1-1-2011

Adaptive Appointment Systems with Patient Preferences

Wen-Ya Wang
San Jose State University, wenta@ie.umn.edu

D. Gupta
University of Minnesota

Recommended Citation

Wen-Ya Wang and D. Gupta. "Adaptive Appointment Systems with Patient Preferences" *Manufacturing & Service Operations Management* (2011): 373-389. https://doi.org/10.1287/msom.1110.0332

M&SOM MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 13, No. 3, Summer 2011, pp. 373–389
ISSN 1523-4614 | EISSN 1526-5498 | 11 | 1303 | 0373

DOI 10.1287/msom.1110.0332

Adaptive Appointment Systems with Patient Preferences

Wen-Ya Wang, Diwakar Gupta
Industrial and Systems Engineering Program, University of Minnesota, Minneapolis, Minnesota 55455
{wenya@ie.umn.edu, guptad@me.umn.edu}

Patients' satisfaction with an appointment system when they attempt to book a nonurgent appointment is affected by their ability to book with a doctor of choice and to book an appointment at a convenient time of day. For medical conditions requiring urgent attention, patients want quick access to a familiar physician. For such instances, it is important for clinics to have open slots that allow same-day (urgent) access. A major challenge when designing outpatient appointment systems is the difficulty of matching randomly arriving patients' booking requests with physicians' available slots in a manner that maximizes patients' satisfaction as well as clinics' revenues. What makes this problem difficult is that booking preferences are not tracked, may differ from one patient to another, and may change over time. This paper describes a framework for the design of the next generation of appointment systems that dynamically learn and update patients' preferences and use this information to improve booking decisions. Analytical results leading to a partial characterization of an optimal booking policy are presented. Examples show that heuristic decision rules, based on this characterization, perform well and reveal insights about trade-offs among a variety of performance metrics important to clinic managers.

Key words: appointment scheduling; health care; probability: stochastic model applications
History: Received: June 1, 2009; accepted: February 10, 2011. Published online in *Articles in Advance* June 8, 2011.

1. Introduction

An outpatient appointment is a contract between a patient and a clinic by which the latter reserves a certain amount of service providers' time and physical assets for the exclusive use of the patient who holds the appointment. Patients' satisfaction with their health-care clinic is affected not only by the perceived quality of medical services that they receive during their visit but also by their appointment booking experiences. Clinic managers care about having high scores on patient satisfaction surveys because that helps them attract new patients and negotiate better rates with insurers. Because the vast majority of medical appointments are booked with physicians working in primary care clinics, we focus in this paper on the design of primary care appointment systems. A detailed description of the primary care service environment is provided in Gupta and Denton (2008). The ensuing abbreviated description focuses on features that are central to this study.

Patients that belong to a health system choose both a preferred clinic and a preferred physician. The latter is commonly referred to as the preferred care provider (PCP) for the patient. The term *panel* is used to denote a group of patients that has chosen the same PCP. Patients usually call in advance to book an appointment. Patients' satisfaction with an appointment system when they attempt to book a nonurgent appointment is affected by their ability to book with their doctor of choice and at a convenient time of day (Cheraghi-Sohi et al. 2008, Gerard et al. 2008). Patients also prefer a sooner rather than a later appointment so long as it meets their time and physician preferences. For urgent medical conditions, patients want quick access to a physician. Clinics plan for such appointment requests and have open slots each day that allow same-day (urgent) access.

Because appointments are booked one at a time without knowledge of the number, sequence, and service requirements of future arrivals, many clinics use a two-step process to design appointment systems, which we call *clinic profile setup* and *appointment booking* steps, respectively. Clinic profile setup refers to the common practice of dividing physicians' available time on each workday into appointment slots. All slots need not be of the same length. For example, whereas a standard slot may be appropriate for the vast majority of routine appointments, physical exams and in-office procedures may require longer slots. In the appointment booking (second) step, the clinic profile is known and the decision concerns which available appointment slot to book for each incoming appointment request. This paper is concerned with the second step. That is, we assume that the number of appointments and the length of each appointment slot have been determined for each physician. Clinic profile setup may take into account a whole host of factors, including physicians' willingness to work overtime, no-show rates, service time variability, and demand for physicians' slots (see LaGanga and Lawrence 2007, Robinson and Chen 2003, Denton and Gupta 2003, Ho and Lau 1992, Weiss 1990).

What makes the appointment booking problem (the focus of this paper) difficult is that booking preferences are different for each patient, and they change over time for the same patient. For example, some patients are willing to see any available doctor if they can have an appointment sooner whereas others prefer to wait until a slot becomes available with their PCP. Some patients are able to visit the clinic only within a short time window because of job-related constraints or personal schedules (Jennings et al. 2005, Olowokure et al. 2006), whereas others can be quite flexible. Finally, changes in work schedule, marital status, and family size can alter a patient's booking pattern.

Evidence shows that clinics benefit by accommodating patients' preferences. First, matching patients with their PCP ensures continuity (quality) of care (Doescher et al. 2004) and allows physicians to provide more value-added services to their patients, which increases clinics' revenues (O'Hare and Corlett 2004). Second, matching patients with their PCP and offering them a convenient appointment time can decrease the number of no-shows and thereby increase operational efficiency (Barron 1980, Carlson 2002, Smith and Yawn 1994). The above arguments provide the motivation for paying attention to patients' physician and time preferences and adapting appointment booking practices as these preferences change. The purpose of this paper is to develop a framework for the design of such adaptive appointment systems. We use patient-PCP match rate, advance-book failure rate, and the total number of patients served as surrogate measures for patients' satisfaction with the appointment booking system.

We assume a known clinic profile, which may include overbooking, and develop algorithms for making appointment booking decisions to maximize clinic revenue. We model each panel as a different revenue class and allow the revenue from each appointment to depend on whether the appointment is with a patient's PCP. Patients have different acceptance probabilities for each physician and time-block combination, and each patient may have several acceptable combinations when he or she attempts to book an appointment. We also model advance-book (nonurgent) and same-day (urgent) demand. Inadequate capacity to serve urgent demand results in a higher cost to the health system. If a patient's service-time class can be ascertained at the time of booking an appointment, then such information can be incorporated in the proposed system by checking that the offered appointment slot is appropriate for the services requested. However, in numerical examples presented in this paper, the availability of such information is not assumed.

Booking decisions do not depend on each patient's individual no-show probability because such probabilities are difficult to estimate from historical data. We comment on this issue in §2 based on an analysis of data from a large health system that had low no-show rates. Thus, our approach is suitable for health systems with low no-show rates. For the problem features mentioned above, we show that certain types of information that may be retrieved from existing Web-based appointment request systems can be used to estimate patients' preferences and improve booking decisions. Our approach may be viewed as an application of the Bayesian learning approach for directly estimating empirical distributions of patient acceptance probabilities (e.g., see Carlin and Louis 2000). Our booking algorithm is a two-step process based on a partial characterization of the optimal booking decisions.

In the remainder of this section, we compare and contrast our approach with other approaches used to design appointment systems. A detailed review of relevant operations research (OR) literature can be found in Gupta and Wang (2008). Commonly used appointment systems can be categorized into four main types: (1) traditional systems that accept any booking request so long as the requested slot is open when the booking request is made, (2) carve-out systems that reserve a certain amount of capacity for specific procedures or urgent services, (3) advanced access (or open access) systems that accommodate patients' appointment requests on the day they call, and (4) hybrid approaches that accommodate both advance-book and same-day appointments. The traditional system allows each open slot to be booked by any patient who happens to be the first person to request it. This approach usually results in large backlogs of appointments for popular physicians as well as a significant spoilage of slots (Savin 2006). Same-day requests are often deflected to urgent care clinics, sent to emergency rooms, or double booked. Because of these shortcomings, some clinics choose a carve-out approach in which a certain amount of capacity is reserved for later-arriving patients. Once available capacity drops to the reservation level, a variety of rules are used to release this capacity for specific procedures or urgent-need patients. The urgency of each patient's needs is determined by a triage nurse. Nonurgent patients generally cannot obtain same-day appointments (Murray and Berwick 2003).

Table 1 Literature Analysis

	1. Objectives				2. Class					3. Assumptions					4. Criteria				
Study	(a)	(b)	(c)	(d)	(a)	(b)	(c)	(d)	(e)	(a)	(a′)	(b)	(c)	(d)	(a)	(b)	(c)	(d)	(e)
Adaptive appointment system	—	v, vi, viii	✓	✓	✓	✓	—	—	✓	1, 2	I	✓	✓	✓	✓	✓	✓	—	—
Gupta and Wang (2008)	—	v, viii	—	✓	✓	✓	—	—	✓	4	—	✓	✓	✓	✓	—	✓	—	—
Rohleder and Klassen (2000)	—	vi	—	✓	—	—	—	✓	—	4	—	✓	✓	✓	✓	—	—	—	—
Liu et al. (2010)	—	vii	—	✓	—	—	✓	—	—	3	II	✓	✓	—	✓	—	—	—	—
Muthuraman and Lawley (2008)	i	vi	—	—	—	—	✓	—	—	2	I	✓	✓	—	✓	—	—	✓	✓
Cayirli et al. (2008)	ii	ix	—	✓	—	—	—	✓	—	1	I	✓	—	—	—	—	—	✓	✓
Klassen and Rohleder (1996)	—	vi	—	—	—	—	—	✓	✓	4	—	✓	✓	—	—	—	—	✓	✓
Robinson and Chen (2010)	i	—	—	✓	—	—	—	—	✓	1	I	—	✓	—	—	—	—	✓	✓
Kim and Giachetti (2006)	i	—	—	—	—	—	—	—	—	1	I	—	—	—	✓	—	—	—	—
Denton and Gupta (2003)	i, ii	—	—	—	✓	—	—	—	—	4	—	✓	✓	—	✓	—	—	✓	✓
Hassin and Mendel (2008)	ii	—	—	✓	—	—	—	—	—	1	I	✓	✓	—	—	—	—	✓	✓
LaGanga and Lawrence (2007)	i	—	—	✓	—	—	—	—	—	1	I	✓	✓	—	—	—	—	✓	✓
Kaandorp and Koole (2007)	ii	—	—	—	—	—	—	—	—	1	I	✓	✓	—	—	—	—	✓	✓
Robinson and Chen (2003)	ii	—	—	—	—	—	—	—	—	4	—	✓	✓	—	—	—	—	✓	✓
Weiss (1990)	ii, iv	—	—	—	—	—	—	—	—	4	—	✓	✓	—	—	—	—	✓	✓
Green and Savin (2008)	iii	—	—	—	—	—	—	—	✓	3	I	—	—	—	—	—	✓	✓	✓
Vanden Bosch and Dietz (2000)	ii, iv	—	—	✓	—	—	✓	✓	—	4	—	✓	✓	—	—	—	—	✓	✓
Wang (1999)	ii, iv	—	—	✓	—	—	—	✓	—	4	—	✓	✓	—	—	—	—	✓	✓

Note. A "✓" (respectively, "—") indicates that the corresponding attribute is included in (respectively, absent from) the study.

An advanced access system is designed to offer each patient an appointment with his or her PCP on the day he or she calls. In many cases, the implementation of an advanced access system allows patients to be seen sooner and improves clinics' operational efficiency (Murray and Tantau 2000). However, physicians are typically unable to cover all appointment requests that arise each day and push some demand to future days. In addition, some patients prefer to book appointments in advance, at a time and day of their choice, rather than call on the day they wish to see a doctor (Gerard et al. 2008, Parente et al. 2005, Salisbury et al. 2007). For reasons such as these, implementations of advanced access systems are not always successful (Murray et al. 2003).

Clinics that implement advanced access systems usually adopt hybrid approaches that allow both advance and same-day bookings. Gupta and Wang (2008) provide a model of a hybrid approach in the presence of patients' preferences upon assuming knowledge of the conditional probability that a patient belonging to physician l's panel, after calling in period t and observing the state of the appointment system s, will request an appointment for slot j of physician i, for each i, j, s, t, and l. The study shows that the optimal policy for a single-physician clinic is a threshold-type policy so long as patient-choice probabilities satisfy a weak condition. The authors also partially characterize the structure of an optimal policy for multiple-doctor clinics. This work provides insights into the importance of modeling patients' choices in the primary care setting. However, patient-choice probabilities are not easily obtained from appointment records, and patients generally do not have complete knowledge of the system state when requesting an appointment. We address both these issues in this paper.

In Table 1, we compare this study with some recent papers in the appointment scheduling (AS) literature in terms of (1) the objectives of the study, (2) patient classification scheme, (3) key model assumptions, and (4) performance criteria that drive parameter selection. Each major attribute is further divided into subattributes, which we describe next. Study objectives may consist of one or more of the following: clinic profile setup (1.a), booking decisions (1.b), learning/adaptive approach for improving booking decisions (1.c), and comparison of different system designs (1.d). Furthermore, clinic profile setup may be static or dynamic and include one or more of the following decisions: number of of appointments per slot/session/day (1.a.i), appointment intervals/start times (1.a.ii), panel sizes (1.a.iii), and sequencing groups of appointments (1.a.iv). The decisions at the appointment booking stage include whether to accept a patient's request (1.b.v), which slot to book (1.b.vi), which appointment day to book (1.b.vii), whether to reserve capacity for same-day/urgent demand (1.b.viii), and how to sequence individual appointments (1.b.ix).

Patient classification may be based on revenue/costs (2.a), patient preferences (2.b), no-show rates (2.c), service time distribution (2.d), and same-day versus advance-book requests (2.e). Classification typically helps improve capacity allocation decisions.

Key modeling assumptions concern no-show patterns (3.a), the decision stage at which no-shows affect AS design (3.a′), service time randomness (3.b), patients' punctuality (3.c), and patients' preferences (3.d). Patterns of no-shows may be homogeneous (3.a.1), patient characteristics dependent (3.a.2), lead time dependent (3.a.3), and zero no-shows (3.a.4). No-shows may be modeled at the clinic profile setup stage (3.a′.I) and/or appointment booking stage (3.a′.II). Performance criteria used to select AS parameters are revenue/cost (4.a), patient-PCP match (4.b), booking failure rate/utilization (4.c), patients' wait (4.d), and physicians' idle/overtime (4.e).

Studies reported in Table 1, except Liu et al. (2010), focus on single session/day appointment problems. Liu et al. (2010) assume that patients have no preference for a particular appointment day and that the clinic decides which day to book after taking into account system state and lead time dependent no-show probabilities. In the proposed adaptive appointment system, advance-book patients first pick a desired appointment date. Booking decisions are made separately for each day and depend on the combinations of physician and appointment time blocks that are deemed acceptable by patients on the chosen date. It also reserves capacity for same-day requests. The proposed approach is novel because it learns (1.c) and utilizes patients' preference information (2.b) in the booking process and because it prioritizes patient-PCP match (4.b).

Because our approach considers patients' preferences and learning, discrete choice models such as probit or logit models that have been studied extensively in economics, marketing, and OR literatures are also relevant. These methods usually derive choice probabilities from the assumed utility-maximizing behavior of individual decision makers. Each decision maker, upon receiving an offer of a choice set, selects one of the alternatives in the set. The individual choices are then aggregated to obtain group-level measures of choice, e.g., the probability that an arbitrary member of the group will choose a particular option in the choice set. McFadden (2001) and Train (2003) present extensive surveys of discrete choice models and Talluri and van Ryzin (2004) present customer-choice models in revenue management. The contrast between revenue management studies and our approach can be explained in terms of the ownership of the choice set and booking decisions. In the former, the choice set is determined by the service provider and customers decide which product to purchase, whereas in our framework each patient (customer) reveals an acceptable set of slots and the clinic (service provider) decides which slot to book.

The remainder of this paper is organized as follows. In §2, we present empirical evidence that supports the proposed model. Model formulation is presented in §3. Then we analyze properties of optimal booking decisions and present two heuristics to help clinics make real-time booking decisions in §4. Section 5 contains an evaluation of the impact of patients' preferences on different performance metrics, including those that are affected by no-shows and service time variability. Section 6 concludes the paper.

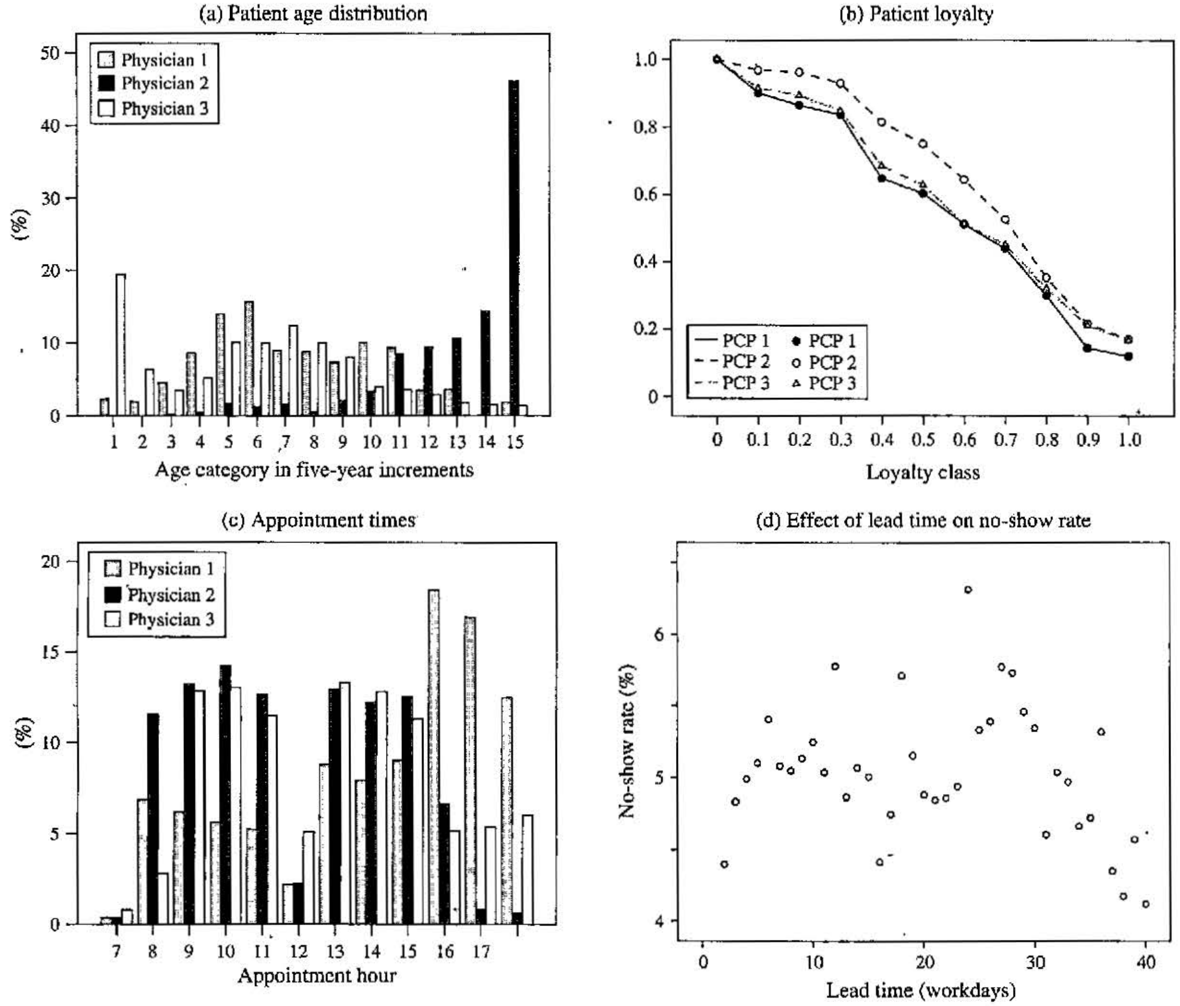
2. Analysis of a Health System's Appointment Data

We studied appointment processes of a large health system and obtained historical appointment data concerning 37 primary care clinics that operate in urban, suburban, and rural areas. We analyzed these data to guide the choice of model features in §3. The data covered appointment times with a range of 13 months that were booked over 18 months. It contained 1,461,948 records pertaining to 377,284 patients. The data elements were blinded medical record number (MRN), date and time of call and appointment, blinded PCP ID and provider ID (provider was the doctor that actually saw the patient for that appointment), age category, insurance status, five-digit zip code for each patient's address on file, and clinic location. Patient ages were divided into five-year intervals to obtain age categories.

The data reveal that both the panel size and its age distribution are different for each physician. Although we did not have access to revenue data, publicly available data support a strong correlation between patients' age and the different types and costs of services they need (U.S. Bureau of Labor Statistics 2008). This implies that both the demand and the expected revenue generated by patients of different panels are different. To make our point, we show distributions of patients' ages, loyalty (determined by the proportion of patient-PCP matched visits among that patient's past visits in 10% increments) for three physicians' panels in our data set in Figures 1(a)–1(c). Chi-square tests showed that the distributions of age, loyalty, and time preferences were significantly different for different panels (p-values were <0.0005 in each case). Moreover, the number of unique MRNs within the 13-month data for the three panels were 495, 719, and 1,631, respectively, which suggests that panel sizes also differ by physician.

We recognize that realized appointment times may not reflect true time preferences because booking success is also affected by the availability of requested slots. For example, it is possible that Physician 2 rarely works after 4 P.M. and that patients in his or her panel have adapted by accepting morning appointments. However, it is also possible that service providers respond to patients' needs. For example, families with teenagers and young adults often prefer appointment times after school hours, so as not to disrupt school attendance. Physician 1 may have chosen his or her work pattern with more availability in

Figure 1 **Evidence from the Analysis of Data from 37 Clinics**


(a) Patient age distribution

(b) Patient loyalty

(c) Appointment times

(d) Effect of lead time on no-show rate

Notes. In Figure 1(a), group 15 includes all patients who are 70 years of age or older. Figure 1(b) shows the proportion of panel patients that belonged to a higher loyalty class for patients with more than three encounters.

the afternoon in response to such demand. Irrespective of the underlying root causes, Figures 1(a)–1(c) serve to highlight that panels provide a reasonable means by which to define revenue classes and aggregate patients' preferences.

Next, we investigate the ability to predict patient-specific no-show probabilities from a data set such as ours. We first excluded canceled appointments from our data because the vast majority of the slots freed up in this fashion are rebooked. This resulted in a 1,171,950 encounters. Two factors that have been identified in previous studies are (1) history of no-shows and (2) appointment lead time (i.e., the time between the appointment request and the appointment date). It has been suggested that patients with a history of no-shows are more likely to be a no-show and that longer appointment lead times increase the likelihood of no-shows (see Dove and Schneider 1981, Lee et al. 2005, Gallucci et al. 2005, Whittle et al. 2008). Figure 1(d) shows that appointment delays are not significantly correlated with no-show rates in our data (Pearson correlation test shows no significant correlation with p-value > 0.4). A similar conclusion is also reached in Snow et al. (2009), Starkenburg et al. (1988), Irwin et al. (1981), Fosarelli et al. (1985), Neinstein (1982), and Dervin et al. (1978).

Turning to the history of no-shows, our data contained appointment times that ranged over 13 months. Therefore, we normalized the number of appointments per patient to a yearly basis and found that more than 75% of the patients in our data had fewer than four appointments per year, which would make it difficult to estimate individuals' no-show probabilities reliably. We believe such estimation problems could arise in many practical settings.

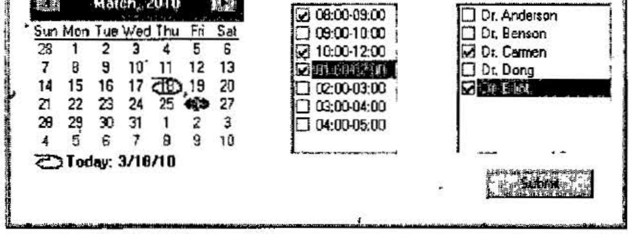
Finally, the overall no-show rate for the 37 clinics is 4.06% for all appointment and 2.97% for patient-PCP matched appointments. The overall patient-PCP match rate was 45.7%. This implies that there may be a substantial opportunity to reduce no-show rates further by increasing patient-PCP match rates, which the adaptive appointment system is designed to do.

3. Model Formulation and Assumptions

The vast majority of large health systems operate call centers where patients call to book appointments. With the adoption of electronic medical record (EMR) systems, however, many health systems are also able to provide a parallel Web-based option to patients for requesting nonurgent appointments. Patients are instructed to call if their needs are urgent. Similar instructions may also apply for special appointments such as physical exams and in-office procedures that take more time and for which physicians reserve specific slots in their daily schedule. It is generally believed that Web-based systems will become the primary means by which patients book nonurgent appointments in the future. Therefore, our model assumes the availability of real-time data from a Web-based system. We illustrate the types of information that can be obtained from existing Web interfaces in a mock-up in Figure 2. This mock-up is fashioned after existing systems familiar to the authors. However, it is not an exact replica of any particular system.

In the mock-up, a patient indicates a preferred appointment date and acceptable combinations of physicians and time blocks. Clinics use time blocks rather than individual time slots because patients tend to have similar acceptance rates for time slots within each half-hour or one-hour time block. Note that our formulation allows clinics to choose arbitrary block size and number of slots in each block. That is, appointment lengths may not be uniform and may depend on anticipated service-time class and no-show rates. Upon receiving a patient's request, the clinic considers any checked combination of the blocks of time and physicians to be acceptable to the patient on the chosen day of appointment request. The clinic either books an appointment in one of the combinations indicated by the patient or responds that none of the requested combinations are available. Patients are encouraged to try a different date if their request is denied.

Figure 2 A Web-Based Patient-Clinic Interface



Sun	Mon	Tue	Wed	Thu	Fri	Sat
28	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	[illegible]	27
28	29	30	31	1	2	3
4	5	6	7	8	9	10

Today: 3/18/10

☑ 08:00-09:00
☐ 09:00-10:00
☑ 10:00-12:00
☑ 01:00-02:00
☐ 02:00-03:00
☐ 03:00-04:00
☐ 04:00-05:00

☐ Dr. Anderson
☐ Dr. Benson
☑ Dr. Carmen
☐ Dr. Dong
☑ Dr. Eliot

To increase clinics' flexibility in scheduling patients in a manner that maximizes patient-PCP match and revenue, patients are asked to provide their acceptable sets but not rank their preferences among the acceptable combinations. If patients were asked to rank their preferences, clinics would be obligated to book appointments in the most preferred and available slots first, which would prevent them from keeping more capacity available in more popular combinations.

The proposed adaptive appointment system has two components—a component that updates estimates of acceptance probabilities and a component that makes booking decisions after receiving patients' requests. Below we describe each component in a separate section. Each section states model assumptions first and then presents a formulation. We show in §3.2 that for making booking decisions, clinics only need to estimate panel-level acceptance probabilities. Therefore, §3.1 deals only with panel-level probabilities. Throughout the paper, we use m to denote the number of physicians and b to denote the number of time blocks on a workday.

3.1. Learning Acceptance Probabilities

Given that Web-based options similar to that in Figure 2 are in existence today, our approach models each patient's preferences in terms of acceptance probabilities. For each physician indexed i and time-block indexed j, the probability that the kth patient in physician l's panel will find combination (i, j) acceptable is denoted by $p_{i,j}^{l,k}$. Furthermore, we assume that physician and time preferences are independently captured by terms $\alpha_i^{l,k}$ and $\beta_j^{l,k}$, with $p_{i,j}^{l,k} = \alpha_i^{l,k}\beta_j^{l,k}$. This is consistent with the implied decomposition of physician and time preferences in Figure 2. From a technical viewpoint, it is possible to generalize our approach to situations where acceptance probabilities do not have the multiplicative form that we assume. However, we did not find any evidence to suggest that the multiplicative form is an unreasonable assumption.

We assume that each patient reveals his or her true acceptable set in each request (prior to receiving an appointment) and that each booking attempt is an independent draw from a patient's preference distribution. The first assumption is based on the argument that if a patient's utility from booking an appointment for a particular physician and time-block combination is higher than the utility from not booking an appointment, then the patient will include that combination in his or her acceptable set. The second assumption is based on anecdotal evidence that patients' time preferences vary by calendar day.

As shown in §3.2, booking decisions depend only on panel-level acceptance probabilities $p^l_{i,j} = \alpha^l_i \beta^l_j$, where α^l_i and β^l_j are the physician i and block j acceptance probabilities for panel l. We propose direct estimation of these probabilities; see §A of the online supplement for details. This is not the only way to estimate patients' choices. A parallel utility-based model can be constructed as well, and subsequently its parameters can be estimated. It can be shown that the strong independence of attributes assumed by clinics (as implied by Figure 2) leads to an equivalent model with similar estimation effort. We provide details of this alternative approach and compare it to the proposed approach in §B of the online supplement.

3.2. Making Appointment Booking Decisions

At the time of booking appointments, clinic profiles exist for all future workdays on which appointments may be booked. The model that is used to obtain a partial characterization of optimal booking decisions also assumes that patients and physicians are punctual, patient no-show rates are negligibly small, and all same-day patients call before the start of the day. The existence of clinics with relatively low no-show rates has been documented in the literature (see, e.g., Cayirli and Veral 2003) and supported by our data (see §2). However, some clinics are also reported to have high no-show rates, and the proposed adaptive appointment system design may not be the best choice for such clinics. Clinics typically count all requests received within 24 hours before the start of a workday as same-day demand. This makes it reasonable to assume that same-day demand is realized just before the start of each workday.

Our model considers each workday's appointment booking problem separately. This is justified, in part, by assumptions in §3.1 that patients' preferences may differ by calendar day and that patients are encouraged to try other dates if earlier booking attempts fail. Because the clinic profile is assumed known, the clinic's objective function considered in this section does not include patient wait times and physician overtime, which are caused by service time variability, and choices of appointment lengths and overbooking. However, when evaluating different heuristics in §5, we also compare these metrics for different approaches.

The following information is needed to make booking decisions: (1) patients' acceptance probabilities and arrival rates at the panel level; (2) clinic's average revenue, by panel, of each PCP matched/mismatched appointment; (3) average costs of delaying an advance-book and same-day appointment; and (4) each physician's same-day demand distribution. We define the inputs to the booking decision model in Table 2 and explain model features below.

Table 2 Inputs of the Booking Decision Model

X_i:	Same-day demand for physician i
X:	Total same-day demand; $X = \sum_{i=1}^m X_i$
$\kappa = (\kappa_{ij})$:	Matrix of capacity of each combination (i, j) of physician and time-block combination
$s = (s_{ij})$:	Matrix of number of appointments that have been booked for (i, j) combination
$\bar{\kappa}_i$:	Physician i's capacity; $\bar{\kappa}_i = \sum_{j=1}^b \kappa_{ij}$
$\bar{\bar{\kappa}}$:	Clinic's capacity; $\bar{\bar{\kappa}} = \sum_{i=1}^m \bar{\kappa}_i$
$\bar{s}_i$:	Physician i's booked appointments; $\bar{s}_i = \sum_{j=1}^b s_{ij}$
$\bar{\bar{s}}$:	Number of booked appointments at the clinic level; $\bar{\bar{s}} = \sum_{i=1}^m \bar{s}_i$
$r^l_{1,l}$:	Average revenue of a PCP matched advance-book panel l appointment
$r^l_{1,i}$:	Average revenue of a PCP mismatched advance-book panel l appointment, $i \neq l$
r_2:	Average revenue of a PCP matched same-day appointment
r_2':	Average revenue of a PCP mismatched same-day appointment
π_t:	Time-dependent average penalty induced by a failure to satisfy an advance-book request
c:	Average cost of insufficient same-day capacity
τ:	Number of potential advance-book appointment request epochs for a particular workday
t:	An arbitrary advance-book appointment request epoch
λ^l_t:	The probability of having an arrival from physician l's panel at epoch t
α^l_I:	The probability that an arbitrary panel l patient's set of acceptable physicians is I
β^l_J:	The probability that an arbitrary panel l patient's set of acceptable time blocks is J
$p^l_{I,J}$:	The probability that a panel l patient's acceptable combinations are (I, J); $p^l_{I,J} = \alpha^l_I \beta^l_J$

In reality, patients' true acceptance probabilities are unknown. Therefore, we propose to use the best available estimates of acceptance probabilities at each decision epoch (from the updating procedure of §3.1). This can be justified because the updating procedure is independent of booking decisions and converges quickly to the true acceptance probabilities. Unit revenues from each booked appointment satisfy the following inequalities: $r^l_{1,l} \geq r^i_{1,l}$ for each l and $i \neq l$, and $r_2 \geq r_2'$; see O'Hare and Corlett (2004) for supporting evidence. Same-day visit revenue does not depend on panel index because these appointments are usually for acute symptoms for which the treatments offered are more likely to be independent of panel characteristics.

The time between the start of advance-book requests for each workday and that workday is divided into τ intervals such that the probability of obtaining more than one arrival in each interval is infinitesimally small. Time is counted backward. Specifically, advance bookings occur from period τ to period 1 and all same-day bookings occur in period 0. Because patients who try to book appointments must have at least one acceptable combination, neither I nor J is an empty set.

In our model, the penalty for denying a patient's appointment requests n time periods before the appointment date is assumed to be different (smaller) from the penalty for doing so $(n+d)$ period before the appointment date, where $d \geq 1$. This makes sense for two reasons. First, patients who call well in advance are generally more particular about their time and physician preferences. The clinic harmonizes its booking practices with this behavior by setting $\pi_n \leq \pi_{n+d}, \forall d \geq 1$. Second, this assumption leads to a fair allocation of slots as we shall show in §4.1.1. In particular, this means that if a physician l patient's request for a particular combination is denied in period $(n+d)$ given a particular system state, then another physician l patient's request for that combination will be denied in period n as well for the same system state.

Researchers have made a variety of attempts in recent years to estimate the cost of patient waiting (terms c and π_t in our models). For example, Yabroff et al. (2005) and Russell (2009) estimate the cost of patient waiting based on wage rates whereas Robinson and Chen (2011) provide an observation-based method for estimating the relative cost of customer waiting time. Clinic administrators can either perform a study similar to those reported in the literature or use the results in these papers to guide their choice of patient waiting costs.

We are now ready to set up the clinic's revenue function for the appointment booking problem. For this purpose, it helps to conceptualize the availability of different levels of information about the arriving patient. Specifically, we identify three levels of information and label them the (1) patient-level, (2) panel-level, and (3) clinic-level. At the patient-level, known information includes the patient label (l, k) (i.e., the arrival in period t is the kth patient in physician l's panel); the system state s; and the patient's acceptable set (I, J). In contrast, panel-level information consists of the arriving patient's panel index and the system state, whereas the clinic-level information includes only the system state.

We use notation $u_t^{l,k}(s)$, $u_t^l(s)$, and $u_t(s)$ to denote the maximum expected revenue from time t onward given patient-level, panel-level, and clinic-level information, respectively. With this notation in hand, the following recursive relationship holds:

$$u_t^{l,k}(s) = \max_{(i,j)\in(I,J)} \{r_{1,l}^i + u_{t-1}(s + e_{i,j}), u_{t-1}(s) - \pi_t\}, \quad (1)$$

where $e_{i,j}$ is an $m \times b$ matrix with the (i, j)th entry equal to 1 and all other entries equal to 0. The first two terms in the curly brackets above capture the benefit of accepting the patient's request for the (i, j) physician and time-block combination, whereas the next two terms capture the benefit of rejecting the patient's requests. Equation (1) suggests that the clinic should accept a slot, say $(i^{\text{opt}}, j^{\text{opt}})$, among the arriving patient's requests (I, J) for which $r_{1,l}^i + u_{t-1}(s + e_{i,j}) \geq u_{t-1}(s) - \pi_t$ and the clinic's revenue is maximized. That is, $(i^{\text{opt}}, j^{\text{opt}}) \in \arg\max_{(i,j)\in(I,J)}\{r_{1,l}^i + u_{t-1}(s+e_{ij})\colon r_{1,l}^i + u_{t-1}(s+e_{i,j}) \geq u_{t-1}(s) - \pi_t]\}$. Ties may be broken arbitrarily.

Using a logic similar to that behind Equation (1), revenue functions with panel- and clinic-level information can be written as follows:

$$u_t^l(s) = \sum_{\text{all }(I,J)} p_{I,J}^l \max_{(i,j)\in(I,J)} \{r_{1,l}^i + u_{t-1}(s+e_{i,j}),\ u_{t-1}(s) - \pi_t\}. \quad (2)$$

$$u_t(s) = \sum_{l=1}^{m} \lambda_t^l u_t^l(s) + \left(1 - \sum_{l=1}^{m} \lambda_t^l\right) u_{t-1}(s). \quad (3)$$

Upon comparing (1) with (2), we observe that the decision rule for accepting or denying a particular (i, j) request is the same for all patients in the same panel. This arises because the arriving patient's information does not affect the clinic's valuation of its benefit from saving each combination for future arrivals. Similarly, upon comparing (2) and (3), we observe that the revenue function with clinic-level information is a weighted sum of revenue functions with panel-level information.

Next, we turn to the revenue function corresponding to same-day requests, which has a different form because all same-day requests are assumed to arrive just before the start of the workday. In the model, we assume that we can optimally match them with available capacity. Therefore, it suffices to define the same-day revenue function with clinic-level information only, as shown below.

$$\begin{aligned} u_0(s) = E\Bigg\{ & r_2 \sum_{i=1}^{m} \min\{X_i, (\bar{\kappa}_i - \bar{s}_i)\} \\ & + r_2' \min\left\{ \sum_{i=1}^{m} (\bar{\kappa}_i - \bar{s}_i - X_i)^+, \sum_{i=1}^{m} (X_i - \bar{\kappa}_i + \bar{s}_i)^+ \right\} \\ & - c\left(\sum_{i=1}^{m} X_i - \sum_{i=1}^{m} (\bar{\kappa}_i - \bar{s}_i) \right)^+ \Bigg\}. \end{aligned} \quad (4)$$

In Equation (4), the first term is the expected revenue from same-day patient-PCP matched visits, the second term is the expected revenue from mismatched visits, and the third term is the expected cost due to excess same-day demand.

4. Analysis

The formulation of the appointment booking decision problem in §3.2 has a high-dimensional state space, which precludes the use of real-time and stored solutions of the stochastic dynamic program for every system state in each period. In what follows, we show

Table 3 An Ordering of Blocks from the Clinic's Perspective

Case 1						Case 2					
t	s	Block 1	Block 2	Block 3	Block 4	t	s	Block 1	Block 2	Block 3	Block 4
13	(3, 0, 0, 0)	1	1	1	2	13	(0, 0, 0, 3)	1	2	3	4
8	(3, 2, 0, 4)	1	1	NB	—	8	(4, 2, 0, 3)	—	1	1	NB
3	(3, 3, 3, 1)	NB	NB	NB	NB	3	(1, 3, 3, 3)	NB	NB	NB	NB

with the help of an example that there is no pattern or structure to booking decisions.

Consider a single physician clinic with four slots each in four time blocks. We omit the physician label for simplicity. The panel-level acceptance probabilities for these blocks are $\beta = (0.1, 0.2, 0.6, 1)$. Other parameters are $(r^1_{1,1}, r_2, c, \pi_t, \lambda, \mu, \tau) = (6, 6, 10, 5, 0.7, 5, 16)$, where $\pi_t = \pi$ and $\lambda_t = \lambda$ for each $t = 1, \ldots, \tau$, and μ is the arrival rate for the same-day demand, which is assumed to be Poisson distributed. The expected total demand is 16.2, whereas the total capacity is 16. Because this problem has a small state space, we are able to solve the underlying stochastic dynamic program to obtain an ordering of slots from the clinic's perspective for each system state and decision epoch. If the optimial decision is to deny the request for time-block j in every decision epoch at and after time t, then we say the system is in a no-book (NB) state for block j. In Table 3, the best slot to book for an arriving patient is the highest ranked available slot that is acceptable to the patient and that is not designated NB.

We use two cases, each with three examples, to illustrate how an optimal decision may depend on the remaining capacity, time preferences of future arrivals, and the acceptable set of the next appointment request (see Table 3). In the first example, $t = 13$, and the total remaining capacity is 13. For Case 1 (state $s = (3, 0, 0, 0)$), the clinic's first choice is to book either block 1, 2, or 3, and the second choice is to book block 4. For Case 2 (state $s = (0, 0, 0, 3)$), the rank order of available time blocks is as follows: $1 \succ 2 \succ 3 \succ 4$. That is, a patient whose acceptable set includes blocks 1 and 3 may be booked into either block 1 or 3 in Case 1, but only in block 1 in Case 2. In the second example, when $t = 8$ and the total remaining capacity is 7, block 3 (respectively, block 4) is a NB block if $s = (3, 2, 0, 4)$ (respectively, $s = (4, 2, 0, 3)$), and a patient whose acceptable set includes blocks 1 and 3 will be booked into block 1 in Case 1 and block 3 in Case 2. In the third example, $t = 3$, the total remaining capacity is 6, and the clinic is in a no-book state for all blocks for both cases.

These examples show that the optimal decision depends in a nontrivial fashion on the vector of remaining capacities, the index of the decision epoch, and acceptable sets. In addition, certain blocks are designated NB, which means that they are reserved for future same-day demand. The complexity of decisions increases when there are multiple physicians. Therefore in the next section, we characterize certain properties of optimal booking decisions, which are subsequently used to construct heuristic solutions.

4.1. Properties of Optimal Booking Decisions

For modeling convenience, we may think of the booking decision as a two-step process. Given that a panel l patient makes a booking request in period t with acceptable set (I, J), the clinic in the first step identifies sets of no-book states $S_t^{i,l}$ for each $i \in I$, i.e., states in which a panel l patient's request for an appointment with physician i is denied irrespective of J. If the current state is in the set of no-book states for all physicians in I, then the requesting patient is asked to try another date. However, if the process proceeds to the second step, then the clinic decides which of the acceptable and available appointments to book. That is, in stage two, the clinic ranks available (i, j) combinations in (I, J). It is also possible at this stage to deny a patient's request. Denials may happen either when the intersection set of patients' requested appointments and available appointments is empty or when the clinic earns a greater expected revenue by protecting specific appointments requested by the patient for future arrivals. This two-step process can be operationalized by developing procedures for identifying no-book states and procedures for rank ordering requested appointments (from the clinic's viewpoint) when the system state is not in the no-book set. We obtain partial solutions for these two tasks in §§4.1.1 and 4.1.2, which form the basis for the heuristics proposed in §4.2.

4.1.1. No-Book States. In this section, we obtain $S_t^{i,l}$ for $t = 1$, and for $t > 1$ we identify a set of states $\hat{S}_t^{i,l}$ such that $\hat{S}_t^{i,l} \subseteq S_t^{i,l}$. We also show that for $t > 2$, when $\pi_t \geq \pi_{t-1}$ (which we assume), $\hat{S}_t^{i,l} \subseteq \hat{S}_{t-1}^{i,l}$. That is, patients who call earlier encounter smaller sets of no-book states.

Consider a time t decision epoch when the kth panel l patient makes a booking request, and assume that there will be no more future advance-book requests after this decision epoch. Let (I, J) denote this patient's acceptable set of appointments. Then the clinic's decision problem is encapsulated in the following revenue function:

$$u_t^{l,k}(s) = \max_{(i,j)\in(I,J)} \{r^i_{1,l} + u_0(s+e_{i,j}),\, u_0(s) - \pi_t\}.$$

The above revenue function is identical to (1) when $t = 1$. For $t > 1$, the difference is that the right-hand side contains u_0 instead of u_{t-1} because we assume no advance-book arrivals after period t. The clinic should consider booking an appointment for a panel l patient if there is at least one (i, j) combination such that $u_0(s) - u_0(s+e_{i,j}) \le r^i_{1,l} + \pi_t$.

Let $F_i(\cdot)$ and $F(\cdot)$ denote the CDF of physician i's and clinic's same-day demand, respectively. Upon rearranging the terms in Equation (4), we obtain $u_0(s) = r_2E(X) - r_2\sum_{l=1}^m E(X_l - \bar{\kappa}_l + \bar{s}_l)^+ + r_2'\sum_{l=1}^m E(\bar{\kappa}_l - \bar{s}_l - X_l)^+ - r_2'E(\bar{\bar{\kappa}} - \bar{\bar{s}} - X)^+ - cE(X - \bar{\bar{\kappa}} + \bar{\bar{s}})^+$. Let $\bar{s}_{-i}$ be the total number of slots booked for all physicians except physician i. After a few more steps of algebra, the marginal benefit for reserving a physician i's slot in the last period can be further simplified to

$$\begin{aligned}\Delta(\bar{s}_i, \bar{s}_{-i}) &\doteq u_0(s) - u_0(s+e_{i,j})\\ &= r_2 + c - (r_2 - r_2')F_i(\bar{\kappa}_i - \bar{s}_i - 1)\\ &\quad - (r_2' + c)F(\bar{\bar{\kappa}} - \bar{s}_i - \bar{s}_{-i} - 1).\end{aligned}$$

Same-day patients do not have time preferences. Therefore, the value of $\Delta(\bar{s}_i, \bar{s}_{-i})$ does not depend on which block j is being considered.

Let $\bar{a}_i$ and $\bar{a}_{-i}$, respectively, be the number of available slots of physician i and the clinic not including i. Because CDF is a nondecreasing function, for any fixed value of $\bar{s}_{-i}$, $\Delta(\bar{s}_i, \bar{s}_{-i})$ increases in $\bar{s}_i$. Therefore, there exists a protection level $a^l_i(s) = \min\{\bar{a}_i\colon \Delta(\bar{\kappa}_i - \bar{a}_i, \bar{s}_{-i}) > r^i_{1,l} + \pi_t$ given $\bar{a}_i \ge 0$ and fixed $\bar{s}_{-i}\}$ such that no physician i slot should be booked for a panel l patient if $\bar{k}_i - \bar{s}_i$ is less than $a^l_i(s)$. Similarly, for any fixed value of $\bar{s}_i$, $\Delta(\bar{s}_i, \bar{s}_{-i})$ increases in $\bar{s}_{-i}$, which implies that there exists a protection level $a^l_{-i}(s) \doteq \min\{\bar{a}_{-i}\colon \Delta(\bar{s}_i, \bar{\kappa}_{-i} - \bar{a}_{-i}) > r^i_{1,l} + \pi_t$ given $\bar{a}_{-i} \ge 0$ and fixed $\bar{s}_i\}$ such that no physician i slot should be booked for a panel l patient if the remaining clinic capacity, not counting physician i, is less than $a^l_{-i}(s)$. Similar protection levels also exist with convex cost of unmet same-day demand (see §C of the online supplement for details).

Proposition 1. *Given a panel l patient's booking request for an appointment with physician i at decision-epoch t and no more advance-book requests after t, the set of no-book states is $\hat{S}^{i,l}_t = \{s\colon \bar{\kappa}_i - \bar{s}_i \le a^l_i(s)\}$.*

An immediate corollary of Proposition 1 is that $S^{i,l}_1 = \hat{S}^{i,l}_1$ for each (i, l) pair because after $t = 1$, there are indeed no more advance-book requests. Also, the booking decision for a type l arrival regarding a physician i's slot depends on the current state of the clinic only through $a^l_i(s)$ and $a^l_{-i}(s)$, which leads to a two-dimensional booking profile. Gupta and Wang (2008) obtain a similar result when advance-book revenue is independent of panel index. However, in their paper, all open slots of a physician are equally valued and are made available to the arriving patient so long as the remaining capacity is higher than the protection level. In our framework the protection level serves only as an availability check in the first step of the booking process. We refer the reader to §D of the online supplement for an example that identifies no-book states for a two-physician clinic.

Proposition 2. *The set of no-book states assuming no more advance-book requests is a subset of the true set of no-book states, i.e. $\hat{S}^{i,l}_t \subseteq S^{i,l}_t$, and if π_t is nondecreasing in t, then $\hat{S}^{i,l}_t \subseteq \hat{S}^{i,l}_{t-1}$.*

A formal proof of Proposition 2 is included in §E of the online supplement. On an intuitive level, the first part of this proposition holds because when there are no more advance-book requests, there are no competing advance-book requests for the same slot. The only demand for a slot is from same-day requests. Therefore, the protection level after making the assumption of no more advance-book requests is never greater than the true protection level when advance-book requests do occur. The second result follows because higher cost of denying a patient's request leads to lower protection levels.

4.1.2. Rank Order of Appointment Slots. Consider a single-physician clinic with block j capacity κ_j and state s_j. In this section, a block is deemed available when $s_j < \kappa_j$ and the current state s is not in the set of no-book states. We analyze this simpler problem instance because in this case an advance-book patient's request is denied only when it is optimal to reserve capacity for same-day patients. This happens because each advance-book appointment results in the same revenue. This means that when there is a single physician labeled l, $\hat{S}^l_t = S^l_t$ for each t. A formal argument is provided in §F of the online supplement.

The clinic faces the problem of deciding which of the requested appointments in the acceptable set J to book. We consider only those instances in which for at least one $j \in J$, $s_j < \kappa_j$. If there is at least one block $j \in J$ such that $\kappa_j - s_j > \tau - t$ and state s is not a no-book state, then it is straightforward to show that the clinic can book the patient in block j without affecting its ability to book future patients because all those patients still have a chance to book block j. Similarly, if the system is not in a no-book state and there is only one $j \in J$ such that $\kappa_j - s_j > 0$, then a slot in block j should be booked. This means that a clinic needs guidance only when $\kappa_j - s_j \le \tau - t$ for all j, and there is more than one acceptable block with remaining capacity. We focus on such cases in the remainder of this section.

Let $\phi(s) = 1 - \prod_{j: s_j < \kappa_j}(1 - \beta_j)$ be the probability that at least one time block is acceptable to an arriving patient and has remaining capacity when system state is s and consider a decision epoch after which the clinic expects at most one additional advance-book arrival. Suppose that the patient's acceptable set includes blocks j and k, both of which have at least one open slot. The clinic may then base its decision on the value of $\phi(s)$. The higher the value of $\phi(s)$, the higher the chance of satisfying a future arrival's request. The clinic may consider the relative magnitudes of $\phi(s + e_j)$ and $\phi(s + e_k)$ when deciding which block to book. When both j and k have exactly one remaining slot, it may also consider the relative magnitudes of β_j and β_k. The above informal arguments are formalized in Proposition 3; a proof of Proposition 3 can be found in §G of the online supplement.

Proposition 3. *When choosing between blocks j and k, the clinic prefers to book in block j so long as $\beta_j < \beta_k$ and $\phi(s + e_j) > \phi(s + e_k)$. Mathematically, if $\beta_j < \beta_k$ and $\phi(s + e_j) > \phi(s + e_k)$, then $u_t(s + e_j) \geq u_t(s + e_k)$ for all $t \geq 1$.*

Proposition 3 gives a partial ordering of acceptable time blocks of a single physician. It suggests that among the available and acceptable combinations, a particular block is more likely to be a clinic's top choice if it has greater remaining capacity and if assigning a slot in that block has a smaller effect on the clinic's ability to meet future demand. It is difficult to show a similar result when multiple physician's slots are being compared because of different time-preference patterns of patients belonging to different panels and because of different revenue rates. However, we use insights from Proposition 3 to develop a metric, $q^t_{i,j}$, to rank order available and acceptable blocks from the clinic's viewpoint. This metric is used in heuristic rules for making booking decisions (see §4.2).

We define $q^t_{i,j}$ as a measure of popularity of each (i, j) combination when $\kappa_{i,j} - s_{i,j} > 0$ in period t as follows:

$$q^t_{i,j} = \sum_{z=1}^{t-1} \lambda^i_z p^i_{i,j} / (\kappa_{i,j} - s_{i,j}). \qquad (5)$$

The numerator of (5) is the expected number of times that (i, j) combination will be included in the acceptable set by panel i patients in the remaining advance-book periods, and the denominator is the remaining capacity of the (i, j) combination. The popularity measure does not account for anticipated demand from nonpanel patients because both heuristics proposed in the next section give priority to achieving high patient-PCP match.

4.2. Heuristic Approaches

We present two heuristics (H1 and H2) that utilize the popularity index in (5) and give priority to matching patients with their PCP. In describing the heuristics below, we assume that a panel l patient has tendered an appointment request with acceptable set (I, J) and that the system state is s. The booking decisions generated by H1 and H2 are appealing on an intuitive level for two reasons. First, because $r^l_{1,l} \geq r^i_{1,l}$ for $i \neq l$, and there are a variety of other benefits of matching patients with their PCP, it is reasonable to strive for a high patient-PCP match. Second, because any combination in (I, J) is acceptable to the patient who tendered that request, it can be beneficial to reserve slots with higher $q^t_{i,j}$ values for future patients.

H1 books an appointment so long as the intersection set of open slots and (I, J) is not empty. That is, H1 assumes that the set of no-book states is empty. It attempts to first book a patient with his or her PCP. If multiple PCP slots are open and included in J, then H1 books a slot with the smallest value of $q^t_{l,j}$. If none of the acceptable PCP slots are available, then H1 books the slot with the smallest value of $q^t_{i,j}$, $i \neq l$, among all non-PCP slots in the acceptable set.

H2 calculates $\hat{S}^{i,l}_t$ and only considers physicians i included in I for which $s \notin \hat{S}^{i,l}_t$. Upon ascertaining that $s \notin \hat{S}^{l,l}_t$, H2 attempts to first book a patient with his or her PCP. If multiple PCP slots are open and included in J, then H2 books a slot with the smallest value of $q^t_{l,j}$. If none of the acceptable PCP slots are available, then H2 books the slot with the smallest value of $q^t_{i,j}$, $i \neq l$, among all non-PCP slots in the acceptable set for which $s \notin \hat{S}^{i,l}_t$. The key difference between H1 and H2 is that H1 does not protect slots for same-day demand.

4.3. Tests of Performance of H1 and H2

For the single physician example presented at the beginning of §4, the expected daily revenue evaluated at the beginning of the advance-book period, when the system starts empty and we use H1 and H2 to make booking decisions, turns out to be 99.76% and 99.81%, respectively, of the optimal expected revenue. This suggests that the performance of H1 and H2 is reasonable in problem instances with a single physician.

However, problems with multiple physicians are not tractable and the corresponding optimal expected revenue cannot be determined exactly. Therefore, we compare the expected revenues obtained from the two heuristics to the expected maximum attainable revenue, which is an upper bound. To calculate this bound, we simulate sequences of advance-book and same-day arrivals and then use an integer program, shown in §4.3.1, to calculate the maximum attainable revenue for each sample path.

4.3.1. Maximum Attainable Revenue. Let K (respectively, K_l) be the set of decision epochs with an arrival from an arbitrary panel (respectively, panel l). In addition, let $a^t_{i,j} = 1$ if (i, j) physician and time-block combination is acceptable to the advance-book patient who arrives in period t and $a^t_{i,j} = 0$ otherwise. Let x_l denote the realized same-day demand from panel l. The decision variables are $y^l_{i,j}$ and $o^t_{i,j}$, where $y^l_{i,j}$ is the number of slots that belong to the (i, j) combination and that are assigned to same-day panel l patients. Furthermore, $o^t_{i,j} = 1$ if the clinic assigns a slot of the (i, j) combination to the patient who arrives in period t, and $o^t_{i,j} = 0$ otherwise. Let $M(l)$ be the set of physicians excluding l. Then the maximum attainable revenue of a sequence of arrivals can be obtained by solving the following integer program.

$$\max \sum_{l=1}^{m}\sum_{j=1}^{b}\sum_{t\in K_l} r^l_{1,l}o^t_{l,j} + \sum_{l=1}^{m}\sum_{i\in M(l)}\sum_{j=1}^{b}\sum_{t\in K_l} r^i_{1,l}o^t_{i,j}$$
$$-\sum_{l=1}^{m}\sum_{t\in K_l}\pi_t\Big(1-\sum_{i=1}^{m}\sum_{j=1}^{b}o^t_{i,j}\Big)+\sum_{l=1}^{m}\sum_{j=1}^{b}r_2y^l_{l,j}$$
$$+\sum_{l=1}^{m}\sum_{i\in M(l)}\sum_{j=1}^{b}r'_2y^l_{i,j}-c\Big(\sum_{l=1}^{m}x_l-\sum_{l=1}^{m}\sum_{i=1}^{m}\sum_{j=1}^{b}y^l_{i,j}\Big)$$

subject to

$$o^t_{i,j}\le a^t_{i,j} \quad \forall\, i=1,\ldots,m;\; j=1,\ldots,b;\; t\in K,$$
$$\sum_{i=1}^{m}\sum_{j=1}^{b}o^t_{i,j}\le 1 \quad \forall\, t\in K,$$
$$\sum_{t\in K}o^t_{i,j}+\sum_{l=1}^{m}y^l_{i,j}\le \kappa_{i,j} \quad \forall\, i=1,\ldots,m;\; j=1,\ldots,b,$$
$$\sum_{i=1}^{m}\sum_{j=1}^{b}y^l_{i,j}\le x_l \quad \forall\, l=1,\ldots,m,$$
$$o^t_{i,j}\in\{0,1\} \quad \forall\, i=1,\ldots,m;\; j=1,\ldots,b;\, t\in K, \text{ and}$$
$$y^l_{i,j}\ge 0 \quad \forall\, i=1,\ldots,m;\; j=1,\ldots,b;\, l=1,\ldots,m.$$

Using CPLEX 8.1 solver, the maximum attainable revenue for each sequence of advance-book and same-day arrivals in the examples reported in §4.3.2 was obtained in less than a second.

4.3.2. Results of Performance Tests. We tested H1 and H2 with the help of a five-factor design of experiments. The factors were (1) four clinic sizes ($m = 2, 4, 6$, and 8); (2) five clinic loads (expected demand/average capacity = 85%, 90%, 100%, 110%, and 115%); (3) two types of panel loads (homogeneous or heterogeneous); (4) four preference types (time dominant in Table 4, physician dominant in Table 5, moderate in Table 6, and no preferences); and (5) two levels of information accuracy (perfect or biased)—for a total of 320 different scenarios. We repeated the evaluation of the 320 scenarios under two cost structures: $c/\pi = 2$ and $c/\pi = 8$. A higher c/π ratio is appropriate for clinics that place a high priority on meeting same-day appointment requests. Results are summarized in Table 7 and Figure 3. They confirm that H1 and H2 are robust under a variety of different clinic environments. However, before discussing the results, we first describe the experimental setup in more detail below.

Table 4 Time Dominant

l	α^l_l	β^l_1	β^l_2	β^l_3	β^l_4
1	1	0.2	0.4	0.6	1
2	1	1	0.5	0.5	0.5
Odd $l \ge 3$	1	1	0.5	0.5	0.3
Even $l \ge 4$	1	0.3	0.5	0.5	1

Table 5 Physician Dominant

l	α^l_{odd}	α^l_{even}	β^l_j
Odd l	1	0.3	1
Even l	0.3	1	1

Table 6 Moderate

l	α^l_{odd}	α^l_{even}	β^l_{odd}	β^l_{even}
Odd l	1	0.4	1	0.4
Even l	0.4	1	0.4	1

A clinic may set up time blocks with different lengths and/or different number of appointment slots within a time block. For example, a clinic may divide physicians' morning sessions into three one-hour blocks, each with two 30-minute slots, and afternoon sessions into two two-hour blocks, each with three 20-minute appointments. On any given day, a particular physician's slots in each block may vary on account of staff meetings, training, variable work schedules, and differences in the number of work-in/overbook slots. To capture this variability, we assume that clinics have on average five slots within each of four daily blocks for each physician, but the actual number of slots within each block for each physician is independently sampled from a uniform $[4, 6]$ distribution.

Each physician's same-day demand is assumed to be independent and Poisson distributed with mean 6 (30% of the average capacity). Different levels of clinic load are simulated by choosing $\tau = 0.7\times$

Table 7 Aggregate Performance

			Relative revenue		PCP match		Advance-book failure		Spoilage	
c/π		Clinic load (%)	Mean	SD	Mean	SD	Mean	SD	Mean	SD
2	H1	85	0.992	(0.017)	0.958	(0.040)	0.007	(0.018)	0.123	(0.088)
		90	0.989	(0.021)	0.952	(0.041)	0.009	(0.022)	0.094	(0.083)
		100	0.981	(0.026)	0.943	(0.042)	0.013	(0.027)	0.047	(0.068)
		110	0.965	(0.033)	0.942	(0.041)	0.022	(0.033)	0.022	(0.055)
		115	0.958	(0.036)	0.943	(0.040)	0.027	(0.038)	0.013	(0.048)
	H2	85	0.992	(0.017)	0.958	(0.040)	0.010	(0.024)	0.124	(0.088)
		90	0.989	(0.020)	0.952	(0.041)	0.016	(0.035)	0.095	(0.082)
		100	0.983	(0.024)	0.942	(0.042)	0.037	(0.054)	0.050	(0.068)
		110	0.974	(0.028)	0.940	(0.039)	0.083	(0.077)	0.028	(0.057)
		115	0.971	(0.028)	0.939	(0.039)	0.108	(0.083)	0.021	(0.051)
8	H1	85	0.968	(0.100)	0.957	(0.040)	0.008	(0.020)	0.119	(0.085)
		90	0.955	(0.106)	0.950	(0.042)	0.010	(0.024)	0.093	(0.084)
		100	0.902	(0.161)	0.943	(0.041)	0.014	(0.026)	0.051	(0.071)
		110	0.807	(0.212)	0.941	(0.040)	0.022	(0.034)	0.021	(0.053)
		115	0.728	(0.250)	0.944	(0.038)	0.027	(0.037)	0.014	(0.047)
	H2	85	0.977	(0.063)	0.957	(0.040)	0.016	(0.037)	0.121	(0.084)
		90	0.969	(0.069)	0.949	(0.042)	0.022	(0.044)	0.095	(0.083)
		100	0.947	(0.081)	0.941	(0.041)	0.049	(0.063)	0.058	(0.071)
		110	0.923	(0.085)	0.936	(0.040)	0.101	(0.083)	0.035	(0.059)
		115	0.909	(0.102)	0.936	(0.039)	0.135	(0.090)	0.031	(0.055)

(20 slots/physician) $\times\, m \times$ (clinic load)/$\sum_{l=1}^{m} \lambda_t^l$, where $\sum_{l=1}^{m} \lambda_t^l = 0.1$. In the homogeneous panel load scenario, the arrival probability for each panel equals $\sum_{l=1}^{m} \lambda_t^l/m$, whereas in the heterogeneous panel load scenario, $\lambda_t^1 = 0.8\sum_{l=1}^{m} \lambda_t^l/m$, $\lambda_t^2 = 1.2\sum_{l=1}^{m} \lambda_t^l/m$, and $\lambda_t^{l'} = 0.8\sum_{l=1}^{m} \lambda_t^l/m$ for $l' \geq 3$. We also varied the clinic load by keeping τ fixed and changing each decision epoch's arrival rate. The performance of H1 and H2 was similar to what we report in Table 7. Therefore those results are not presented in the interest of brevity.

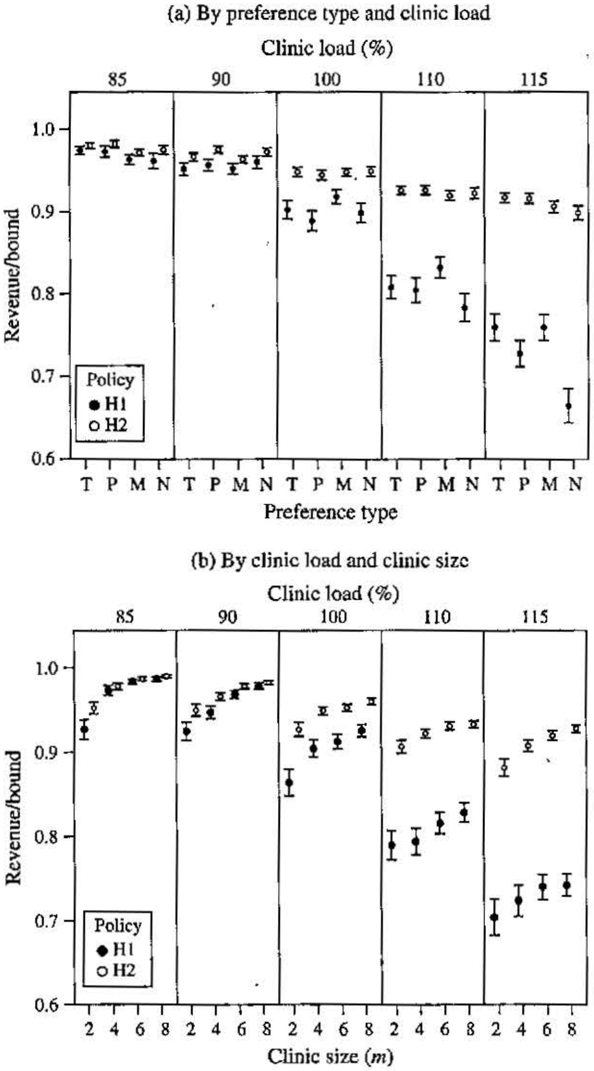
We assume that information bias results in inaccurate estimates of β_j^l. Let $\beta_{j(d)}^l$ be panel l's acceptance probability for time-block j, where (d) indicates that block j has the dth highest probability among the b time blocks for panel l. In the biased case, the clinic's estimate is assumed to be sufficiently inaccurate that it reverses the ordering for each panel's time-block acceptance probabilities. That is, the clinic uses $\hat{\beta}_{j(d)}^l = \beta_{(b-d+1)}^l$ when making booking decisions. For example, the clinic would use $\hat{\beta}^1 = (1, 0.6, 0.4, 0.2)$ as the clinic's biased estimates for β^1 in Table 4.

To focus attention on the impact of patient preferences and to not confound this effect with the effect of different revenue classes, we assumed that all panels had the same expected revenue. In particular, $(r_{1,l}^l, \pi, r_2, r_2^i) = (100, 35, 100, 85)$, $r_{1,l}^i = 85$ for $i \neq l$, and $\pi_t = \pi$ for all t. We generated 50 sample paths for each scenario and tracked the performance of H1 and H2 by average relative revenue (as compared to the bound discussed in §4.3.1), average patient-PCP match rate, average advance-book failure rate as a percentage of nonurgent requests for a particular appointment date, and average spoilage rate as a percentage of slots unused.

We first compared H1's and H2's performance with accurate and biased acceptance probabilities for each sample path. Neither H1's nor H2's average performance is affected much by using inaccurate acceptance probabilities—relative revenue on average increased by 0.53% for H1 and decreased by 0.49% for H2; all other metrics were on average affected less than 1.8% and 0.7% for H1 and H2, respectively. Note that the improvement in H1's performance is because of the higher advance-book failure rate induced by biased estimates of acceptance probabilities, which increases availability of slots for same-day demand.

Table 7 reports average performance measures sorted by c/π ratio and clinic load. For each combination, the reported performance metrics are aggregated over all possible scenarios of panel load, preference type, and information accuracy. Both H1 and H2 have high average patient-PCP match rates (94.7% and 94.5%), low average advance-book failure rates (1.6% and 5.7%), and low average spoilage rates (6.0% and 6.6%). The relative revenue performance of H1 and H2 is respectable with low variability when $c/\pi = 2$—the relative revenue is on average 97.7% and 98.2% of the bound for H1 and H2, respectively. When $c/\pi = 8$, the relative revenue performance is worse (on average 87.2% and 94.5% of the bound for H1 and H2, respectively), but H2 performs better. This suggests that when a health system has other options for taking care of urgent requests (e.g., urgent clinics), its cost of turning away same-day requests is smaller

Figure 3 **Average Relative Revenue and 95% Confidence Intervals When $c/\pi = 8$**


(a) By preference type and clinic load

(b) By clinic load and clinic size

Note. T, time dominant; P, physician dominant; M, moderate; N, no preferences.

(low c/π) and it may be justified in using H1, which is much simpler to implement.

Next, we report more detailed results in Figure 3 for the case when $c/\pi = 8$. As seen in Figure 3(a), for each clinic load, the relative revenue performance of H1 and H2 is robust across preference types. Similar results were also observed for panel loads, which are not reported here in the interest of brevity. Note that the means and confidence intervals are based on all sample paths generated across different clinic environments conditioned on the levels of factors presented in each subgraph. The relative performance of H1 and H2 deteriorates when clinic load exceeds capacity (Figure 3(b)) but improves as the clinic size increases (Figure 3(b)). The latter happens because each panel's same-day demand is assumed to be independent, and clinics with more physicians benefit from pooling available capacity to take care of same-day demand. The size effect may disappear when same-day demand patterns are correlated across physician panels.

5. Insights

In this section, we first compare the performance of H1 and H2 to a straw policy that does not utilize patients' preference information when making booking decisions. The straw policy attempts to book each arriving patient with the earliest available and acceptable patient-PCP matched slot. If none of the matched acceptable slots is available, the straw policy then books an appointment in the earliest available non-PCP slot, paying no attention to remaining capacity and time preferences. Next, we compare H1 (or H2) to itself when using true acceptance probabilities and naive acceptance probabilities. The purpose of this comparison is to tease out the value of information if a clinic decides to adopt either H1 or H2 booking heuristic. Finally, we evaluate the effect of low levels of no-show rates and service time variability by comparing H1, H2, and the straw policy.

All examples of this section use the following common parameters: $(r^l_{1,l}, \pi_t, r_2, r'_2, c) = (100, 35, 100, 85, 280)$, $r^i_{1,l} = 85$ for $i \neq l$, Poisson same-day demand with $E(X_l) = 6 = 30\%$ of each physician's capacity of 20 appointments per day. If desired, each panel's advance-book arrival rate can be varied to realize different workloads for different physicians. Total advance-book periods equal (0.7 × clinic capacity)$/(\sum_{l=1}^{m} \lambda^l_t)$, which ensures expected clinic demand equals clinic capacity. We report results when $\sum_{l=1}^{m} \lambda^l_t = 0.1$. Each experimental setup is simulated for 200 sample paths, and all booking strategies are evaluated for the same sample paths.

The first set of comparisons consider a six-PCP and four-time-block clinic whose patients always show up and find all physicians acceptable, but these patients have the following time preferences: $\beta^l = (\bar\beta, \bar\beta, 1, 1)$ for $l = 1, 2$; $\beta^l = (1, \bar\beta, \bar\beta, 1)$ for $l = 3, 4$; $\beta^l = (1, 1, \bar\beta, \bar\beta)$ for $l = 5, 6$. Each physician's clinic profile has more slots in blocks that are more acceptable to his or her panel patients (i.e., $\kappa_{l,j} = 6$ if $\beta^l_j = 1$, and $\kappa_{l,j} = 4$ otherwise). We vary $\bar\beta$ from 0.2 to 0.8 in 0.1 increments and study two arrival patterns: (1) constant arrival rates: $\lambda^l_t = 0.1/m$ for all $l = 1, \ldots, 6$ and $t = 1, \ldots, \tau$ and (2) varying arrival rates: when $t \leq (1/3)\tau$, $\lambda^k_t = 3\lambda^i_t$ for $k = 1, 2$ and $i = 3, 4, 5, 6$; when $(1/3)\tau < t \leq (2/3)\tau$, $\lambda^k_t = 3\lambda^i_t$ for $k = 3, 4$ and $i = 1, 2, 5, 6$; and when $t > (2/3)\tau$, $\lambda^k_t = 3\lambda^i_t$ for $k = 5, 6$ and $i = 1, 2, 3, 4$.

H1 and H2 on average result in about 1% and 8% higher revenue as compared to the straw policy regardless of the value of $\bar\beta$ and the arrival pattern. We report only the aggregate results in Table 8.

Table 8 **Performance of H1 and H2 Compared to the Straw Policy**

Policy	PCP match (%)	Advance-book failure rate (%)	Spoilage rate (%)	No. scheduled/ served	Rev. improv. compared to straw (%)
Straw	92.88	0.00	3.41	115.9	—
H1	94.34	0.04	3.50	115.8	1
H2	94.06	0.43	4.17	115.0	8

H1 achieves a higher PCP match, relative to the straw policy, by reserving more popular slots for future advance-book arrivals. In contrast, H2 with higher spoilage and advance-book failure rates achieves a high PCP match rate for different reasons. By reserving slots for same-day patients, it allows more of those patients to have an appointment with their PCP. H2 has much higher advance-book failure rate and slightly smaller number of patients served because some advance-book requests are denied when no-book states are reached. However, overall revenue is higher because it is costlier to turn away same-day patients.

Next, we evaluate the performance of H1 and H2 with two levels of information: (1) true acceptance probabilities and (2) naive acceptance probabilities. The latter assumes that every physician and time combination is acceptable to every patient. For each heuristic, we calculate the value of preference information by comparing that heuristic's average daily revenue to itself when the clinic uses true versus naive acceptance probabilities as inputs. We also monitor changes in patient-PCP match rates, advance-book failure rates, and number of patients served. In the results reported here, the clinic has eight full-time physicians, four time blocks, and five slots per block. Arrival pattern is time homogeneous, but expected demand rates can vary by panel, resulting in imbalanced workload across physicians.

We use a full factorial design of three factors, each with two levels. For a panel with strong (respectively, weak) time preferences, we allow one block to be always acceptable and the remaining blocks to be accepted with probability 0.3 (respectively, 0.7). For a panel with strong (respectively, weak) physician preferences, the PCP is always acceptable and each non-PCP is acceptable with probability 0.3 (respectively, 0.7). For a panel with adequate (respectively, inadequate) capacity, we let the expected demand be 66.7% (respectively, 133.3%) of the capacity. These combinations lead to eight stylized panel types.

Accurate preference information on average increases daily revenue by $20.35 and $93.15 for H1 and H2, respectively. H2 reduces advance-book failure rate by 1.3% and serves on average 1.16 more patients per day when using true acceptance probabilities as inputs. That is, our example clinic would be able to serve on average 423.4 more patients per year by updating patients' acceptance probabilities and using H2. We also studied a different scenario (results not reported for brevity) in which physician workloads were balanced and found that in such cases, knowledge of accurate acceptance probabilities does not significantly affect average daily revenues, patient-PCP match rates, advance-book failure rates, and the number of patients served (paired sample tests were not significant in all comparisons). This suggests that clinics whose physicians' workloads are imbalanced are more likely to benefit from accurate preference information when using H1 or H2 booking schemes. Imbalanced workloads are a common occurrence in practice.

In the last set of examples, we evaluate the effect of no-shows and service time variability, assuming punctual physicians and patients, independent and identically distributed service times and equal-length appointment slots. We test two levels of average no-show rates: 5% and 10%, with each patient's no-show probability drawn independently from Beta(0.05, 0.95) and Beta(0.1, 0.9) distribution, respectively. To test the impact of service time variability, we sample five distributions (see Table 9) that have the same mean but different coefficients of variation (0.33, 0.58, 0.58, 0.71, and 1, respectively). These distributions cover the range of service time variability observed in empirical studies (0.3–0.85); see Cayirli and Veral (2003). Finally, $\bar{\beta} = 0.5$ and $\lambda_t^l = 0.1/m$ for each l and t. We report performance comparisons in terms of paired sample t statistics for the average difference in revenue, average patient wait, and average physician overtime between H1 (or H2) and the straw policy in Table 9.

H1 and H2 on average have a statistically higher revenue than the straw policy has at each level of no-shows. The difference in average patient wait and physician overtime time between H1 (or H2) and the straw policy is statistically insignificant in most cases. However, when the difference is significant, H1 and H2 perform better. If the length of the appointment time slot is 30 minutes and average service time is 27.3 minutes, then patients' average wait ranges from 6.8 to 45 minutes while physicians' average overtime ranges from 6 to 40 minutes across these 10 scenarios. These results show that it is reasonable to use H1 and H2 when a clinic's no-show probability is not too high (≤10%) and the service time variability is not more extreme than the variability observed in empirical studies.

6. Concluding Remarks

This paper presents a framework for using appointment request data to update patients' preferences

Table 9 Performance Comparison in Terms of t Statistics with Degree of Freedom = 199 Under Paired Sample Tests

No-show rate (%)	Service time distribution	Revenue		PCP match		Avg. wait		Avg. OT	
		H1	H2	H1	H2	H1	H2	H1	H2
5	Unif[0.4, 1.6]	6.5*	8.1*	14.6*	13.5*	−1.8	−1.7	−1.6	−1.9
	Unif[0, 2]	5.9*	8.2*	14.4*	12.4*	−1.0	−1.3	−0.7	−1.9
	Gamma(3, 1/3)	7.5*	8.4*	14.7*	11.1*	0.5	−0.9	−0.8	−2.7*
	Gamma(2, 0.5)	5.7*	7.9*	15.4*	12.7*	−0.8	−1.2	−0.3	−0.9
	Exp(1)	6.9*	8.8*	14.8*	12.5*	−1.1	−1.8	−0.3	−0.5
10	Unif[0.4, 1.6]	8.3*	3.9*	15.1*	11.7*	0.4	−0.5	−1.9	−3.4*
	Unif[0, 2]	7.6*	7.8*	15.8*	12.5*	−0.1	−1.0	−1.6	−3.8
	Gamma(3, 1/3)	6.7*	9.2*	15.1*	13.3*	−0.6	−0.1	−1.2	−1.9
	Gamma(2, 0.5)	6.7*	7.9*	15.2*	11.9*	−0.1	−0.2	−0.8	−3.2*
	Exp(1)	7.3*	8.9*	13.8*	10.4*	1.7	1.4	0.4	0.3

*Significant at the 0.05 level.

and to subsequently use this information to improve clinics' revenues, serve more patients, and increase patient-PCP match rates. This approach can be implemented by utilizing data that can be retrieved from existing Web-based appointment request systems. However, it may not be suitable for clinics with high no-show rates that cannot be controlled by the use of a reminder system and patient education or by better matching patients' preferences with available slots. Such clinics may benefit from using approaches that explicitly consider no-shows when making booking decisions.

Our model is limited because it considers each workday's booking problem separately. A clinic may benefit from knowing all acceptable dates and the physician and time-block combinations that are acceptable to each arriving patient on each date before making a booking decision. However, that will make the booking process tedious for the patients and the state space of the appointment system will become unmanageable because acceptable dates may span an arbitrarily large period of time. It is perhaps for this reason that common Web-based booking request systems accept requests for one day at a time.

Patient-centered service models have attracted much attention in recent health policy literature. For example, a medical home model is a one-stop model that matches each patient with a team of providers based on the patient's needs. This team monitors a patient's health status and coordinates appointments for acute, chronic, and preventive services. Similarly, many health systems allow patients to see several service providers in a day or within a short period of time so that out-of-town patients do not need to travel to the service facility multiple times. Both models require matching patients' needs and preferences to multiple providers' availability. An interesting avenue of future research along the lines presented in this paper is the development of a model-based design of an adaptive appointment system for integrated medical services.

Electronic Companion

An electronic companion to this paper is available on the *Manufacturing & Service Operations Management* website (http://msom.pubs.informs.org/ecompanion.html).

Acknowledgments

This material is based upon work supported, in part, by the National Science Foundation under Grant CMMI-0620328. The authors are grateful to the referees, an associate editor, and Editor-in-Chief Steve Graves for their constructive comments on earlier versions of this paper.

References

Barron, W. M. 1980. Failed appointments. Who misses them, why they are missed, and what can be done. *Primary Care* 7(4) 563–574.

Carlin, B. P., T. A. Louis. 2000. *Bayes and Empirical Bayes Methods for Data Analysis*, 2nd ed. Chapman and Hall, New York.

Carlson, B. 2002. Same-day appointments promise increased productivity. *Managed Care* 11(12) 43–44.

Cayirli, T., E. Veral. 2003. Outpatient scheduling in health care: A review of literature. *Production Oper. Management* 12(4) 519–549.

Cayirli, T., E. Veral, H. Rosen. 2008. Assessment of patient classification in appointment system design. *Production Oper. Management* 17(3) 338–353.

Cheraghi-Sohi, S., A. Hole, N. Mead, R. McDonald, D. Whalley, P. Bower, M. Roland. 2008. What patients want from primary care consultations: A discrete choice experiment to identify patients' priorities. *Ann. Family Medicine* 6(2) 107–115.

Denton, B., D. Gupta. 2003. A sequential bounding approach for optimal appointment scheduling. *IIE Trans.* 35(11) 1003–1016.

Dervin, J. V., D. L. Stone, C. H. Beck. 1978. The no-show patient in the model family practice unit. *J. Family Practice* 7(6) 1177–1180.

Doescher, M. P., B. G. Saver, K. Fiscella, P. Franks. 2004. Preventive care: Does continuity count? *J. General Internal Medicine* 19(6) 632–637.

Dove, H. G., K. C. Schneider. 1981. The usefulness of patients' individual characteristics in predicting no-shows in outpatient clinics. *Medical Care* 19(7) 734–740.

Fosarelli, P., C. DeAngelis, A. Kaszuba. 1985. Compliance with follow-up appointments generated in a pediatric emergency room. *Amer. J. Preventive Medicine* 1(3) 23–29.

Gallucci, G., W. Swartz, F. Hackerman. 2005. Brief reparts: Impact of the wait for an initial appointment on the rate of kept appointments at a mental health center. *Psychiatric Services* **56**(3) 344–346.

Gerard, K., C. Salisbury, D. Street, C. Pope, H. Baxter. 2008. Is fast access to general practice all that should matter? A discrete choice experiment of patients' preferences. *J. Health Services Res. Policy* **13**(2) 3–10.

Green, L. V., S. Savin. 2008. Reducing delays for medical appointments: A queueing approach. *Oper. Res.* **56**(6) 1526–1538.

Gupta, D., B. Denton. 2008. Appointment scheduling in health care: Challenges and opportunities. *IIE Trans.* 40(9) 800–819.

Gupta, D., L. Wang. 2008. Revenue management for a primary care clinic in the presence of patient choice. *Oper. Res.* **56**(3) 576–592.

Hassin, R., S. Mendel. 2008. Scheduling arrivals to queues: A single-server model with no-shows. *Management Sci.* **54**(3) 565–572.

Ho, C.-J., H.-S. Lau. 1992. Minimizing total cost in scheduling outpatient appointments. *Management Sci.* **38**(12) 1750–764.

Irwin, C. E., Jr., S. G. Millstein, M.-A. B. Shafer. 1981. Appointment-keeping behavior in adolescents. *J. Pediatrics* **99**(5) 799–802.

Jennings, B. M., L. A. Loan, S. L. Heiner, E. A. Hemman, K. M. Swanson. 2005. Soldiers' experiences with military health care. *Military Medicine* **170**(12) 999–1004.

Kaandorp, G. C., G. Koole. 2007. Optimal outpatient appointment scheduling. *Health Care Management Sci.* **10**(3) 217–229.

Kim, S.; R. E. Giachetti. 2006. A stochastic mathematical appointment overbooking model for healthcare providers to improve profits. *IEEE Trans. on Systems, Man and Cybernetics–Part A: Systems and Humans* **36**(6) 1211–19.

Klassen, K. J., T. R. Rohleder. 1996. Scheduling outpatient appointments in a dynamic environment. *J. Oper. Management* **14**(2) 83–101.

LaGanga, L. R., S. R. Lawrence. 2007. Clinic overbooking to improve patient access and increase provider productivity. *Decision Sci.* 38(2) 251–276.

Lee, V. J., A. Earnest, M. I. Chen, B. Krishnan. 2005. Predictors of failed attendances in a multi-specialty outpatient centre using electronic databases. *BMC Health Services Res.* **5**, Article 51, http://www.biomedcentral.com/1472-6963/5/51.

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

McFadden, D. 2001. Economic choices. *Amer. Econom. Rev.* **91**(3) 351–378.

Murray, M., D. M. Berwick. 2003. Advanced access: Reducing waiting and delays in primary care. *J. Amer. Medical Assoc.* **289**(8) 1035–1040.

Murray, M., C. Tantau. 2000. Same-day appointments: Exploding the access paradigm. *Family Practice Management* **7**(8) 45–50.

Murray, M., T. Bodenheimer, D. Rittenhouse, K. Grumbach. 2003. Improving timely access to primary care: Case studies of the advanced access model. *J. Amer. Medical Assoc.* **289**(8) 1042–1046.

Muthuraman, K., M. Lawley. 2008. A stochastic overbooking model for outpatient clinical scheduling with no-shows. *IIE Trans.* **40**(9) 820–837.

Neinstein, L. S. 1982. Lowering broken appointment rates at a teenage health center. *J. Adolescent Health Care* 3(2) 110–113.

O'Hare, C. D., J. Corlett. 2004. The outcomes of open-access scheduling. *Family Practice Management* **11**(1) 35–38.

Olowokure, B., M. Caswell, H. V. Duggal. 2006. What women want: Convenient appointment times for cervical screening tests. *Eur. J. Cancer Care* **15**(5) 489–492.

Parente, D. H., M. B. Pinto, J. C. Barber. 2005. A pre-post comparison of service operational efficiency and patient satisfaction under open access scheduling. *Health Care Management Rev.* **30**(3) 220–228.

Robinson, L. W., R. R. Chen. 2003. Scheduling doctors' appointments: Optimal and empirically-based heuristic policies. *IIE Trans.* **35**(3) 295–307.

Robinson, L. W., R. R. Chen. 2010. A comparison of traditional and open-access policies for appointment scheduling. *Manufacturing Service Oper. Management* **12**(2) 330–346.

Robinson, L. W., R. R. Chen. 2011. Estimating the implied value of the customer's waiting time. *Manufacturing Service Oper. Management.* **13**(1) 53–57.

Rohleder, T. R., K. J. Klassen. 2000. Using client-variance information to improve dynamic appointment scheduling performance. *Omega* **28**(3) 293–302.

Russell, L. B. 2009. Completing costs: Patients' time. *Medical Care* **47**(7) S89–S93.

Salisbury, C., A. A. Montgomery, L. Simons, F. Sampson, S. Edwards, H. Baxter, S. Goodall, H. Smith, V. Lattimer, D. M. Pickin. 2007. Impact of advanced access on access, workload, and continuity: Controlled before-and-after and simulated-patient study. *British J. General Practice* **57**(541) 608–614.

Savin, S. 2006. Managing patient appointments in primary care. S. Hillier, F. R. W. Hall, eds. *Patient Flow: Reducing Delay in Healthcare Delivery, International Series in Operations Research & Management Science*, Vol. 91. Springer, New York, 123–150.

Smith, C. M., B. P. Yawn. 1994. Factors associated with appointment keeping in a family practice residency clinic. *J. Family Practice* **38**(1) 25–29.

Snow, B. W., P. C. Cartwright, S. Everitt, M. Ekins, W. Maudsley, S. Aloi. 2009. A method to improve patient access in urological practice. *J. Urology* **182**(2) 663–667.

Starkenburg, R. J., F. Rosner, K. Crowley. 1988. Missed appointments among patients new to a general medical clinic. *New York State J. Medicine* **88**(9) 437–435.

Talluri, K., G. van Ryzin. 2004. Revenue management under a general discrete choice model of consumer behavior. *Management Sci.* **50**(1) 15–33.

Train, K. E. 2003. *Discrete Choice Methods with Simulation*. Cambridge University Press, New York.

U.S. Bureau of Labor Statistics. 2008. Consumer expenditures in 2006. Technical Report 1010, U.S. Department of Labor, Washington, DC.

Vanden Bosch, P. M., D. C. Dietz. 2000. Minimizing expected waiting in a medical appointment system. *IIE Trans.* **32**(9) 841–848.

Wang, P. P. 1999. Sequencing and scheduling n customers for a stochastic server. *Eur. J. Oper. Res.* **119**(3) 729–738.

Weiss, E. N. 1990. Models for determining estimated start times and case orderings in hospital operating rooms. *IIE Trans.* 22(2) 143–150.

Whittle, J., G. Schectman, N. Lu, B. Baar, M. F. Mayo-Smith. 2008. Relationship of scheduling interval to missed and cancelled clinic appointments. *J. Ambulatory Care Management* **31**(4) 290–302.

Yabroff, K. R., J. L. Warren, K. Knopf, W. W. Davis, M. L. Brown. 2005. Estimating patient time costs associated with colorectal cancer care. *Medical Care* **43**(7) 640–648.