Appointment Scheduling under Patient Preference and No-Show Behavior

Jacob Feldman
School of Operations Research and Information Engineering,
Cornell University, Ithaca, NY 14853
jbf232@cornell.edu

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

Huseyin Topaloglu
School of Operations Research and Information Engineering,
Cornell University, Ithaca, NY 14853
topaloglu@orie.cornell.edu

Serhan Ziya
Department of Statistics and Operations Research,
University of North Carolina, Chapel Hill, NC 27599
ziya@unc.edu

May 21, 2013

Abstract

Motivated by the rising popularity of electronic appointment booking systems, we develop appointment scheduling models that take into account the patient preferences regarding when they would like to be seen. The service provider dynamically decides which appointment days to make available for the patients. Patients arriving with appointment requests may choose one of the days offered to them or leave without an appointment. Patients with scheduled appointments may cancel or not show up for the service. The service provider collects "revenues" from each patient who shows up and incurs a service "cost" that depends on the number of scheduled appointments. The objective is to maximize the expected net "profit" per day. We begin by developing a static model that does not consider the current state of the scheduled appointments. We give a characterization of the optimal policy under the static model and bound its optimality gap. Building on the static model, we proceed to develop a dynamic model that considers the current state of the scheduled appointments. In our computational experiments, we test the performance of our models under the patient preferences estimated through a discrete choice experiment that we conduct in a large community health center. We find that our proposed models, especially the dynamic one, can significantly outperform other benchmarks.

1 Introduction

Enhancing patient experience of care has been set as one of the "triple aims" to improve the healthcare system in many developed countries including the United States, Canada, and the United Kingdom. This aim is considered as equally, if not more, important as the other aims of improving the health of the population and managing per capita cost of care; see Berwick et al. (2008) and Institute for Healthcare Improvement (2012). An important component of enhancing patient experience of care is to provide more flexibility to patients regarding how, when and where to receive treatment. In pursuit of this objective, the National Health Service in the United Kingdom launched its electronic booking system, called Choose and Book, for outpatient appointments in January 2006; see Green et al. (2008). In the United States, with the recent Electronic Health Records and Meaningful Use Initiative, which calls for more and better use of health information technology, online scheduling is being adopted by an increasingly larger percentage of patients and clinics; see US Department of Health and Human Services (2011), Weiner et al. (2009), Silvestre et al. (2009), Wang and Gupta (2011), Zocdoc (2012).

In contrast to the traditional appointment scheduling systems where patients are more or less told by the clinic when to come and whom to see or are given limited options on the phone, electronic appointment booking practices make it possible to better accommodate patient preferences by providing patients with more options. Giving patients more flexibility when scheduling their appointments have benefits that can go beyond simply having more satisfied patients. More satisfied patients lead to higher patient retention rates, which potentially allow providers to negotiate better reimbursement rates with payers; see Rau (2011). More satisfied patients can also lead to reduced no-show rates, helping maintain the continuity of care and improve patient health outcomes; see Bowser et al. (2010) and Schectman et al. (2008). An important issue when providing flexibility to patients is that of managing the operational challenges posed by giving more options. In particular, one needs to carefully choose the level of flexibility offered to the patients while taking into account the operational consequences. It is not difficult to imagine that giving patients complete flexibility in choosing their appointment times would lead to high variability in the daily load of a clinic. Thus, options provided to the patients need to be restricted in a way that balances the benefits with the costs. While such decisions have been studied in some industries, such as airlines, hospitality and manufacturing, scheduling decisions with consideration of patient preferences has largely been ignored in appointment scheduling literature; see Cayirli and Veral (2003) and Gupta and Denton (2008) Rohleder and Klassen (2000) . The goal of this paper is to develop models that can aid in appointment scheduling process while considering the patient preferences.

In this paper, with electronic appointment booking systems in mind, we develop models to decide which appointment days to offer in response to incoming appointment requests. Specifically, we consider a single service provider receiving appointment requests every day. The service provider offers a menu of appointment dates within the scheduling window to choose from. During the day, patients arrive with the intention of scheduling appointments and they either choose to book

an appointment on one of the days made available to them or leave without scheduling any. We assume that patient choice behavior is governed by a multinomial logit choice model; see McFadden (1974). In the mean time, patients with appointments may decide not to show up and those with appointments on a future day may cancel. The service provider generates a "revenue" from each patient who shows up for her appointment and incurs a "cost" that depends on the number of patients scheduled to be seen on a day. The objective is to maximize the expected net "profit" per day by choosing the offer set, the set of days offered to patients who demand appointments. We begin by developing a static model that makes its decisions without using the information about the appointments that are currently booked. Building on the static model, we develop a dynamic one that takes the current state of the booked appointments into consideration.

Our static model is based on solving a mathematical program in which the decision variables are the probabilities with which a particular subset of appointment days will be offered to the patients, independent of the state of the booked appointments. One difficulty with this formulation is that since we have one decision variable for each subset of days in the scheduling window, the number of decision variables increases exponentially with the length of the scheduling window. To overcome this, we exploit the special structure of the multinomial logit model to reformulate the static model in a more compact form. The number of decision variables in the compact formulation increases only linearly with the number of days in the scheduling window, making it tractable to solve. We show that if the no-show probability, conditional on the event that the patient has not canceled before her appointment, does not depend on patient delays, then there exists a simple and easy to implement optimal policy from the static model, which randomizes between only two adjacent offer sets. To assess the potential performance loss as a result of the static nature of the model, we provide a bound on the optimality gap of the static model. The bound on the optimality gap approaches zero as the average patient load per day increases, indicating that the static model may work well in systems handling large amount of demand.

Our dynamic model improves on the static one by taking the state of the booked appointments into consideration when making its decisions. The starting point for the dynamic model is the Markov decision process formulation of the appointment scheduling operations. Unfortunately, this formulation cannot be solved by using standard dynamic programming tools since the state space is too large. We propose an approximate method based on the Markov decision process formulation and this approximate method can be seen as applying a single step of the standard policy improvement algorithm on an initial "good" policy. For the initial "good" policy, we employ the policy provided by the static model. Implementing the policy provided by the dynamic model allows us to make decisions in online fashion by solving a mathematical program that uses the current state of the booked appointments. The structure of this mathematical program is similar to one that we solve for the static model. Thus, the structural results obtained for the static model problem, at least partially, apply to the dynamic model as well. We carry out a simulation study to investigate how our proposed models perform. We generated numerous problem instances by varying model parameters so that we can compare the performance of our

policies with benchmarks over a large set of possible settings regarding the clinic capacity and the patient demand. Furthermore, we base the preference structure of the patients on a discrete choice experiment that we conduct in a large community health center. Our simulation study suggests that the proposed dynamic model significantly outperforms benchmarks.

The studies in Ryan and Farrar (2000), Rubin et al. (2006), Cheraghi-Sohi et al. (2008), Gerard et al. (2008) and Hole (2008) all point out that patients do have preferences regarding when to visit the clinic and which physician to see. In general, patients prefer appointments that are sooner than later, but they may prefer a later day over an earlier one if the appointment on the later day is at a more convenient time; see Sampson et al. (2008). Capturing these preferences in their full complexity while keeping the formulation sufficiently simple can be a challenging task, but our static and dynamic models yield tractable policies to implement in practice. The models we propose can be particularly useful for clinics that are interested in a somewhat flexible implementation of open access, the policy of seeing today's patient today; see Murray and Tantau (2000). While same-day scheduling reduces the access time for patients and helps keep no-show rates low, it is a feasible practice only if demand and service capacity are in balance; see Green and Savin (2008). Furthermore, some patients highly value the flexibility of scheduling appointments for future days. A recent study in Sampson et al. (2008) found that a 10% increase in the proportion of same-day appointments was associated with an 8% reduction in the proportion of patients satisfied. This is a somewhat surprising finding, attributed to the decreased flexibility in booking appointments that comes with restricting access to same-day appointments. Our models help find the "right" balance among competing objectives of providing more choices to patients, reducing appointment delays, increasing system efficiency, and reducing no-shows.

While past studies have found that patients have preferences regarding when they would like to be seen, to our knowledge, there has been limited attempt in quantifying the relative preferences of patients using actual data. To develop our understanding of these patient preferences in practice and to use this insight in populating our model parameters in our computational study, we collected data from a large urban health center in New York City and used these data to estimate the parameters of the patient choice model. For the patients of this particular clinic, we give a choice model that estimates how patients would choose one day over the other. The estimated model parameters confirm that when patients have a health condition that does not require an emergency but still needs fast attention, they do prefer to be seen earlier, even after taking their work and other social obligations into account.

In addition to capturing the preferences of the patients on when they would like to be seen, a potentially useful feature of our models is that they capture the patient no-show process in a quite general fashion. In particular, the probability that a customer cancels her appointment on a given day depends on when the appointment was scheduled and how many days there are left until the appointment day arrives. In this way, we can model the dependence of cancellation probabilities on appointment delays. Similarly, we allow the probability that a patient shows up for

her appointment to depend on how much in advance the appointment was booked. Past studies are split on how cancellation and no-show probabilities depend on appointment delays. A number of articles find that patients with longer appointment delays are more likely to not show up for their appointments; see Grunebaum et al. (1996), Gallucci et al. (2005), Dreiher et al. (2008), Bean and Talaga (1995) and Liu et al. (2010). On the other hand, there have been other studies that find no such relationships; see Wang and Gupta (2011) and Gupta and Wang (2011). In this paper, we keep our formulation general so that no-show and cancellation probabilities can possibly depend on the delays experienced by patients.

The rest of the paper is organized as follows. Section 2 reviews the relevant literature. Section 3 describes the model primitives we use. Section 4 presents our static model, gives structural results regarding the policy provided by this model and provides a bound on the optimality gap. Section 5 develops our dynamic model. Section 6 discusses our discrete choice experiment on patient preferences, explains our simulation setup and presents our numerical findings. Section 7 provides concluding remarks.

2 Literature Review

The appointment scheduling literature has been growing rapidly in recent years and the articles by Cayirli and Veral (2003) and Gupta and Denton (2008) provide a broad and coherent overview of this literature. Most work in this area focuses on intra-day scheduling and is typically concerned with timing and sequencing patients with the objective of finding the "right" balance between in-clinic patient waiting times and physician utilization. Recent examples of this literature include Wang (1999), Rohleder and Klassen (2000), Denton and Gupta (2003), Robinson and Chen (2003), Klassen and Rohleder (2004), Green et al. (2006), Kaandorp and Koole (2007), Hassin and Mendel (2008), Jouini and Benjaafar (2010), Cayirli et al. (2012), and Luo et al. (2012). Other researchers, including Kim and Giachetti (2006), LaGanga and Lawrence (2007), Muthuraman and Lawley (2008), Zeng et al. (2010), Huang and Zuniga (2012) and LaGanga and Lawrence (2012), have studied overbooking strategies to mitigate the impact of no-shows on clinic operations. Along with the increasing popularity of open access in practice, there are a number of articles on operational implications of the open access policy; see, for example, Kopach et al. (2007), Qu et al. (2007), Green and Savin (2008) and Robinson and Chen (2010).

In contrast to the literature on intra-day scheduling, our work focuses on inter-day scheduling and does not explicitly address how intra-day scheduling needs to be done. In that regard, our models can be viewed as daily capacity management models for a medical facility. Similar to our work, Gerchak et al. (1996), Patrick et al. (2008), Liu et al. (2010), Ayvaz and Huh (2010) and Huh et al. (2012) deal with the allocation and management of daily service capacity. Liu et al. (2010) is particularly relevant since it also considers a formulation that keeps track of appointments over a number of days and uses the idea of applying a single step of the policy improvement algorithm to

develop a heuristic method. However, Liu et al. (2010) do not take patient preferences into account in any way and assume that patients always accept the appointment day offered to them. This assumption makes the formulation and analysis in Liu et al. (2010) significantly simpler than ours. Also, we give a characterization of the structure of the optimal state-independent policy obtained by solving our static model and provide a bound on the optimality gap of this policy.

To our knowledge, Rohleder and Klassen (2000), Gupta and Wang (2008) and Wang and Gupta (2011) are the only three articles that consider patient preferences in the context of appointment scheduling. These three articles deal with appointment scheduling within a single day, whereas our work focuses on decisions over multiple days. By focusing on a single day, Rohleder and Klassen (2000), Gupta and Wang (2008) and Wang and Gupta (2011) develop policies regarding the specific timing of the appointments, but they do not incorporate the fact that the appointment preferences of patients may not be restricted within a single day. Furthermore, their proposed policies do not smooth out the daily load of the clinic. In contrast, since our policies capture preferences of patients over a number of days into the future, they can help in using the capacity of the clinic efficiently by distributing demand over multiple days. However, our focus on scheduling appointments over multiple days precludes us from providing any specific prescription as to how the appointments should be scheduled within each day. Another interesting point of departure between our work and the previous literature is the assumptions regarding how patients interact with the clinic while scheduling their appointments. Specifically, Gupta and Wang (2008) and Wang and Gupta (2011) assume that patients first reveal a set of their preferred appointment slots to the clinic, which then decides whether to offer the patient an appointment in the set or to reject the patient. In our model, the clinic offers a set of appointment dates to the patients and the patients either choose one of the offered days or decline to make an appointment.

While there is limited work on customer choice behavior in healthcare appointment scheduling, there are numerous related papers in the broader operations literature, particularly in revenue management and assortment planning. Typically, these papers deal with problems where a firm chooses a set of products to offer its customers based on inventory levels and remaining time in the selling horizon. In response to the offered set of products, customers make a choice within the offered set. Talluri and van Ryzin (2004*a*) consider a revenue management model on a single flight leg, where customers make a choice among the fare classes made available to them. The authors show that if the choices of the customers are governed by the multinomial logit model, then it is optimal to offer a set that includes a certain number of fare classes with the highest fares. Gallego et al. (2004), Liu and van Ryzin (2008), Kunnumkal and Topaloglu (2008) and Zhang and Adelman (2009) extend this model to a flight network. The fundamental approach in the latter papers is to formulate deterministic approximations to the revenue management problem by assuming that the customer demand takes on its expected value. Bront et al. (2009) and Rusmevichientong et al. (2010) study product offer decisions when there are multiple customer types, each type choosing according to multinomial logit models with different parameters and show that the corresponding optimization problem is NP-hard. In contrast, Gallego et al. (2011) show that the product offer

decisions under the multinomial logit model can be formulated as a linear program when there is a single customer type. Topaloglu (2010) extends the work in Gallego et al. (2011) to a setting that accounts for stocking decisions for the products. We refer the reader Talluri and Van Ryzin (2004*b*) and Kök et al. (2009) for detailed overview of the literature on revenue management and product offer decisions.

3 Model Primitives

We are interested in scheduling patient appointments over time. On each day, we observe the state of the appointments that were already scheduled and decide which appointment days to make available for the patients requesting appointments on the current day. During the day, arriving patients may choose to schedule appointments among the days made available to them and patients with scheduled appointments may cancel their appointments. Furthermore, patients with scheduled appointments for the current day may not show up. We generate a revenue from each patient that shows up for service and we incur a capacity cost as a function of the number of patients we plan to serve on a day. The objective is to decide which set of days to make available for appointments to maximize the expected net profit per day.

Throughout the paper, we assume that the following sequence of events occur on a particular day. First, we observe the state of the appointments that were already scheduled and decide which subset of days in the future to make available for the patients making appointment requests on the current day. Second, patients arrive with appointment requests and choose an appointment day among the days that are made available to them. Third, some of the patients with appointments scheduled for future days may cancel their appointments and we observe the cancellations. Also, some of the patients with appointments scheduled on the current day may not show up and we observe those who do. In reality, appointment requests, cancellations and show-ups occur throughout the day with no particular time ordering among them, but our sequence of events is simply a modeling choice and our policies continue to work as long as the set of days made available for appointments are chosen at the beginning of a day. The revenue we generate is determined by the number of patients that we serve on a day, whereas the cost we incur is determined by the number of appointments scheduled for a day before we observe the show-ups. The cost is assumed to be driven by the number of appointments scheduled for a day before we observe the show-ups mainly because staffing is the primary cost driver and staffing decisions have to be made before we observe which appointments show up.

The number of appointment requests on each day has Poisson distribution with mean λ. Each patient calling in for an appointment can be scheduled either for the current day or for one of the τ future days. Therefore, the scheduling horizon is $\mathcal{T} = \{0, 1, \ldots, \tau\}$, where day 0 corresponds to the current day and the other days correspond to a day in the future. The decision we make on each day is the subset of days in the scheduling horizon that we make available for appointments. If we make

the subset $S \subset \mathcal{T}$ of days available for appointments, then a patient schedules an appointment j days into the future with probability $P_j(S)$. We assume that the choice probability $P_j(S)$ is governed by the multinomial logit choice model; see McFadden (1974). Under this choice model, each patient associates the preference weight of v_j with the option of scheduling an appointment j days into the future. Furthermore, each patient associates the nominal preference weight of 1 with the option of not scheduling an appointment at all. In this case, if we offer the subset S of days available for appointments, then the probability that an incoming patient schedules an appointment $j \in S$ days into the future is given by

$$P_j(S) = \frac{v_j}{1 + \sum_{k \in S} v_k}. \tag{1}$$

With the remaining probability, $N(S) = 1 - \sum_{j \in S} P_j(S) = 1/(1 + \sum_{k \in S} v_k)$, a patient leaves the system without scheduling an appointment at all.

If a patient called in i days ago and scheduled an appointment j days into the future, then this patient cancels her appointment on the current day with probability r'_{ij}, independent of the other appointments scheduled. For example, if the current day is October 15, then r'_{13} is the probability a patient who called in on October 14 and scheduled an appointment for October 17 cancels her appointment on the current day. We let $r_{ij} = 1 - r'_{ij}$ so that r_{ij} is the probability that we retain a patient who called in i days ago and scheduled an appointment j days into the future. If a patient called in i days ago and scheduled an appointment on the current day, then this patient shows up for her appointment with probability s_i, conditional on the event that she has not canceled until the current day. We assume that r_{ij} is decreasing in j for a fixed value of i so that the patients scheduling appointments further into the future are less likely to be retained.

For each patient served on a particular day, we generate a nominal revenue of 1. We have a regular capacity of serving C patients per day. After observing the cancellations, if the number of appointments scheduled for the current day exceeds C, then we incur an additional staffing and overtime cost of θ per patient above the capacity. The cost of the regular capacity of C patients is assumed to be sunk and we do not explicitly account for this cost in our model. With this setup, the profit per day is linear in the number of patients that show up and piecewise-linear and concave in the number of appointments that we retain for the current day, but our results are not tied to the structure of this cost function and it is straightforward to extend our development to cover the case where the profit per day is a general concave function of the number of patients that show up and the number of appointments that we retain.

We can formulate the problem as a dynamic program by using the status of the appointments at the beginning of a day as the state variable. In particular, given that we are at the beginning of day t, we can let $X_{ij}(t)$ be the number of patients that called in i days ago (on day $t-i$), scheduled an appointment j days into the future (for day $t-i+j$) and are retained without cancellation until the current day t. In this case, the vector $X(t) = \{X_{ij}(t) : 1 \leq i \leq j \leq \tau\}$ describes the state of the scheduled appointments at the beginning of day t. This state description has $\tau\,(\tau+1)/2$

dimensions and the state space can be very large even if we bound the total number of scheduled appointments on a given day by some realistic finite number. In the next section, we begin with developing a static model that makes its decisions without considering the current state of the booked appointments. Later in the paper, we build on the static model to construct a dynamic model that indeed considers the state of the booked appointments when making its decisions.

4 Static Model

In this section, we consider a static model that makes each subset of days in the scheduling horizon available for appointments with a fixed probability. Since the probability offering a particular subset of days is fixed, this model does not account for the current state of appointments when making its decisions. We solve a mathematical program to find the best probability with which each subset of days should be made available.

To formulate the static model, we let $h(S)$ be the probability with which we make the subset $S \subset \mathcal{T}$ of days available. If we make the subset S of days available with probability $h(S)$, then the probability that a patient schedules an appointment j days into the future is given by $\sum_{S \subset \mathcal{T}} P_j(S)\, h(S)$. Given that a patient schedules an appointment j days into the future, we retain this patient until the day of the appointment with probability $\bar{r}_j = r_{0j}\, r_{1j} \ldots r_{jj}$ and this patient shows up for her appointment with probability $\bar{s}_j = r_{0j}\, r_{1j} \ldots r_{jj}\, s_j$. Thus, noting that the number of appointment requests on a day has Poisson distribution with parameter λ, if we make the subset S of days available with probability $h(S)$, then the number of patients that schedule an appointment j days into future and that are retained until the day of the appointment is given by a Poisson random variable with mean $\lambda\, \bar{r}_j \sum_{S \subset \mathcal{T}} P_j(S)\, h(S)$. Similarly, we can use a Poisson random variable with mean $\lambda\, \bar{s}_j \sum_{S \subset \mathcal{T}} P_j(S)\, h(S)$ to capture the number of patients that schedule an appointment j days into the future and that show up for their appointments. In this case, using $\mathsf{Pois}(\alpha)$ to denote a Poisson random variable with mean α, on each day, the total number of patients whose appointments we retain until the day of the appointment is given by $\mathsf{Pois}\big(\sum_{j \in \mathcal{T}} \sum_{S \subset \mathcal{T}} \lambda\, \bar{r}_j\, P_j(S)\, h(S)\big)$ and the total number of patients that show up for their appointments is given by $\mathsf{Pois}\big(\sum_{j \in \mathcal{T}} \sum_{S \subset \mathcal{T}} \lambda\, \bar{s}_j\, P_j(S)\, h(S)\big)$. To find the subset offer probabilities that maximize the expected profit per day, we can solve the problem

$$\max \quad \sum_{j \in \mathcal{T}} \sum_{S \subset \mathcal{T}} \lambda\, \bar{s}_j\, P_j(S)\, h(S) - \theta\, \mathbb{E}\Big\{\Big[\mathsf{Pois}\Big(\sum_{j \in \mathcal{T}} \sum_{S \subset \mathcal{T}} \lambda\, \bar{r}_j\, P_j(S)\, h(S)\Big) - C\Big]^+\Big\} \tag{2}$$

$$\text{subject to} \quad \sum_{S \subset \mathcal{T}} h(S) = 1 \tag{3}$$

$$h(S) \geq 0 \qquad S \subset \mathcal{T}, \tag{4}$$

where we use $[\cdot]^+ = \max\{\cdot, 0\}$. In the problem above, the two terms in the objective function correspond to the expected revenue and the expected cost per day. The constraint ensures that the total probability with which we offer a subset of days is equal to 1. Noting $\emptyset \subset \mathcal{T}$, the problem above allows not offering any appointment slots to an arriving patient.

4.1 Reformulation

We observe that problem (2)-(4) has $2^{|\mathcal{T}|}$ decision variables, which can be too many in practical applications. For example, if we have a scheduling horizon of a month, then the number of decision variables in this problem exceeds a billion. However, it turns out that we can give an equivalent formulation for problem (2)-(4) that has only $|\mathcal{T}| + 1$ decision variables, which makes the solution of this problem quite tractable. In particular, Proposition 1 below shows that problem (2)-(4) is equivalent to the problem

$$\max \quad \sum_{j \in \mathcal{T}} \lambda \, \bar{s}_j \, x_j - \theta \, \mathbb{E}\Big\{ \Big[\mathsf{Pois}\Big(\sum_{j \in \mathcal{T}} \lambda \, \bar{r}_j \, x_j \Big) - C \Big]^+ \Big\} \tag{5}$$

$$\text{subject to} \quad \sum_{j \in \mathcal{T}} x_j + u = 1 \tag{6}$$

$$\frac{x_j}{v_j} - u \leq 0 \qquad j \in \mathcal{T} \tag{7}$$

$$x_j, u \geq 0 \qquad j \in \mathcal{T}. \tag{8}$$

In the problem above, we interpret the decision variable x_j as the probability that a patient schedules an appointment j days into the future. The decision variable u corresponds to the probability that a patient does not schedule an appointment. The objective function accounts for the expected profit per day as a function of the scheduling probabilities. Constraint (6) captures the fact that each patient either schedules an appointment on one of the future days or does not. To see an interpretation of constraints (7), we note that if we offer the subset S of days, then the probability that a patient schedules an appointment j days into the future is given by $P_j(S) = v_j/(1 + \sum_{k \in S} v_k)$ if $j \in S$ and 0 otherwise. On the other hand, the probability that a patient does not schedule an appointment is given by $N(S) = 1/(1 + \sum_{k \in S} v_k)$. Therefore, we have $P_j(S)/v_j - N(S) \leq 0$. Surprisingly constraints (7) are the only place where the parameters of the multinomial logit choice model appears in problem (5)-(8) and these constraints turn out to be adequate to capture the choices of the patients as stipulated by the multinomial logit choice model. In the next proposition, we show that problems (2)-(4) and (5)-(8) are equivalent to each other. The proof uses the approach followed by Topaloglu (2010). We give the main ideas of the proof here, but defer the details to the appendix.

Proposition 1. *Problems* (2)-(4) *is equivalent to problem* (5)-(8) *in the sense that given an optimal solution to one problem, we can generate a feasible solution to the other one providing the same objective value.*

Proof. Assume that $h^* = \{h^*(S) : S \subset \mathcal{T}\}$ is an optimal solution to problem (2)-(4). Letting $x_j^* = \sum_{S \subset \mathcal{T}} P_j(S) \, h^*(S)$ and $u^* = \sum_{S \subset \mathcal{T}} N(S) \, h^*(S)$, we show in the appendix that (x^*, u^*) with $x^* = (x_0^*, \ldots, x_\tau^*)$ is a feasible solution to problem (5)-(8) providing the same objective value as the solution h^*. On the other hand, assume that (x^*, u^*) with $x^* = (x_0^*, \ldots, x_\tau^*)$ is an optimal solution to problem (5)-(8). We reorder and reindex the days in the scheduling horizon so that

we have $x_0^*/v_0 \geq x_1^*/v_1^* \geq \ldots \geq x_\tau^*/v_\tau^*$. Constraints (7) in problem (5)-(8) also ensure that $u^* \geq x_0^*/v_0 \geq x_1^*/v_1 \geq \ldots \geq x_\tau^*/v_\tau^*$. We define the subsets $S_0, S_1, \ldots, S_\tau$ as $S_j = \{0, 1, \ldots, j\}$. For notational convenience, we define $x_{\tau+1}^* = 0$. In this case, letting

$$h^*(\emptyset) = u^* - \frac{x_0^*}{v_0} \quad \text{and} \quad h^*(S_j) = \Big[1 + \sum_{k \in S_j} v_k\Big] \Big[\frac{x_j^*}{v_j} - \frac{x_{j+1}^*}{v_{j+1}}\Big] \tag{9}$$

for all $j = 0, 1, \ldots, \tau$ and letting $h^*(S) = 0$ for all other subsets of $\mathcal{T}$, we show in the appendix that $\{h^*(S) : S \subset \mathcal{T}\}$ ends up being a feasible solution to problem (2)-(4) providing the same objective value as the solution (x^*, u^*). □

The proof above indicates that we can obtain an optimal solution to problem (2)-(4) by solving problem (5)-(8) and using (9). Furthermore, (9) shows that there are at most $|\mathcal{T}| + 1$ subsets for which the decision variables $\{h(S) : S \subset \mathcal{T}\}$ take strictly positive values in an optimal solution. In the next section, under certain assumptions on the show-up probabilities, we further reduce the number of subsets for which the decision variables $\{h(S) : S \subset \mathcal{T}\}$ can take strictly positive values. Problem (5)-(8) has a manageable number of decision variables and constraints, but its objective function is nonlinear. Noting that the complicating term in the objective function is of the form $\mathbb{E}\{[\mathsf{Pois}(\alpha) - C]^+\}$ for $\alpha \in \Re_+$, a simple calculation, given in Lemma 7 in the appendix, shows that $F(\alpha) = \mathbb{E}\{[\mathsf{Pois}(\alpha) - C]^+\}$ is a differentiable and convex function of α.

Given that the objective function of problem (5)-(8) is convex and its constraints are linear, we can use a variety of convex optimization approaches to solve this problem. One approach is to use a general-purpose cutting-plane method for convex optimization, which represents the objective function of problem (5)-(8) with a number of cuts; see Ruszczynski (2006). A more direct approach for solving this problem is to observe that the function $F(\cdot)$ as defined in the previous paragraph is a scalar and convex function. So, it is straightforward to build an accurate piecewise linear and convex approximation to $F(\cdot)$ by evaluating this function at a finite number of grid points. We denote the approximation constructed in this fashion by $\hat{F}(\cdot)$. In this case, to obtain an approximate solution to problem (5)-(8), we can maximize the objective function $\sum_{j \in \mathcal{T}} \lambda\, \bar{s}_j\, x_j - \theta \hat{F}(\alpha)$ subject to the constraint that $\alpha = \sum_{j \in \mathcal{T}} \lambda\, \bar{r}_j\, x_j$ and constraints (6)-(8). Since this optimization problem has a piecewise linear objective function and linear constraints, it can be solved as a linear program. Furthermore, if we choose $\hat{F}(\cdot)$ as a lower bound or an upper bound on $F(\cdot)$, then we obtain a lower or upper bound on the optimal objective value of problem (5)-(8). When the lower and upper bounds are close to each other, we can be confident about the quality of the solution we obtain from our approximation approach.

Another algorithmic strategy that we can use to solve problem (5)-(8) is based on dynamic programming. For a fixed value of u, we observe that the constraints in this problem correspond to the constraints of a knapsack problem. In particular, the items correspond to the days in the scheduling horizon, the capacity of the knapsack is $1 - u$ and we can put at most $v_j \times u$ units of item j into the knapsack. So, for a fixed value of u, we can solve problem (5)-(8) by using

dynamic programming. In this dynamic program, the decision epochs correspond to days in the scheduling horizon $\mathcal{T}$. The state at decision epoch $j \in \mathcal{T}$ has two components. The first component corresponds to the value of $\sum_{k=0}^{j-1} x_k$, capturing to the portion of the knapsack capacity consumed by the decisions in the earlier decision epochs. The second component corresponds to the value of $\sum_{k=0}^{j-1} \lambda \bar{r}_k x_k$, capturing the accumulated value of the argument of the second term in the objective function of problem (5)-(8). Thus, for a fixed value of u, we can solve problem (5)-(8) by computing the value functions $\{\Theta_j(\cdot,\cdot \,|\, u) : j \in \mathcal{T}\}$ though the optimality equation

$$\Theta_j(b, c \,|\, u) = \max \quad \lambda \bar{s}_j x_j + \Theta_{j+1}(b + x_j, c + \lambda \bar{r}_j x_j \,|\, u) \tag{10}$$

$$\text{subject to} \quad 0 \leq x_j / v_j \leq u, \tag{11}$$

with the boundary condition that $\Theta_{\tau+1}(b, c \,|\, u) = -\theta \, \mathbb{E}\{[\mathsf{Pois}(c) - C]^+\}$ when $b = 1 - u$ and $\Theta_{\tau+1}(b, c \,|\, u) = -\infty$ when $b \neq 1 - u$. In the dynamic program above, we accumulate the first component of the objective function of problem (5)-(8) during the course of the decision epochs, but the second component of the objective function through the boundary condition. Having $\Theta_{\tau+1}(b, c \,|\, u) = -\infty$ when $b \neq 1 - u$ ensures that we always consume the total capacity availability of $1 - u$ in the knapsack. It is straightforward to accurately solve the dynamic program for all $u \in \Re_+$ by discretizing the state variable (b, c) and u over a fine grid in $\Re_+^3$. Solving the dynamic program for all $u \in \Re_+$, $\max_{u \in \Re_+} \Theta_1(0, 0 \,|\, u)$ gives optimal objective value of problem (5)-(8).

4.2 Static Model under Delay-Independent Show-up Probabilities

As we describe at the end of the previous section, it is not difficult to obtain good solutions to problem (5)-(8). Nevertheless, although we can obtain good solutions to problem (5)-(8), the structure of the optimal subset offer probabilities obtained from problem (2)-(4) is still not obvious. Furthermore, many of the subset offer probabilities $\{h(S) : S \subset \mathcal{T}\}$ may take positive values in the optimal solution, indicating that the static policy may be highly randomized, which may be undesirable for practical implementation. In this section, we consider the special case where the show-up probability of an appointment does not depend on how many days ago the patient called in. We show two results regarding the structure of the optimal subset availability probabilities. First, we show that the subsets of days that we make available always consists of a certain number of consecutive days. In particular, we show that there exist optimal offer probabilities $\{h^*(S) : S \subset \mathcal{T}\}$ such that $h^*(S) > 0$ if and only if S is a subset of the form $\{0, 1, \ldots, j\}$ for some $j \in \mathcal{T}$. Second, we show that the optimal subset availability probabilities randomize between only two subsets and these two subsets differ only in one day. These results indicate that the randomized nature of the static model is not a huge practical concern.

Throughout this section, we assume that s_j is independent of j and we use s_0 to denote the common value of $\{s_j : j \in \mathcal{T}\}$. Thus, noting the definitions $\bar{r}_j = r_{0j} \, r_{1j} \ldots r_{jj}$ and $\bar{s}_j = r_{0j} \, r_{1j} \ldots r_{jj} \, s_j$, we obtain $\bar{s}_j = \bar{r}_j \, s_0$. We emphasize that although the show-up probabilities are assumed to be independent of how many days ago an appointment was scheduled, the

probability of retaining an appointment r_{ij} can still be arbitrary. Define the scalar function $R(\cdot)$ as $R(\alpha) = s_0\,\alpha - \theta\,\mathbb{E}\big\{[\mathsf{Pois}(\alpha) - C]^+\big\}$. Using the fact that $\bar{s}_j = \bar{r}_j\,s_0$, we can write the objective function of problem (5)-(8) succinctly as $R\big(\sum_{j\in\mathcal{T}}\lambda\,\bar{r}_j\,x_j\big)$ and problem (5)-(8) becomes

$$\max \quad R\Big(\sum_{j\in\mathcal{T}}\lambda\,\bar{r}_j\,x_j\Big) \tag{12}$$

$$\text{subject to} \quad \sum_{j\in\mathcal{T}} x_j + u = 1 \tag{13}$$

$$\frac{x_j}{v_j} - u \le 0 \qquad j\in\mathcal{T} \tag{14}$$

$$x_j, u \ge 0 \qquad j\in\mathcal{T}. \tag{15}$$

In the next proposition, we show that there exists an optimal solution to the problem above where at most one of the decision variables $(x_0, x_1, \ldots, x_\tau)$ satisfy $0 < x_j/v_j < u$. The other decision variables satisfy either $x_j/v_j = u$ or $x_j/v_j = 0$. This result ultimately becomes useful to characterize the structure of the optimal subset offer probabilities. Before giving this result, we make one observation that becomes useful in the proof. In particular, noting the assumption that r_{ij} is decreasing in j for fixed i, the definition of $\bar{r}_j$ implies that $\bar{r}_j = r_{0j}\,r_{1j}\ldots r_{jj} \ge r_{0,j+1}\,r_{1,j+1}\ldots r_{j,j+1}\,r_{j+1,j+1} = \bar{r}_{j+1}$, establishing that $\bar{r}_0 \ge \bar{r}_1 \ge \ldots \ge \bar{r}_\tau$.

Proposition 2. *Assume that $\{s_j : j\in\mathcal{T}\}$ share a common value. Then, there exists an optimal solution (x^*, u^*) with $x^* = (x_0^*, \ldots, x_\tau^*)$ to problem (12)-(15) that satisfies*

$$u^* = \frac{x_0^*}{v_0} = \ldots = \frac{x_{k-1}^*}{v_{k-1}} \ge \frac{x_k^*}{v_k} \ge \frac{x_{k+1}^*}{v_{k+1}} = \ldots = \frac{x_\tau^*}{v_\tau} = 0 \tag{16}$$

for some $k\in\mathcal{T}$.

Proof. We let (x^*, u^*) be an optimal solution to problem (12)-(15) and $y^* = \sum_{j\in\mathcal{T}}\lambda\,\bar{r}_j\,x_j^*$ so that the optimal objective value of the problem is $R(y^*)$. For a fixed value of u, consider the problem

$$\zeta(u) = \min \quad \sum_{j\in\mathcal{T}} x_j \tag{17}$$

$$\text{subject to} \quad \sum_{j\in\mathcal{T}}\lambda\,\bar{r}_j\,x_j = y^* \tag{18}$$

$$0 \le x_j \le v_j\,u \qquad j\in\mathcal{T}, \tag{19}$$

whose optimal objective value is denoted by $\zeta(u)$. For the moment, assume that there exists $\hat{u}$ satisfying $\zeta(\hat{u}) = 1 - \hat{u}$. We establish the existence of such $\hat{u}$ later in the proof. We let $\hat{x} = (\hat{x}_0, \ldots, \hat{x}_\tau)$ be an optimal solution to the problem above when we solve this problem with $u = \hat{u}$. In this case, we have $\sum_{j\in\mathcal{T}}\hat{x}_j = \zeta(\hat{u}) = 1 - \hat{u}$, $\sum_{j\in\mathcal{T}}\lambda\,\bar{r}_j\,\hat{x}_j = y^*$, $\hat{x}_j/v_j \le \hat{u}$ for all $j\in\mathcal{T}$, which implies that $(\hat{x}, \hat{u})$ is a feasible solution to problem (12)-(15) providing an objective function value of $R(y^*)$. Thus, the solution $(\hat{x}, \hat{u})$ is also optimal to problem (12)-(15).

Problem (17)-(19) is a knapsack problem where the items are indexed by $\mathcal{T}$, the disutility of each item is 1 and the space requirement of item j is $\lambda\,\bar{r}_j$. We can put at most $v_j\,u$ units of item j into the

knapsack. We can solve this knapsack problem by starting from the item with the smallest disutility to space ratio and filling the knapsack with the items in that order. Noting that $\bar{r}_0 \geq \bar{r}_1 \geq \ldots \geq \bar{r}_\tau$, the disutility to space ratios of the items satisfy $1/(\lambda\,\bar{r}_0) \leq 1/(\lambda\,\bar{r}_1) \leq \ldots \leq 1/(\lambda\,\bar{r}_\tau)$ so that it is optimal to fill the knapsack with the items in the order $0, 1, \ldots, \tau$. Therefore, if we solve problem (17)-(19) with $u = \hat{u}$, then the optimal solution $\hat{x}$ satisfies $\hat{x}_0 = v_0\,\hat{u}$, $\hat{x}_1 = v_1\,\hat{u}, \ldots, \hat{x}_{k-1} = v_{k-1}\,\hat{u}$, $\hat{x}_k \in [0, v_k\,\hat{u}]$, $x_{k+1} = 0, \ldots, x_\tau = 0$ for some $k \in \mathcal{T}$. Therefore, the solution $(\hat{x}, \hat{u})$, which is optimal to problem (12)-(15), satisfies (16) and we obtain the desired result.

It remains to show that there exists $\hat{u}$ with $\zeta(\hat{u}) = 1 - \hat{u}$. Note that $\zeta(\cdot)$ is continuous. By the definitions of (x^*, u^*) and y^*, if we solve problem (17)-(19) with $u = u^*$, then x^* is a feasible solution providing an objective value of $\sum_{j\in\mathcal{T}} x^*_j = 1 - u^*$, where the equality is by (13). Since the solution x^* may not be optimal to problem (17)-(19), we obtain $\zeta(u^*) \leq 1 - u^*$. Also, we clearly have $\zeta(1) \geq 0$. Letting $g(u) = 1 - u$, we obtain $\zeta(u^*) \leq g(u^*)$ and $\zeta(1) \geq g(1)$. Since $\zeta(\cdot)$ and $g(\cdot)$ are continuous, there exists $\hat{u}$ such that $\zeta(\hat{u}) = g(\hat{u}) = 1 - \hat{u}$. □

We emphasize that the critical assumption in Proposition 2 is that $\bar{s}_j = \bar{r}_j\,s_0$ for all $j \in \mathcal{T}$ and it is possible to modify this Proposition 2 to accommodate the cases where we do not necessarily have the ordering $\bar{r}_0 \geq \bar{r}_1 \geq \ldots \geq \bar{r}_\tau$. The key observation is that whatever ordering we have among the probabilities $\{\bar{r}_j : j \in \mathcal{T}\}$, the disutility to space ratios of the items in the problem (17)-(19) satisfy the reverse ordering as long as $\bar{s}_j = \bar{r}_j\,s_0$. In this case, we can modify the ordering of the decision variables $(x_0, \ldots, x_\tau)$ in the chain of inequalities in (16) in such a way they follow the ordering of $\{r_j : j \in \mathcal{T}\}$ and the proof of Proposition 2 still goes through.

We can build on Proposition 2 to solve problem (12)-(15) through bisection search. In particular, for fixed values of k and u, (16) shows that the decision variables $x_0, x_1, \ldots, x_{k-1}$ can be fixed at $x_j = u\,v_j$ for all $j = 0, 1, \ldots, k-1$, whereas the decision variables $x_{k+1}, \ldots, x_\tau$ can be fixed at 0. So, for a fixed value of k, to find the best values of u and x_k, we can solve the problem

$$\begin{aligned} Z_k = \max \quad & R\Big(\sum_{j=0}^{k-1} \lambda\,\bar{r}_j\,v_j\,u + \lambda\,\bar{r}_k\,x_k\Big) \\ \text{subject to} \quad & \sum_{j=0}^{k-1} v_j\,u + x_k + u = 1 \\ & \frac{x_k}{v_k} - u \leq 0 \\ & x_k, u \geq 0, \end{aligned}$$

where we set the optimal objective value Z_k of the problem above to $-\infty$ whenever the problem is infeasible. To find the best value of k, we can simply compute Z_k for all $k \in \mathcal{T}$ and pick the value that provides the largest value of k. Thus, $\max\{Z_k : k \in \mathcal{T}\}$ corresponds to the optimal objective value of problem (12)-(15). Although the problem above, which provides Z_k for a fixed value of k, appears to involve the two decision variables x_k and u, we can solve one of the decision variables in terms of the other one by using the first constraint, in which case, the problem above becomes

a scalar optimization problem. Furthermore, noting the definition of $R(\cdot)$ and the discussion in the last two paragraphs of Section 4.1, the objective function of the problem above is concave. Therefore, we can solve the problem above by using bisection search.

The next corollary shows two intuitive properties of the optimal subset availability probabilities. First, the optimal subset offer probabilities from problem (2)-(4) makes only subsets of the form $\{0, 1, \ldots, j\}$ available for some $j \in \mathcal{T}$. So, it is optimal to make a certain number of days into the future available without skipping any days in between. Second, the optimal subset offer probabilities randomize at most between two possible subsets and these two subsets differ from each other only in one day. These results indicate that the randomized nature of the static policy is not a huge concern, as we randomize between only two subsets that are not too different from each other.

Corollary 3. *Assume that $\{s_j : j \in \mathcal{T}\}$ share a common value. Then, there exists an optimal solution $h^* = \{h^*(S) : S \subset \mathcal{T}\}$ to problem (2)-(4) with only two of the decision variables satisfying $h^*(S_1) \geq 0$, $h^*(S_2) \geq 0$ for some $S_1, S_2 \subset \mathcal{T}$ and all of the other decision variables are equal to 0. Furthermore, the two subsets S_1 and S_2 are either of the form $S_1 = \emptyset$ and $S_2 = \{0\}$, or of the form $S_1 = \{0, 1, \ldots, j\}$ and $S_2 = \{0, 1, \ldots, j+1\}$ for some $j \in \mathcal{T}$.*

Proof. By Proposition 2, there exists an optimal solution x^* to problem (5)-(8) that satisfies (16). We define the subsets $S_0, S_1, \ldots, S_\tau$ as in the proof of Proposition 1 and construct an optimal solution h^* to problem (2)-(4) by using x^* as in (9). In this case, since x^* satisfies (16) for some $k \in \mathcal{T}$, only two of the decision variables $\{h(S) : S \subset \mathcal{T}\}$ can take on nonzero values and these two decision variables are $h^*(S_{k-1})$ and $h^*(S_k)$. Thus, the desired result follows by observing that S_k is a subset of the form $\{0, 1, \ldots, k\}$. □

4.3 Performance Guarantee

The static model in problem (2)-(4) makes each subset of days available with a fixed probability and does not consider the current state of the scheduled appointments when making its decisions. A natural question is what kind of performance we can expect from such a static model. In this section, we develop a performance guarantee for our static model. In particular, we study the policy obtained from the simple deterministic approximation

$$Z_{DET} = \max \quad \sum_{j \in \mathcal{T}} \lambda\, \bar{s}_j\, x_j - \theta \Big[\sum_{j \in \mathcal{T}} \lambda\, \bar{r}_j\, x_j - C \Big]^+ \tag{20}$$

$$\text{subject to} \quad \sum_{j \in \mathcal{T}} x_j + u = 1 \tag{21}$$

$$\frac{x_j}{v_j} - u \leq 0 \qquad j \in \mathcal{T} \tag{22}$$

$$x_j, u \geq 0 \qquad j \in \mathcal{T}. \tag{23}$$

The objective function above is the deterministic analogue of the one in problem (5)-(8), where the Poisson random variable is replaced by its expectation. Similar to problem (5)-(8), problem

(20)-(23) characterizes a static model, but this static model is obtained under the assumption that all random quantities take on their expected values. In this section, we show that the policy obtained from problem (20)-(23) has a reasonable performance guarantee, even though this problem ignores all uncertainty. Since problem (5)-(8) explicitly addresses the uncertainty, the policy obtained from this problem is trivially guaranteed to perform better than the one obtained from problem (20)-(23). It is common in revenue management literature to develop performance bounds for policies obtained from deterministic approximations. Gallego and van Ryzin (1994) give a performance bound for the policy obtained from a deterministic approximation in a dynamic pricing problem with a single product. Gallego and van Ryzin (1997) extend this work to multiple products. Levi and Radovanovic (2010) focus on resource allocation problems with reusable resources and provide a performance bound for the policy obtained from a deterministic approximation. Topaloglu et al. (2011) bound the performance of a deterministic approximation for overbooking over a single flight leg. Our work has a similar flavor to these papers but it ultimately allows us to point out the problem parameters that affect the performance of the policies from a static model.

In the next lemma, we begin by showing that the optimal objective value of problem (20)-(23) provides an upper bound on the expected profit per day generated by the optimal policy, which can be a policy that depends on the state of the system. The proof of this result follows from a standard argument that uses Jensen's inequality and is given in the appendix. Throughout this section, we use V^* to denote the expected profit per day generated by the optimal policy.

Lemma 4. *We have $Z_{DET} \geq V^*$.*

We let $\Pi(x)$ with $x = (x_0, \ldots, x_\tau)$ be the objective function of problem (5)-(8). If (x^*, u^*) is an optimal solution to this problem, then the static policy that uses the subset offer probabilities $h^* = \{h^*(S) : S \subset \mathcal{T}\}$ defined as in (9) generates an expected profit of $\Pi(x^*)$ per day. Since V^* is the expected profit per day generated by the optimal policy, we have $\Pi(x^*)/V^* \leq 1$. In the next proposition, we give a lower bound on $\Pi(x^*)/V^*$, which bounds the optimality gap of the policy that we obtain by solving problem (5)-(8).

Proposition 5. *Letting (x^*, u^*) be an optimal solution to problem (5)-(8), we have*

$$\frac{\Pi(x^*)}{V^*} \geq 1 - \frac{\theta \sqrt{\frac{\bar{r}_0}{2\pi}} \sqrt{\frac{\lambda \bar{r}_0}{C}}}{\bar{s}_0 \min\left\{\frac{v_0}{1+v_0}, \frac{C}{\lambda \bar{r}_0}\right\} \sqrt{\lambda}}. \tag{24}$$

Proof. Letting $(\hat{x}, \hat{u})$ be an optimal solution to problem (20)-(23), we have $\Pi(x^*) \geq \Pi(\hat{x})$. Since we can always offer the empty set with probability one, the optimal objective value of problem (5)-(8) is nonnegative and we obtain $\Pi(x^*)/V^* = [\Pi(x^*)]^+/V^* \geq [\Pi(\hat{x})]^+/V^*$. Using Lemma 4, we continue this chain of inequalities as $[\Pi(\hat{x})]^+/V^* \geq [\Pi(\hat{x})]^+/Z_{DET} \geq \Pi(\hat{x})/Z_{DET} = 1 - (Z_{DET} - \Pi(\hat{x}))/Z_{DET}$. So, it is enough to show that the second term on the right side of (24)

upper bounds $(Z_{DET} - \Pi(\hat{x}))/Z_{DET}$. For a Poisson random variable with mean α, Lemma 8 in the appendix shows that $\mathbb{E}\{[\mathsf{Pois}(\alpha) - C]^+\} \leq [\alpha - C]^+ + \alpha/\sqrt{2\,\pi\,C}$. In this case, letting $\beta = \sum_{j=1}^{T} \lambda\,\bar{s}_j\,\hat{x}_j$ and $\alpha = \sum_{j=1}^{T} \lambda\,\bar{r}_j\,\hat{x}_j$ for notational brevity, we obtain

$$\frac{Z_{DET} - \Pi(\hat{x})}{Z_{DET}} = \frac{[\beta - \theta\,[\alpha - C]^+] - [\beta - \theta\,\mathbb{E}\{[\mathsf{Pois}(\alpha) - C]^+\}]}{Z_{DET}} \leq \frac{\frac{\theta\,\alpha}{\sqrt{2\,\pi\,C}}}{Z_{DET}} \leq \frac{\frac{\theta\,\lambda\,\bar{r}_0}{\sqrt{2\,\pi\,C}}}{Z_{DET}}, \tag{25}$$

where the second inequality is by noting that $\sum_{j \in \mathcal{T}} \hat{x}_j \leq 1$, $\bar{r}_0 \geq \bar{r}_1 \geq \ldots \geq \bar{r}_\tau$ so that $\alpha \leq \lambda\,\bar{r}_0$.

We proceed to constructing a lower bound on Z_{DET}. The solution $(\tilde{x}, \tilde{u})$ we obtain by setting $\tilde{x}_0 = \frac{v_0}{1+v_0}$, $\tilde{u} = \frac{1}{1+v_0}$ and all other decision variables to zero is feasible to problem (20)-(23). Thus, if $\lambda\,\bar{r}_0\,\frac{v_0}{1+v_0} \leq C$, then we can lower bound Z_{DET} as $Z_{DET} \geq \lambda\,\bar{s}_0\,\tilde{x}_0 - \theta\,[\lambda\,\bar{r}_0\,\tilde{x}_0 - C]^+ = \lambda\,\bar{s}_0\,\frac{v_0}{1+v_0}$. On the other hand, if $\lambda\,\bar{r}_0\,\frac{v_0}{1+v_0} > C$, then the solution $(\tilde{x}, \tilde{u})$ we obtain by setting $\tilde{x}_0 = \frac{C}{\lambda\,\bar{r}_0}$, $\tilde{u} = 1 - \frac{C}{\lambda\,\bar{r}_0}$ and all other decision variables to zero is feasible to problem (20)-(23). Thus, if $\lambda\,\bar{r}_0\,\frac{v_0}{1+v_0} > C$, then we can lower bound Z_{DET} as $Z_{DET} \geq \lambda\,\bar{s}_0\,\tilde{x}_0 - \theta\,[\lambda\,\bar{r}_0\,\tilde{x}_0 - C]^+ = \bar{s}_0\,\frac{C}{\bar{r}_0}$. Collecting the two cases together, we lower bound Z_{DET} by $\bar{s}_0 \min\left\{\lambda\,\frac{v_0}{1+v_0}, \frac{C}{\bar{r}_0}\right\}$. Continuing the chain of inequalities in (25) by using the lower bound on Z_{DET}, we obtain

$$\frac{\frac{\theta\,\lambda\,\bar{r}_0}{\sqrt{2\,\pi\,C}}}{Z_{DET}} \leq \frac{\frac{\theta\,\lambda\,\bar{r}_0}{\sqrt{2\,\pi\,C}}}{\bar{s}_0 \min\left\{\lambda\,\frac{v_0}{1+v_0}, \frac{C}{\bar{r}_0}\right\}}.$$

Arranging the terms in the last expression above yields the desired result. □

The performance guarantee in (24) has useful practical implications. Noting that $\lambda\,\bar{r}_0$ is an upper bound on the expected number of appointments we retain on a particular day, if C and λ satisfy $C \geq \lambda\,\bar{r}_0$, then we essentially have a situation where the capacity exceeds the expected demand. In this case, the quantity $\sqrt{\lambda\,\bar{r}_0/C}$ in the numerator on the right side of (24) is upper bounded by 1 and the min operator in the denominator evaluates to $\frac{v_0}{1+v_0}$, which implies that $\frac{\Pi(x^*)}{V^*}$ is lower bounded by $1 - \theta\,\sqrt{\bar{r}_0/(2\,\pi)}/(\bar{s}_0\,\frac{v_0}{1+v_0}\,\sqrt{\lambda})$. Therefore, the performance guarantee improves with rate $1/\sqrt{\lambda}$, indicating that as long as capacity exceeds the expected demand, the performance of static policies improves with the demand volume. In other words, we expect static policies to perform well in systems handling large demand volumes. Similarly, if the capacity and the expected demand increase with the same rate so that the ratio $\frac{C}{\bar{r}_0\,\lambda}$ stays constant, then the performance of state independent policies still improves with rate $1/\sqrt{\lambda}$, even when $\frac{C}{\lambda\,\bar{r}_0}$ evaluates to a number less than 1. These observations support using policies from static models for large systems with high demand volumes.

It turns out that λ does not have to get too large to get a sensible performance guarantee from (24) for practical applications. To get a feel for the performance guarantee in (24), we consider a system with $C = 12$, $\lambda = 16$, $v_0 = 1$, $\bar{r}_0 = 1$, $\bar{s}_0 = 0.9$ and $\theta = 1.5$. These parameters correspond to a case where if we offer only the current day for an appointment, then a patient leaves without

scheduling any appointments with probability $N(\{0\}) = 1/(1+v_0) = 1/2$. The probability that we retain a patient with a same day appointment is 1 and the probability that this patient shows up for her appointment is 0.9. Considering the fact that we generate a nominal revenue of 1 from each served patient, the overtime cost of 1.5 is reasonably high. The capacity and the arrival rate are reasonable for a small clinic. For these problem parameters, the right side of (24) evaluates to about 62%, indicating that there exists a static policy that generates at least 62% of the optimal expected profit per day. For a larger clinic with $C = 36$ and $\lambda = 48$, the performance guarantee comes out to be about 78%.

4.4 Extensions to Static Model

Our results in Section 4.2 demonstrate that if the show-up probability of an appointment does not depend on how many days ago the patient called in, then the randomized nature of the static model is not a huge practical concern. Under this assumption, there exists an optimal solution to problem (2)-(4) that randomizes between only two subsets of days and these subsets include a certain number of consecutive days, without skipping any days in between. A natural question is whether we can alleviate the concerns about the randomized nature of the static model when the show-up probabilities do depend on how many days ago the patient called in.

If we want to eliminate randomization between multiple subsets of days in problem (2)-(4), then we can impose the constraints $h(S) \in \{0,1\}$ for all $S \subset \mathcal{T}$ in this problem, in which case, problem (2)-(4) looks for one set of days to offer to patients. It turns out that a transformation similar to the one in Section 4.1 still holds even when we have the constraints $h(S) \in \{0,1\}$ for all $S \subset \mathcal{T}$ in problem (2)-(4). In particular, if we have the constraints $h(S) \in \{0,1\}$ for all $S \subset \mathcal{T}$ in problem (2)-(4), then we can add the constraints $x_j/v_j \in \{0,u\}$ for all $j \in \mathcal{T}$ in problem (5)-(8). Under this modification to problem (5)-(8), problem (5)-(8) becomes equivalent to problem (2)-(4) with the additional constraints that $h(S) \in \{0,1\}$ for all $S \subset \mathcal{T}$. We record this result in the next proposition. The proof follows from an argument similar to that of Proposition 1 and we defer it to the appendix.

Proposition 6. *Consider problem* (2)-(4) *with the additional constraints* $h_t(S) \in \{0,1\}$ *for all* $S \subset \mathcal{T}$ *and problem* (5)-(8) *with the additional constraints* $x_j/v_j \in \{0,u\}$ *for all* $j \in \mathcal{T}$*. These two problems are equivalent in the sense that given an optimal solution to one problem, we can generate a feasible solution to the other one providing the same objective value.*

Therefore, to find one set of days to offer to customers, we can solve problem (5)-(8) with the additional constraints $x_j/v_j \in \{0,u\}$ for all $j \in \mathcal{T}$. When we have the constraints $x_j/v_j \in \{0,u\}$ for all $j \in \mathcal{T}$ in problem (5)-(8), the algorithmic strategies that we can use to solve this problem closely mirror those described at the end of Section 4.1. In particular, since the nonlinearity in the objective function of problem (5)-(8) involves the scalar function $F(\alpha) = \mathbb{E}\{[\mathsf{Pois}(\alpha) - C]^+\}$, it is straightforward to construct accurate piecewise linear approximations to this scalar function. In

this case, problem (5)-(8) with the additional constraints that $x_j/v_j \in \{0, u\}$ for all $j \in \mathcal{T}$ can be formulated as a mixed integer linear program. Another option is to use the dynamic program in (10)-(11). In particular, if we impose the constraint $x_j/v_j \in \{0, u\}$ in each decision epoch, then this dynamic program solves problem (5)-(8) for a fixed value of u and with the additional constraints $x_j/v_j \in \{0, u\}$ for all $j \in \mathcal{T}$.

We emphasize that the reformulation in Section 4.1 is still instrumental to solving problem (2)-(4) under the additional constraints $h_t(S) \in \{0, 1\}$ for all $S \subset \mathcal{T}$. Without this reformulation, problem (2)-(4) has $2^{|\mathcal{T}|}$ decision variables, whereas our reformulation yields a problem with $|\mathcal{T}|+1$ decision variables, which can either directly be solved as a mixed integer linear program or through dynamic programming.

Another issue in problem (2)-(4) is that this problem is constructed under the assumption that the arrival rates are stationary and the preference of a patient for different appointment days depends on how many days into the future the appointment is made, rather than the particular day of the week of the appointment. Under this assumption, we were able to focus on the expected profit per day in problem (2)-(4). In practice, however, the arrival rates may be nonstationary, depending on the day of the week. Furthermore, the preference of a patient for different days of the week may be more pronounced than the preference for different appointment delays. It is possible to enrich our static model to capture the case where the arrival rates are nonstationary and a patient has different preferences for different days of the week. If nonstationarities follow, say, a weekly pattern, then we can choose the scheduling horizon τ to be a multiple of a week. In this case, we need to work with the extended set of decision variables by letting $h_t(S)$ be the probability with which we offer the subset S of days given that we are on day t of the week. Using these decision variables, we can construct a model analogous to problem (2)-(4), but we account for the total expected profit per a week in the objective function, rather than expected profit per day. For this extended version of problem (2)-(4), we can still come up with a transformation similar to the one in Section 4.1 that reduces the number of decision variables from exponential in the length of the scheduling horizon to linear. Finally, our model assumes that the patients do not have a preference between different hours of the day. Just as we can incorporate preferences between different days of the week, we can incorporate preferences between different hours of the day, as long as we incur overtime cost whenever the aggregate capacity availability in a day is violated.

5 Dynamic Model

The static model in the previous section identifies a fixed set of probabilities with which to offer each subset of days, independent of the currently booked appointments. In other words, the policy from this static model does not take the state of the system into consideration. Clearly, there is potential to improve such a policy if decisions can take into account the current system state information. In this section, we begin by giving a Markov decision process formulation of the

appointment scheduling operation. By building on this formulation, we develop a dynamic model that makes its decisions by considering the current state of the system.

5.1 Dynamic Program

As mentioned at the end of Section 3, for $1 \leq i \leq j \leq \tau$, we can use $X_{ij}(t)$ to denote the number of patients that called in i days ago (on day $t-i$) and scheduled an appointment j days into the future (for day $t-i+j$) given that we are at the beginning of day t. We call the vector $X(t) = \{X_{ij}(t) : 1 \leq i \leq j \leq \tau\}$ as the appointment schedule at the beginning of day t. For a given policy π, let $h^\pi(X(t), S)$ denote the probability that the subset S of days is offered to the patients when the system state is $X(t)$ under policy π. In this case, the evolution of the appointment schedule $X^\pi(t)$ under policy π is captured by

$$X_{ij}^\pi(t+1) = \begin{cases} \mathsf{Pois}(\sum_{S \subset \mathcal{T}} \lambda \, r_{0j} \, P_j(S) \, h^\pi(X^\pi(t), S)) & \text{if } 1 = i \leq j \leq \tau \\ \mathsf{Bin}(X_{i-1,j}^\pi(t), r_{i-1,j}) & \text{if } 2 \leq i \leq j \leq \tau, \end{cases}$$

where we use $\mathsf{Bin}(n,p)$ to denote a Binomial random variable with parameters n and p. On the other hand, given that we are on day t, we let $U_i^\pi(t)$ denote the number of patients who called on day $t-i$ to make an appointment for the current day and show up for their appointment on day t. As a function of the appointment schedule on day t, we can characterize $U_i^\pi(t)$ by

$$U_i^\pi(t) = \begin{cases} \mathsf{Pois}(\sum_{S \subset \mathcal{T}} \lambda \, s_0 \, P_0(S) \, h^\pi(X^\pi(t), S)) & \text{if } i = 0 \\ \mathsf{Bin}(X_{ii}^\pi(t), s_i) & \text{if } 1 \leq i \leq \tau. \end{cases}$$

Finally, we let $X_{00}^\pi(t)$ be the number of patients who call on day t and choose to schedule their appointments on the same day, in which case, $X_{00}^\pi(t)$ is characterized by

$$X_{00}^\pi(t) = \mathsf{Pois}\Big(\sum_{S \subset \mathcal{T}} \lambda \, P_0(S) \, h^\pi(X^\pi(t), S) \Big).$$

Define $\phi^\pi(x)$ to be the long-run average expected reward under policy π given the initial state $x = X^\pi(0)$, which is to say that

$$\phi^\pi(x) = \lim_{k \to \infty} \frac{\mathbb{E}\Big\{ \sum_{t=0}^{k} \Big(\sum_{i=0}^{\tau} U_i^\pi(t) - \theta \Big[\sum_{i=1}^{\tau} X_{ii}^\pi(t) + X_{00}(t) - C \Big]^+ \Big) \Big| \, X^\pi(0) = x \Big\}}{k}.$$

A scheduling policy π^* is optimal if it satisfies $\phi^{\pi^*}(x) = \sup_\pi \phi^\pi(x)$ for all x.

We can try to bound the number of appointments on a given day by a reasonably large number to ensure that there is a finite number of possible values for the appointment schedule $X(t)$. However, even if we bound the number of appointments, the appointment schedule $X(t)$ is a high-dimensional vector and the number of possible values for the appointment schedule would get extremely large even for small values for the length of the scheduling horizon τ. This precludes us from using conventional dynamic programming algorithms, such as value iteration or policy iteration, to

compute the optimal scheduling policy. Therefore, it is of interest to develop computationally efficient approaches to obtain policies that consider the status of the booked appointments. We describe one such approach in the next section.

5.2 Policy Improvement Heuristic

In this section, we develop a dynamic model that makes its decisions by taking the current appointment schedule into consideration. The idea behind our dynamic model is to start with a static policy that ignores the current appointment schedule and apply the single step of the policy improvement algorithm. Thus, throughout the paper, we refer to our dynamic model as the policy improvement heuristic. It is important to note that our policy improvement heuristic does not call for applying the policy improvement step and computing the improved policy for every possible state of the appointment schedule. Rather, we apply the policy improvement step only for the current state of the appointment schedule. In this way, we compute the decisions of our policy improvement heuristic as needed, only for the states of the appointment schedule we visit. It turns out that finding the decisions made by our policy improvement heuristic requires solving a mathematical program similar to the one solved for our static model.

Our policy improvement heuristic is developed by building on a static policy mainly because we can, in this case, identify closed-form expressions for the value functions in the policy-improvement step, as described below. Ideally, when implementing the policy improvement heuristic, one uses the optimal static policy obtained by solving problem (2)-(4), but if there are reasonable alternative static policies, then one can simply use them instead. Our derivation of the policy improvement heuristic does not depend on whether or not the initially chosen static policy is obtained through problem (2)-(4) and we keep our presentation general.

We let $h(S)$ for $S \subseteq \mathcal{T}$ be the probability that the subset S of days is offered to each appointment request under a static policy, which, as mentioned above, may or may have not been obtained through problem (2)-(4). Given that we are at the beginning of a particular day, the current appointment schedule is given by $x = \{x_{ij} : 1 \leq i \leq j \leq \tau\}$ before we observe the appointment requests on the current day. Similar to $X_{ij}(t)$, the component x_{ij} of the appointment schedule x represents the number of appointments made i days ago for j days into the future. Consider a policy π that makes the subset S of days available with probability $f(S)$ on the current day and switches to the static probabilities $h = \{h(S) : S \subset \mathcal{T}\}$ from tomorrow on.

Given that we start with the appointment schedule x, define $\Delta(x, f, h)$ to be the difference in the long-run total expected rewards obtained by following policy π rather than the static policy that uses the subset offer probabilities h all along. To conduct a one step of the policy improvement algorithm, we need to maximize $\Delta(x, f, h)$ with respect to $f = \{f(S) : S \subset \mathcal{T}\}$. Since policy π and the static policy that uses h make identical decisions after the current day and appointments cannot be scheduled beyond τ days into the future, the appointment schedules under the two

policies are stochastically identical beyond the first $\tau+1$ days. Thus, we can write $\Delta(x,f,h) = Q_\pi(x,f,h) - Q_{SP}(x,h)$, where $Q_\pi(x,f,h)$ and $Q_{SP}(x,h)$ are respectively the total expected rewards accumulated over the next $\tau+1$ days under policies π and the static policy.

When determining f that maximizes $\Delta(x,f,h)$, the function $Q_{SP}(x,h)$ is simply a constant. Thus, our objective is equivalent to maximizing $Q_\pi(x,f,h)$. We proceed to deriving an expression for $Q_\pi(x,f,h)$. For $1 \le i \le j \le \tau$, we let $V_{ij}(x)$ denote the number of patients who called for appointments i days ago and will not cancel by the morning of their appointment, which is $j-i$ days from today. We let $W_{ij}(x)$ denote the number of these patients who show up for their appointments. Similarly, for $0 \le j \le \tau$, we let $\bar{V}_j(f)$ denote the number of patients who call to make an appointment today, schedule an appointment for day j in the scheduling horizon and do not cancel by the morning of their appointment. We let $\bar{W}_j(f)$ denote the number of these patients who show up for their appointments. Finally, for $1 \le k \le j \le \tau$, we define $\hat{V}_{kj}(h)$ to be the number of patients who will call to make an appointment on day k in the scheduling horizon and will not have canceled their appointment by the morning of their appointment, which is on day j of the scheduling horizon. We define $\hat{W}_{kj}(h)$ to be the number of these patients who will show up for their appointments. Under the assumption that the cancellation and no-show behavior of the patients are independent of each other and the appointment schedule, we characterize the six random variables defined above as $V_{ij}(x) = \mathsf{Bin}(x_{ij}, \tilde{r}_{ij})$, $W_{ij}(x) = \mathsf{Bin}(x_{ij}, \tilde{s}_{ij})$,

$$\bar{V}_j(f) = \mathsf{Pois}\Big(\sum_{S\subset\mathcal{T}} \lambda\, \bar{r}_j\, P_j(S)\, f(S)\Big), \qquad \bar{W}_j(f) = \mathsf{Pois}\Big(\sum_{S\subset\mathcal{T}} \lambda\, \bar{s}_j\, P_j(S)\, f(S)\Big),$$

$$\hat{V}_{kj}(h) = \mathsf{Pois}\Big(\sum_{S\subset\mathcal{T}} \lambda\, \bar{r}_{j-k}\, P_{j-k}(S)\, h(S)\Big), \quad \hat{W}_{kj}(h) = \mathsf{Pois}\Big(\sum_{S\subset\mathcal{T}} \lambda\, \bar{s}_{j-k}\, P_{j-k}(S)\, h(S)\Big),$$

where $\tilde{r}_{ij} = r_{ij}\, r_{i+1,j} \ldots r_{jj}$ is the probability that a patient who scheduled her appointment i days ago will not have canceled her appointment by the morning of the day of her appointment, which is $j-i$ days from today and $\tilde{s}_{ij} = s_j\, \tilde{r}_{ij}$ is the probability that a patient who scheduled her appointment i days ago will show up for her appointment, which is $j-i$ days from today.

Using the random variables defined above, we can capture the reward obtained by the policy that uses the subset offer probabilities f today and switches to the static probabilities h from tomorrow on. In particular, letting $q_j(x,f,h)$ denote the total expected reward obtained by this policy j days from today, we have

$$q_j(x,f,h) = \mathbb{E}\Big\{ \sum_{k=1}^{\tau-j} W_{k,k+j}(x) + \bar{W}_j(f) + \sum_{k=1}^{j} \hat{W}_{kj}(h) \\ - \theta\Big[\sum_{k=1}^{\tau-j} V_{k,k+j}(x) + \bar{V}_j(f) + \sum_{k=1}^{j} \hat{V}_{kj}(h) - C\Big]^+ \Big\},$$

where we set $\sum_{k=1}^{j} \hat{W}_{kj}(h) = 0$ and $\sum_{k=1}^{j} \hat{V}_{kj}(h) = 0$ when we have $j=0$. Thus, it follows that $Q_\pi(x,f,h) = \sum_{j=0}^{\tau} q_j(x,f,h)$ and we can implement the policy improvement heuristic by maximizing $\sum_{j\in\mathcal{T}} q_j(x,f,h)$ subject to the constraint that $\sum_{S\subset\mathcal{T}} f(S) = 1$. The optimal solution

$f^* = \{f^*(S) : S \subset \mathcal{T}\}$ to the last problem yields the probability with which each subset of days should be offered on the current day. We note that since $q_j(x, f, h)$ depends on the current state x of the appointment schedule, the offer probabilities f we obtain in this fashion also depend on the current appointment schedule. Finally, we observe that solving the last optimization problem simplifies when we approximate the binomial random variables $V_{ij}(x)$ and $W_{ij}(x)$ for $1 \leq i \leq j \leq \tau$ with Poisson random variables with their corresponding means. In that case, one can check that the last optimization problem is structurally the same as problem (2)-(4) and we can use arguments similar to those in Section 4 to obtain the decisions made by the policy improvement heuristic.

6 Computational Results

In this section, we begin by giving the findings from our discrete choice experiment conducted to elicit patient preferences and to obtain data for our computational work. We proceed to describing our benchmark policies and experimental setup. We conclude with our computational results.

6.1 Discrete Choice Experiment

Our experiment took place in the Farrell Community Health Center, a large urban primary care health center in New York City. This center serves about 26,000 patient visits per year. A self-report survey instrument was created in both English and Spanish to collect data. Adult patients waiting were approached and informed about this study. After verbal consent was obtained, each patient was given a questionnaire in her native language, which described a hypothetical health condition entailing a need for a medical appointment and listed a randomly generated set of two possible choices of appointment days (for example, the same day and 2 days from now) as well as the option of seeking care elsewhere. The patient was asked to consider her current schedule and mark the appointment choice she preferred. Since the nature of the health problem is likely to influence the preference of the patient, each patient was asked to make a choice for two different problems, one that can be characterized as "ambiguous" and the other as "urgent." When describing the health conditions, we used wording similar to one in Cheraghi-Sohi et al. (2008).

Data were collected from December 2011 to February 2012. To reduce sampling bias, we conducted surveys across different times of day and different days of week. Overall, 240 patients were eligible to participate, out of which 161 of them agreed to participate, yielding a response rate of 67%. Each patient were provided appointment slots over the next five days including the current day, corresponding to a scheduling horizon length of $\tau = 5$. We separated the data under the ambiguous health condition and the urgent health condition. For each health condition, we let $\mathcal{K}$ be the set of patients who picked an appointment day within the options offered and $\bar{\mathcal{K}}$ be the set of patients who picked the option of seeking care elsewhere. For each patient $k \in \mathcal{K}$, the data provide the set of appointment days S^k made available to this patient and the appointment day j^k chosen by this patient, whereas for each patient $k \in \bar{\mathcal{K}}$, the data provide the set of appointment

days S^k made available to the patient. As a function of the parameters $v = (v_1, \ldots, v_\tau)$ of the choice model, the likelihood function is given by

$$L(v) = \Big(\prod_{k \in \mathcal{K}} \frac{v_{j^k}}{1 + \sum_{j \in S^k} v_j} \Big) \Big(\prod_{k \in \bar{\mathcal{K}}} \frac{1}{1 + \sum_{j \in S^k} v_j} \Big)$$

We maximized the likelihood function above by using a nonlinear programming package to obtain the estimator for the preference weights v.

We use v^A and v^U to respectively denote the preference weights estimated by using the data obtained under the ambiguous and urgent health conditions. Our results yield the estimates $v^A = (4.99, 4.66, 3.84, 4.58, 2.54, 1.83)$ and $v^U = (2.19, 1.95, 1.09, 0.33, 0.71, 0.19)$. These estimates conform with intuition. The preference weights derived from the patients with an ambiguous health condition reveal that these patients are less sensitive to appointment delays. They appear to favor any appointment within four day window more or less equally. The patients with urgent symptoms desire faster access. They prefer appointments on the day during which they call and on the following day significantly stronger than other appointment day options. Also, we observe that the magnitudes of the preference weights are smaller when patients face an urgent health condition compared to when patients face an ambiguous health condition. Noting that the preference weight of seeking care elsewhere is normalized to one, this observation suggests that patients facing an urgent health condition have a higher tendency to seek care elsewhere.

6.2 Benchmark Policies

We compare the performance of static and dynamic policies through a series of computational experiments and test them against other benchmark policies. We work with a total of four policies in our computational experiments. The first policy we consider is a static policy based on the model in Section 4. We refer to this policy as SP, standing for static policy. SP solves problem (5)-(8) and transforms the optimal solution to this problem to an optimal solution to problem (2)-(4) by using (9). Using h^* to denote an optimal solution to problem (2)-(4) obtained in this fashion, SP offers the subset S of days with probability $h^*(S)$.

Our second policy is a dynamic policy based on the model in Section 5. We refer to this policy as DP, standing for dynamic policy. Letting h^* be the subset offer probabilities obtained by SP, given that the appointment schedule on the current day is given by x, DP maximizes $\sum_{j \in \mathcal{T}} q_j(x, f, h^*)$ subject to the constraint that $\sum_{S \subset \mathcal{T}} f(S) = 1$. In this case, using f^* to denote the optimal solution to this problem, on the current day, we offer the subset S of days with probability $f^*(S)$. We note that DP recomputes the subset offer probabilities at the beginning of each day by using the current appointment schedule, but the value of h^* stays constant throughout.

The third policy can be interpreted as a capacity controlled implementation of open access. We refer to this policy as CO, standing for controlled open access. CO makes only the current day available for appointments, but to balance the demand with the available capacity, it also allows the

possibility of offering no appointment days to an arriving patient. To achieve this, we solve problem (2)-(4) only by considering the sets $S = \{0\}$ and $S = \emptyset$, in which case, problem (2)-(4) becomes a convex optimization problem with two decision variables $h(\{0\})$ and $h(\emptyset)$. Letting $(h^*(\{0\}), h^*(\emptyset))$ be the optimal solution we obtain, CO offers the current day for appointments with probability $h^*(\{0\})$ and does not offer any appointment days with probability $h(\emptyset)$.

The fourth policy is a complement of CO, offering all days in the scheduling horizon. We refer to this policy as AN, standing for all or nothing. To be specific, AN either makes all days in the scheduling horizon available for appointments or does not offer an appointment day at all. The probabilities of offering these two options is obtained by using a similar approach to CO. In particular, we solve problem (2)-(4) only by considering the sets $S = \{0, 1, \ldots, \tau\}$ and $S = \emptyset$.

6.3 Experimental Setup

In all of our computational experiments, we set the daily arrival rate $\lambda = 16$. To come up with the capacity of the clinic C, we choose a nominal capacity $\bar{c}$ and use three different capacity levels of $0.75\,\bar{c}$, $\bar{c}$ and $1.25\,\bar{c}$. The nominal capacity $\bar{c}$ is computed as the minimum of the expected number of appointments that we retain on a particular day given that we open only the current day or all days in the scheduling horizon for appointments. In other words, we have $\bar{c} = \lambda \min\{\sum_{j \in \mathcal{T}} \bar{r}_j \, P_j(\{0\}), \sum_{j \in \mathcal{T}} \bar{r}_j \, P_j(\mathcal{T})\}$. If we open all days for appointments, then a greater fraction of patients book appointments, but these appointments might be scheduled further into the future and a significant portion of the appointments may be canceled before the appointment day. On the other hand, if we open only the current day for appointments, then a smaller fraction of patients book appointments, but they are less likely to cancel. We choose the nominal capacity value as the minimum of the patient load on a particular day under these two cases. We calibrate the probability of retaining an appointment such that $\bar{r}_j = 1 - 0.04\,j$. Recalling that we retain a patient with a j day appointment delay until the day of her appointment with probability $\bar{r}_j$, our choice of retaining probabilities models the situation where the probability of retaining a patient decreases with the appointment delay. We set $s_j = 1$ so that all patients that have not canceled by the day of their appointments show up.

Recalling that the revenue of serving a patient is normalized to one, we vary the overtime cost θ over three different values, 1.25, 1.5 and 1.75. We use two different values for the length of the scheduling horizon τ. In one set of test problems, we set $\tau = 5$. For these test problems, we work with two different preference weights, v^A and v^U, which are the preference weights estimated under the ambiguous and urgent health conditions in Section 6.1. In another set of test problems, we set $\tau = 15$. For these test problems, we use two different sets of preference weights that we generate ourselves. The first one is $v^I = (1, 1, \ldots, 1)$. Since all preference weights are equal, this situation corresponds to the case where a patient is indifferent over the days offered to her. The second set of preference weights is $v^D = \{1.5, 1.4, \ldots, 0.1\}$, capturing the situation where patients prefer later appointments less. Thus, our experimental setup varies the capacity of the clinic C

over $\{0.75\,\bar{c}, \bar{c}, 1.25\bar{c}\}$, the overtime cost θ over $\{1.25, 1.5, 1.75\}$ and the preference weights v over $\{v^A, v^U, v^I, v^D\}$, in which case, we obtain a total of 36 test problems.

We use a simulation study to test the performance of the four policies described in the previous section. On each day, we sample the number of appointment requests that arrive. Letting $h = \{h(S) : S \subset \mathcal{T}\}$ be the probability with which each subset of days is offered to a patient under the policy in consideration, each appointment request is offered the subset S of days with probability $h(S)$, independent of the others. The patient making the appointment request chooses among the offered days according to the multinomial logit model or leaves without booking an appointment. Once we collect the booked appointments on the current day in this fashion and update the appointment schedule, we sample the cancellations and no shows. Based on the no-shows, we account for the revenue and the overtime cost on the current day, after which, we can move on to the next day. We simulate each one of the four policies for 135 days and for 100 replications. The first 45 days of each replication is used as a warm up period, during which we do not collect any revenue or cost information.

6.4 Computational Results

Our computational results are given in Tables 1 and 2. Table 1 focuses on the test problems with $\tau = 15$, having the preference weights v^I or v^D. Table 2 focuses on the test problems with $\tau = 5$, having the preference weights v^A or v^U. The first three columns in these tables show the characteristics of the test problems by using the triplet (v, C, θ). The following four columns show the expected profit per day obtained by the four policies. The last three columns show the percent gap between the expected profit per day obtained by DP and the remaining three policies.

Our results in Table 1 indicate that DP can provide substantial improvements over the other three policies that do not consider the state of the appointment schedule when making their decisions. For the test problems with $v = v^I$, DP performs better than SP, CO and AN respectively by 8.96%, 12.59% and 16.64% on average. The corresponding gaps are 6.68%, 9.92% and 13.27% for the test problems with $v = v^D$. In all of the test problems in Table 1, there is a statistically significant gap between the performance of DP and the performance of the best of the remaining three policies. Overall, the performance gaps between DP and the other three policies tend to increase as the overtime cost gets larger. For the test problems with $\theta = 1.75$, the performance gap between DP and all other three policies can exceed 10%. Indeed, when the overtime cost is high, it becomes more important to make the scheduling decisions more carefully and the improvement provided by DP over the static policies, SP, CO and AN, become more noticeable. Interestingly, the test problems with $C = 1.25\,\bar{c}$ have relatively ample capacity, but this does not necessarily mean that simpler static policies provide satisfactory performance.

The improvement provided by DP over the other policies can be due to both the increased number of patients served and the decreased overtime cost. In particular, consider a policy such as

Prob. Char.			Exp. Prof. Per Day				% Gap with DP		
v	C	θ	SP	DP	CO	AN	SP	CO	AN
v^I	$0.75\bar{c}$	1.25	5.06	5.40	5.04	4.85	6.30	6.67	10.19
v^I	$0.75\bar{c}$	1.5	4.60	5.11	4.61	3.72	9.98	9.78	27.20
v^I	$0.75\bar{c}$	1.75	4.32	4.87	4.30	2.69	11.29	11.70	44.76
v^I	$\bar{c}$	1.25	6.93	7.36	6.52	6.93	5.84	11.41	5.84
v^I	$\bar{c}$	1.5	6.40	7.09	6.25	6.26	9.73	11.85	11.71
v^I	$\bar{c}$	1.75	6.06	6.81	6.02	5.61	11.01	11.60	17.62
v^I	$1.25\bar{c}$	1.25	7.81	8.27	6.85	7.66	5.56	17.17	7.38
v^I	$1.25\bar{c}$	1.5	7.27	8.04	6.66	7.16	9.58	17.16	10.95
v^I	$1.25\bar{c}$	1.75	6.89	7.77	6.53	6.67	11.33	15.96	14.16
Average							8.96	12.59	16.64
v^D	$0.75\bar{c}$	1.25	6.02	6.31	5.99	5.75	4.60	5.07	8.87
v^D	$0.75\bar{c}$	1.5	5.49	5.93	5.5	4.62	7.42	7.25	22.09
v^D	$0.75\bar{c}$	1.75	5.15	5.60	5.16	3.49	8.04	7.86	37.68
v^D	$\bar{c}$	1.25	7.86	8.27	7.57	7.87	4.96	8.46	4.84
v^D	$\bar{c}$	1.5	7.31	7.91	7.26	7.16	7.59	8.22	9.48
v^D	$\bar{c}$	1.75	6.96	7.61	6.92	6.47	8.54	9.07	14.98
v^D	$1.25\bar{c}$	1.25	9.68	10.18	8.53	9.50	4.91	16.21	6.68
v^D	$1.25\bar{c}$	1.5	9.14	9.62	8.36	9.09	4.99	13.10	5.51
v^D	$1.25\bar{c}$	1.75	8.73	9.60	8.25	8.71	9.06	14.06	9.27
Average							6.68	9.92	13.27

Table 1: Performance of the four policies on the test problems with $v = v^I$ or $v = v^D$.

AN, which does not impose a control on the days offered and either offers all days in the scheduling horizon or does not offer anything. Due to the wide availability of options, the patients may be more likely to book appointments under AN, but since AN offers all days, patients have the option of booking appointments further into the future. Noting that the cancellation probabilities are dependent on the appointment delay, more patients cancel. As a result, a policy like AN, may end up serving a smaller number of patients when compared to DP, despite the fact that AN makes a wide range of days available for appointments. SP alleviates this shortcoming of AN to a certain extent since it chooses the sets of days to offer by balancing the tendency of the patients to book and cancel appointments. On the other hand, from the perspective of overtime cost, due to their static nature, policies such as SP, CO and AN cannot immediately react to capacities being filled up and may schedule unnecessary overtime. Overall, DP chooses the set of days to offer by considering the fact that a large set of days increases the likelihood that a patient books an appointment, but appointments further into the future are more likely to be canceled. Furthermore, DP adjusts the set of offered days in response to the current appointment schedule. Thus, DP can effectively increase the number of patients served, while decreasing the overtime cost.

Table 2 gives our computational results for the case where the preference weights are based on our discrete choice experiment. We observe that DP continues to provide noticeable improvements over the other three static policies. Over the test problems with the ambiguous health condition, DP improves the performance of SP, CO and AN respectively by 3.04%, 4.84% and 5.41% on average. The same performance gaps come out to be 4.12%, 8.13% and 9.09% for the test problems with

Prob. Char.			Exp. Prof. Per Day				% Gap with DP		
v	C	θ	SP	DP	CO	AN	SP	CO	AN
v^A	$0.75\bar{c}$	1.25	8.82	8.83	8.82	8.75	0.11	0.11	0.91
v^A	$0.75\bar{c}$	1.5	8.22	8.25	8.23	7.78	0.36	0.24	5.70
v^A	$0.75\bar{c}$	1.75	7.80	7.80	7.81	6.80	0.00	-0.13	12.82
v^A	$\bar{c}$	1.25	11.42	$^{\times}$11.86	11.15	11.41	3.71	5.99	3.79
v^A	$\bar{c}$	1.5	10.95	$^{\times}$11.31	10.79	10.94	3.18	4.60	3.27
v^A	$\bar{c}$	1.75	10.53	10.52	10.47	10.51	-0.10	0.48	0.10
v^A	$1.25\bar{c}$	1.25	12.95	$^{\times}$13.92	12.34	12.84	6.97	11.35	7.76
v^A	$1.25\bar{c}$	1.5	12.80	$^{\times}$13.72	12.22	12.73	6.71	10.93	7.22
v^A	$1.25\bar{c}$	1.75	12.62	$^{\times}$13.48	12.10	12.52	6.38	10.24	7.12
Average							3.04	4.87	5.41
v^U	$0.75\bar{c}$	1.25	6.92	6.93	6.93	6.71	0.14	0.00	3.17
v^U	$0.75\bar{c}$	1.5	6.41	$^{\times}$ 6.65	6.38	5.52	3.61	4.06	16.99
v^U	$0.75\bar{c}$	1.75	6.04	$^{\times}$ 6.30	6.05	4.37	4.13	3.97	30.63
v^U	$\bar{c}$	1.25	9.70	$^{\times}$10.01	9.21	9.70	3.10	7.99	3.10
v^U	$\bar{c}$	1.5	9.13	$^{\times}$ 9.34	8.93	9.12	2.25	4.39	2.36
v^U	$\bar{c}$	1.75	8.70	8.69	8.65	8.55	-0.12	0.46	1.61
v^U	$1.25\bar{c}$	1.25	11.52	$^{\times}$12.71	10.23	11.53	9.36	19.51	9.28
v^U	$1.25\bar{c}$	1.5	11.36	$^{\times}$12.31	10.16	11.34	7.72	17.47	7.88
v^U	$1.25\bar{c}$	1.75	11.09	$^{\times}$11.91	10.08	11.10	6.88	15.37	6.80
Average							4.12	8.13	9.09

Table 2: Performance of the four policies on the test problems with $v = v^A$ or $v = v^U$.

the urgent health condition. SP provides noticeably better than CO and AN, indicating that it is worthwhile to solve a mathematical program to find the best probability with which each subset of days should be offered. The gap between SP and DP can still be significant and a policy, such as DP, that makes its decisions by considering the current appointment schedule can be attractive. In Table 2, we use a "$\times$" to mark the test problems where the performance gap between DP and the best performing static policy is statistically significant at 5% level. In 13 out of 18 test problems, the performance gap is statistically significant.

Our computational experiments indicate that there are significant benefits from using dynamic policies that explicitly consider the state of the appointment schedule. Inevitably, there is additional computational burden brought out by dynamic policies, but for the test problems we considered, the decision made by DP can be computed in less than 0.01 seconds. Considering that the subset offer probabilities used by DP are recomputed at the beginning of each day, such run times are reasonable. Furthermore, run times that are on the order of a few milliseconds are suitable even if we recompute the subset offer probabilities for each individual patient. Online scheduling systems, which are becoming more prevalent in practice, offer immediate access to current appointment schedules and provide a particularly suitable environment for implementing dynamic policies.

7 Conclusion

Efficient use of healthcare resources is an important challenge. There is high demand for limited resources and the demand is expected to increase in the near future, particularly in the United States with the recent passage of the Patient Protection and Affordable Care Act. Appointment scheduling is one way of smoothing out the variable daily patient demand, which not only helps in increasing the utilization of resources, but also provides the patients with the flexibility of choosing when they will be seen. Although there is a growing body of work on various aspects of appointment scheduling, with few exceptions, the fact that patients have their own preferences regarding when they would like to be seen has been assumed away. This paper is directed to fill this gap.

The models we propose incorporate no-shows and cancellations that are dependent on the appointment delays, embed customer preferences among the different appointment days and take the current appointment schedule into consideration when making decisions. We provide a number of analytical characterizations that help simplify the way both the static and dynamic models are solved. We give a bound on the performance of the static model. Given that the dynamic model is obtained by applying the policy iteration algorithm on the static model, the dynamic model is trivially guaranteed to improve the performance of the static one. Our computational results, which make use of the patient utility estimates obtained from actual clinic data, reveal that our dynamic model can significantly outperform other benchmark policies.

In this paper, we use the multinomial logit model to capture patient preferences. Multinomial logit model has found many applications in economics, decision analysis and operations research to model choice behavior and there are well-established methods for estimating the parameters of this choice model from data. Recently, Gallego et al. (2011) describe an extension of the multinomial logit model where the preference weight of not booking an appointment increases if too few days are offered to patients. Under their choice model, each day in the scheduling horizon has two parameters, denoted by v_j and w_j. If we offer the subset S of days, then a patient chooses day j in the scheduling horizon with probability $P_j(S) = v_j/(1 + \sum_{k \notin S} w_k + \sum_{k \in S} v_k)$, where the quantity $\sum_{k \notin S} w_k$ captures the increase in the preference weight of not booking an appointment as a function of the days that are not offered. It is possible to check that all of our results in the paper continue to hold under this more general form of the multinomial logit model.

Our approach implicitly assumes that the choice behavior represented by the multinomial logit model captures the overall patient population preferences. This approach is reasonable when we do not have access to additional information about patient conditions. When additional information is available, subset offer decisions may depend on the condition of patients. For example, considering our computational study, if patients having urgent and ambiguous conditions arrive simultaneously into the system and we have information about the condition of the patient before offering the available appointment days, then we can define two sets of decision variables $h^U = \{h^U(S) : S \subset \mathcal{T}\}$ and $h^A = \{h^A\{S) : S \subset \mathcal{T}\}$ to respectively capture the probability that each subset of days is

offered to a patient with urgent and ambiguous conditions. Our approach can be extended to jointly optimize the subset offer probabilities (h^U, h^A).

There are a number of research directions to pursue further. First, although the multinomial logit model is a widely used way of modeling choice, it is of interest to generalize our results to other choice models that can capture a variety of patient preferences. The key is to keep the corresponding optimization problems tractable, while using more intricate choice models. Second, the current paper focuses on the aggregate capacity on each day and charges the overtime cost by comparing the available daily capacity with the total number of appointments retained on the current day. If the patients are scheduled for different times of the day, then it might be possible to shift appointments within a day and still avoid overtime costs. Such an extension may result in more intricate overtime costs as it requires keeping track of the detailed appointment dynamics within a day and it is certainly worth investigation.

References

Ayvaz, N. and Huh, W. (2010), 'Allocation of hospital capacity to multiple types of patients', *Journal of Revenue & Pricing Management* **9**(5), 386–398.

Bean, A. G. and Talaga, J. (1995), 'Predicting appointment breaking.', *Journal of Health Care Marketing* **15**(1), 29–34.

Berwick, D., Nolan, T. and Whittington, J. (2008), 'The triple aim: Care, health, and cost', *Health Affairs* **27**(3), 759–769.

Bowser, D., Utz, S., Glick, D. and Harmon, R. (2010), 'A systematic review of the relationship of diabetes mellitus, depression, and missed appointments in a low-income uninsured population', *Archives of psychiatric nursing* **24**(5), 317–329.

Bront, J. J. M., Diaz, I. M. and Vulcano, G. (2009), 'A column generation algorithm for choice-based network revenue management', *Operations Research* **57**(3), 769–784.

Cayirli, T. and Veral, E. (2003), 'Outpatient scheduling in health care: A review of literature', *Production and Operations Management* **12**(4), 519–549.

Cayirli, T., Yang, K. K. and Quek, S. A. (2012), 'A universal appointment rule in the presence of no-shows and walk-ins', *Production and Operations Management* . Forthcoming.

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

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

Dreiher, J., Froimovici, M., Bibi, Y., Vardy, D., Cicurel, A. and Cohen, A. (2008), 'Nonattendance in obstetrics and gynecology patients', *Gynecologic and Obstetric Investigation* **66**(1), 40–43.

Gallego, G., Iyengar, G., Phillips, R. and Dubey, A. (2004), Managing flexible products on a network, Computational Optimization Research Center Technical Report TR-2004-01, Columbia University.

Gallego, G., Ratliff, R. and Shebalov, S. (2011), A general attraction model and an efficient formulation for the network revenue management problem, Technical report, Columbia University, New York, NY.

Gallego, G. and van Ryzin, G. (1994), 'Optimal dynamic pricing of inventories with stochastic demand over finite horizons', *Management Science* **40**(8), 999–1020.

Gallego, G. and van Ryzin, G. (1997), 'A multiproduct dynamic pricing problem and its applications to yield management', *Operations Research* **45**(1), 24–41.

Gallucci, G., Swartz, W. and Hackerman, F. (2005), 'Brief reports: 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., Salisbury, C., Street, D., Pope, C. and Baxter, H. (2008), 'Is fast access to general practice all that should matter? A discrete choice experiment of patients' preferences', *Journal of Health Services Research & Policy* **13**(suppl 2), 3–10.

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

Green, J., McDowall, Z. and Potts, H. (2008), 'Does choose & book fail to deliver the expected choice to patients? a survey of patients' experience of outpatient appointment booking', *BMC Medical Informatics and Decision Making* **8**(1), 36.

Green, L., Savin, S. and Wang, B. (2006), 'Managing patient service in a diagnostic medical facility', *Operations Research* pp. 11–25.

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

Grunebaum, M., Luber, P., Callahan, M., Leon, A., Olfson, M. and Portera, L. (1996), 'Predictors of missed appointments for psychiatric consultations in a primary care clinic', *Psychiatric Services* **47**(8), 848–852.

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

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

Gupta, D. and Wang, W. (2011), Patient appointments in ambulatory care, *in* R. W. Hall, ed., 'Handbook of Healthcare System Scheduling: Delivering Care When and Where It is Needed', Springer, New York.

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

Hole, A. (2008), 'Modelling heterogeneity in patients' preferences for the attributes of a general practitioner appointment', *Journal of Health Economics* **27**(4), 1078–1094.

Huang, Y. and Zuniga, P. (2012), 'Dynamic overbooking scheduling system to improve patient access', *Journal of the Operational Research Society* . Forthcoming.

Huh, T., Liu, N. and Truong, V. (2012), 'Multi-resource allocation scheduling in dynamic environments.'. Working paper. Sauder School of Busniess, University of British Columbia.

Institute for Healthcare Improvement (2012), 'The IHI triple aim.'. Retrieved May 8, 2012, from http://www.ihi.org/offerings/Initiatives/TripleAim/Pages/default.aspx.

Jouini, O. and Benjaafar, S. (2010), 'Queueing systems with appointment-driven arrivals, non-punctual customers, and no-shows.'. Technical Report. University of Minnesota, Minneapolis, MN.

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

Kim, S. and Giachetti, R. E. (2006), 'A stochastic mathematical appointment overbooking model for healthcare providers to improve profits', *IEEE Transactions on Systems, Man and Cybernetics, Part A* **36**(6), 1211–1219.

Klassen, K. J. and Rohleder, T. R. (2004), 'Outpatient appointment scheduling with urgent clients in a dynamic, multi-period environment', *International Journal of Service Industry Management* **15**(2), 167–186.

Kök, A., Fisher, M. and Vaidyanathan, R. (2009), 'Assortment planning: Review of literature and industry practice', *Retail supply chain management* pp. 1–55.

Kopach, R., DeLaurentis, P. C., Lawley, M., Muthuraman, K., Ozsen, L., Rardin, R., Wan, H., Intrevado, P., Qu, X. and Willis, D. (2007), 'Effects of clinical characteristics on successful open access scheduling', *Health Care Management Science* **10**(2), 111–124.

Kunnumkal, S. and Topaloglu, H. (2008), 'A refined deterministic linear program for the network revenue management problem with customer choice behavior', *Naval Research Logistics Quarterly* **55**(6), 563–580.

LaGanga, L. and Lawrence, S. R. (2012), 'Appointment overbooking in health care clinics to improve patient service and clinic performance', *Production and Operations Management* . Forthcoming.

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

Levi, R. and Radovanovic, A. (2010), 'Provably near-optimal LP-based policies for revenue management in systems with reusable resources', *Operations Research* **58**(2), 503–507.

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

Liu, Q. and van Ryzin, G. (2008), 'On the choice-based linear programming model for network revenue management', *Manufacturing & Service Operations Management* **10**(2), 288–310.

Luo, J., Kulkarni, V. G. and Ziya, S. (2012), 'Appointment scheduling under patient no-shows and service interruptions', *Manufacturing and Service Operations Management* . Forthcoming.

McFadden, D. (1974), Conditional logit analysis of qualitative choice behavior, *in* P. Zarembka, ed., '*Frontiers in Economics*', Academic Press, pp. 105–142.

Murray, M. and Tantau, C. (2000), 'Same-day appointments: Exploding the access paradigm.'. *Family Practice Management*. September, 2000.

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

Patrick, J., Puterman, M. L. and Queyranne, M. (2008), 'Dynamic multi-priority patient scheduling for a diagnostic resource', *Operations Research* **56**(6), 1507–1525.

Qu, X., Rardin, R. L., Williams, J. A. S. and Willis, D. R. (2007), 'Matching daily healthcare provider capacity to demand in advanced access scheduling systems', *European Journal of Operational Research* **183**(2), 812–826.

Rau, J. (2011), 'Medicare to begin basing hospital payments on patient-satisfaction scores', *Kaiser Health News* . Retrieved on Jan 19, 2012, `http://www.kaiserhealthnews.org/Stories/2011/April/28/medicare-hospital-patient-satisfaction.aspx`.

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

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

Rohleder, T. and Klassen, K. (2000), 'Using client-variance information to improve dynamic appointment scheduling performance', *Omega* **28**(3), 293–302.

Rubin, G., Bate, A., George, A., Shackley, P. and Hall, N. (2006), 'Preferences for access to the GP: A discrete choice experiment', *The British Journal of General Practice* **56**(531), 743.

Rusmevichientong, P., Shmoys, D. B. and Topaloglu, H. (2010), Assortment optimization with mixtures of logits, Technical report, Cornell University, School of Operations Research and Information Engineering.
Available at `http://people.orie.cornell.edu/huseyin/publications/publications.html`.

Ruszczynski, A. (2006), *Nonlinear Optimization*, Princeton University Press, Princeton, New Jersey.

Ryan, M. and Farrar, S. (2000), 'Using conjoint analysis to elicit preferences for health care', *Bmj* **320**(7248), 1530.

Sampson, F., Pickin, M., O'Cathain, A., Goodall, S. and Salisbury, C. (2008), 'Impact of same-day appointments on patient satisfaction with general practice appointment systems', *The British Journal of General Practice* **58**(554), 641–643.

Schectman, J., Schorling, J. and Voss, J. (2008), 'Appointment adherence and disparities in outcomes among patients with diabetes', *Journal of general internal medicine* **23**(10), 1685–1687.

Silvestre, A., Sue, V. M. and Allen, J. Y. (2009), 'If you build it, will they come? the kaiser permanente model of online health care', *Electronic Health Records* **28**, 334–344.

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

Talluri, K. and Van Ryzin, G. (2004*b*), 'Revenue management under a general discrete choice model of consumer behavior', *Management Science* pp. 15–33.

Topaloglu, H. (2010), Joint stocking and product offer decisions under the multinomial logit model, Technical report, Cornell University, School of Operations Research and Information Engineering.
Available at `http://people.orie.cornell.edu/huseyin/publications/publications.html`.

Topaloglu, H., Birbil, S., Frenk, J. and Noyan, N. (2011), 'Tractable open loop policies for joint overbooking and capacity control over a single flight leg with multiple fare classes', *Transportation Science* (to appear).

US Department of Health and Human Services (2011), 'Electronic health records and meaningful use'. Retrieved on Jan 19, 2012, `http://healthit.hhs.gov/`.

Wang, P. (1999), 'Sequencing and scheduling N customers for a stochastic server', *European journal of operational research* **119**(3), 729–738.

Wang, W. and Gupta, D. (2011), 'Adaptive appointment systems with patient preferences', *Manufacturing & Service Operations Management* **13**(3), 373–389.

Weiner, M., El Hoyek, G., Wang, L., Dexter, P. R., Zerr, A. D., Perkins, A. J., James, F. and Juneja, R. (2009), 'A web-based generalist-specialist system to improve scheduling of outpatient specialty consultations in an academic center', *Journal of General Internal Medicine* **6**, 710–715.

Zeng, B., Turkcan, A., Lin, J. and Lawley, M. (2010), 'Clinic scheduling models with overbooking for patients with heterogeneous no-show probabilities.', *Annals of Operations Research* **178**(1), 121–144.

Zhang, D. and Adelman, D. (2009), 'An approximate dynamic programming approach to network revenue management with customer choice', *Transportation Science* **42**(3), 381–394.

Zocdoc (2012), 'About us.'. Retrieved on Jan 19, 2012, `http://www.zocdoc.com/aboutus`.

A Appendix: Omitted Results

In this section, we give the proofs of the results that are omitted in the paper.

A.1 Proof of Proposition 1

We complete the proof of Proposition 1 in two parts. First, assume that $h^* = \{h^*(S) : S \subset \mathcal{T}\}$ is an optimal solution to problem (2)-(4). Letting $x_j^* = \sum_{S \subset \mathcal{T}} P_j(S)\, h^*(S)$ and $u^* = \sum_{S \subset \mathcal{T}} N(S)\, h^*(S)$, we need to show that (x^*, u^*) is a feasible solution to problem (5)-(8) providing the same objective value as the solution h^*. We have

$$\sum_{j \in \mathcal{T}} x_j^* + u^* = \sum_{S \subset \mathcal{T}} h^*(S)[\sum_{j \in \mathcal{T}} P_j(S) + N(S)] = \sum_{S \subset \mathcal{T}} h^*(S) = 1,$$

where the second equality follows by the definition of the multinomial logit model and the third equality follows since h^* is feasible to problem (2)-(4). Thus, the solution (x^*, u^*) satisfies the first constraint in problem (5)-(8). On the other hand, using $\mathbf{1}(\cdot)$ to denote the indicator function, we have $x_j^*/v_j = \sum_{S \subset \mathcal{T}}[\mathbf{1}(j \in S)\, h^*(S)/(1 + \sum_{k \in S} v_k)]$ by the definition of $P_j(S)$ in the multinomial logit model. Noting that $u = \sum_{S \subset \mathcal{T}}[h^*(S)/(1 + \sum_{k \in S} v_k)]$, it follows that $x_j^*/v_j \leq u$, indicating that the solution (x^*, u^*) satisfies the second set of constraints in problem (5)-(8). Since $x_j^* = \sum_{S \subset \mathcal{T}} P_j(S)\, h^*(S)$, comparing the objective functions of problems (2)-(4) and (5)-(8) shows that the solutions h^* and (x^*, u^*) provide the same objective values for their problems.

Second, assume that (x^*, u^*) is a feasible solution to problem (5)-(8). We construct the solution h^* as in (9). To see that the solutions (x^*, u^*) and h^* provide the same objective values for their respective problems, we observe that

$$\sum_{S \subset \mathcal{T}} P_j(S)\, h^*(S) = \sum_{i=0}^{\tau} P_j(S_i)\, h^*(S_i) = \sum_{i=j}^{\tau} P_j(S_i)\, h^*(S_i) = \sum_{i=j}^{\tau} v_j \Big[\frac{x_i^*}{v_i} - \frac{x_{i+1}^*}{v_{i+1}}\Big] = x_j^*,$$

where the first equality is by the fact that $h^*(S)$ takes positive values only for the sets $S_0, \ldots, S_\tau$ and $\emptyset$, the second equality is by the fact that $j \in S_i$ only when $j \leq i$ and the third equality follows by the definition of $h^*(S_i)$ and noting that $x_{\tau+1}^* = 0$. Using the equality above and comparing the objective functions of problems (2)-(4) and (5)-(8) show that the solutions h^* and (x^*, u^*) provide the same objective values for their problems. To see that the solution h^* is feasible to problem (2)-(4), we let $V_i = 1 + \sum_{k \in S_i} v_k$ for notational brevity and write $\sum_{S \subset \mathcal{T}} h^*(S)$ as

$$u^* - \frac{x_0^*}{v_0} + \sum_{j=0}^{\tau} V_j \Big[\frac{x_j^*}{v_j} - \frac{x_{j+1}^*}{v_{j+1}}\Big] = u^* + (V_0 - 1)\frac{x_0^*}{v_0} + (V_1 - V_0)\frac{x_1^*}{v_1} + \ldots + (V_\tau - V_{\tau-1})\frac{x_\tau^*}{v_\tau} = 1,$$

where the first equality follows by rearranging the terms and using the convention that $x_{\tau+1}^* = 0$ and the second equality is by noting that $V_i - V_{i-1} = v_i$ and using the fact that (x^*, u^*) is feasible to problem (5)-(8) so that $\sum_{j=0}^{\tau} x_j^* + u^* = 1$. □

A.2 Proof of Proposition 6

The proof follows from an argument similar to the one in the proof of Proposition 1. Assume that $h^* = \{h^*(S) : S \subset \mathcal{T}\}$ is an optimal solution to problem (2)-(4) with the additional constraints $h_t(S) \in \{0, 1\}$ for all $S \subset \mathcal{T}$. Letting $x^*_j = \sum_{S \subset \mathcal{T}} P_j(S)\, h^*(S)$ and $u^* = \sum_{S \subset \mathcal{T}} N(S)\, h^*(S)$, we can follow the same argument in Section A.1 to show that (x^*, u^*) with $x^* = (x^*_0, \ldots, x^*_\tau)$ is a feasible solution to problem (5)-(8) with the additional constraints $x_j/v_j \in \{0, u\}$ for all $j \in \mathcal{T}$. Furthermore, the objective values provided by the two solutions for their respective problems are identical. On the other hand, assume that (x^*, u^*) with $x^* = (x^*_0, \ldots, x^*_\tau)$ is an optimal solution to problem (5)-(8) with the additional constraints $x_j/v_j \in \{0, u\}$ for all $j \in \mathcal{T}$. We reorder and reindex the days in the scheduling horizon so that we have $u^* = x^*_0/v_0 = x^*_1/v^*_1 = \ldots = x^*_{j-1}/v^*_{j-1} \geq x^*_j/x_j = x^*_{j+1}/v_{j+1} = \ldots = x^*_\tau/v^*_\tau = 0$. We define the subsets $S_0, S_1, \ldots, S_\tau$ as $S_j = \{0, 1, \ldots, j\}$. For notational convenience, we define $x^*_{\tau+1} = 0$. In this case, letting

$$h^*(\emptyset) = u^* - \frac{x^*_0}{v_0} \quad \text{and} \quad h^*(S_j) = \Big[1 + \sum_{k \in S_j} v_k\Big] \Big[\frac{x^*_j}{v_j} - \frac{x^*_{j+1}}{v_{j+1}}\Big]$$

for all $j = 0, 1, \ldots, \tau$ and letting $h^*(S) = 0$ for all other subsets of $\mathcal{T}$, we can follow the same argument in Section A.1 to show that $\{h^*(S) : S \subset \mathcal{T}\}$ is a feasible solution to problem (2)-(4) with the additional constraints $h_t(S) \in \{0, 1\}$ for all $S \subset \mathcal{T}$. Furthermore, we can check that the two solutions provide the same objective value for their respective problems. □

A.3 Lemma 7

The following lemma is used in Section 4.

Lemma 7. *Letting $F(\alpha) = \mathbb{E}\big\{[\mathsf{Pois}(\alpha) - C]^+\big\}$, $F(\cdot)$ is differentiable and convex.*

Proof. The proof uses elementary properties of the Poisson distribution. By using the probability mass function of the Poisson distribution, we have

$$\begin{aligned} F(\alpha) &= \sum_{i=C+1}^{\infty} \frac{e^{-\alpha}\,\alpha^i}{i!}\,(i - C) = \sum_{i=C+1}^{\infty} \frac{e^{-\alpha}\,\alpha^i}{(i-1)!} - \sum_{i=C+1}^{\infty} \frac{e^{-\alpha}\,\alpha^i}{i!}\,C \\ &= \alpha \sum_{i=C}^{\infty} \frac{e^{-\alpha}\,\alpha^i}{i!} - C \sum_{i=C+1}^{\infty} \frac{e^{-\alpha}\,\alpha^i}{i!} = \alpha\,\mathbb{P}\{\mathsf{Pois}(\alpha) \geq C\} - C\,\mathbb{P}\{\mathsf{Pois}(\alpha) \geq C + 1\}. \quad (26) \end{aligned}$$

Thus, the differentiability of $F(\cdot)$ follows by the differentiability of the cumulative distribution function of the Poisson distribution with respect to its mean. For the convexity of $F(\cdot)$, we have

$$\begin{aligned} \frac{d\,\mathbb{P}\{\mathsf{Pois}(\alpha) \geq C\}}{d\alpha} &= -\frac{d\,\mathbb{P}\{\mathsf{Pois}(\alpha) \leq C - 1\}}{d\alpha} = -\sum_{i=0}^{C-1} \frac{d\Big(\frac{e^{-\alpha}\,\alpha^i}{i!}\Big)}{d\alpha} \\ &= \sum_{i=0}^{C-1} \frac{e^{-\alpha}\,\alpha^i}{i!} - \sum_{i=1}^{C-1} \frac{e^{-\alpha}\,\alpha^{i-1}}{(i-1)!} = \mathbb{P}\{\mathsf{Pois}(\alpha) = C - 1\}. \end{aligned}$$

In this case, if we differentiate both sides of (26) with respect to α and use the last chain of equalities, then we obtain

$$\frac{dF(\alpha)}{d\alpha} = \mathbb{P}\{\mathsf{Pois}(\alpha) \geq C\} + \alpha\,\mathbb{P}\{\mathsf{Pois}(\alpha) = C-1\} - C\,\mathbb{P}\{\mathsf{Pois}(\alpha) = C\}$$
$$= \mathbb{P}\{\mathsf{Pois}(\alpha) \geq C\} + \alpha\,\frac{e^{-\alpha}\,\alpha^{C-1}}{(C-1)!} - C\,\frac{e^{-\alpha}\,\alpha^{C}}{C!} = \mathbb{P}\{\mathsf{Pois}(\alpha) \geq C\}.$$

To see that $F(\cdot)$ is convex, we use the last two chains of equalities to observe that the second derivative of $F(\alpha)$ with respect to α is $\mathbb{P}\big\{\mathsf{Pois}(\alpha) = C-1\big\}$, which is positive. □

A.4 Proof of Lemma 4

We let $\pi^*(S)$ be the steady state probability with which we offer the subset S of days under the optimal, possibly state-dependent, policy. So, if we consider a particular day in steady state, then the number of patients that are scheduled for this day j days in advance is given by a Poisson random variable with mean $\sum_{S\subset\mathcal{T}} \lambda\,P_j(S)\,\pi^*(S)$. Therefore, if we use the random variable A_j^* to denote the number of patients that we schedule for a particular day j days in advance in steady state, then A_j^* has mean $\sum_{S\subset\mathcal{T}} \lambda\,P_j(S)\,\pi^*(S)$. We note that $A_1^*, A_2^*, \ldots, A_\tau^*$ are not necessarily independent of each other, since the decisions under the optimal state-dependent policy on different days can be dependent. Similarly, in steady state, we let S_j^* be the number of patients that we schedule for a particular day j days in advance and that show up under the optimal state-dependent policy. Finally, we let R_j^* be the number of patients that we schedule for a particular day j days in advance and that we retain until the morning of the appointment under the optimal state dependent policy. Noting that the show-up and cancellation decisions of the patients are independent of how many patients we schedule for a particular day, we have $\mathbb{E}\{S_j^*\} = \bar{s}_j\,\mathbb{E}\{A_j^*\}$ and $\mathbb{E}\{R_j^*\} = \bar{r}_j\,\mathbb{E}\{A_j^*\}$. In this case, the average profit per day generated by the optimal state-dependent policy satisfies

$$V^* = \mathbb{E}\Big\{\sum_{j\in\mathcal{T}} S_j^*\Big\} - \theta\,\mathbb{E}\Big\{\Big[\sum_{j\in\mathcal{T}} R_j^* - C\Big]^+\Big\} \leq \sum_{j\in\mathcal{T}} \mathbb{E}\{S_j^*\} - \theta\Big[\sum_{j\in\mathcal{T}} \mathbb{E}\{R_j^*\} - C\Big]^+$$
$$= \sum_{j\in\mathcal{T}}\sum_{S\subset\mathcal{T}} \lambda\,\bar{s}_j\,P_j(S)\,\pi^*(S) - \theta\Big[\sum_{j\in\mathcal{T}}\sum_{S\subset\mathcal{T}} \lambda\,\bar{r}_j\,P_j(S)\,\pi^*(S) - C\Big]^+ \leq Z_{DET}.$$

In the chain of inequalities above, the first inequality is by the Jensen's inequality. The second equality is by $\mathbb{E}\{S_j^*\} = \bar{s}_j\,\mathbb{E}\{A_j^*\}$ and $\mathbb{E}\{R_j^*\} = \bar{r}_j\,\mathbb{E}\{A_j^*\}$. To see the second inequality, we note that $\{\pi^*(S) : S\subset\mathcal{T}\}$ is a feasible but not necessarily an optimal solution to the problem

$$\begin{aligned} \max \quad & \sum_{j\in\mathcal{T}}\sum_{S\subset\mathcal{T}} \lambda\,\bar{s}_j\,P_j(S)\,w(S) - \theta\Big[\sum_{j\in\mathcal{T}}\sum_{S\subset\mathcal{T}} \lambda\,\bar{r}_j\,P_j(S)\,w(S) - C\Big]^+ \\ \text{subject to} \quad & \sum_{S\subset\mathcal{T}} w(S) = 1 \\ & w(S) \geq 0 \qquad S\subset\mathcal{T} \end{aligned}$$

and the optimal objective values of the problem above and problem (20)-(23) are equal to each other, which can be verified by using the argument in the proof of Proposition 1. □

A.5 Lemma 8

The following result is used in the proof of Proposition 5.

Lemma 8. *It holds that* $\mathbb{E}\{[\mathsf{Pois}(\alpha) - C]^+\} \leq [\alpha - C]^+ + \alpha/\sqrt{2\,\pi\,C}$.

Proof. For $k \geq C + 1$, we observe that $[k - C]^+ - [\alpha - C]^+ \leq k - \alpha$. In particular, for $k \geq \alpha$, this inequality follows by the Lipschitz continuity of the function $[\cdot - C]^+$. For $k < \alpha$, we have $C + 1 \leq k < \alpha$ and it follows $[k - C]^+ - [\alpha - C]^+ = k - \alpha$, establishing the desired inequality. In this case, the result in the lemma follows by noting that

$$\begin{aligned}\mathbb{E}\{[\mathsf{Pois}(\alpha) - C]^+\} &= \sum_{k=C+1}^{\infty} [k - C]^+ \frac{e^{-\alpha}\alpha^k}{k!} \leq [\alpha - C]^+ + \sum_{k=C+1}^{\infty} \left[[k - C]^+ - [\alpha - C]^+\right] \frac{e^{-\alpha}\alpha^k}{k!} \\ &\leq [\alpha - C]^+ + \sum_{k=C+1}^{\infty} (k - \alpha)\frac{e^{-\alpha}\alpha^k}{k!} = [\alpha - C]^+ + \frac{e^{-\alpha}\alpha^C}{C!}\,\alpha \leq [\alpha - C]^+ + \frac{e^{-C}C^C}{C!}\,\alpha,\end{aligned}$$

where the first inequality follows by adding and subtracting $[\alpha - C]^+$ to the expression on the left side of this inequality, the second inequality follows by the inequality derived at the beginning of the proof, the last equality is by arranging the terms in the summation on the left side of this inequality and the third inequality is by noting that the function $f(\alpha) = e^{-\alpha}\alpha^C$ attains its maximum at $\alpha = C$. In this case, the result follows by noting that $C! \geq \sqrt{2\,\pi\,C}(C/e)^C$ by Stirling's approximation and using this bound on the right side of the chain of inequalities above. □