in-between the two extremes described above. They serve a mix of acute- and chronic-care patients and do not charge more for same-day service. Clinic directors recognize that patients’ perceptions of urgency of need are an important factor in determining their overall satisfaction with the timeliness of access. Put differently, patients who perceive their need to be urgent are more likely to be dissatisfied if the clinic cannot offer an appointment on the day of call. Until recently, clinics employed a triage nurse to assess the urgency of medical need whenever a caller requested a same-day appointment. If the patient’s need was judged urgent, then (s)he was offered one of several slots reserved each day for...
urgent requests. Otherwise, the patient was offered a future appointment. This allowed clinics to reserve capacity for acute-care patients based on historical averages.

In recent years, many clinics have implemented an advanced or open-access system, credited to Murray and Tantau (1999, 2000), according to which patients choose between an available same-day slot or a future appointment slot at a convenient time. The impetus for this change comes from the desire to make clinic practices more patient focused, to alleviate persistent complaints from patients who could not obtain a timely appointment, and to gain competitive advantage. This approach also eliminates the need for a triage nurse, but introduces new challenges for clinic directors who need to decide how many slots to leave open for patients requesting a same-day appointment. The demand for same-day service is now a function of patient choice.

Same-day appointments are a part of a more comprehensive system of managing clinic capacity known as the advanced-access system; see Murray and Berwick (2003) for details. In this paper, we focus only on the problem of allocating available slots to appointment requests to maximize clinic revenue. The presence of patient choices makes this problem difficult. We begin with a stylized example below to highlight this point. More-realistic examples are presented later in the paper.

Suppose that a clinic is served by one doctor who has four available slots on a particular day. It is known that three patients will call in advance to schedule appointments and that one patient will call at the start of that day to request a same-day appointment. This patient will accept any available slot. All appointments generate the same revenue for the clinic. If patients were willing to accept any available slot, there is no decision to be made. Because demand and supply are perfectly balanced, all requests should be accepted. This is, however, not true in presence of patient choice. To see this, consider the following plausible choice pattern. Suppose that the first caller will randomly pick either slot 1 or slot 2, with equal probability. The second caller prefers slot 2 if slot 1 is open at the time of booking (because this implies she is more likely to be seen on time by the physician), and slot 1 otherwise. The third caller will always take slot 1 if that slot is available. However, if slot 1 is unavailable, she will book one of the available slots as long as no more than 50% of the doctor’s slots are open. This behavior can be explained as follows. When the third caller’s first choice is available, nothing else matters. However, if her first choice is unavailable, other factors do matter. If more than 50% of the doctor’s slots are available, she takes it as a cue that the doctor is inexperienced and chooses not to book. Upon simulating each patient-choice realization and comparing all possible clinic actions, it is easy to confirm that the optimal action is to deny a request for slot 1, if this slot is requested by the first caller. All other requests should be accepted.

In the example above, accepting a request for slot 1 from the first caller will turn away two advance callers. In all other instances, every appointment request can be accommodated, earning the maximum revenue for the clinic. This shows that a booking-limit policy is not optimal under arbitrary patient choices. However, it is entirely possible that for certain types of choice probability models, the optimal policy has a simple structure. The goal of this paper is to identify such scenarios and to suggest implementable heuristics when no simple structure exists. We also investigate the impact of demand variability, physician workload, patient loyalty, and revenue/cost parameters on appointment decisions and clinic revenues. We shall describe our approach and main results later in this section, after discussing related literature in the next several paragraphs.

The problem of allocating service capacity among several competing customer classes has been studied in diverse applications, including airlines, hotels, and car rentals. In particular, airline revenue management (RM) has been studied particularly well; see McGill and van Ryzin (1999) and Talluri and van Ryzin (2004b) for detailed reviews. Of the various models suggested for airline RM, comparisons with the expected-marginal-seat-revenue (EMSR) model (see Belobaba 1989) help highlight the complexity of our problem scenario. In the two-fare class EMSR model, lower-fare class customers purchase tickets first. The optimal policy in this instance reserves a certain number of seats exclusively for the higher fare class customers. The lower-fare class customers are not allowed to book once a certain number of seats (equal to the booking limit) have been sold. Our problem is related, but more complicated. We have multiple physicians, which is akin to having multiple parallel flights. Furthermore, each slot (equivalently, seat) is different in the eyes of the patient and, as we explain in §2, physician/clinic revenue is smaller when patients are not matched with their PCPs (O’Hare and Corlett 2004).

Articles focusing on the allocation of medical service capacity among distinct demand streams are limited. Gerchak et al. (1996) consider the problem of reserving surgical capacity for emergency cases when the same operating rooms are also used for elective surgeries and surgery durations are random. They formulate the problem as a stochastic dynamic program and characterize the structure of the optimal policy. Patients do not choose a particular time slot for surgery; instead, they are scheduled according to a first-in-first-out (FIFO) rule. In contrast, we model patients’ choice for each physician-time slot combination. Green et al. (2006) analyze a related problem for a diagnostic medical facility that is shared by outpatients, inpatients, and emergency patients. Their analysis focuses on establishing optimal service priority rules for admitting patients into service, given an arbitrary outpatient appointment schedule. In contrast, our focus is on determining which appointment requests should be accepted by the clinic, given patient choices.
Dobson et al. (2006) study the impact on patient waiting times of using a fixed booking limit for nonurgent patients. They assume that nonurgent patients will accept any appointment on or after their preferred date. They do not model patients’ preferences for a physician or a particular time slot. Under similar modeling assumptions, Gerchak et al. (1996) have earlier shown that a booking-limit policy is not optimal. Green and Savin (2005) propose methods for determining the right panel sizes for physicians using a newsvendor-like formulation. However, they also do not model patient choice, clinic dynamics (particularly patients seeing physicians other than their PCP), and the appointment-booking process.

There are several recent papers that model customer choice explicitly. Talluri and van Ryzin (2004a) analyze a single-leg RM problem with customer choice among the open-fare classes. The airline decides which subset of fare products to offer at each point in time, and the customers’ choices are a function of the set offered. They prove certain structural properties that can simplify the calculation of an optimal policy. Zhang and Cooper (2005) consider a seat-allocation problem in which there are multiple flights between the same origin and destination, and customers choose among the open flights. The patient choice function in our setting is more complicated than both these models because patients may have preferences for both the physician and the time of the appointment slot.

In addition to the papers mentioned above, there is a stream of research in which authors assume a threshold policy and propose models for performance evaluation as well as policy parameter optimization. Specifically, the goal is to either study the impact of a given \( N_i \) or to compute the optimal value of \( N_i \), where \( N_i \) is the threshold level for class-\( i \) customers. A class-\( i \) customer is served if and only if fewer than \( N_i \) servers are occupied. Taylor and Templeton (1980) and Schack and Larson (1986) use queueing models to obtain performance measures such as average utilization and overflow rates for a given set of threshold levels. Kolesar (1970) and Esogbue and Singh (1976) focus on the problem of finding the optimal threshold levels under a linear cost structure. In our setting, owing to the presence of patient choice and patient-PCP mismatch costs, the optimal booking policy is not a threshold-type policy.

Our paper is also related to the literature on the hospital appointment-scheduling problem in outpatient settings; see Denton and Gupta (2003) for details. These papers focus on the problem of setting appointment start times for outpatients to minimize the sum of physicians’ and patients’ waiting costs, service facility idling costs, and tardiness costs with respect to the session length (usually a day). All appointment requests are known at the time of setting appointment times, and service time uncertainty is the source of model complexity. In contrast, we focus on dynamic scheduling in which appointments are booked one at a time, without knowing how many additional requests will arrive before the start of the session. Complexity comes from patient choices and random demand arrivals, but each patient can be served within the allotted time slot.

We model the clinic’s problem of choosing which appointment requests to accept for a particular workday as a Markov decision process (MDP). The formulation, presented in §2, allows us to show that when the clinic is served by a single physician, the optimal booking policy is a threshold policy under a normal form patient-choice model. (The class of normal-form models is defined in §3.) When the clinic has multiple doctors, patients may choose a more convenient time with a doctor other than their PCP. This makes the optimal policy more complicated because patient-PCP mismatch lowers physician and clinic revenues (O’Hare and Corlott 2004). We show that for each appointment request with a particular physician, there exists an appointments’ threshold beyond which it is not optimal to book additional appointments. This threshold depends on the total number of booked slots in the clinic.

We carry out a number of numerical experiments to compare the performance of straw policies, such as the first-come-first-served policy, and heuristic policies based on the partial characterization of the optimal policy. The examples show that the latter approaches achieve nearly optimal clinic revenues for each workday. In our model, when patients’ requests are denied, or they choose not to book, there is an implied penalty to the clinic, but we do not model subsequent booking requests by these patients. In reality, patients do typically try to book on other days, which makes the booking decisions dependent across time. Unfortunately, a mathematical model that optimizes accept/deny decisions over multiple booking attempts has proven to be intractable. Therefore, additional tests of the validity of the booking controls suggested by our heuristics are carried out via computer simulation model. Parameters used in this model are based on real data from a medium-sized clinic in the Twin Cities area. These experiments confirm that the proposed heuristics perform very well in the more realistic setting with multiple booking attempts. We also carry out other experiments that shine light on the effect of data values on the policy parameters and the clinic’s expected profit.

The remainder of this paper is organized as follows. In §2, we formulate the clinic’s appointments management problem. Section 3 introduces three commonly used choice models. We also specify a condition that defines the normal-form choice probabilities, and show that all three models are, in fact, normal form. Section 4 contains structural results, bounds, and heuristics. Section 5 reports analytical and numerical comparisons, as well as tests of accuracy of various heuristic policies. Insights are presented in §6, and we conclude the paper in §7.

2. Notation and Model Formulation

We focus on a primary-care clinic’s appointment management problem for a particular day, which we call “workday,” and model it as a discrete-time, finite-horizon MDP.
The clinic is served by \( n \) physicians. Physician \( i \) has \( \kappa_i \) total slots on the workday, and \( \kappa = \sum_{i=1}^{n} \kappa_i \) is the total clinic capacity. The booking horizon is the length of time between the opening of the physicians’ slots for reservations (usually no more than three months prior) and the beginning of the workday. It is divided into \( \tau \) discrete time periods, indexed by \( t = 1, 2, \ldots, \tau \), such that there is at most one appointment request in each period. Time is counted backwards, from period \( \tau \) to period 1. A physician’s panel consists of all patients for whom (s)he is the PCP. The probability that a patient in physician \( m \)’s panel calls to book an appointment in period \( i \) is \( \lambda^m_i \), where \( \lambda^m_1 \geq 0 \), and \( \lambda_i = \sum_{m=1}^{n} \lambda^m_i \leq 1 \) is the call rate in period \( i \). Call arrivals are independent across time periods.

We use boldface font to denote a matrix, character accent ‘\( \cdot \)’ to denote a vector, superscripts to denote vector components, and subscripts to denote time. Vector dimensions are not explicitly indicated, but should be evident from the context. Throughout the paper, all random variables are nonnegative and integer valued with finite means. Notation \( (\cdot)^+ \) stands for \( \max\{-, 0\} \), and \( \land b \) for \( \min\{a, b\} \).

Patients who call at \( t \geq 1 \) are called regular patients. In addition, the clinic receives calls from same-day patients that arrive at the start of the workday. We denote this arrival epoch by \( t = 0 \). Same-day patients consider their need to be of an acute nature and accept any available slot on the workday. This is consistent with appointments data we obtained from a clinic and observations based on interviewing several appointment schedulers (Gupta et al. 2006). In contrast, regular patients have a preference for a particular physician and a particular time of day. If they are unable to obtain the desired physician-slot combination, they attempt to book an appointment for a later date. Our model includes a penalty, \( \pi_t \), when a regular patient either chooses not to book, or his/her request is denied. However, multiple booking attempts are not considered because then the formulation becomes intractable due to the curse of dimensionality. We test the booking policies obtained from our formulation in simulation experiments that use data from an actual clinic and allow multiple booking attempts. The heuristics perform well. This is partly because clinics can affect the fraction of regular patients who are turned away by appropriately choosing \( \pi_t \).

Superscript \( ij \) denotes the \( j \)th slot of physician \( i \). The reservation state of physician \( i \)’s workday is denoted by \( \vec{s}_{i} = (s^{1i}, s^{2i}, \ldots, s^{ni}) \), where \( s^{ji} = 1 \) if slot \( ij \) is previously booked or unavailable, and \( s^{ji} = 0 \) otherwise. (Slots are typically marked unavailable when physicians are absent due to illness, professional development leave, or on hospital rounds.) Similarly, \( s = (\vec{s}_{1}, \vec{s}_{2}, \ldots, \vec{s}_{n}) \) denotes the reservation state of the entire clinic, and \( \mathcal{A}(s) = \{ij; s^{ji} = 0\} \) the index set of unreserved slots. \( P^i_t(s) \) is the conditional probability that a patient of physician \( m \), upon calling in period \( t \) and observing \( s \), requests an appointment for slot \( j \) of physician \( i \). We let \( ij = 00 \) denote the no-request choice and \( P_t^{0i}(s) \) the no-request probability. Note that \( P_t^{0i}(s) = 0 \) if \( ij \not\in \mathcal{A}(s) \) and \( \sum_{ij \in \mathcal{A}(s)} P_t^{ji}(s) + P_t^{0i}(s) = 1 \). In summary, for each \( m, t, \) and \( s \), patient choices are captured by the probability distribution \( (P_t^{ij}(s)) \).

In each period, the clinic makes an accept/deny decision upon receiving a caller’s request—i.e., after knowing the caller’s designated physician and the preferred time slot from the set \( \mathcal{A}(s) \). The clinic may choose to deny requests to protect certain slots for later-arriving same-day patients, or for patients belonging to the requested physician’s panel, both of which may generate a greater benefit to the clinic.

The booking process described above is used to gain insights into the optimal clinic actions. It may be difficult to implement directly because patients observe \( s \) and may complain when their request for an open slot is denied. We will show in this paper that when \( n = 1 \)—i.e., the clinic is a single physician’s private practice—and patients’ choices satisfy a weak condition, the optimal policy allows regular patients to book any one of the unreserved slots up to a booking limit. This policy can be implemented simply by marking all slots in \( \mathcal{A}(s) \) as reserved once the booking limit is reached. When \( n > 1 \), the optimal policy is complicated and depends on the state \( s \). However, we show that a simple heuristic with two booking profiles for each physician, which for the physician indexed \( i \) depend only on \( s_j = \sum_{i=1}^{n} s^{ji} \) and \( s = \sum_{i=1}^{n} s_{ij} \), is near optimal. One booking profile is for each physician’s panel patients and the second profile is for nonpanel patients. Such a heuristic can be implemented if each caller is first asked to reveal his/her identity, which lets the clinic know the identity of his/her PCP. This is already a common practice when booking appointments. Additional discussion of implementation issues is included in §7.

The clinic’s average contribution from each accepted regular appointment is \( r_i \) when the patient visits his/her PCP and \( r_i' \) when the patient visits another physician. Corresponding parameters for same-day patients are \( r_2 \) and \( r_2' \). A natural ordering of the revenue parameters is \( r_i \geq r_i' \) and \( r_2 \geq r_2' \). In fact, a recent study has shown that patient-PCP mismatch reduces physicians’ and clinic’s revenues by approximately 15% per visit (O’Hare and Corlett 2004). This happens because physicians spend valuable time reading medical histories of unfamiliar patients and have little time left to perform billable extra services. At the same time, patients are both hesitant to ask for, and less receptive of, recommended additional services. In a prospective payment environment, patient-PCP mismatch increases costs because patients often book follow-up appointments with their PCPs (see, e.g., Murray and Berwick 2003). A clinic that operates on a fee-for-service basis and has overall excess capacity may benefit from increased visits generated by patient-PCP mismatch. However, this is not common because most physicians are able to build patient panels that lead to near 100% utilization of their available slots. In this environment, physician/clinic revenue is lower when nonpanel patients occupy the available slots (as documented in O’Hare and Corlett 2004). Finally, patient-PCP
mismatch lowers patient satisfaction, costing the clinic in the long run.

All same-day patients are accommodated. If necessary, this is achieved by working patients in, working overtime, or directing same-day patients to urgent-care clinics or hospital emergency departments, all of which cost more to the health-service provider. We denote the cost of insufficient same-day capacity by \(c\) per patient. Parameters \(\pi_1\) and \(c\) are often difficult to estimate from clinic data because health-care providers do not routinely track patients who are unable to book a regular or a same-day appointment. (Recall that \(\pi_1\) is incurred when a regular patient’s request is denied, or (s)he does not find any available slot suitable.) However, these parameters can be used as management levers. Larger values of the ratio \(c/\pi_1\) result in a greater number of slots being reserved for same-day patients. Finally, \(\pi_2\) denotes the penalty for an unutilized appointment slot. This can be calculated by taking the ratio of the average monthly fixed cost of running the clinic (facilities, equipment, plus staff salaries) and the average number of clinic appointments in a month.

For reasons listed below, it is appropriate to assume that \(r_1 + \pi_1 < r_1 + c\). Clinics and physicians place priority on being able to satisfy their acute-care clients to improve patient satisfaction and grow physicians’ panels. Moreover, established physicians with stable panels sometimes prefer to reserve capacity for same-day patients who provide variety in their practices. In the absence of these considerations, and assuming that choice probabilities are normal form, the appointment-booking problem becomes trivial—a revenue-maximizing physician/clinic should accept all regular patient requests and not reserve any slots for later-arriving same-day patients.

Let \(v_t(s)\) be the maximum expected revenue obtainable from period \(t\) onwards, given that the clinic’s reservation state at time \(t\) is \(s\). The clinic achieves \(v_t(s)\) by choosing the best accept/deny decision for each \(t\) and each \(s\). Its maximum expected total revenue is \(v_\tau(s_\tau)\), where \(s_\tau\) is a known starting state when booking opens. The appointment-booking problem can be reduced to recursively solving the following optimality equations for \(t = \tau, \tau - 1, \ldots, 1\):

\[
v_t(s) = \sum_{m=1}^{n} \lambda^n_m \sum_{i \in A(s)} P_{ij}^{\text{m}}(s) \max\{v_{t-1}(s + e_{ij}) + r_{mj}, v_{t-1}(s) - \pi_1\} + \sum_{m=1}^{n} \lambda^n_m p^{\text{m}}(s) (v_{t-1}(s) - \pi_1) + (1 - \sum_{m=1}^{n} \lambda^n_m) v_{t-1}(s)
\]

\[
= \sum_{m=1}^{n} \lambda^n_m \sum_{i \in A(s)} P_{ij}^{\text{m}}(s) \max\{\Delta v_{t-1}(s), 0\} + v_{t-1}(s) - \lambda_t \pi_1. \tag{1}
\]

In (1), the term \(\Delta v_{t-1}(s) = r_{mj} + \pi_1 + v_{t-1}(s + e_{ij}) - v_{t-1}(s)\) denotes the incremental benefit of booking slot \(ij \in A(s)\) at time \(t\) when the clinic’s reservation state is \(s\). An appointment request for slot \(ij\) at time \(t\) is accepted if and only if \(\Delta v_{t-1}(s) \geq 0\). Other parameters can be explained as follows. Revenue \(r_{mj} = r_i\), if \(m = i\), and \(r'\) otherwise; \(e_{ij} = (e_{ij1}, e_{ij2}, \ldots, e_{ijm}, \ldots, e_{i0})\); \(e_{0m}\), for each \(m = 1, \ldots, i - 1, i + 1, \ldots, n\), is a \(\kappa_m\)-dimensional vector with the \(j\)th component one and all other components zero.

Let \(X_m\) denote the random number of same-day appointment requests for physician \(m\), and \(X = \sum_{m=1}^{n} X_m\) denote the total same-day demand. For each \(m\), we use \(F_m\) and \(F_m^{-1}\) to denote the c.d.f. of \(X_m\) and its inverse. Additionally, we use \(G\) and \(G^{-1}\) to denote the c.d.f. of \(X\) and its inverse. With this notation in hand, the clinic’s reward from same-day demand, \(v_0(s)\), is

\[
v_0(s) = E\left\{r_2 \sum_{m=1}^{n} \left(\kappa_m - s_m\right) + c\left[X - (\kappa - s)\right] \right\} - \pi_2\left[(\kappa - s) - X\right] + r_2' \sum_{m=1}^{n} \left(\kappa_m - s_m - X_m\right) + \sum_{m=1}^{n} (X_m + s_m - \kappa_m)^+
\]

\[
- \pi_2 + r_2' + cE\left(\left((\kappa - X) - s\right)^+\right). \tag{2}
\]

In (2), \(s_m = \sum_{m=1}^{n} s^m\) denotes the total number of physician \(m\)’s reserved slots, and \(s = \sum_{m=1}^{n} s_m\) denotes the corresponding metric for the clinic.

Using the backward induction algorithm to solve (1)–(2) can be computationally challenging due to the curse of dimensionality. For example, for a typical urban clinic with 10 doctors and 30 slots per doctor per day, there are \(2^{300}\) states.

3. Modeling Patients’ Choices

Discrete choice models have attracted increasing attention in economics, marketing, and operations literature in recent years; see McFadden (2001) for an excellent survey and Train (2003) for methods to simulate choice probabilities. Here, we present three models that have been used extensively in previous articles. We also present a weak condition that defines what we call the class of normal-form choice models. The condition is weaker than the requirement of independence of irrelevant alternatives (IIA), which is an implied assumption in models based on the expected utility-maximization principle (Luce and Raiffa 1957, §2.5). All three examples presented below are normal form.

When the patient-choice model is normal form and \(n = 1\), we are able to characterize the optimal policy fully; when \(n > 1\), we obtain certain structural properties (details are in §4). For clarity of exposition, we describe the choice models only when \(n = 1\). It is straightforward to specify equivalent models for \(n > 1\). Note that with \(n = 1\), we do
not need superscripts $i$ and $m$, which are used in §2 to identify, respectively, the physician chosen by the caller and his/her PCP.

The Independent Demand Model. A patient calling in period $t$ chooses the $j$th slot with probability $q^j$. If slot $j$ is not available, the patient does not book an appointment, resulting in the following choice probabilities:

$$P^j(\tilde{s}) = \begin{cases} q^j & \text{if } j \in \mathcal{A}(\tilde{s}), \\ 1 - \sum_{k \in \mathcal{A}(\tilde{s})} q^k & \text{if } j = 0, \\ 0 & \text{otherwise.} \end{cases}$$

(3)

Independent demand is a common approach used in airline RM studies (Talluri and van Ryzin 2004b).

The Multinomial Logit Model (MNL). A patient's benefit from visiting the physician during the $t$th slot is captured by a utility function of the form $U^j_t = u^j_t + \xi^j_t$, where $u^j_t$ is the nominal (or mean) utility of choice $j$, and $\xi^j_t$ accounts for unobservable heterogeneity. Patients choose a utility-maximizing slot from the set of available slots. The terms $\xi^j_t$ are i.i.d. random variables with a Gumbel distribution. In particular, this means $P(\xi^j_t \leq x) = \exp[-\exp(-(x/\mu + y))]$, $\gamma \approx 0.5772$ is the Euler’s constant, $\mu$ is a parameter of the distribution, $E(\xi^j_t) = 0$, and $\text{Var}(\xi^j_t) = (\mu^2 \pi^2)/6$. The Gumbel distribution is used primarily because it is closed under maximization (Gumbel 1958). Similarly, for a no-request choice, the corresponding utility is $U^0_t = u^0_t + \xi^0_t$, where $\xi^0_t$ is also Gumbel distributed with mean zero.

For the MNL model, it can be shown (details are in Ben-Akiva and Lerman 1985) that for each $t$ and $\tilde{s}$, the choice probabilities are given by

$$P^j(\tilde{s}) = \begin{cases} \frac{\exp(u^j_t/\mu)}{\sum_{j \in \mathcal{A}(\tilde{s})} \exp(u^j_t/\mu) + \exp(u^0_t/\mu)} & \text{if } j \in \mathcal{A}(\tilde{s}) \text{ or } j = 0, \\ 0 & \text{otherwise.} \end{cases}$$

(4)

The MNL model is popular in the economics and marketing literature (see Ben-Akiva and Lerman 1985 and Train 2003).

The Random Utility-Maximization Model. More general utility-maximization choice mechanisms are used to model dynamic consumer substitution among product variants in some articles (see, for example, Mahajan and van Ryzin 2001). In our notation, this means that a caller in period $t$ assigns a random utility $U^j_t$ to slot $j$, which is independent of everything else. We assume for each $j$ and $k$ that $U^j_t$ and $U^k_t$ have continuous joint density $f_{jk}(x, y)$. This eliminates the need for tie-breaking rules because $P(U^j_t \neq U^k_t) = 1$ for all $j$ and $k$. Patients are utility maximizers; that is, they choose either an available slot or no-request with the highest utility, giving rise to the following choice probabilities:

$$P^j(\tilde{s}) = \begin{cases} P(U^j_t > U^k_t \text{ for all } k \in \mathcal{A}(\tilde{s}) & \text{or } k = 0, \text{ and } k \neq j) \text{ if } j \in \mathcal{A}(\tilde{s}) \text{ or } j = 0, \\ 0 & \text{otherwise.} \end{cases}$$

(5)

DEFINITION 1. For each $t$ and $\tilde{s}$, the choice probabilities $P^j(\tilde{s})$ are normal form if

$$P^j(\tilde{s}) \leq P^j(\tilde{s} + \vec{e}_k), \quad \forall j \in \mathcal{A}(\tilde{s} + \vec{e}_k),$$

(6)

where $k \in \mathcal{A}(\tilde{s})$ is an arbitrary open slot and $\vec{e}_k$ is a $k$-dimensional vector whose $k$th component is one and all other components are zero.

Inequality (6) states that when fewer slots are available, the probability that a patient chooses an available slot does not decrease. Note that in the example in §1, because $P^2(1, 0, 0, 0) < P^2(0, 0, 0, 0)$ (because the second caller books slot 2 only when the first slot is open), this patient’s choice model is, in fact, not normal form. Although verification of (6) for arbitrary choice probabilities requires checking all possible states, verifying it for a specific model can be simpler. For example, we show in online Appendix A that all three choice models described above are normal form. This result is stated as a proposition below.

An electronic companion to this paper is available as part of the online version that can be found at http://or.journal.informs.org/.

PROPOSITION 1. The choice probabilities according to the independent demand, the MNL, and the random utility-maximization models are normal form.

On an intuitive level, the choice probabilities in the independent demand model satisfy (6) trivially because they are independent of the reservation state. The choice probabilities in the remaining two models are based on the expected utility-maximization principle, where IIA is an implied assumption (Luce and Raiffa 1957, §2.5). Under IIA, the preference over any two slots should be independent of which other slots are available. This property implies the normal-form choice model. We will next argue, with the help of an example, that the defining inequality of the normal-form choice model is, in fact, a weak condition.

Suppose that a clinic is served by two physicians, each having two slots. Each patient strictly prefers to book with his/her own PCP when such an appointment is available. Patients are indifferent between the two slots of their PCP. For example, if a patient of physician 1 observes that both slots of physician 1 are available, then (s)he will choose either slot with probability 1/2. If only one slot is available, then the patient will choose that slot with probability one. When both slots of his/her PCP are unavailable, the patient’s actions depend on the reservation state of the other
doctor. If both slots of the other physician are available, the patient will choose to leave without booking an appointment. If one slot of the other physician is reserved, then the patient will take the open slot with probability one. This behavior can be explained by the following heuristic reasoning. When both slots of a non-PCP are available, the patient interprets this as a cue that this physician is either not good or is relatively inexperienced. Therefore, (s)he chooses not to book. However, if only one slot of the non-PCP is available, the patient does not make such an inference.

Using patients’ choices as described above, it is possible to write choice probabilities for each arrival for every possible reservation state of the clinic that show that patients’ choices violate IIA. Specifically, the preference between an available slot, they are more likely to book. In contrast, for reservation state changes from \( [(1, 1), (0, 0)] \) to either \( [(1, 1), (1, 0)] \) or \( [(1, 1), (0, 1)] \). That is, when patients face fewer available slots, they are more likely to book. In contrast, for the three choice models we described earlier in this section, the likelihood of booking cannot increase if fewer slots are available.

Recent experimental evidence suggests that in some cases, extensive choice can reduce the chance that consumers will buy a product (Iyengar and Lepper 2000). However, when the number of options is moderate, greater choice increases the probability of sale. The above example illustrates that the normal-form choice models allow both of these possibilities.

### 4. Structural Properties

In this section, we assume, for the most part, that the choice model is normal form. Our analysis is presented in two parts. When the clinic is served by a single physician, we show that the optimal policy is a booking-limit policy and that the optimal booking limit can be obtained by solving a newsvendorlike problem. In contrast, the structure of the optimal policy is much more complicated when the clinic has more than one physician. We demonstrate this with some examples first. Then, we obtain upper bounds on the optimal booking actions and the optimal value function, which are used to develop implementable heuristics.

#### 4.1. Single-Physician Clinic

When \( n = 1 \), there can be no patient-PCP mismatch and there is no need to carry physician indices \( i \) and \( m \). The maximum expected revenue from period \( t \) onwards is denoted by \( v_t(\bar{s}) \), which can be obtained as the following simplification of Equation (1):

\[
v_t(\bar{s}) = \lambda_t \sum_{j \in \bar{s}} P_j(\bar{s}) \max\{\Delta v_{j-1}(\bar{s}), 0\} + v_{j-1}(\bar{s}) - \lambda_t \pi_t.
\]  (7)

The incremental benefit of booking the \( j \)th slot at time \( t \) when the reservation state is \( \bar{s} \) equals \( \Delta v_{j-1}(\bar{s}) = r_1 + \pi_1 + v_{j-1}(\bar{s}) + \tilde{e}_j - v_{j-1}(\bar{s}) \) for each \( j \in \bar{s}(\bar{s}) \). In (2), we noticed that the clinic’s revenue from same-day patients depends only on \( s = \sum_{j=1}^{s} s_j \) and not on which slots are actually booked. Therefore, we define a real-valued function \( h(s) \), such that \( v_0(\bar{s}) = h(s) \) and

\[
h(s) = -c E(X) + (r_2 + c)(\kappa - s) - (r_2 + \pi_2 + c) E\{(|X - s|^+)}.
\]  (8)

Observe that (8) is similar to the newsvendor expected profit function (see, e.g., Porteus 2002, p. 9) and concave in \( s \). Next, we define \( \Delta h(b) = h(b+1) - h(b) + r_1 + \pi_1 \) and \( \eta = (r_2 - r_1 - \pi_1)/(r_2 + \pi_2 + c) \), and obtain the following key property of the function \( h(s) \).

**Lemma 1.** There exists a critical number \( b^* \) such that if \( s < b^* \), then \( \Delta h(s) \geq 0 \); otherwise, \( \Delta h(s) < 0 \). Specifically, \( b^* = \lfloor \kappa - F^{-1}(\eta) \rfloor \).

Lemma 1 is presented without proof. On an intuitive level, the existence of \( b^* \) follows from the concavity of \( h(s) \). Furthermore, \( b^* \) can be calculated using the following marginal analysis. Suppose that \( \kappa - b \) slots are available for same-day patients. If the same-day demand exceeds \( \kappa - b \), then the expected benefit of having one more slot available for same-day patients is \( (r_2 + c - r_1 - \pi_1)P(X > \kappa - b) \). On the other hand, if the same-day demand is no more than \( \kappa - b \), then the expected loss associated with the same action is \( (r_1 + \pi_1 + \pi_2)P(X \leq \kappa - b) \). Critical number \( b^* \) is the maximum value of \( b \) for which the expected benefit of reserving the \( (b+1) \)th slot for same-day patients is smaller than the associated expected loss. The range of parameters for which the trade-off is nontrivial is given by the inequalities \( r_2 + c > r_1 + \pi_1 \) and \( r_1 + \pi_1 + \pi_2 \geq 0 \); the former ensures \( \eta > 0 \), whereas the latter ensures \( \eta < 1 \). Parameters are assumed to lie in the desired range in the remainder of the paper.

When choice probabilities are normal form and optimal action is taken at each decision epoch, the property of the terminal value function proved in Lemma 1 also holds at an arbitrary \( t \). This result is presented in Theorem 1 and proved in online Appendix B. An immediate corollary of Theorem 1 establishes the structure of the optimal acceptance policy.

**Theorem 1.** Given a normal-form patient-choice model, \( j \in \bar{s}(\bar{s}) \), and \( s = \sum_{j=1}^{s} s_j \) at an arbitrary decision epoch \( t \),

1. if \( s \geq b^* \), then \( \Delta v_{j-1}(\bar{s}) < 0 \).
2. if \( s = b^* - 1 \), then \( \Delta v_{j-1}(\bar{s}) \geq 0 \) and \( \Delta v_{j-1}(\bar{s}) = \Delta v_{k-1}(\bar{s}) \) for all \( k \in \bar{s}(\bar{s}) \).
3. if \( s < b^* - 1 \), then \( \Delta v_{j-1}(\bar{s}) \geq 0 \).

**Corollary 1.** Given a normal-form patient-choice model, and \( s = \sum_{j=1}^{s} s_j \), the optimal acceptance policy at any decision epoch is a booking-limit policy, with the optimal booking limit \( b^* = \lfloor \kappa - F^{-1}(\eta) \rfloor \).
Suppose that possible reservation states at the beginning of period 2 are $s = \{(1, 1, 0), (1, 0, 0)\}$ and $s' = \{(1, 1, 0), (0, 1, 0)\}$. Note that, in both cases, physician 1's reservation state is identical and that the total number of available slots in the clinic is also the same. The clinic knows that in period 2, a patient of physician 1 will call who requests only the third slot of his/her PCP. Also, in period 1, a patient of physician 2 will call who requests only the second slot of his/her PCP.

The clinic can accept both requests, deny both requests, or else accept one and deny the other. The best actions can be determined by comparing clinic revenue in each instance. When the clinic's state is $s$, its revenue is maximized by rejecting a slot (1, 3) request of the earlier-arriving caller and accepting the second caller’s request. On the other hand, if the state is $s'$ and $r_1 + \pi_1 + r'_2 > r_2$, it is better to accept the earlier caller’s request. (Note that the second caller chooses no-request upon observing $s'$, and that the inequality above can be easily satisfied by plausible problem parameters.) This example shows that the optimal action for a particular slot of a physician (slot (1, 3) above), may depend on the reservation state of other physicians.

Next, we obtain an upper bound on the optimal booking actions. Let $\eta(s)$ and $\eta'(s)$ be critical-ratio functions defined as follows. If $r_2 > r'_2$, then $\eta(s) = \frac{r_2 + c - r_1 - \pi_1 - (r'_2 + c + \pi_2)G(\kappa - s - 1)}{(r_2 - r'_2)}$ and $\eta'(s) = \frac{r_2 + c - r'_1 - (r'_2 + c + \pi_2)G(\kappa - s - 1)}{(r_2 - r'_2)}$ for each fixed $s$. When $r_2 = r'_2$, these ratios are independent of $s$. In particular, $\eta(s)$ is the ratio $\eta$ defined in §4.1, whereas the analog of $\eta'(s)$ is the constant $\eta'(r) = \frac{r_2 + c - r'_1 - (r'_2 + c + \pi_2)G(\kappa - s - 1)}{(r_2 - r'_2)}$. Similarly, for each physician-index $i$, we define $\hat{b}_i(s) = [\kappa - F_i^{-1}(\eta'(s))]^+$ and $\hat{b}_i'(s) = [\kappa - F_i^{-1}(\eta'(s))]^+$. Similarly, if $r_2 > r'_2$, and $\hat{b}_i(s) = [\kappa - F_i^{-1}(\eta'(s))]^+$ and $\hat{b}_i'(s) = [\kappa - F_i^{-1}(\eta'(s))]^+$ when $r_2 > r'_2$.

Because $r_1 \geq r'_1$, $\eta(s) \leq \eta'(s)$, and consequently $\hat{b}_i(s) \geq \hat{b}_i'(s)$ for each $s$. Similar inequalities also hold when $r_2 = r'_2$.

Moreover, $\eta(s)$ and $\eta'(s)$ are nondecreasing in $s$, and functions $F_i^{-1}(\cdot)$ are nondecreasing in their arguments. Therefore, functions $\hat{b}_i(s)$ and $\hat{b}_i'(s)$ are nonincreasing in $s$. In the sequel, $s_i = \sum_{j=1}^{n_i} s^{ij}$ and $s = \sum_{i=1}^{n} s_i$ will denote, respectively, the number of physician $i$ slots and clinic slots booked at a decision epoch. With these definitions on hand, we obtain Lemma 2. Note that the results in Lemma 2 do not require a normal-form choice probability model. However, we test the accuracy of the two heuristics that rely on this lemma only for normal-form choices.

**Lemma 2.** The value function $v_0(s)$, defined in (2), has the following properties for each $ij \in \mathcal{S}$:

- **Given $m = i$:** If $r_2 > r'_2$ and $s \geq \hat{b}_i(s)$, then $\Delta v_{ij}^0(s) < 0$; else $\Delta v_{ij}^0(s) \geq 0$. Similarly, if $r_2 = r'_2$ and $s \geq \hat{b}$, then $\Delta v_{ij}^0(s) < 0$; else $\Delta v_{ij}^0(s) \geq 0$.

- **Given $m \neq i$:** If $r_2 > r'_2$ and $s \geq \hat{b}_i(s)$, then $\Delta v_{ij}^0(s) < 0$; else $\Delta v_{ij}^0(s) \geq 0$. Similarly, if $r_2 = r'_2$ and $s \geq \hat{b}$, then $\Delta v_{ij}^0(s) < 0$; else $\Delta v_{ij}^0(s) \geq 0$.

The function $v_0(s)$ depends only on $s_m$ and has monotone decreasing differences, which give rise to the properties in Lemma 2 (see the proof in online Appendix C). On an intuitive level, the trade-off at time 0 is between assigning an available slot to the regular patient’s request or reserving it for a future same-day caller. This results in a news-vendor-like trade-off with the marginal benefit of accepting the
regular patient’s request, depending on whether or not this caller belongs in the PCP’s panel. Note that because no additional regular patient calls are anticipated, the possibility of using the same slot for a possible higher-paying regular patient, i.e., one that belongs in the PCP’s panel, does not arise, and we obtain the two cases described in Lemma 2. Some properties of \( v_0(s) \) are preserved through value iterations leading to a partial characterization of the optimal policy. Theorem 2, proved in online Appendix D, summarizes the key results. It shows that when a physician has fewer than a threshold number of slots available, no additional booking requests should be accepted for that physician. This threshold depends only on the total number of booked slots in the clinic.

**Theorem 2.** When \( r_2 > r_2' \) and \( s_i \geq \hat{b}_i(s) \), additional requests for appointments with physician \( i \) should be declined. Similarly, when \( r_2 = r_2' \) and \( s \geq \hat{b} \), then additional requests for appointments with any one of the clinic’s physicians should be turned down.

Having an upper bound on the optimal acceptance action reduces computational burden because the optimal action is known a priori for certain states. Still, the number of clinic states in which at least one \( s_i < \hat{b}_i(s) \) (or \( s < \hat{b} \) when \( r_2 = r_2' \)) can be very large. Therefore, in the next subsection we develop easy-to-implement heuristics as well as an upper bound on the optimal clinic revenue.

### 4.3. Bounds and Heuristics

In what follows, we present one upper bound, labeled UB, and five heuristics. The first four heuristics are labeled H1 through H4, whereas the fifth heuristic is labeled FCFS because it accepts all appointment requests, provided the slot is available, in a first-come-first-served sequence. Heuristics H1 and H2 utilize the structural results in §§4.1 and 4.2. Heuristic H3 is based on the solution of a multidimensional newsvendor problem and heuristic H4 reserves capacity equal to mean same-day demand for each physician. Each heuristic also gives rise to a lower bound on the optimal value function. In §5.2, we shall explore the tightness of these bounds and the accuracy of the heuristics via numerical experiments.

Upper bound (UB) is obtained by setting \( r_2' = r_1 \) and \( r_2' = r_2 \), i.e., by eliminating the patient-PCP mismatch costs. There is mathematically no difference between the problem with \( n = 1 \) and the problem with \( n > 1 \) when there are no mismatch costs. The clinic’s optimal booking limit is \( b^* = [\kappa - G^{-1}(\eta)]^+ \). Note that the bound can perform poorly when, in reality, patient-physician mismatch costs are high.

Heuristic H1 uses the booking-limit bounds suggested by Lemma 2. Given \( r_2 > r_2' \), if \( s_i < \hat{b}_i(s) \), then any request for an appointment with physician \( i \) is accepted; if \( \hat{b}_i(s) \leq s_i < \hat{b}_i(s) \), then only requests from patients in physician \( i \)’s panel are accepted and other requests are declined; if \( s_i \geq \hat{b}_i(s) \), then all physician \( i \) appointment requests are rejected. Because \( \hat{b}_i(s) \) and \( \hat{b}_i(s) \) depend on \( s \), the booking decisions for all physicians are coordinated to a certain degree. This heuristic is much simpler when \( r_2 = r_2' \). In that case, bounds \( \hat{b} \) and \( \hat{b}' \) are used as booking limits for all physicians, regardless of the clinic’s reservation state.

We illustrate the heuristic policy for an example in Figure 1, which shows physician 1’s booking decisions for each possible state. In this example, \( n = 2, X_1 = X_2 \) are Poisson distributed with \( E(X_1) = E(X_2) = 12, \kappa_1 = \kappa_2 = 20, r_1 = 30, r'_1 = 24, r_2 = 30, r'_2 = 24, c = 10, \pi_1 = 5, \text{ and } \pi_2 = 20 \).

Heuristic H2 specifies two critical numbers for each physician that are independent of the clinic’s reservation state. Specifically, for physician \( i \) there is a booking limit, \( b_i = [\kappa_i - F_i^{-1}(\eta)]^+ \), for his/her panel patients, and a booking limit, \( b'_i = [\kappa_i - F_i^{-1}(\eta)]^+ \), for nonpanel patients. As before, \( b_i \geq b'_i \) for all \( i \). H2 is much easier to implement, but does make no attempt to coordinate booking decisions for different physicians. For the example shown in Figure 1, the H2 policy parameters for physician 1 are \( b_1 = 12 \) and \( b'_1 = 10 \).

Heuristic H3 uses booking limits \( b^* \), where \( b^* = \arg\max\{\Pi(b)\} \) is an optimal solution to the multidimensional newsvendor problem described below:

\[
\Pi(b) = E \left\{ (r_2 + c) \sum_{i=1}^{m} (c_i - b_i) - c \sum_{i=1}^{m} X_i \right. \\
\left. - (r_2 - r'_2) \sum_{i=1}^{m} (c_i - b_i - X_i)^+ \\
\left. - (r'_2 + c + \pi_2) \left[ \sum_{i=1}^{m} (c_i - b_i - X_i)^+ \right] \right\} .
\]

\( \Pi(b) \) is jointly concave in \( b \), which makes it straightforward to find \( b^* \).
Heuristic H4 specifies physician $i$’s booking limit as $b_i = \kappa_i - E(X_i) - 1$, thereby reserving exactly $E(X_i)$ slots for physician $i$’s same-day demand.

5. Analytical and Numerical Comparisons

This section has two parts. In the first part, we study the effect of physicians allowing work-ins (flexing capacity) on the optimal booking policy. Work-ins are designated slots in which two appointments can be booked simultaneously when same-day demand exceeds available capacity. Thus, work-ins correspond to having some on-demand capacity for same-day patients. The second part contains tests of accuracy of the five heuristics. The tests are first carried out assuming that the patients attempt to book only once. In the second set of tests, we simulate the operation of a typical primary-care clinic with patients making up to three attempts to book an appointment on consecutive days.

We also studied the effect of the size and the variability of same-day demand through analytical comparisons. These comparisons show that the demand affects the optimal booking policy and the clinic’s revenues in predictable ways. In particular, stochastically larger same-day demand reduces the availability of appointment slots for regular patients, greater variability of same-day demand reduces the clinic’s optimal revenues, and greater dependence among same-day demands of different physicians in the clinic reduces clinic revenues. For brevity, we have presented these results in online Appendix E.

5.1. The Effect of Flex Capacity

Flex capacity refers to the availability of work-ins to accommodate excess same-day demand. We consider the problem in which the clinic has a single physician who allows up to $y$ work-in slots for same-day patients when demand exceeds the workday’s capacity. Below, we let $\tilde{v}(s)$ denote the optimal value function of the MDP with flexible capacity, and define $\tilde{h}(s)$ such that $\tilde{v}_0(\tilde{s}) = \tilde{h}(s)$. Moreover,

$$\tilde{h}(s) = E\{r_2X - \pi_2(\kappa - s - X) \mid X \leq \kappa - s\} + r_1E\{X \mid \kappa - s < X \leq \kappa - s + y\} + E\{r_2(\kappa - s + y) - c(X + s - \kappa - y) \mid X > \kappa - s + y\}. \tag{9}$$

We also define $\Delta \tilde{h}(b) = \tilde{h}(b + 1) - \tilde{h}(b) + r_1 + \pi_1$. Now it is possible to show that $\tilde{h}$ possesses a property similar to the property of the function $h$ in Lemma 1. Specifically, $\tilde{h}(s + 1) - \tilde{h}(s) = -(r_2 + c) + \pi_2P(\kappa - s + X < \kappa - s) + (r_2 + c)P(\kappa - s + y < X < \kappa - s + y) + E(X) + E(X) > 0$ is nonincreasing in $s \in [0, \kappa]$. Therefore, upon defining $b^* = \min\{b \geq 0 \mid \Delta \tilde{h}(b) < 0\}$, the following results are immediate.

**Lemma 3.** There exists a critical number $\tilde{b}^*$ such that if $s < \tilde{b}^*$, then $\Delta \tilde{h}(s) > 0$; otherwise, $\Delta \tilde{h}(s) < 0$.

The work-in appointment slots are used exclusively for excess same-day demand. Therefore, $y$ affects the value function $\tilde{v}(s)$ only through the terminal function $\tilde{v}_0(\tilde{s})$. Using Lemma 3 and similar steps as in the proof of Theorem 1, we can show that a booking-limit policy is optimal when the physician allows work-ins. Moreover, the relationship between $b^*$, the optimal booking limit when $y > 0$, and $\bar{b}^*$, the optimal booking limit when $y = 0$, is described in the following proposition.

**Proposition 2.** Suppose that the physician allows up to $y > 0$ work-in appointment slots for same-day patients. Then, $\bar{b}^* \leq b^* \leq b^* + y$.

The proof is provided in online Appendix F. The above analysis suggests that having flexible capacity increases the number of slots available for regular patients; however, the magnitude of increase is less than the number of extra slots. Still, physicians prefer to allow only restricted use of work-in slots to maintain a better control over their daily work schedules (Gupta et al. 2006).

5.2. Numerical Comparisons

We use two different comparisons to test the accuracy of the heuristics. When most patients find a suitable slot on their first attempt, it is appropriate to compare the clinic operations for a single day. These comparisons are consistent with our mathematical model and in this case, both upper and lower bounds are available. When multiple booking attempts are common, a simulation model is used to test how well the heuristics perform relative to each other.

In examples that consider a single booking attempt, we assume that a majority of the booking requests arrive a short time before the start of the day. In particular, arrival probabilities increase linearly from zero to $\lambda$, in going from decision epoch $\tau$ to $0.2 \times \tau$, and then decrease to zero at decision epoch 1. (Recall that time is indexed backwards. Therefore, decision epoch 1 is the last potential arrival epoch of a regular patient’s call.) Given a nominal workload of $\rho_i$ for physician $i$, $\lambda_i$ can be determined from the following relationship: $(1/2) \tau \lambda_i + E(X_i) = \rho_i \lambda_i$. Note that $\rho_i$ would be the $i$th physician’s average workload if patients were to accept any available appointment.

In all examples, the choice probabilities are determined by the MNL model in (4). Patients typically have time-of-day preferences, which are captured as follows. The work day is divided into five intervals. We assume that regular patients assign a stochastically larger utility to slots in the first, third, and fifth intervals. Within each interval, patients assign the same utility to any slot. Specifically, the mean utility $u_{m,m}^{i,m}$ of a physician $m$’s patient assigns to each slot of this PCP in a high-preference interval is obtained by sampling from a uniform distribution over $[0, 1]$. Mean utility of a PCP slot in the low-preference intervals is generated by sampling from a uniform distribution over $[0, 0.5]$. Mean utility associated with a no-request choice, $u_0^{i,0,m}$, is also randomly generated from a uniform
distribution over $[0, 0.5]$. The mean utilities assigned to another physician’s slots are discounted by the factor $\beta_m$, $0 \leq \beta_m \leq 1$, if the patient is in physician $m$’s panel. Put differently, $u_{ijm} = \beta_m u_{i,j}$ for each $i \neq m$. Smaller values of $\beta_m$ imply greater patient loyalty because patients in physician $m$’s panel assign a smaller utility to an appointment with a different provider. Therefore, $\beta_m$ is also called the patient loyalty index.

5.2.1. Single Booking Attempt. We test bounds and heuristics by comparing the expected profits obtained from these approaches with the optimal expected profit. In these examples, the clinic is served by two doctors, there are $\tau = 200$ decision epochs, each physician has 10 appointment slots, and all slots of the clinic are available at the beginning of the booking horizon. Physician workday is divided into five equal-length intervals of two slots each. As stated above, patients assign a higher utility to intervals 1, 3, and 5. Same-day demands are specified via a bivariate Poisson distribution; see Johnson et al. (1997, pp. 124–139) for more on multivariate Poisson distributions. Specifically, $X_1 = Y_1 + Y_2$ and $X_2 = Y_2 + Y_1$, where $Y_1, Y_2,$ and $Y_2$ are mutually independent Poisson random variables with means $\theta_1, \theta_2,$ and $\theta_1, \theta_2,$ respectively. The joint probability mass function is

$$P(X_1 = x_1, X_2 = x_2) = \exp[-(\theta_1 + \theta_2 + \theta_1) \sum_{i=0}^{\min(x_1, x_2) - 1} \frac{\theta_1^{x_1-i} \theta_2^{x_2-i} \theta_1 \theta_2}{(x_1-i)! (x_2-i)!}], \quad (10)$$

$E(X_i) = \theta_i + \theta_1$, and the correlation coefficient of $X_1$ and $X_2$ is $\gamma(X_1, X_2) = \theta_1/[(\theta_1 + \theta_2)(\theta_1 + \theta_2)]^{1/2}$. Note that in this case same-day demands $X_1$ and $X_2$ are always positively correlated. This makes sense in practice because same-day demand is affected by common underlying events.

We compute the optimal as well as the lower-and upper-bounding value functions via the backward induction algorithm. Results of the experiments are summarized in Tables 1 and 2 and in Figure 2. In Table 1, the base case is listed in the first row. For each case, we provide the optimal value function and the percent difference between the optimal and each heuristic’s value function. The latter is reported as $\%\Delta_k$ for each $H_k$, where $k = 1, \ldots, 4$. Similar quantities for the FCFS and the UB value functions are reported as $\%\Delta_k$ and $\%\Delta_k$, respectively. We systematically vary the input parameters of our models, and each time one of these parameters is set first lower, and then higher, than the base case, with all other parameters remaining at their base values. In cases 4 and 6, where mismatch cost is high, UB is loose as expected. When the physicians’ nominal workload is imbalanced (see cases 12 and 13), H1 performs much better than H2. This is due to the fact that H1 coordinates booking decisions across clinic physicians, but H2 does not. Overall, both H1 and H2 perform well. However, H3, H4, and FCFS can be significantly worse. An interesting observation is that FCFS performs better than H3 and H4. This is, in part, due to the fact that $c$ is small, relative to $\pi_2$. Therefore, the clinic implicitly assumes a higher cost of slots going unused, as opposed to having

Table 1. Effect of model parameters on the accuracy of the five heuristics and the upper bound.

<table>
<thead>
<tr>
<th>No.</th>
<th>Variation</th>
<th>Optimum</th>
<th>$%\Delta_1$</th>
<th>$%\Delta_2$</th>
<th>$%\Delta_3$</th>
<th>$%\Delta_4$</th>
<th>$%\Delta_5$</th>
<th>$%\Delta_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Base case</td>
<td>446.89</td>
<td>0.019</td>
<td>0.084</td>
<td>4.60</td>
<td>4.87</td>
<td>1.60</td>
<td>5.95</td>
</tr>
<tr>
<td>2</td>
<td>$r_1 = 27, r_1 = 21.6$</td>
<td>420.09</td>
<td>0.022</td>
<td>0.202</td>
<td>3.96</td>
<td>4.39</td>
<td>2.84</td>
<td>5.84</td>
</tr>
<tr>
<td>3</td>
<td>$r_1 = 33, r_1 = 26.4$</td>
<td>470.42</td>
<td>0.008</td>
<td>0.076</td>
<td>5.10</td>
<td>5.21</td>
<td>0.78</td>
<td>6.35</td>
</tr>
<tr>
<td>4</td>
<td>$r_1 = 21, r_1 = 21$</td>
<td>433.15</td>
<td>0.036</td>
<td>0.105</td>
<td>4.23</td>
<td>4.67</td>
<td>3.00</td>
<td>8.91</td>
</tr>
<tr>
<td>5</td>
<td>$r_1 = 27, r_1 = 27$</td>
<td>460.02</td>
<td>0.003</td>
<td>0.097</td>
<td>4.50</td>
<td>5.08</td>
<td>0.77</td>
<td>3.04</td>
</tr>
<tr>
<td>6</td>
<td>$r_1 = 21$</td>
<td>434.48</td>
<td>0.039</td>
<td>0.114</td>
<td>4.39</td>
<td>4.70</td>
<td>3.06</td>
<td>8.63</td>
</tr>
<tr>
<td>7</td>
<td>$r_1 = 27$</td>
<td>459.12</td>
<td>0.004</td>
<td>0.087</td>
<td>4.55</td>
<td>5.02</td>
<td>0.73</td>
<td>3.26</td>
</tr>
<tr>
<td>8</td>
<td>$c = 7.5$</td>
<td>450.03</td>
<td>0.010</td>
<td>0.088</td>
<td>5.00</td>
<td>5.17</td>
<td>0.87</td>
<td>6.02</td>
</tr>
<tr>
<td>9</td>
<td>$c = 12.5$</td>
<td>443.92</td>
<td>0.022</td>
<td>0.099</td>
<td>4.26</td>
<td>4.59</td>
<td>2.69</td>
<td>5.93</td>
</tr>
<tr>
<td>10</td>
<td>$\pi_1 = 10$</td>
<td>468.46</td>
<td>0.021</td>
<td>0.075</td>
<td>3.21</td>
<td>3.57</td>
<td>2.05</td>
<td>5.69</td>
</tr>
<tr>
<td>11</td>
<td>$\pi_1 = 30$</td>
<td>423.95</td>
<td>0.010</td>
<td>0.107</td>
<td>6.01</td>
<td>6.27</td>
<td>1.30</td>
<td>6.37</td>
</tr>
<tr>
<td>12</td>
<td>$\rho_1 = 0.8$</td>
<td>403.76</td>
<td>0.002</td>
<td>0.037</td>
<td>2.83</td>
<td>2.90</td>
<td>0.25</td>
<td>5.50</td>
</tr>
<tr>
<td>13</td>
<td>$\rho_1 = 1.2$</td>
<td>473.07</td>
<td>0.004</td>
<td>0.179</td>
<td>6.25</td>
<td>6.76</td>
<td>6.29</td>
<td>6.45</td>
</tr>
<tr>
<td>14</td>
<td>$EX_1 = 4 (\theta_1 = 2)$</td>
<td>422.25</td>
<td>0.007</td>
<td>0.078</td>
<td>4.11</td>
<td>4.25</td>
<td>0.85</td>
<td>6.38</td>
</tr>
<tr>
<td>15</td>
<td>$EX_1 = 6 (\theta_1 = 4)$</td>
<td>464.53</td>
<td>0.029</td>
<td>0.155</td>
<td>6.20</td>
<td>6.29</td>
<td>2.75</td>
<td>5.80</td>
</tr>
<tr>
<td>16</td>
<td>$\beta_1 = 0.3$</td>
<td>446.34</td>
<td>0.018</td>
<td>0.080</td>
<td>4.38</td>
<td>4.83</td>
<td>1.47</td>
<td>5.76</td>
</tr>
<tr>
<td>17</td>
<td>$\beta_1 = 0.9$</td>
<td>445.06</td>
<td>0.019</td>
<td>0.088</td>
<td>4.42</td>
<td>4.85</td>
<td>1.67</td>
<td>6.37</td>
</tr>
<tr>
<td>18</td>
<td>$\theta_1 = 0 (\theta_1 = \theta_2 = 5)$</td>
<td>456.42</td>
<td>0.023</td>
<td>0.096</td>
<td>4.52</td>
<td>4.53</td>
<td>2.40</td>
<td>6.06</td>
</tr>
<tr>
<td>19</td>
<td>$\theta_1 = 4 (\theta_1 = \theta_2 = 1)$</td>
<td>436.21</td>
<td>0.011</td>
<td>0.088</td>
<td>4.62</td>
<td>5.07</td>
<td>1.07</td>
<td>6.05</td>
</tr>
</tbody>
</table>

Note. Base case: $r_1 = 30, r_2 = 24, r_2 = 30, r_2 = 24, \pi_1 = 5, c = 10, \pi_2 = 20, n = 2, \rho_1 = \rho_2 = 1, \kappa_1 = \kappa_2 = 10, EX_1 = EX_2 = 5 (\theta_1 = \theta_2 = 3$ and $\theta_1 = 2), \beta_1 = \beta_2 = 0.6$. 

Gupta and Wang: Revenue Management for a Primary-Care Clinic in the Presence of Patient Choice
insufficient capacity for same-day demand. The FCFS policy does not reserve any slots for future same-day demand and performs better in these examples.

Next, we examine the effect of physician workload. Nominal workloads are set equal \((p = \rho_1 = \rho_2)\) and increased from 0.6 to 1.7. Other relevant parameters are kept fixed and are identical to the base case in Table 1. In Figure 2, we plot \(\%\Delta_i\) as functions of \(p\). Figure 2 shows that when the clinic workload is low, all five heuristics perform well. In this event, acceptance rules do not matter much because it makes sense to accept virtually all requests. When clinic workload is high, the effectiveness of all heuristics decreases; however, H1 and H2 perform significantly better than others, with H1 being superior to H2. When the clinic is significantly overloaded, the FCFS heuristic becomes very bad. However, for low values of \(p\), it is better than H3 and H4. We also note that the UB is closest to the optimal value when \(p\) is either low or high. It is less accurate in the middle range. This happens because at both extremes, the proportion of appointments that are mismatched is relatively small. Given that most clinics operate at or near full capacity, heuristics H1 and H2 can significantly improve clinic revenues.

The effect of same-day demand variability is reported in Table 2. In these experiments, same-day demands are mutually independent binomial random variables with parameters \([n, p = E(X_m)/n]\), where \(n\) takes values 25, 50, and 100. Note that for a binomial distribution, the variability is increasing in \(n\). In cases 1–3, we set \(E(X_m) = E(X) = 5\); in cases 4–6, we set \(E(X_m) = 4\) and \(E(X) = 5\); whereas in cases 7–9, we set \(E(X_m) = 6\) and \(E(X) = 5\). In each case, we provide the corresponding coefficient of variation (CoV) for \(X_m\) and \(X\), respectively. Other relevant parameters are kept fixed and are identical to the base case in Table 1. We observe that the optimal revenue is decreasing in demand variability. This result is consistent with Proposition 4 in the online appendix, which proves the same result formally for a single-physician clinic. Note that H1 performs better than H2 and that both heuristics perform better, relative to the optimal policy, as the same-day demand variability increases. The former is a consequence of H1’s ability to coordinate booking decisions of different physicians, whereas the latter comes from the fact that booking decisions have a limited effect on the expected profit when variability is high. A similar observation does not hold for H3, H4, and the FCFS heuristics.

The aggregate performance of the five heuristics and the upper bound over all problem instances reported above (total 40 cases) can be summarized by examining the average and maximum \(\%\Delta_i\).s. These values are 0.30% and 4.09% for \(\%\Delta_1\); 0.39% and 4.35% for \(\%\Delta_2\); 5.14% and 13.93% for \(\%\Delta_3\); 5.44% and 14.17% for \(\%\Delta_4\); 4.08% and 30.94% for \(\%\Delta_5\); and 5.82% and 8.91% for \(\%\Delta_6\). H1 performs best in all cases. The surprisingly good performance of H1 can be explained by a variety of factors. First, H1 prescribes the right action in large regions of the reservation state space. For example, both when a large number of slots are already booked or when many slots are open, its deny/accept actions are likely to match optimal actions. This leaves a relatively small region in which actions may be incorrect. As we see through the examples reported here, the overall impact of such suboptimal actions is relatively small. Second, H1 opens more slots for patients belonging to a doctor’s panel, which is also consistent with the actions of an optimal policy. Finally, the trade-off underlying the decisions to accept/deny booking requests have a structure

---

**Table 2.** Effect of same-day demand variability on the accuracy of the five heuristics and the upper bound.

<table>
<thead>
<tr>
<th>No.</th>
<th>Coefficient of variation</th>
<th>Optimum</th>
<th>(%\Delta_1)</th>
<th>(%\Delta_2)</th>
<th>(%\Delta_3)</th>
<th>(%\Delta_4)</th>
<th>(%\Delta_5)</th>
<th>(%\Delta_6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CoV₁ = 0.40, CoV₂ = 0.40</td>
<td>462.29</td>
<td>0.031</td>
<td>0.111</td>
<td>3.74</td>
<td>4.24</td>
<td>2.92</td>
<td>5.80</td>
</tr>
<tr>
<td>2</td>
<td>CoV₁ = 0.42, CoV₂ = 0.42</td>
<td>455.87</td>
<td>0.025</td>
<td>0.095</td>
<td>3.92</td>
<td>4.26</td>
<td>2.36</td>
<td>6.05</td>
</tr>
<tr>
<td>3</td>
<td>CoV₁ = 0.44, CoV₂ = 0.44</td>
<td>455.97</td>
<td>0.026</td>
<td>0.095</td>
<td>4.17</td>
<td>4.25</td>
<td>2.36</td>
<td>6.06</td>
</tr>
<tr>
<td>4</td>
<td>CoV₁ = 0.46, CoV₂ = 0.40</td>
<td>435.67</td>
<td>0.016</td>
<td>0.092</td>
<td>3.65</td>
<td>3.77</td>
<td>1.60</td>
<td>6.26</td>
</tr>
<tr>
<td>5</td>
<td>CoV₁ = 0.48, CoV₂ = 0.42</td>
<td>433.14</td>
<td>0.016</td>
<td>0.094</td>
<td>3.87</td>
<td>4.01</td>
<td>1.43</td>
<td>6.36</td>
</tr>
<tr>
<td>6</td>
<td>CoV₁ = 0.49, CoV₂ = 0.44</td>
<td>432.11</td>
<td>0.015</td>
<td>0.087</td>
<td>3.91</td>
<td>4.08</td>
<td>1.35</td>
<td>6.34</td>
</tr>
<tr>
<td>7</td>
<td>CoV₁ = 0.36, CoV₂ = 0.40</td>
<td>492.00</td>
<td>0.025</td>
<td>0.179</td>
<td>4.96</td>
<td>5.37</td>
<td>4.38</td>
<td>5.66</td>
</tr>
<tr>
<td>8</td>
<td>CoV₁ = 0.38, CoV₂ = 0.42</td>
<td>476.60</td>
<td>0.044</td>
<td>0.164</td>
<td>5.15</td>
<td>5.57</td>
<td>4.19</td>
<td>5.71</td>
</tr>
<tr>
<td>9</td>
<td>CoV₁ = 0.40, CoV₂ = 0.44</td>
<td>475.18</td>
<td>0.040</td>
<td>0.161</td>
<td>5.13</td>
<td>5.68</td>
<td>4.00</td>
<td>5.73</td>
</tr>
</tbody>
</table>
similar to the newsvendor problem, which is known to have a relatively flat expected-profit function near the optimum.

5.2.2. Multiple Booking Attempts. We use a computer simulation model to assess the effectiveness of the five heuristics when patients try up to three times to book an appointment. Data for this model comes from an actual medium-sized clinic in the Twin Cities area. These data pertained to all appointments served by 10 physicians, whose tenure ranged from 3.5 to over 9 years. All physicians used 15-minute slots.

Daily demand and physician availability revealed a significant day-of-week seasonality in the supply-demand match patterns, with largest capacity shortages on Mondays and Fridays. On Tuesdays and Wednesdays, the demand at the clinic level was roughly matched with the supply, whereas there was some slack capacity on Thursdays. There were also significant variations in supply and demand by physician. Two of the 10 physicians were very busy (i.e., their nominal workload exceeded one), six were operating with their availability roughly matched with demand (nominal workload equaled one), and two physicians had slack capacity (nominal workload was smaller than one).

We estimated the average number, by PCP ID, of calls each weekday that result in same-day appointments. This is used as a surrogate for mean daily demand for same-day slots. Note that all these callers may not be able to book an appointment with their PCP. We varied physician availability by the day of the week as follows. For doctors with slack capacity, we set their available slots equal to the mean demand from their panel patients on peak-demand days (Monday and Friday) and 10% more than their mean demands on other weekdays. For doctors with balanced capacity, the available slots are 5% less than mean demand on peak-demand days, 5% more than mean demand on Tuesday and Wednesday, and equal to mean demands on Thursday. For busy doctors, the available slots are 10% less than their mean demands on peak days, and 5% less than mean demands on other weekdays. The overall clinic workload was approximately the same as the available capacity. Specifically, the clinic’s theoretical utilization was 0.998 with a weekly mean demand of 733 slots and availability of 734 slots.

At the beginning of each day, we simulate the number of regular patient arrivals for each physician according to a Poisson process with the mean rates calculated above. These arrivals are then assigned to each future workday within a two-week window. We assume that 80% of the arriving patients try to book in week 1 of the two-week period, with the remaining trying to book in week 2. This is based on the clinic data, which showed that 70% of all patients actually book an appointment within one week of their call. We use a slightly higher percentage in our numerical experiment to account for the fact that a certain proportion of patients would want to book earlier, but are not able to do so because a convenient time slot is not available. Within each one-week period, we assume that a patient is equally likely to choose any day. Thus, associated with each arrival is a PCP ID and a preferred day on which the patient wishes to book the appointment. For each future day in the two-week window, we call these arrivals virtual regular patients because not all of them will be able to find a desirable slot that day.

After the patient makes his/her appointment-day selection, he/she chooses an appointment slot on that day. There are at most 30 available slots per day for each physician. The actual number of available slots depends on the physician’s workload and the day of the week. The workday is divided into five intervals containing 5, 7, 5, 8, and 5 slots, respectively. As explained at the beginning of §5.2, regular patients prefer the first, third, and fifth intervals.

Same-day requests arise according to a Poisson process with the mean of daily same-day demand corresponding to the particular weekday and physician. We assume that each same-day patient will take any available PCP slot. Among those who cannot be served by their PCP, there is a 0.10 probability that they will choose to take an available slot with their PCP the next day. (This is based on anecdotal evidence. Clinics do not track booking behavior of same-day patients who find that their own PCP has no available slots. Other values of this probability were also tested. Results are omitted in the interest of brevity. The relative performance of different heuristics, reported in Table 3, remains unaltered.) We treat these spillovers exactly the same as additional same-day requests that arrive

<table>
<thead>
<tr>
<th>Table 3. Performance of heuristics with multiple booking attempts.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Policy</strong></td>
</tr>
<tr>
<td>----------------</td>
</tr>
<tr>
<td>Average weekly revenue</td>
</tr>
<tr>
<td>Unused slots (%)</td>
</tr>
<tr>
<td>Excess same-day demand (%)</td>
</tr>
<tr>
<td>Standard deviation of excess same-day demand (%)</td>
</tr>
<tr>
<td>Regular appointments accepted and lost (%)</td>
</tr>
<tr>
<td>First try</td>
</tr>
<tr>
<td>Second try</td>
</tr>
<tr>
<td>Third try</td>
</tr>
<tr>
<td>Lost</td>
</tr>
</tbody>
</table>
on the next day. The remaining same-day patients first take any available slots of other physicians. If some demand still remains unmet, then it is met either by stretching capacity (working extra hours, working in patients, double booking) or by making use of urgent clinics/hospital emergency departments.

We run the simulation for 1,000 replications to calculate average statistics. Each replication represents a 10-year operation of the primary-care clinic. We discard data from the first six months of operations, which is treated as the warm-up period. (Note that the system starts from empty.) We keep track of the weekly clinic revenue as well as the following operational performance measures: the proportion of unused slots (spoilage), the mean and the standard deviation of excess same-day demand (measured as a percentage of the realized same-day demand), the percentage of regular patients who book an appointment on the first, second, and third attempts, and the percentage of regular patients who exit the system without appointments. Our results are summarized in Table 3. Heuristic H1 achieves the highest weekly revenue by balancing capacity allocation to the regular and same-day demands. It results in about 3% of the slots going unused, about 10% of the same-day patients being accommodated outside normal working hours, and about 0.36% of the regular patients not booking an appointment after three attempts. The standard deviation of excess same-day demand is 11.5%. All other heuristics perform worse in terms of overall revenue. However, they may perform better on one of the operational performance measures mentioned above. For example, FCFS favors regular callers, but it leads to a significantly higher percentage of same-day patients (mean = 18.6%, standard deviation = 26.2%) having to be accommodated by working extra hours.

6. Insights

We solved several examples that mimic the operations of a typical medium-sized urban clinic to gain operational insights. In the first experiment, we used the simulation model described in §5.2.2 to study how regular patients’ load affects clinic profits. We multiplied each physician’s regular workload on each work day by a common factor \( \nu \), while keeping his/her same-day demand fixed. Results of this experiment are shown in Figure 3.

The clinic maximizes profit by operating with a slight capacity shortage when bookings are controlled by either H1, H2, or the FCFS policy. H3 and H4 achieve maximum profit at \( \nu = 1 \). Recall from §5.2.2 that when \( \nu = 1 \), clinic-level demand and capacity are nearly equal. The reason why some extra demand is beneficial is driven by the fact that regular patient choices are random. Therefore, some patients invariably choose not to book. Having extra demand leads to a better utilization of available slots. It may also increase the amount of same-day demand that is accommodated by working overtime, but overall profit is higher. Specifically, when policy H1 is used, together with the optimal \( \nu \), 1.5% of slots go unused, 13% of same-day patients are handled either via overtime or an urgent-care clinic, and 0.81% of regular patients are not able to book after three attempts. The standard deviation of excess same-day demand is 11.7%. If 13% of same-day demand is deemed too high to handle by flexing capacity, the clinic can obtain a more desirable booking policy by increasing \( c \). H3 and H4 reserve slots based only on the level of same-day demand. Therefore, they do not benefit from increased regular demand.

Next, we study the effect of patient loyalty and physician workload imbalance on clinic revenue. In these experiments, we used the best heuristics (H1 and H2), and evaluated the value function assuming a single-booking attempt. We also obtain the upper bound UB. Because H1 and H2 yield similar patterns with either single or multiple booking attempts, we restrict our attention to problems with a single booking attempt to keep computational effort manageable.

The problem data are as follows. The clinic is served by 12 physicians, each physician has 30 appointment slots, and there are 20,000 decision epochs. Because of the large size of this problem, the value function cannot be evaluated from the backward induction algorithm. Therefore, we use Monte Carlo methods to evaluate the expected profits under H1, H2, and UB. Call arrival probabilities and patient utilities for each available slot are generated as described at the start of §5.2. Same-day demands are generated by sampling from a multivariate Poisson distribution. Specifically, \( X_i = Y_i + Y \), where \( Y_i \) and \( Y \) are mutually independent Poisson random variables with means \( \theta_i \) and \( \theta \), respectively. The simulation is run until either the lengths of 95% confidence intervals for the expected profits under H1, H2, and UB are within at most 0.1% of their expected profits, or the number of simulation runs reaches 10,000. In all experiments
we conducted, the size of the 95% confidence interval was ≤ 0.1% of the expected profit.

We carry out two experiments to illustrate the effect of changing patient loyalty. In these experiments, all physicians have the same patient loyalty index \( \beta \). We change \( \beta \) from zero to one, which corresponds to decreasing patient loyalties. We plot the expected profit realized under \( H_1 \) as a function of \( 1 - \beta \). In Figure 4, physician workloads are balanced; \( \rho = 1.0, 18 \) [\( \theta = 6 \)], and all other parameters are identical to the base case in Table 1. In Figure 5, physicians’ nominal workloads and mean same-day demands are \( (1.0, 1.0, 1.0, 1.0, 1.25, 1.25, 1.25, 0.75, 0.75, 0.75, 0.75) \) and \( (1.4, 14, 14, 14, 18, 18, 18, 18, 10, 10, 10, 10) \) [\( \theta = 6 \)], respectively. Other relevant parameters are identical to the base case in Table 1. Figure 5 also plots the proportion of unmet regular patients’ demand at the clinic level. These experiments serve to bring out the effect of changing patient loyalty under conditions of balanced and imbalanced workload distributions. For the example in Figure 4, the percentage of unmet regular patients’ demand is 3.96%.

Figure 4 shows that greater patient loyalty is beneficial for the clinic when physician workloads are balanced. It is interesting to note that loyalty level need not be high in absolute terms to realize its full benefit. Furthermore, with imbalanced workloads (a common occurrence), greater loyalty is beneficial only up to a point. Figure 5 shows that very high patient loyalty can lower clinic profits. This happens because greater loyalty limits the clinic’s ability to use different physicians as substitutes. Note that the proportion of unmet regular patients demand also increases with patient loyalty. Gupta et al. (2006) discuss several practical methods that clinic directors can use to improve clinic performance when loads are imbalanced and patients have high loyalty, particularly for physicians with high workload.

In the final experiment, we study the effect of workload imbalance on the clinic’s expected profit. Below, we use \( \hat{\rho}_i = (\rho_{i1}, \ldots, \rho_{in}) \) to denote the \( i \)th vector in a series, where the nominal workload of physician \( m \) in this vector is \( \rho_i^m \). Aggregate clinic workload for the \( i \)th series is \( \sum_{m=1}^n \rho_i^m \). To produce workload vectors with different degrees of imbalance (or dissimilarity), we use the concept of majorization. Basics are reviewed below and further details can be found in Müller and Stoyan (2002, pp. 31–33).

For any \( n \)-dimensional vector \( \bar{x} = (x_1, \ldots, x_n) \) of reals, let \( x_1 \geq \cdots \geq x_n \) denote the components of \( \bar{x} \) in decreasing order. For \( \bar{x} \) and \( \bar{y} \) in \( \mathbb{R}^n \), we say that \( \bar{x} \) is majorized by \( \bar{y} \) (written \( \bar{x} \preceq_M \bar{y} \)) if \( \sum_{i=1}^k x_i \leq \sum_{i=1}^k y_i \) for each \( k = 1, \ldots, n - 1 \), and \( \sum_{i=1}^n x_i = \sum_{i=1}^n y_i \). In economics, majorization is used to compare the inequality of income distributions. In our setting, if \( \bar{x} \) and \( \bar{y} \) denote two distributions of the clinic’s aggregate workload among physicians, then the relation \( \bar{x} \preceq_M \bar{y} \) means that \( \bar{x} \) has a more even workload distribution than \( \bar{y} \).

Let \( A = (a_{ij}) \) be a nonnegative \( n \times n \) matrix such that \( \sum_{j=1}^n a_{ij} = 1 \) for each \( j \), and \( \sum_{i=1}^n a_{ij} = 1 \) for each \( i \). Note that matrices of this type are called doubly stochastic matrices. It can be shown that if \( \bar{x} = A \bar{y} \), then \( \bar{x} \preceq_M \bar{y} \). Thus, we can obtain a series of majorized workload distributions with the same aggregate load by multiplying an initial workload vector successively with the matrix \( A \). Specifically, if the initial workload vector is \( \rho_{0w} \), and for a particular choice of \( A \) we recursively obtain \( \hat{\rho}_i \) according to \( \hat{\rho}_i = A^i \rho_{0w} \), then \( \hat{\rho}_i \preceq_M \rho_{i-1} \) for each \( i \geq 1 \).

For the experimental results reported in Figure 6, we start with three different initial vectors, which correspond, respectively, to low, medium, and high aggregate workload. Specifically, the initial vectors are \( (1.34, 1.26, 1.18, 1.10, 1.02, 0.94, 0.86, 0.78, 0.70, 0.62, 0.54, 0.46) \) for low workload, \( (1.44, 1.36, 1.28, 1.20, 1.12, 1.04, 0.96, 0.88, 0.80, 0.72, 0.64, 0.56) \) for medium workload, and \( (1.64, 1.56, 1.48, 1.40, 1.32, 1.24, 1.16, 1.08, 1.00, 0.92, 0.84, 0.76) \) for high workload. The matrix \( A \) is a \( 12 \times 12 \) matrix.
Figure 6. The effect of physicians’ workload imbalance.

![Graph showing the effect of physicians’ workload imbalance on expected profit.](image)

Note. Higher $I$ ⇒ more balanced workload.

such that $a_{ij} = 0.9$ and $a_{ij} = 0.1$ for $i = 1, \ldots, 12$. All other $a_{ij}$s are set equal to zero. The choice of $A$ ensures that at each iteration, the physician with the highest workload will transfer some portion of his/her workload to the physician with the lowest workload; the physician with the second-highest workload will transfer some portion to the physician with the second-lowest workload, and so on. The portion transferred between any pair of physicians is 10% of the workload difference between them. The mean same-day demand is $EX_i = 14 (\theta = 6)$. Other relevant parameters are identical to the base case in Table 1. Figure 6 shows the expected profit realized when heuristic H1 is used as a function of $I$. Observe that the clinic’s expected profit increases when physicians’ workloads become more balanced. The relative benefit of balanced workload is higher when the clinic’s average workload is smaller.

7. Concluding Remarks

This paper proposes a methodology for a clinic to decide how to manage access to its slots when patients choose between a same-day slot (which may not be at the most convenient time) or a future appointment at a convenient time. We show that for a large class of patient-choice models, which includes some of the most commonly used models in economics, marketing, and operations, relatively simple heuristic approaches for managing patients’ access to appointment slots are near optimal. The clinic does not need to assess patient-choice probabilities. It only needs to ascertain that these probabilities belong to the class of normal-form probabilities. Membership in this class is not a restrictive assumption, and many choice models, including those based on the expected utility-maximization principle, belong to this class.

To implement H1 and H2, a clinic will need to estimate the same-day demand for each physician. Demand estimates are required in the calculation of the physicians’ booking profiles. Such calculations can be performed offline, which makes it particularly attractive to apply this approach for managing Web-based appointment bookings in real time. Upon receiving a call for a future appointment with a particular physician, and upon learning that the caller’s identity and the identity of his/her PCP, the clinic will either make all open slots available to the caller, or show all slots as being occupied. This decision depends on the number of booked slots of the physician, the total number of booked slots in the clinic, and the identity of the patient’s PCP. Because the implementation of this approach may change demand patterns until a new equilibrium is reached, clinics may need to periodically collect same-day demand data to fine-tune booking policies. Such methodology should be relatively easy to incorporate in a variety of electronic medical record systems used by clinics and health maintenance organizations. Practitioners should note that the FCFS policy produces smaller revenues and a greater variability in excess same-day demand. However, it does not require same-day demand data.

Key observations from analytical and numerical comparisons presented in this paper are as follows. Greater variability of same-day demand and greater positive dependence among same-day demands of clinic physicians deteriorates performance. Increasing patient loyalty is beneficial when physician workloads are balanced. However, it is more common for physician workloads to be imbalanced, with a few physicians struggling to keep up with demand, a few others with relatively light workload (usually those trying to establish new practices), and a majority with workloads that approximately match available capacity. In this environment, increased loyalty is beneficial only up to a point. If patients become extremely loyal, it reduces the clinic’s ability to benefit from pooling physicians’ capacities, which lowers overall clinic revenue. Finally, attempts by clinic directors to even workload distribution among physicians increase revenue.

Knowing when to add capacity is the key to making advanced-access systems succeed. If physicians are hired too soon, there may not be enough work for them to build full practices. In such cases, physicians either leave or cannibalize demand from colleagues, both of which can be detrimental for the clinic. On the other hands, if the clinic waits too long to add a new physician, the goals of advanced access may be eroded and patients may switch to other clinics. Our models can be used to study the impact of change in workload, and to identify the point of time when hiring a new physician is appropriate.

8. Electronic Companion

An electronic companion to this paper is available as part of the online version that can be found at http://or.journal.informs.org/.
Acknowledgments

This research was supported in part by a research grant from the Agency for Healthcare Quality and Research. The authors are grateful to an associate editor and two anonymous referees for their constructive feedback on earlier versions of this paper.

References


Murray, M., C. Tantau. 1999. Redefining open access to primary care. Managed Care Quart. 7(3) 45–55.


