POMS PRODUCTION AND OPERATIONS MANAGEMENT

PRODUCTION AND OPERATIONS MANAGEMENT POMS

Vol. 0, No. 0, xxxx–xxxx 2016, pp. 1–22
ISSN 1059-1478 | EISSN 1937-5956 | 16 | 00 | 0001
DOI 10.1111/poms.12659

Are Patients Patient? The Role of Time to Appointment in Patient Flow

Nikolay Osadchiy, Diwas KC
Goizueta Business School, Emory University, Atlanta, Georgia 30322, USA, nikolay.osadchiy@emory.edu, diwas.kc@emory.edu

The current state of outpatient healthcare delivery is characterized by capacity shortages and long waits for appointments, yet a substantial fraction of valuable doctors' capacity is wasted due to no-shows. In this study, we examine the effect of wait to appointment on patient flow, specifically on a patient's decision to schedule an appointment and to subsequently arrive to it. These two decisions may be dependent, as appointments are more likely to be scheduled by patients who are more patient and are thereby more likely to show up. To estimate the effect of wait on these two decisions, we introduce the willingness to wait (WTW), an unobservable variable that affects both bookings and arrivals for appointments. Using data from a large healthcare system, we estimate WTW with a state-of-the-art non-parametric method. The WTW, in turn, allows us to estimate the effect of wait on no-shows. We observe that the effect of increased wait on the likelihood of no-shows is disproportionately greater among patients with low WTW. Thus, although reducing the wait to an appointment will enable a provider to capture more patient bookings, the effects of wait time on capacity utilization can be non-monotone. Our counterfactual analysis suggests that increasing wait times can sometimes be beneficial for reducing no-shows.

Key words: queuing behavior; empirical study; healthcare operations
History: Received: June 2015; Accepted: September 2016 by Sergei Savin, after 3 revisions.

1. Introduction

Timely access to healthcare has been a topic of significant and growing concern for patients, healthcare providers, and policymakers in the United States (Millman 1993). According to several estimates (e.g., Bodenheimer and Pham 2010), as many as 65 million individuals lacked access to primary care in 2010. In a 2011 national survey, 57% of patients who were sick and needed medical attention could not obtain timely access to care (Commonwealth 2011). Compounding this problem, a substantial fraction of valuable doctors' capacity is wasted due to patients failing to arrive to their scheduled appointments. For example, Moore *et al*. (2001) found that 31% of scheduled patients at a family clinic did not arrive to their appointment.

In this study, we examine patients' no-show behavior. Previous research has identified that the wait time to an appointment is a significant controllable factor affecting no-shows (Cohen et al. 2007, Gallucci et al. 2005, Sherman et al. 2009). Based on observed data for scheduled appointments, these studies find that a longer wait can reduce the odds of a patient arriving for a previously booked appointment; collectively, this literature suggests that patient no-shows can be minimized by offering shorter waits.

In practice, patients undergo a two-stage process prior to being seen by a physician. First, the patient initiates an appointment scheduling process and only makes an appointment if the wait time to see a physician is acceptable. Second, after the waiting period, the patient either arrives to the appointment or not. To examine the no-show behavior in the second stage, it is important to also consider the patient's appointment scheduling decision in the first stage. This is because the first stage effectively filters out impatient patients (those with a low willingness to wait), resulting in a mix of scheduled patients who are generally more patient. This means that a reduction in wait times to appointments not only increases the likelihood of arrival for scheduled patients, but also has the effect of altering the mix of patients who are scheduled to see the physician. In other words, no-show behavior cannot be examined independent of scheduling behavior.

Ours is the first paper to combine both sets of patient decisions in order to estimate the net impact of wait time reduction on patient no-shows. Ignoring the patient selection process in the first stage can lead to biased estimates between wait times and the likelihood of a no-show. To examine these issues in detail, we conduct an extensive empirical study to quantify the effect of wait times on patient scheduling and arrival decisions and, consequently, overall patient flow. In particular, we ask the following questions:

Please Cite this article in press as: Osadchiy, N., D. KC. Are Patients Patient? The Role of Time to Appointment in Patient Flow. *Production and Operations Management* (2016), doi 10.1111/poms.12659

- How long are patients willing to wait to see a doctor? In other words, what is their willingness to wait (WTW, for short)?
- How many patients forego scheduling an appointment (balk) due to a disparity between offered wait times and WTW? In other words, what is the extent of lost sales, or the volume of patients that could be seen, had the waits been shorter?
- What is the effect of wait time on no-shows? Specifically, what are the implications of reduced waits to appointment on the mix of patients who choose to book an appointment, as well as capacity utilization and revenue?

To address these questions, we undertake a large-scale empirical study at a leading US East Coast medical center featuring over 100 clinical specialties. We collect a unique and comprehensive dataset consisting of providers' availability calendar, and patients' bookings and arrivals for appointments with the providers over a 2-month period. We also observe patient-level sources of heterogeneity that could influence the queuing behavior, including patient demographic factors such as age, gender, race, as well as their medical condition and payor information.

Empirically estimating WTW for an appointment is challenging. This is because providers only observe waits for scheduled patients, but not for those who forego an appointment. To overcome this inherent data limitation, we employ a general non-parametric model of patient choice. The model allows us to empirically estimate the distribution of patient WTW in the entire patient population (including those who balk) for any given provider. Our analysis further allows us to quantify the extent to which a given wait time reduction improves patient booking.

Likewise, estimating the impact of wait times on the no-show rate is confounded by the underlying endogenous selection described earlier: patients who choose to schedule an appointment may have an inherently different sensitivity to wait (we define sensitivity to wait as the effect of wait time on the likelihood of a no-show) compared to those who do not. The standard approach for generating consistent estimates in the presence of sample selection is based on Heckman (1979). However, the Heckman approach requires the researcher to observe all of the patients, including those who balk. Since this approach is not feasible in many settings, including our own, we instead develop an alternative method that combines the estimates of WTW and the sensitivity of arrival probability to wait. We test the performance of the method on simulated data and show that it helps to correct for selection bias.

Our analysis uncovers a number of interesting findings. First, we show that a large number of providers in our study experience lost sales of 30% or more due to excessive wait times. In other words, these providers could generate significant improvements in patient bookings by reducing their waits. We further quantify the relationship between wait time for an appointment and lost sales, and find a significant non-linearity in this relationship. In particular, at a higher lost sales rate, a greater reduction in waits time is needed to achieve the same reduction in lost sales.

Second, although the wait time to an appointment can appreciably increase the patient no-show rate, the magnitude of the effect depends on patients' WTW and can vary substantially. Specifically, for those patients with an estimated WTW in the top 10%, increasing wait to appointment by 1 day increases the odds of a no-show by 2%, whereas for patients with an estimated WTW in the bottom 10%, an extra day of wait increases the odds of a no-show by as much as 8%. We run multiple robustness tests under alternative model specifications, for distinct specialties, and over additional data periods, and obtain similar results.

Third, we conduct a counterfactual analysis of the impact of wait to appointment on the patient mix and capacity utilization. Surprisingly, and contrary to prior findings in the literature, we find that the effect can be non-monotone: while for long waits, reducing wait results in reduced no-shows, the effect is reversed for moderate to short waits. The intuition for this result is the following: offering appointments with shorter waits changes the mix of patients. In particular, it attracts more patients who are averse to wait and are more likely to become no-shows. The interaction between WTW and sensitivity to wait suggest an interesting interplay in queuing systems between offline wait, balking, and reneging: waits screen customers (balking) who would have been less likely to arrive for their scheduled services (reneging).

Our estimates allow us to analyze the marginal effect of wait time on patient throughput. Our analysis shows that a 1-day reduction in wait time with respect to the current state would increase throughput by 5.7% on average (assuming ample capacity). In general, a reduction in balking (or lost sales) would account for 79% of the throughput increase, whereas reduction in no-shows would account for the remaining 21%. We also identify those specialties where wait reductions can have the highest impact on revenue.

Finally, although our study is motivated by a healthcare setting, there are numerous other settings that involve a similar two-stage flow process with attrition in the first stage. In particular, the problem of waits and no-shows is especially pressing across a wide range of service contexts that use appointments

Please Cite this article in press as: Osadchiy, N., D. KC. Are Patients Patient? The Role of Time to Appointment in Patient Flow. *Production and Operations Management* (2016), doi 10.1111/poms.12659

including governmental, professional, and counseling services. Our empirical method can be applied to correct for the sample selection bias when attrition loss in the appointment stage is not observed.

The rest of the paper is organized as follows: section 2 provides a literature review; we describe the data and industry context in section 3. Section 4 develops our models and econometric methods. We present our findings and describe the managerial implications of our study in section 5. Section 6 presents the counterfactual analysis of the effect of wait on capacity utilization, and section 7 reports the results of robustness tests. Section 8 concludes with a summary and opportunities for future research.

2. Literature Review

Patient flow management and capacity planning in healthcare have been active and fruitful areas of research in the fields of management science and operations research. Much of the prior and ongoing work addresses the sizing of care capacity (Huang 1995, Kwak and Lee 1997), scheduling of workforce (Gerchak et al. 1996), or the design of service delivery systems (Hall 2013). Methods and tools derived from queuing analysis have been helpful in understanding the performance of healthcare delivery systems, such as wait times, utilization levels, and turn-away probabilities. In particular, the results from queuing analysis have been used in formulating various capacity allocation decisions (Green and Nguyen 2001).

However, how individuals perceive and respond to wait (e.g., through balking and reneging) (e.g., Maister 1985) can have significant implications for system performance, including utilization and throughput. In general, individuals display an aversion to wait. For example, Leclerc et al. (1995) estimate the monetary-equivalent disutility associated with waiting. Notable exceptions to this trend are the recent papers by Buell and Norton (2011) and Debo and Kremer (2014), who show that the wait times can increase the perceived value of the service and demand via signaling quality. Recent work has begun to empirically examine how consumers behave in the presence of queues in various field settings, and how such behavior may impact operational performance. For example, Allon et al. (2011) study a system in which fast-food customers make a trade-off between the expected wait times in the queue vs. the price of food, and estimate the implied cost of wait. Aksin et al. (2013) model customer waiting in a call center as an optimal stopping problem and estimate the cost associated with waiting. In the healthcare setting, Batt and Terwiesch (2015) examine the effect of queue length and wait times on queue abandonment in an emergency department (ED) and explore the impact of available information on queuing behavior.

An important distinction between the present paper and much of the empirical queuing literature is that in prior work, arriving customers are presented with an estimate of the wait time, such as the average wait time, or their position in the queue (e.g., see Brown et al. 2005). In other words, customers anticipate a distribution of potential waits, and then infer the corresponding disutility of waiting. In contrast, patients in the present paper are offered a precise measure of the actual wait that will be incurred. In that sense, patients in our setting can "take it or leave it." This distinction is important if customer behavior is described by a Von Neumann–Morgenstern utility function (e.g., see Von Neumann and Morgenstern 1953). Specifically, an increase in the uncertainty associated with waiting (e.g., a higher standard deviation of the wait time) leads to a reduction in the utility for the customer. Therefore, a precise measure of the offered wait allows us to infer the preference for wait, separate from the individual's own level of risk aversion, and estimate the effect of wait on the decision to book an appointment or balk.

To empirically analyze patient balking, we develop a generalizable, non-parametric model of patient WTW. Our approach draws on recently developed models of customer choice from the revenue management and retail literatures. We employ a modified version of a non-parametric choice model described by Mahajan and van Ryzin (2001) and most recently modified by Farias et al. (2013) and van Ryzin and Vulcano (2014). The model assumes that choice is driven by customer type, which is characterized by a ranked preference list. Given a set of options, a customer chooses the option with the highest preference rank. From the observed customer choice data, which includes no-purchases, the model allows to impute the demand from each customer type. In the context of our study, patients vary in their preferences for wait; the model thus yields estimates of patients' WTW.

The resulting analysis allows us to estimate lost sales, which has been a subject of recent empirical work. For example, using observed sales data from a retailer, Musalem et al. (2010) estimate the effect of product stockouts on lost sales. Similarly, Lu et al. (2013) find that both customer sensitivity to wait and purchase price negatively impact the purchase decision. In the context of our study, we generate the estimates of lost sales as a function of wait time to an appointment.

Crucially, the present study establishes a linkage between balking and reneging behavior and estimates the extent to which these two decisions are inextricably linked. Specifically, a patient's behavior to renege

Please Cite this article in press as: Osadchiy, N., D. KC. Are Patients Patient? The Role of Time to Appointment in Patient Flow. *Production and Operations Management* (2016), doi 10.1111/poms.12659

is contingent on their decision to not balk. This suggests that patients who are confronted with the choice to renege are inherently different from those who chose to balk.

Our work also contributes to a related stream of research that has focused on improving access to healthcare. Reducing no-shows in particular has been identified as an important factor for improving overall throughput. Gallucci et al. (2005), Cohen et al. (2007), and Sherman *et al*. (2009) identify the wait time to appointment as a key driver of no-shows. Several researchers have examined the effect of scheduling policies on reducing wait times and no-shows (Green et al. 2006, Luo et al. 2012, Wang and Gupta 2011). A stream of this literature has developed policies for appointment scheduling, accounting for the possibility that patients may cancel or not show (Hassin and Mendel 2008, Liu 2016, Liu et al. 2010). Green and Savin (2008) propose a queuing model to quantify the effect of wait time on no-shows and examine the implications of patient patience on panel size. LaGanga and Lawrence (2012) develop an overbooking strategy that maximizes capacity utilization and patient service. Collectively, this work has shown that scheduling policies can significantly improve access (and reduce no-shows) if patients' wait times are taken into account.

We extend the literature on no-shows by incorporating a linkage between the no-show probability and the first-stage balking decision. Typically, no-shows are analyzed using a sample of booked appointments, which may be subject to the sample selection problem, leading to biased estimates. However, despite its potentially serious consequences, sample selection is often overlooked because the individuals who self-select out of the sample are not always observable, rendering the use of classical sample selection methods problematic. Furthermore, a straightforward application of the Heckman (1979) correction is often not possible when selection is based on an unobservable variable, such as WTW. In our paper, we correct for this type of *latent* selection process by explicitly modeling the sensitivity to wait as a function of WTW. We also control for the sample attrition associated with the selection of low-WTW individuals out of the sample using a method recently developed by Wooldridge (2007). The method is based on weighting observations inversely to the probability of being selected into the sample and has been proven to be consistent for a wide class of estimators, including the non-linear models.

Finally, our work contributes to the ongoing public health debate concerning the increased demand for limited healthcare resources, a phenomenon of significant relevance in the current healthcare landscape (Ghorob and Bodenheimer 2012, Sack 2008). The shortage of primary care has already been shown to increase ED visits, an indicator of a systemic failure in the healthcare delivery system (e.g., see Werner et al. 2012). By quantifying the impact of wait on patient access and overall throughput, we hope to provide accurate operational performance estimates that can be used in policy formulation and managerial decision making.

3. Data and Clinical Context

We obtained our data from a large US East Coast multi-specialty healthcare system, which serves an estimated catchment area of upto 20 million people. The dataset provided to us, which was abstracted from their patient scheduling IT system, includes more than 237,000 records for distinct patient appointments scheduled for a 2-month period beginning in January 2012. Among them, 37,688 appointments were flagged as "new patient visits" (NPV) with the healthcare system—the focus of our analysis. The wait to appointment is particularly important for new patients' timely access to care; in contrast, waits for repeat visits are often driven by treatment plan or continuity of care needs. In addition, with the ongoing adoption of the bundled payments reimbursement model, NPV directly translate into higher revenue for providers.

Each of the providers in our study is a healthcare professional, either a physician or a high-level nurse practitioner. Our study site is a research facility; therefore, providers tend to be highly specialized, offering a narrower range of services compared to a more general community hospital. After further clean up (see Table A1 and notes), there are 29,089 appointments for 587 distinct providers in 107 specialties in our final dataset. The specialties with the largest number of NPV appointments are orthopedics, spine care, neurology, and otolaryngology. The specialties with the largest number of providers are cardiology, neurology, psychiatry, hematology, and orthopedics.

A new patient looking to schedule a visit with a given provider begins the booking process by placing a telephone call to a centralized call center. For the majority of providers (about 70%), appointments can be booked as far as one year in advance. The call center handles only non-emergency outpatient visits; call center operators are advised to direct any emergency patients to the local ED. As per our interview with the patient access management team, the call center receives approximately 12,000 calls per day, with day-to-day variations of under 10%. For each call request, a decision support tool implemented in the call center's IT system automatically provides a list of the three next available appointments. If the patient is not

Please Cite this article in press as: Osadchiy, N., D. KC. Are Patients Patient? The Role of Time to Appointment in Patient Flow. *Production and Operations Management* (2016), doi 10.1111/poms.12659

satisfied with any of these, the call center operator looks to find additional dates further out in the future. If none of the provided dates is acceptable, the caller completes the call without scheduling an appointment. If, however, one of the dates provided by the call center operator is acceptable, the patient is scheduled for an appointment. We only observe scheduled appointments, as incidences of caller balking (or lost sales) are never recorded. For scheduled appointments, the wait time is measured in full days.

A scheduled patient may arrive, cancel, or simply not show up for an appointment. The hospital classifies all cancellations made within 24 hours from the scheduled appointment time as no-shows. The hospital representatives rationalized this classification by indicating to us that 24 hours is generally not adequate to rebook the newly available slot. Therefore, from the perspective of throughput, a late cancelation is equivalent to a no-show because both result in unutilized capacity.

However, cancellations made more than 24 hours in advance of the appointment time can also lead to unutilized capacity. Using information on capacity utilization and patient arrivals, we were able to identify cancellations that resulted in unutilized capacity. We find that, on average, clinics are able to fully rebook about 72% of their cancellations. These rebookings can either be due to new or repeat patient visits. Rebooked cancellations do not hurt capacity utilization or patient throughput and are thus excluded from our analysis of no-show behavior (we find that our results are robust to this exclusion). On the other hand, the *un*rebooked cancellations are treated as no-shows because they result in unutilized capacity. The records for rescheduled and pending appointments are excluded from the no-show analysis.

A unique feature of our dataset is the allotted capacity for each provider during the observation period. Specifically, we observe the maximum number of patients that a given physician can see on any given date. Combined with the ordered sequence of appointment bookings, this allows us to fully reconstruct the process of capacity usage for each provider at any given point in time. Since there can be multiple callers who book appointments in one calendar day, the time is discretized into periods of equal length, so that there is at most one booked appointment per period. Thus, one calendar day may correspond to multiple consecutive arrival periods.

For any patient who calls at period i to schedule an appointment with a provider, we recreate the calendar view of both full and available appointments that would have been visible to the call center operator at period i. All patients who called prior to i are assigned to the slots for which they were scheduled. This means that all calls before period i that resulted in a booked appointment were accounted for in their respective scheduled slot. We note here that both first-time and revisit patient arrivals were used to recreate the schedule, since the available capacity can be used by both types of patients. Since we also know the maximum number of patients who can be scheduled on a given day, we can take the difference between the maximum capacity and the number of patients already scheduled, to impute the number of slots that are still vacant as of period i. For the caller at period i, the first date with one or more free slots starting from i is the earliest date that can be provided. We thus generate an entire menu of possible slots that can be offered to the caller at period i, ordered by the date of availability. These choice sets, in particular, the shortest available wait, together with the actual booking decisions made by calling patients allow us to generate a WTW distribution in the patient population.

Wait times for appointments vary across clinical specialties (Table A1). The average wait in the sample is 20.8 days. Among those patients who arrived to their scheduled appointments, the average wait is 18.1 days. In contrast, the average wait is 32.4 days among no-show patients. Despite the prevalent aversion to wait, a noticeable number of patients are willing to wait upto 1 year, even though earlier appointments are available. We run our analysis of arrivals and no-shows in aggregate, as well as by specialty. For that, we choose the four largest clinical specialties: orthopedics, spine center, neurology, and otolaryngology. These specialties represent a diverse set of medical conditions and offer average waits ranging from approximately 12 to 68 days (Table A1).

Over the 2-month observation period, providers on average scheduled 49.6 first-time patients, of which 40.1 patients arrived. The average no-show rate is 20.2%, with unrebooked cancellations accounting for 10.6% of no-shows. Table A2 provides information on the capacity availability for the set of providers in our sample. We find that, on average, providers have allotted capacity that can serve around 179 patients during this 2-month period, including first-time and follow-up visits. We also find significant variation across the providers in terms of their allotted capacity for the total number of patients served during this period.

To account for various sources of patient-level heterogeneity that could influence patient behavior, we obtained a set of patient-level controls abstracted from a clinical database, which we merged with the scheduling data using a unique patient identifier. The clinical data provide us with a range of patient level controls, including demographic factors such as age,

Please Cite this article in press as: Osadchiy, N., D. KC. Are Patients Patient? The Role of Time to Appointment in Patient Flow. *Production and Operations Management* (2016), doi 10.1111/poms.12659

gender, and ethnicity, as well as the provider's specialty and the general area of medical concern for the patient. We also observe payor information including the type of payor (e.g., Medicare, Medicaid, private insurance, self-pay, etc.). The dataset also includes the stated reason for the patient's visit. This information is entered at the time the appointment is scheduled. This data field allows us to infer if the scheduled visit is routine (e.g., annual physical, consultation, etc.) or urgent (e.g., fever, bleeding, severe, or urgent condition, etc.). The keywords identifying urgency level are listed in Table A3. Overall, we identify 2601 urgent and 9784 routine appointments. The median wait is 7 days for urgent appointments and 13 days for routine ones.

In addition to these patient-level factors, exogenous considerations such as travel distance, or weather on the day of the visit could influence the patient's decision to arrive for a scheduled appointment. Our dataset includes the zip code of the patient's residence, which we use to impute the travel distance to the care facility. To examine the role of weather on the day of the visit, we obtained the historical weather data for the zip codes where our study facilities are located. The historical weather data, obtained from the Applied Climate Information System, includes the amount of rainfall, as well as the temperature on the date of the patient's scheduled visit. The data also contain deviations from historic averages which allow us to examine whether abnormal deviations from historic weather patterns (e.g., unusually cold weather on January 30) influenced a patient's decision to keep the appointment.

To examine the effect of waits on provider revenues, we obtained the average per-patient revenue charged by a given provider from the Centers for Medicare and Medicaid Services (CMS). We were able to merge the CMS per-patient revenue data with the appointment data for those providers. Given the prospective payment system in US hospitals, where reimbursement is often fixed depending on the patients diagnosis, revenue is therefore directly proportional to the overall patient throughput. This allows us to examine how the queuing behavior of patients (including scheduling an appointment and arriving for that appointment) impacts provider revenues. The summary statistics on the patient-, appointment-, and provider-level heterogeneity controls is presented in Table A4.

In sum, we assemble a novel and rich dataset consisting of detailed provider-level capacity availability, queuing decisions by individual patients, as well as a comprehensive list of patient-, provider-, and appointment-level sources of heterogeneity. In the following sections, we explore this dataset to analyze the drivers of patient queuing behavior.

4. Model Formulation

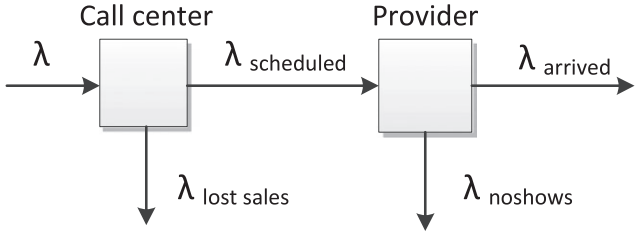
Patient flow is modeled as a two-stage process, depicted in Figure 1. There are two sources of attrition that affect throughput. First, patients who call to schedule an appointment may balk if the time to appointment is too long. Second, increased waits may deter patients from showing up once they have scheduled an appointment. Section 4.1 presents a model of patients' choice to schedule an appointment. It allows us to estimate patients' WTW and providers' lost sales. Section 4.2 presents the model of no-shows and the link between the decision to book an appointment and the decision to arrive for it. Section 4.3 tests the proposed methods on simulated data.

4.1. Patient Choice Model

We assume that patients requesting appointments for a specific provider arrive according to a time homogeneous, discrete time Bernoulli process. Subscript p denotes the provider, $p = 1, \ldots, P$. Subscript i denotes the demand period, which can have at most one arrival, $i = 1, \ldots, I_p$. In each demand period, an arrival occurs with probability λ_p, independent of other periods. The clinic is part of a large medical center that serves a calling population of approximately 20 million individuals. The clinic call center receives approximately 12,000 calls every day. Given the large patient population, and the steady overall call volume, the assumption of the time-invariant arrival probability is justified. The Bernoulli process is analytically convenient and serves as a good approximation of the Poisson process if $\lambda \ll 1$. A patient arriving in period i can make at most one appointment. We index patients by their arrival period.

Patients have a preference for shorter waits, and are characterized by WTW τ. Patient i is said to be of type k, if $\tau_i = k$. In other words, the patient arriving in period i is prepared to wait at most k days for an appointment. Upon calling, the patient is notified by the call center operator of the wait until the next available appointment, $\tilde{w}_i$. If $\tilde{w}_i > \tau_i$, the patient balks and leaves the system without scheduling an appointment. If $\tilde{w}_i \leq \tau_i$, the patient books the appointment and the appointment is recorded with the wait

Figure 1 Patients Flow: Demand, Scheduled Appointments, Lost Sales, Arrivals, and No-Shows



Please Cite this article in press as: Osadchiy, N., D. KC. Are Patients Patient? The Role of Time to Appointment in Patient Flow. *Production and Operations Management* (2016), doi 10.1111/poms.12659

$w_i = \tilde{w}_i$. We assume that τ_i is a discrete random variable with a probability mass function (p.m.f.) f, i.e., $f(k) = Pr(\tau_i = k)$. Given the heterogeneity across clinical specialties, f is assumed to be specialty-specific, and the variation within f is attributed to individual patient heterogeneity. An advantage of defining f as specialty-specific is that it can be estimated more accurately using data for multiple providers. For each provider in a specialty, λ_p, $p = 1, \ldots, P$ is a provider-specific arrival probability, which partially captures heterogeneity across providers in a specialty, such as differences in quality or experience.

Suppose that the maximum booking horizon for all providers in the specialty is S. For a patient arriving in period i, the set of available appointments may contain appointments with wait times ranging from 0 to S days, where a wait time of 0 corresponds to a same-day appointment. We define the support of f as the subset of integers $\{0, 1, \ldots, S\}$ that includes all distinct booked waits w_i, $i = 1, \ldots, I_p$, and all distinct best available waits $\tilde{w}_i$, $i = 1, \ldots, I_p$, i.e., $\mathcal{S} \triangleq supp(f) = \{k : k \in \{0\} \bigcup \{w_i, i = 1, \ldots, I_p\} \bigcup \{\tilde{w}_i, i = 1, \ldots, I_p\}\}$. A booked appointment with wait time w_i may correspond to an arrival of type $\tau_i \in \{\mathcal{S} : \tau_i \geq w_i\}$. If no appointment is booked in period i, this corresponds to either an arrival of type $\tau_i \in \{\mathcal{S} : \tau_i < \tilde{w}_i\}$ or no arrival at all. The following example illustrates how patient bookings are generated.

EXAMPLE 1. *Suppose the maximum booking horizon $S = 5$, and that there are two patient types $\tau_i \in \{2, 5\}$ that arrive over 6 time periods (I). A patient with $\tau_i = 2$ will prefer an appointment with $w_i = 0$ over $w_i = 1$ which, in turn, will be preferred over $w_i = 2$; the patient will not book an appointment if $w_i > 2$. A patient with $\tau_i = 5$ will also prefer appointments with shorter wait times, but will book an appointment if $w_i \leq 5$. For this illustrative example, Table 1 provides a set of available waits (from 0 to 5 days) for a set of 6 distinct periods. The row corresponding to "Wait" indicates the chosen wait if a patient booked an appointment.*

Table 1 Illustrative Example: Offered Waits and Booked Appointments

Observable data: Availability and appointment bookings						
Wait times	Period					
	1	2	3	4	5	6
0	Yes	Yes	No	Yes	No	No
1	Yes	Yes	No	Yes	No	No
2	Yes	Yes	No	No	No	No
3	Yes	Yes	No	No	Yes	No
4	Yes	Yes	Yes	No	No	Yes
5	Yes	No	Yes	Yes	No	Yes
Wait (days)	0	–	4	0	–	–
Customer types that match the observed data						
$\{k\}$	{2, 5}	{ }	{5}	{2, 5}	{2}{ }	{2}{ }
Unobservable data						
τ_i	2	No arrival	5	2	No arrival	2

In period $i = 1$, all appointments are open and we see that a same day appointment (wait = 0) is booked. A wait time of 0 means that the appointment could have been booked by either of the two patient types, that is, the set of customer types that match the observed data $\{\tau\} = \{2, 5\}$. In reality, the patient who actually arrived may have been type $\tau_1 = 2$, however, this information is not observable to the econometrician. In period $i = 2$, appointments with wait times of 0 through 4 days are available. Again, this set of wait times is acceptable to either patient type. However, since no appointments were booked, we conclude that there were no arrivals. In period $i = 5$, the only available appointment has a wait time of 3 days. However, no appointments were booked. This observation is consistent with either "no arrival" or an arrival of a patient of type $\tau_5 = 2$.

Let $\mathcal{B}_p \subseteq \mathcal{I}_p = \{1, \ldots, I_p\}$ denote the set of periods with booked appointments for provider p. Then, $\mathcal{I}_p \backslash \mathcal{B}_p$ is the set of periods with no booked appointments. If $i \in \mathcal{B}_p$, w_i is the wait time for a booked appointment. If $i \in \mathcal{I}_p \backslash \mathcal{B}_p$, then $\tilde{w}_i$ is the best wait time available for an arrival in period i that did not result in a booking.

Recall that λ_p is the probability of any patient's arrival to provider p. Let $x_k = f(k)$, $k \in \mathcal{S}$ denote the probability of arrival of a patient of type $\tau_i = k$. Our objective is to estimate λ_p, $p = 1, \ldots, P$ and x_k, $k \in \mathcal{S}$, based on the observed bookings and appointment availabilities. We do not make any assumptions on the parametric form of the distribution of patient types f. Thus, our maximum likelihood method results in a non-parametric estimate of f. The log likelihood function can be expressed as:

$$\mathcal{L}(w|x, \lambda) = \sum_p \left\{ \sum_{i \in \mathcal{B}_p} \left(\log \lambda_p + \log \sum_{k \in \{\mathcal{S}: k \geq w_i\}} x_k \right) + \sum_{i \in \mathcal{I}_p \backslash \mathcal{B}_p, \tilde{w}_i \geq 1} \log \left(\lambda_p \sum_{k \in \{\mathcal{S}: k < \tilde{w}_i\}} x_k + (1 - \lambda) \right) + \sum_{i \in \mathcal{I}_p \backslash \mathcal{B}_p, \tilde{w}_i = 0} \log(1 - \lambda_p) \right\}. \quad (1)$$

The first term under the summation over p corresponds to observed booked appointments with a provider, such as in periods 1, 3, and 4 of Table 1. The second term accounts for periods with offered wait $\tilde{w}_i \geq 1$, in which there were no bookings due to either balking or non-arrival. The example of balking is represented in period 6, and the example of non-arrival in period 5. The third term accounts for periods with

Please Cite this article in press as: Osadchiy, N., D. KC. Are Patients Patient? The Role of Time to Appointment in Patient Flow. *Production and Operations Management* (2016), doi 10.1111/poms.12659

no bookings and zero offered wait, i.e., the cases where no patient would prefer to balk. This specifically corresponds to non-arrivals, such as in period 2 of Table 1.

The constrained optimization problem for the maximum likelihood estimation of x and λ is

$$\max_{x,\lambda} \mathcal{L}(w|x,\lambda) \tag{2}$$

$$\text{s.t.} \sum_{k\in\mathcal{S}} x_k = 1,$$

$$x_k \geq 0, \text{for all } k \in \mathcal{S},$$

$$0 \leq \lambda_p \leq 1, \text{for all } p.$$

The identification of Equation (2) comes from the time invariance of λ_p, the booked waits for the periods with scheduled appointments, the available waits for the periods with no scheduled appointments, and the periods where same-day appointments are available. The model is identified (in the sense that there is a unique set of parameters corresponding to the global extremum of the likelihood function) if the number of periods is greater than the number of types (for a discussion, see van Ryzin and Vulcano 2014). This condition is satisfied in our setup. In that case, estimator (2) possesses all the desired properties of the MLE estimator, i.e., it is consistent, asymptotically unbiased, and asymptotically efficient.

4.1.1. Effect of Time to Appointment on Lost Sales. The solution to Equation (2) provides an empirical distribution of WTW in the entire patient population (including those who choose not to book an appointment) for a given provider. This empirical distribution allows us to estimate the fraction of patients who could be satisfied with a given level of wait. Specifically, for a given wait time of w, the fraction of patients who would schedule an appointment is simply equal to $\Sigma_{k\in\mathcal{S},k\geq w} x_k$. The lost sales corresponding to a wait time of w is $LostSales_p(w) = \Sigma_{k\in\mathcal{S},k<w} x_k$.

$LostSales_p(w)$ allows each provider to explicitly evaluate the trade-off between wait times and patient volume. In addition, the estimate of arrival probability $\lambda_p \times$ (No. of demand periods per calendar day) represents the maximum potential size of their market, i.e., the patient population for provider p.

4.2. Appointment Bookings and Patient No-Shows

Our next goal is to estimate the effect of the wait time to an appointment, on the likelihood that a patient shows up for an appointment. Several papers have documented a negative association between the time to appointment and the probability that a patient arrives for an appointment. For example, an increased time to appointment could increase the likelihood that the patient's medical problem has disappeared, that the patient was able to find another provider, or that the patient quite simply forgot about the appointment (Bodenheimer and Pham 2010). An important characteristic of the above findings is that they were obtained from booked appointment data, and may not be representative of the entire patient population.

There are two potential issues caused by patients self-selecting to make appointments. First, patients who choose to book appointments may be less averse to wait and therefore more likely to show up for an appointment. This could bias the estimates from the selective sample. Second, ignoring the appointment booking stage in the counterfactual analysis of the effect of wait time on the arrival probability would imply that the mix of patients who book appointments does not change with wait time. In practice, however, offering a short wait for an appointment would open the slot to less patient patients, who may have a higher probability of no-shows. Thus, estimating the interaction between patients' WTW and their arrival probability is important for the counterfactual analysis of the effect of wait on no-shows and, consequently, capacity utilization, as illustrated below.

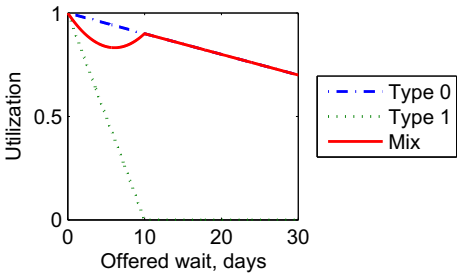
EXAMPLE 2. *Let Type* 0 *represent wait-insensitive patients who will book an appointment with any wait* w*, and will subsequently arrive for it with probability* $1 - c_0 w$ *for* $w \leq 1/c_0$ *and zero otherwise. Let Type* 1 *represent wait-sensitive patients. If they are offered an appointment with wait between* 0 *and* 10 *days, they will book it with probability* $1 - 0.1w$ *(thus, no Type* 1 *patient will book an appointment with a wait time longer than* 10 *days), and will arrive for it with probability* $1 - c_1 w$*, for* $w \leq 1/c_1$*. Since Type* 0 *patients are less sensitive to waits when compared to Type* 1*,* $c_1 \geq c_0$*. At wait* $w = 0$*, both Type* 0 *and Type* 1 *patients are equally likely to book. Assume that the total demand always exceeds capacity.*

As the wait times increase, the fractions of patients of Type 0 and Type 1 in the mix of booked appointments become $m_0 = \min(\frac{1}{2}(1 + 0.1w), 1)$ and $m_1 = \max(\frac{1}{2}(1 - 0.1w), 0)$, $w \leq 10$, respectively. Thus, the probability of a patient's arrival for a booked appointment (capacity utilization) is $m_0(1 - c_0 w) + m_1(1 - c_1 w)$. If $c_1 > c_0$, capacity utilization can be a non-monotone function of w.

Figure 2 illustrates the arrival probabilities of Type 0 and Type 1 patients ($c_0 = 0.01$, $c_1 = 0.1$), as well as the overall capacity utilization. As the result of the changing mix of patients who book appointments, the overall capacity utilization is a non-monotone function of wait. In particular, for moderate waits, increasing wait for an appointment can increase capacity

Please Cite this article in press as: Osadchiy, N., D. KC. Are Patients Patient? The Role of Time to Appointment in Patient Flow. *Production and Operations Management* (2016), doi 10.1111/poms.12659

Figure 2 Capacity Utilization as a Function of Offered Wait Time by Customer Type



utilization. The heterogeneity in sensitivity to wait (i.e., the difference between c_0 and c_1) is critical for the existence of the local minimum of the utilization function. For a given c_0, a larger c_1 corresponds to a larger dip in utilization. In other words, the non-monotone relationship is more pronounced when there is greater heterogeneity in patients' sensitivity to wait.

The key parameter of interest to be estimated is the sensitivity to wait. As in Example 2, the sensitivity to wait is the derivative of the arrival probability function with respect to wait time (in the case of a non-linear probability model, such as logit, by the arrival probability function, we mean the log odds of that function). Higher, in absolute value, sensitivity to wait implies a greater impact of an additional day of wait on arrival probability. The sensitivity to wait may depend on WTW. For example, patients with a low WTW may have a higher likelihood of no-show due to an extra day of wait when compared to their high-WTW counterparts. Individual WTW is unobservable; however, it is possible to compute its expected value, EWTW, given the distribution of WTW and the booked wait. Each EWTW is then assigned to a decile: the bottom 10% of the EWTW are assigned to the first decile, the next 10% to the second, and so on.

The dependent variable, the binary indicator of patients' arrivals, is linked to wait time and other explanatory variables using the logistic function. To capture the dependency between WTW and sensitivity to wait, we include the interaction between wait time and the deciles of the expected WTW. That is, we estimate the sensitivity to wait for each decile of the expected WTW. To control for the specialty effect on WTW, the deciles are computed within each specialty. The model is estimated by the maximum likelihood estimator:

$$\max_{\theta\in\Theta} \sum_i \frac{1}{p_i}\{y_i \log l_i + (1-y_i)\log(1-l_i)\}, \text{ where}$$
$$l_i = \frac{e^{\boldsymbol{\theta}\mathbf{c}_i}}{1+e^{\boldsymbol{\theta}\mathbf{c}_i}}, \tag{3}$$

$$\boldsymbol{\theta}\mathbf{c}_i = \theta_0 + \theta_1 w_i + \theta_2 w_i^2 + \sum_{j=1}^{9} \theta_D(j) w \mathbb{1}\{Decile(E(\tau_i|\tau_i \geq w_i), Specialty_i) \leq j\} + \theta_c controls_i.$$

The summation is over all observations in the selected sample; y_i is the arrival indicator variable for patient i, and l_i is the logistic probability of arriving for a scheduled appointment. We estimate two versions of the model: one with $p_i = 1$ for all i, i.e., when all observations are given equal weight. We call that model SS. The other model uses $p_i = \Pr(\tilde{w} \leq \tau|\tau \geq w_i) = \Sigma_{t=w_i}^{\infty} \Pr(\tilde{w} \leq t)\Pr(\tau = t|\tau \geq w_i)$, which is the estimated probability of inclusion into the sample (recall, τ and $\tilde{w}$ are WTW and the offered wait, respectively). We call that model SS IPW. Weighting observations inversely to the inclusion probability controls for potential attrition among the low-WTW patients in the sample, and the resulting underestimation of sensitivity to wait. We note that as long as the sensitivity to wait is driven by WTW alone, and the interaction between the two is controlled for, the selection into the sample is random and the estimates are not biased, even when the unweighted estimator is used (Wooldridge 2007).

The covariate vector $\mathbf{c}_i$ includes the piecewise linear specification for wait time based on the decile of the EWTW corresponding to each observation. Self-selection at the booking stage can result in a varying mix of patients and their sensitivity to wait (see Example 2). Including the interaction between wait and deciles of EWTW captures that effect. Note that EWTW affects arrival probability only through wait time. If wait time is zero, the sample selection due to wait and hence EWTW is irrelevant: all patients will book appointments and arrive with probability one.

We also include the $wait^2$ term to control for a potential intrinsic non-linearity in the log odds of arrival probability to wait times among patients. Additional controls at the patient level are gender, ethnicity, marital status, age, payor; and at the appointment level—scheduled time, whether the appointment had the shortest wait time among appointments available to a the patient, distance from the patient's home zip code to the provider's location, indicators of the urgent or routine nature of the appointment, and the weather on the day of the appointment. To account for unobserved heterogeneity across providers and time, we use provider fixed effects and day of week fixed effects for the booking and scheduled dates of appointments. The estimation results are reported in section 5.2, followed by numerous robustness tests, including alternative model

Please Cite this article in press as: Osadchiy, N., D. KC. Are Patients Patient? The Role of Time to Appointment in Patient Flow. *Production and Operations Management* (2016), doi 10.1111/poms.12659

specifications and estimates by individual specialties in section 7.

4.3. Performance on Simulated Data

Before proceeding with the empirical analysis on actual patient data, we first evaluate performance of the methods described in Sections 4.1–4.2 on simulated data. Our objective is to verify that: (i) on a dataset representative of the actual data, the WTW estimation method is able to recover the true parameters of the arrival process, including the arrival rate and the WTW distribution, and (ii) recover the dependence between WTW and sensitivity to wait.

We simulate the patient arrival as a Bernoulli process. Each arrival represents a certain patient type, characterized by a WTW. Upon arrival, the patient is presented with the earliest available appointment. If the wait until appointment is less than or equal to the patient's WTW, the appointment is scheduled and the information is recorded. Otherwise, no booking is recorded in that period. For booked appointments, the probability of arrival for an appointment is modeled as a logistic function of the wait.

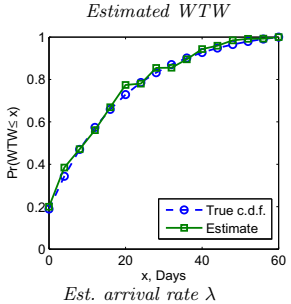
In our dataset, after discretizing the arrival periods, the average number of observations per clinical specialty is approximately 7700. In the majority of these periods, there were no bookings from new patients. However, the choice sets can change due to the passage of calendar time and bookings from repeat patients. Thus, for the simulation, we generate 100 sets of 7700 periods each. We set the arrival rate for new patients' calls to $\lambda = 0.25$. To capture the decreasing fraction of patients with higher WTW, we assume that patient WTW takes one of the following values: 0, 4, 8, …, upto 60 days, with a discrete exponential probability. Offered waits follow the discrete uniform distribution with the same support. WTW and offered waits are independent across time and from each other. The probability of arrival is affected by WTW as follows: patients with WTW below median (upto 12) have a greater sensitivity to wait (−0.05), whereas patients with WTW above median (15 and above) have a lower sensitivity to wait (−0.03). At the booking stage, we observe a significant attrition of the low WTW patients: among arrivals, approximately 57% of patients have low WTW, whereas among scheduled appointments, only 26% have low WTW.

Our first task is to recover the WTW distribution and the arrival rates from simulated data using the methods described above. Figure 3 presents an example of the estimates and the true WTW distribution. We note that the recovered distribution parameters are very close to the true values. For all 100 samples, the estimate is indistinguishable from the true distribution according to the Kolmogorov–Smirnov test at $p > 0.05$. In addition, we find that we were able to recover the simulated arrival rate parameter with a high degree of precision (true value of 0.25 vs. an average estimate 0.2517, SD 0.0195).

We next demonstrate that ignoring the sample selection process and resultant interaction between WTW and sensitivity to wait can bias estimates of the sensitivity to wait. Recall that the true sensitivity to wait is −0.05 for the first five deciles of WTW, and −0.03 for deciles 6–10. Fitting a naive logistic model with wait time as the only predictor variable and a constant significantly underestimates the sensitivity to wait: the estimate is −0.0295 (SD 0.0009). The average McFadden pseudo R^2 is 0.021.

The results for model (2) that accounts for the changing mix of patients due to selection at the booking stage are presented in the right panel of Figure 3. We report the results for the unweighted estimator SS and the weighted estimator SS IPW (see section 4.2). Since true WTW is unobservable, we compute deciles based on the expected WTW, which is computed from the estimated distribution of WTW and the booked wait. The average McFadden pseudo R^2 is

Figure 3 **Left: Estimate of WTW Distribution and the Arrival Rate Using Equation (2). Right: Estimation results for the sensitivity to wait using Equation (2) on 100 simulated datasets, each containing approximately 550 booked appointments. Means and standard deviations over the 100 simulation runs are reported**



Est. arrival rate λ	
True	0.25
Est.	0.2517 (0.0195)

	SS	SS IPW
Decile(E(WTW))	Sensitivity to wait	
1	-0.0490 (0.0118)	-0.0474 (0.0118)
2	-0.0490 (0.0118)	-0.0474 (0.0118)
3	-0.0501 (0.0109)	-0.0489 (0.0109)
4	-0.0478 (0.0071)	-0.0474 (0.0071)
5	-0.0426 (0.0055)	-0.0426 (0.0055)
6	-0.0385 (0.0049)	-0.0385 (0.0049)
7	-0.0387 (0.0037)	-0.0389 (0.0037)
8	-0.0350 (0.0026)	-0.0353 (0.0026)
9	-0.0319 (0.0018)	-0.0321 (0.0018)
10	-0.0300 (0.0011)	-0.0303 (0.0011)
Constant	1.9747 (0.0259)	1.9805 (0.0259)

Please Cite this article in press as: Osadchiy, N., D. KC. Are Patients Patient? The Role of Time to Appointment in Patient Flow. *Production and Operations Management* (2016), doi 10.1111/poms.12659

approximately 0.033 for both modes. We observe that the estimates are close to the true values, with a notable increase (in absolute value) from the sixth to fifth deciles, corresponding to the true parameters. Note that models SS and SS IPW produce very similar estimates. In the simulation, the selection and sensitivity to wait are driven by WTW, and the interaction between WTW and sensitivity to wait is explicitly controlled for. Therefore, attrition does not cause bias as the model remains correctly specified. For each percentile of expected WTW, the sample selection is random.

5. Results

5.1. Willingness to Wait

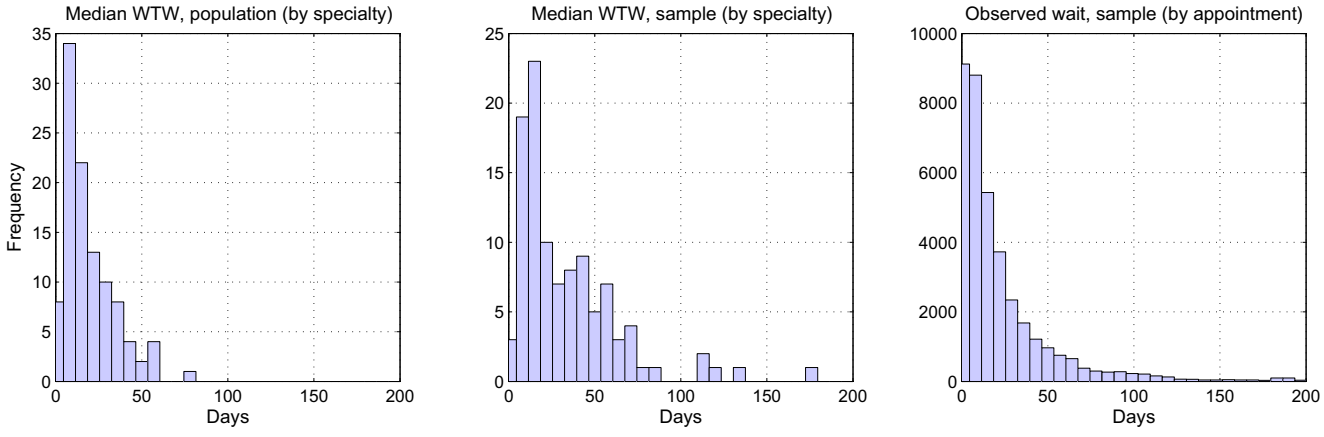
We first examine the results for the maximum likelihood estimation of the distribution of patients' WTW. We estimate model (2) to recover the empirical distribution of the patient WTW for each specialty, as well as the arrival rates for each provider in that specialty. This information can be used by providers to better understand their patient population and, in particular, to examine the relationship between lost sales and wait times. Given the large number of specialties, it is infeasible to provide empirical distributions for all of them. Instead, we compute the mean of the empirical distribution for each specialty and report the summary statistics for this measure in Table A5. We find that for patients who choose to schedule an appointment, the average of the mean WTW across providers is 39.7 days. In contrast, the same measure in the calling patient population at large (including those who balk) is significantly smaller, at 27.0 days. This reflects the fact that low-WTW patients choose not to schedule appointments, and therefore are excluded from the sample of patients with appointments.

We observe a significantly large standard deviation for the mean WTW in both the patient population (18.6), as well as in the selective sample (35.1), which is indicative of heterogeneity across patient populations. The three specialties with the highest WTW are pediatric ophthalmology, neurology, and glaucoma ophthalmology. The specialties with the lowest WTW are general pathology, general internal medicine (GIM) (a corporate location), and diabetes education.

To further elaborate on the WTW distribution, Figure 4 shows the distribution of the median WTW across all specialties in the calling population, and in the selective sample of scheduled appointments. We observe that the latter stochastically dominates the former (empirically). The difference in the WTW in the calling population and in the scheduled appointments suggests the presence of endogenous selection.

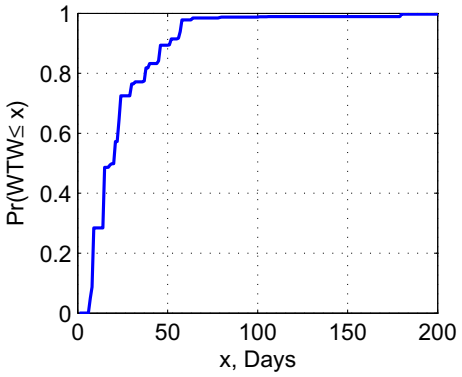
We next examine the effect of wait times on lost sales. Figure 5 shows a fraction of lost sales as a function of the appointment wait time for a single specialty (orthopedics). The pattern of lost sales is similar across specialties (Table A6 reports the summary statistics of lost sales associated with wait across specialties). Specifically, increased wait times to an appointment are associated with increased lost sales (or patient balking). We also uncover a significant non-linearity in the relationship between wait times and lost sales. For example, on average, at a booking rate of 50%, a provider can reduce lost sales by 10% by reducing average wait times by approximately 1.5 days. In contrast, at a booking rate of 30% (i.e., when 70% of patients balk), an approximate wait time reduction of 5.4 days is required to generate a corresponding 10% reduction in lost sales. In other words, at low levels of wait, one more day of wait has a bigger impact on lost sales; at longer waits, the marginal impact of 1 day of wait on lost sales is lower. At the current average wait of ≈21 days, the estimated lost

Figure 4 Histograms of Median WTW across Specialties in the Population (left), in the Sample (center), and the Histogram of Actual Observed Wait Times across All Specialties (right)



Please Cite this article in press as: Osadchiy, N., D. KC. Are Patients Patient? The Role of Time to Appointment in Patient Flow. *Production and Operations Management* (2016), doi 10.1111/poms.12659

Figure 5 Willingness to Wait (WTW) Distribution (fraction of lost sales as a function of the appointment wait time) for Specialty ID 400 (orthopedics)



sales rate is ≈53%. The patient access team of the healthcare system (with whom we shared these results) concurred that it is plausible that half of the calling patients leave without scheduling an appointment.

The "Wait-Lost sales" relationship can be estimated for any clearly defined patient sub-population. For example, Table A6 also illustrates this for urgent patients; as expected, we find that urgent patients are more likely to balk when presented with an increased time to appointment.

5.2. Sensitivity to Wait

Next, we discuss the effect of wait times on the no-show rate. We find that an increase in wait time to an appointment decreases the likelihood of arrival for that appointment (Table 2). Both unweighted (SS) and weighted estimators (SS IPW) produce similar results; the discussion below is based on the SS IPW estimator, correcting for the attrition of low-WTW patients from the sample. Note that the coefficients on the deciles reflect the incremental change in sensitivity to wait between deciles. That is, the sensitivity to wait in the first decile is the sum of the decile coefficients, and in the tenth decile, it is simply the coefficient on the Wait variable. The effect of the quadratic term is small for practically observed wait times. At wait time of 50 days, the contribution of the quadrating term is an order of magnitude smaller than the effect of the linear terms. At smaller wait times, the effect of linear terms is even stronger. The inflection point, where the effect of the quadratic term is equal in magnitude to the effect of the linear term, is established at approximately 165 days.

We find that the sensitivity to wait is lower (in absolute value) for higher deciles of EWTW. Figure 6 shows how the sensitivity to wait changes over the deciles of EWTW. In particular, if EWTW is within the first decile of the distribution, one extra day of waiting reduces the odds of arrival for an appointment by approximately 8%, whereas if EWTW is in the tenth decile, the odds are reduced by just 2%. This suggests that patients who choose to book appointments with long wait times are more tolerant of possible additional wait times. The right panel of Figure 6 shows the estimated arrival probabilities for each EWTW decile. At an average wait time of 21 days, increasing the wait time by 1 day reduces the arrival probability from 66.5% to 64.6% for a patient in the first EWTW decile, and from 85.3% to 85.0% for a patient in the tenth EWTW decile. At the specialty level, we observe similar results (see Table A7 and section 7 for a discussion).

The likelihood of being a no-show is significantly affected by various patient- and appointment-level characteristics. No-shows are lower among older patients and among routine (but not urgent) appointments. No-shows are also higher if the patient's visit is paid by Medicare, Medicaid, or by the patient him/herself; afternoon (PM) appointments have higher likelihoods of no-shows. If the temperature on the day of the appointment happens to be warmer than the historical average for that day, the likelihood of no-shows decreases.

5.3. Marginal Effect of Wait

The effect of wait on lost sales and no-shows has direct implications for overall throughput (Table 3). By definition, any incidence of a patient from the calling population being treated at a provider depends on the probability that a random patient from the calling population schedules an appointment (Pr(*Scheduled*)) and arrives for it (Pr(*Arrived*)). To compute the effect on throughput (*Throughput*), we thus use the population-wide estimates of WTW, and the likelihood of a patient arriving for a scheduled appointment. We account for patient heterogeneity in sensitivity to wait, and assume that the mix of patients does not change in response to the marginal change in wait. If a patient is characterized by a vector of covariates $\mathbf{c}_i$, including time to appointment w_i, then $\Pr(Scheduled) = 1 - \hat{F}(w_i)$, where $\hat{F}(\cdot)$ is an estimate of the distribution of patients' WTW for the respective specialty defined by Equation (2), and $\Pr(Arrived) = \frac{e^{\hat{\boldsymbol{\theta}}\mathbf{c}_i}}{1+e^{\hat{\boldsymbol{\theta}}\mathbf{c}_i}}$, where $\hat{\boldsymbol{\theta}}$ is given by Equation (2). The effect of wait on throughput is given by: $\frac{\partial Throughput}{\partial W} = \Pr(Scheduled) \times \frac{\partial \Pr(Arrived)}{\partial Wait} + \Pr(Arrived) \times \frac{\partial \Pr(Scheduled)}{\partial Wait}$, where the first term represents the marginal increase in no-shows and the second represents the marginal increase in lost sales. Note that the computation of marginal effects focuses on throughput and revenue, disregarding the costs associated with the change in wait time, and applies coefficients estimated in Table 2 for all providers with the

Please Cite this article in press as: Osadchiy, N., D. KC. Are Patients Patient? The Role of Time to Appointment in Patient Flow. *Production and Operations Management* (2016), doi 10.1111/poms.12659

Table 2 Effect of Wait Times on Arrival Probability

	SS (1)	SS IPW (2)
Wait	−0.0267 (0.0018)***	−0.0292 (0.0023)***
Wait^2	0.00007 (0.00001)***	0.00009 (0.00001)***
Wait*$\mathbb{1}\{Decile(EWTW) \leq 1\}$	−0.0061 (0.003)**	−0.0086 (0.0037)**
Wait*$\mathbb{1}\{Decile(EWTW) \leq 2\}$	−0.0208 (0.0089)**	−0.0223 (0.0092)**
Wait*$\mathbb{1}\{Decile(EWTW) \leq 3\}$	0.0013 (0.0031)	0.0026 (0.0035)
Wait*$\mathbb{1}\{Decile(EWTW) \leq 4\}$	−0.0043 (0.0058)	−0.0054 (0.0083)
Wait*$\mathbb{1}\{Decile(EWTW) \leq 5\}$	−0.0033 (0.0048)	−0.0047 (0.0056)
Wait*$\mathbb{1}\{Decile(EWTW) \leq 6\}$	−0.0066 (0.0019)***	−0.0087 (0.0022)***
Wait*$\mathbb{1}\{Decile(EWTW) \leq 7\}$	−0.0014 (0.0013)	−0.002 (0.0015)
Wait*$\mathbb{1}\{Decile(EWTW) \leq 8\}$	−0.0033 (0.0031)	−0.0032 (0.0038)
Wait*$\mathbb{1}\{Decile(EWTW) \leq 9\}$	−0.0002 (0.0019)	−0.001 (0.0027)
FirstAvailable	0.1356 (0.0441)***	0.1537 (0.0505)***
Patient-level controls		
Gender (male)	0.0805 (0.0366)**	−0.0114 (0.0514)
Ethnicity (non-hispanic or latino)	0.0392 (0.1429)	0.1679 (0.2135)
Ethnicity (not reported)	−0.0346 (0.1432)	0.1235 (0.212)
Marital status (life partner)	−0.1978 (0.5421)	−0.5327 (0.5577)
Marital status (married)	−0.0645 (0.5388)	−0.3359 (0.5521)
Marital status (not reported)	−0.2394 (0.5397)	−0.6189 (0.5535)
Marital status (separated)	−0.399 (0.5632)	−0.8454 (0.598)
Marital status (single)	−0.2808 (0.5392)	−0.6081 (0.553)
Marital status (widow/er)	−0.2945 (0.5448)	−0.7017 (0.5667)
Age	0.0081 (0.0013)***	0.0092 (0.002)***
Payer (HMO)	−0.066 (0.0593)	−0.0319 (0.0802)
Payer (Medicaid)	−0.6908 (0.0758)***	−0.7369 (0.1062)***
Payer (Medicare)	−0.418 (0.0682)***	−0.4562 (0.0929)***
Payer (Other)	−0.0456 (0.1457)	−0.0324 (0.1829)
Payer (Outsourced)	−0.3973 (0.609)	0.628 (0.9062)
Payer (PPO)	−0.0705 (0.0792)	−0.0984 (0.1112)
Payer (self-pay)	−0.688 (0.0949)***	−0.7144 (0.1335)***
Appointment-level controls		
Scheduled time (PM)	−0.08 (0.0338)**	−0.1275 (0.0446)***
Distance to provider	0 (0.0002)	−0.0001 (0.0002)
Medical condition controls		
Urgent	0.0448 (0.0666)	0.0249 (0.0889)
Routine	0.0983 (0.0492)**	0.1055 (0.059)*
Weather controls		
Average temperature departure	0.0069 (0.0023)***	0.0066 (0.0031)**
Precipitation	−0.0982 (0.1373)	0.0467 (0.1772)
Constant	2.1577 (0.7158)***	2.2789 (0.7663)***
Provider & DOW reg/sched. F.E.	Yes	Yes
McFadden Pseudo R^2	0.131	0.128

27,555 observations over 476 providers. 111 providers are excluded due to no variation in the dependent variable. Baseline: Gender (female), Ethnicity (hispanic or latino), Marital status (divorced), Payer (Blue Cross), Scheduled time (AM). Robust standard errors are in parentheses. *10% statistical significance; **5% statistical significance; ***1% statistical significance. Deciles of EWTW are computed by specialty. Sample medians of EWTW for deciles 1-10 are: 25.4, 28.9, 30.9, 35.8, 40.6, 42.2, 58.0, 59.8, 72.8, and 114.4 days, respectively. In model (2) the baseline probability of arrival at *Wait* = 0 is 0.9451 (female, hispanic or latino, divorced, Blue Cross, age 49, distance 35 miles, Monday AM appointment with a radiation oncologist, typical dry weather). Sixty-seven out of 476 providers have statistically significant fixed effects (FE), mean(FE) = −0.022, std(FE) = 0.999. The following modifications of model (2) with added interaction terms were also estimated for robustness: (i) with *Wait* × *Urgent*, (ii) *Wait* × *Routine*, and (iii) *Wait* × *Age*. All interactions were found to be insignificant.

appropriate fixed effect, disregarding potential heterogeneity in these coefficients.

Since the patient volume differs across providers, we compute the effect for each provider separately, and report the summary statistics. The derivatives are computed as average marginal effects across all observations for a given provider (as opposed to the value of the derivative at the mean wait). This takes into account the distribution of waits for a provider. We find (row 9 of Table 3) that on average a 1 day increase in wait corresponds to a reduction in throughput by 0.13 NPV per day; 79% of this reduction is due to balking at scheduling an appointment, and 21% is due to reneging (no-shows). The average relative impact of 1 extra day of wait on throughput is a 5.7% decrease in new patients seen per day (row 12

Please Cite this article in press as: Osadchiy, N., D. KC. Are Patients Patient? The Role of Time to Appointment in Patient Flow. *Production and Operations Management* (2016), doi 10.1111/poms.12659

Figure 6 **Left: Sensitivity to Wait as a Function of Offered Wait. Offered waits are the median waiting times for the respective expected willingness to wait (WTW) deciles. Error bars represent standard errors of the estimates. Right: Arrival probability as a function of booked wait, by expected WTW decile**

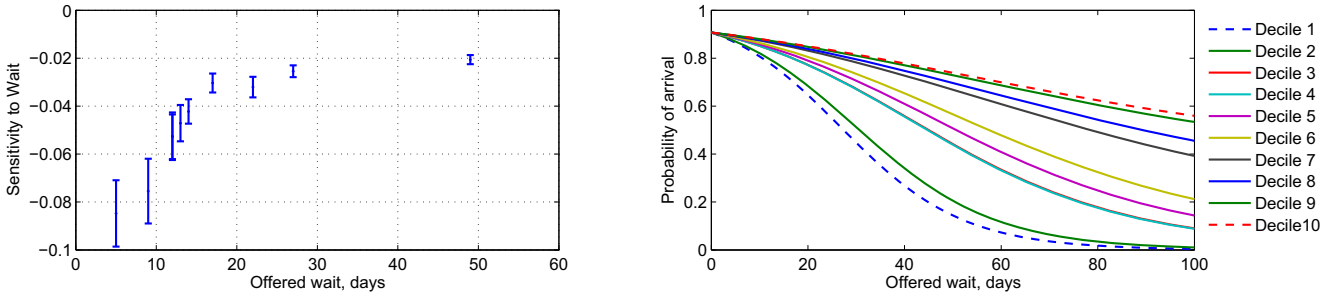


Table 3 **Average Marginal Effect (AME) of Wait Time *w* on the Probability of Booking an Appointment, Arrival for an Appointment, Throughput, and Revenue (AMEs computed for each of 476 providers. Mean, SD, median, min, max are reported)**

		Mean	SD	Median	Min	Max
1	*w* (days)	25.25	26.89	15.31	0.63	169.79
2	No. of requests for appointments (calls/day)	4.01	5.15	2.64	0.00	73.00
3	Pr(*Sched*)	0.6907	0.2315	0.7521	0.0425	1.0000
4	Pr(*Arr*)	0.7774	0.1464	0.8080	0.0529	0.9865
5	Throughput (%)	55.01	20.55	55.47	1.16	94.44
6	Throughput (patients/day)	2.1765	2.7556	1.3892	0.0000	27.9537
7	*d*Pr(*Sched*)/*dw*	−0.0310	0.0396	−0.0228	−0.3724	0.0000
8	*d*Pr(*Arrived*)/*dw*	−0.0042	0.0022	−0.0039	−0.0102	0.0003
9	*dThroughput*/*dw* (patients/day)	−0.1316	0.2879	−0.0489	−3.3333	0.0000
10	Included from increased lost sales	78.97%	18.39%	81.84%	2.24%	100.00%
11	Included from increased noshows	21.03%	18.39%	18.16%	0.00%	97.76%
12	*dThroughput*(%)/*dw* (% patients/day)	−5.71%	5.94%	−4.33%	−63.01%	−0.11%
13	Average medicare payment per visit, $	74.59	42.30	65.12	15.76	279.86
14	Daily revenue, $	142.27	153.61	90.12	0.00	956.69
15	*dRevenue*/*dw* ($ per day/day of wait)	−9.34	20.40	−3.38	−168.92	0.00
16	*dRevenue* (%)/*dw* (% daily revenue/day of wait)	−5.71%	5.94%	−4.33%	−63.01%	−0.11%

of Table 3). Note that some providers could actually improve throughput by increasing wait times to an appointment; for them, the increase in lost sales due to wait is compensated by a reduction in no-shows.

To explore the revenue implications of wait times, we augment the throughput analysis above with the provider revenue data described in section 3. The impact on revenue is then simply the throughput gains multiplied by the average reimbursement rate. For scenarios with reduced waits, we assume that there is adequate capacity available to accommodate the increased demand. We find that on average across providers, a 1-day reduction in wait times is associated with an increase of approximately $9.34 in daily revenue from NPV (Table 3, row 15). However, we note that there is significant heterogeneity in the impact of wait reduction on revenues, as the impact ranges from $169 per day, to small negative amounts. The provider with the highest impact is a GIM physician, and the one with the lowest impact is a cardiac electro physiologist. In relative terms, about 5% of the providers could achieve an improvement in revenues in excess of 15% by reducing their waits by 1 day. In particular, one provider (GIM) is estimated to achieve a 41% increase in NPV revenues by reducing waits by 1 day on average. However, for a substantial portion (15%) of providers, the impact on revenues from new patients is marginal (<1%). Those include a medical oncologist (a revenue impact of 0.94%), a neurologist (0.89%), and others.

Our analysis shows that the impact of wait times on revenue is moderated by physician specialty. In particular, the top three providers with the greatest relative impact on revenues (and throughput) resulting from a reduction in wait times were general internal medicine providers. In contrast, the three physicians with the lowest impact of a wait time reduction on revenues were an interventional cardiologist, a cardiac electro-physiologist, and a neuro-ophthalmologist, respectively. This result is consistent with our expectations: patients seeking to see a primary care or GIM physician often have a medical condition that requires immediate attention (e.g., cold, fever, bleeding, etc.), whereas their visits to a highly specialized

Please Cite this article in press as: Osadchiy, N., D. KC. Are Patients Patient? The Role of Time to Appointment in Patient Flow. *Production and Operations Management* (2016), doi 10.1111/poms.12659

provider for chronic conditions often involve longer waits.

6. Wait and Capacity Utilization

In this section, we investigate a counterfactual effect of wait reduction on capacity utilization. Specifically, we study the capacity utilization (or the no-show rate) under the condition when all patients are offered the same shorter wait. We assume that there is ample demand, and the mix of patients is driven solely by the WTW distribution and the offered wait cut-off. Capacity is held constant. The effect of offering a shorter wait to all patients on no-shows is comprised of two components: on one hand, offering shorter wait times will increase the arrival probability for patients who would have booked appointments even if the wait is long. On the other hand, offering shorter wait times will attract low-WTW patients (the ones who would not have booked otherwise) to book appointments. Those new patients can have higher sensitivity to wait and, therefore, may be less likely to show up. The resultant effect will thus be a superposition of the increased probability of arrival and the change in the mix of patients. Thus, the total effect of wait on capacity utilization can be non-monotone (e.g., see Example 2).

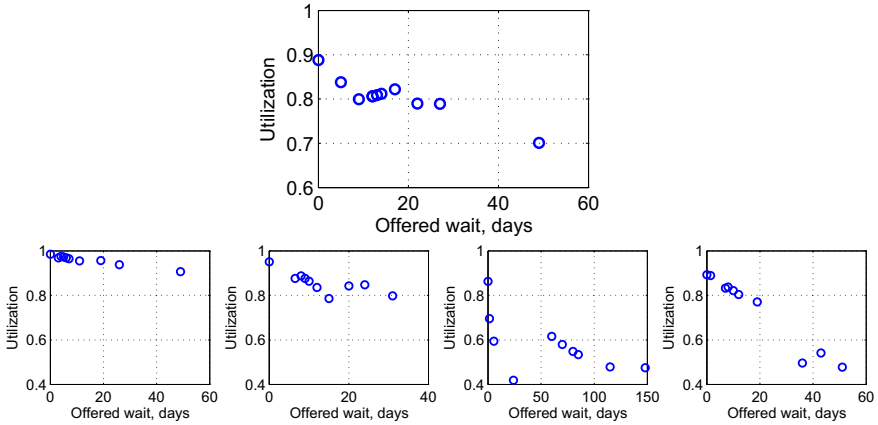
We construct the counterfactual experiment as follows. We take a median wait corresponding to each decile of the expected WTW and compute the probability of arrival using the coefficient estimates from Table 2 (SS IPW) corresponding to that decile. An additional point on the graph corresponds to the zero wait, where the probability of arrival is determined by the intercept in the regression. The fixed effect for provider is at its mean level, and all others are at zero.

Figure 7 presents the impact of wait on capacity utilization pooled across clinical specialties, as well as for distinct specialties. We find that offering shorter wait times does not necessarily reduce no-shows. In fact, the capacity utilization increases only when wait times are reduced to about 20–25 days. A further reduction in wait time changes the mix of patients so that the overall sensitivity to wait increases. Consequently, the arrival probability or capacity utilization decrease. Across specialties, a reduction of wait times from 20 to 10 days decreases capacity utilization from 83% to 80%. Given that on average there are 2.17 NPV appointments per day (see Table 3, row 6), this reduction in capacity utilization decreases NPV throughput from approximately 1.8 to 1.7 patients per day. The associated average revenue loss from NPV is $7.5 per day.

We continue to observe the non-monotone effect of wait on capacity utilization at the specialty level. The regression results at the specialty level (Table A7) show a steeper decrease in sensitivity to wait across WTW deciles for the spine center, neurology, and otolaryngology specialties, contributing to the non-monotone effect (Figure 7). This suggests that wait serves an important function in deterring appointments with a low probability of arrival.

It is important to contrast the results of this section with the marginal effects of section 5.3. The marginal effects are computed under the assumption that the mix of patients does not change with wait time. Consequently, they indicate that reducing wait times increases capacity utilization, a finding consistent with those previously reported in the literature (e.g., Gallucci et al. 2005). The counterfactual analysis accounts for the changing mix of patients with an offered wait times and suggests that reducing wait times may have a more muted or even negative effect on capacity utilization. This has important implications for capacity planning and improving access in healthcare and other services

Figure 7 The of Offered Wait on Capacity Utilization for a New Patient Visit for All Specialties (top), and by Specialty (bottom L to R): Orthopedics, Spine Center, Neurology, Otolaryngology



Please Cite this article in press as: Osadchiy, N., D. KC. Are Patients Patient? The Role of Time to Appointment in Patient Flow. *Production and Operations Management* (2016), doi 10.1111/poms.12659

(e.g., counseling, professional, or government). If demand is unchanged, a capacity investment generally reduces wait times and makes the service more accessible. However, it can generate less revenue than predicted, as the rate of no-shows can remain unchanged or even increase. Thus, in evaluating the effect of capacity investments, it is important to jointly consider booking and arrival decisions, as the reduced wait time can affect the mix of customers drawn into the system.

7. Validity of Modeling Assumptions and Robustness Tests

In this section, we check the validity of the assumptions made for the WTW and no-show estimation and report the results of several robustness tests. Our estimation methods assume that patients' calls arrive according to a time-homogeneous Bernoulli process, that patients prefer shorter waits, and have preferences for a specific provider.

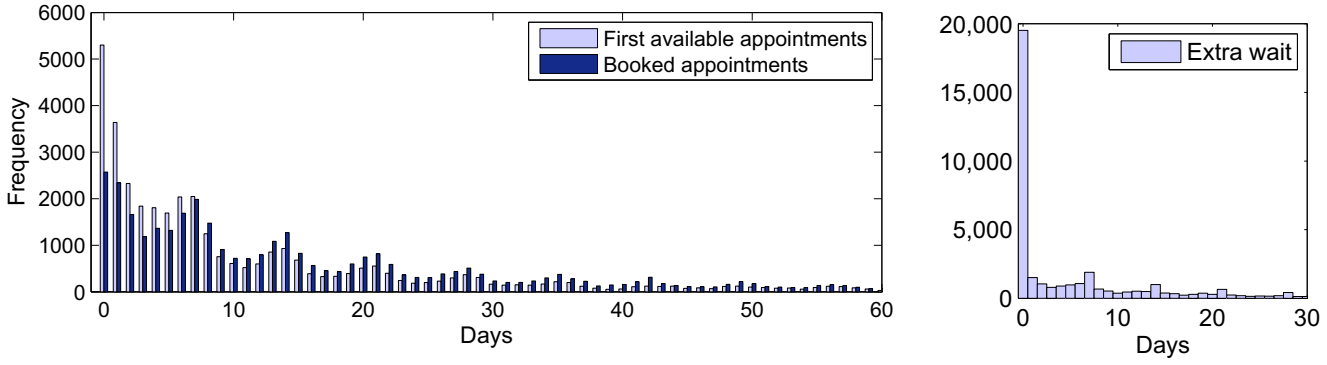
To test the *inter-temporal heterogeneity in arrival* of patient requests, we first examined the capacity availability data to see whether there were statistically significant differences in available appointment waits by calling day of week. By analyzing the variance of the available waits, we found that there was no statistically significant difference in wait time availability across days of week. Controlling for provider effects, calling day of week in the ANOVA analysis is not significant. This means that we can test for arrival homogeneity by simply looking at the booking rates as proxies for the arrival rates across different days of week. We perform ANOVA analysis and find that the estimates of booking rates are not significantly different across providers by calling day of week: calling day of week is a significant factor for only 22 out of 587 providers (<4%).

To test whether *patients prefer shorter waits* or exhibit preferences for specific wait times or days of week, we perform ANOVA analysis for the number of appointments with a provider by scheduled day of week. Day of week is a significant factor only for 44 out of 587 providers (<8%). We observe a close match between the closest available wait, and the incurred wait, indicating that patients choose time slots that are not far off from the shortest waits. In particular, the correlation between the available wait and the one actually chosen by the patient is high, at 96.1%, and the booked wait is equal to the first available wait for 52% of appointments.

Figure 8 shows the distribution of the closest available appointment wait offered to patients, as well as the actual wait time that they chose. We find that patients book the first available appointment in approximately 52% of cases. Of note is the periodicity in the histograms' peaks at multiples of 7. This is not due to the preference for specific waits (of multiples of 7), but rather due to the call center being open on week days only. For this reason, appointments with waits other than multiples of 7 may fall on a weekend and become unavailable. Periodicity disappears if wait is measured in business days.

The specialized nature of providers' services, as well as our interviews with call center representatives, suggest that few providers are substitutes, and patients rarely call multiple providers for an appointment. That is, patients exhibit a *preference for a specific provider*, despite other providers having shorter wait. We examine the data to validate this account. In the list of providers, we were able to identify a set of three family medicine doctors (who were more likely to offer similar services). These doctors all have about 15 years of experience, high-quality ratings, share the same location, and see similar patients. For this set of providers, we wanted to determine whether a patient

Figure 8 Left: First Available Waiting Times vs. Waiting Times for Booked Appointments (truncated at 60 days). Right: Voluntary extra wait incurred by patients with respect to the first available appointment (truncated at 30 days). Booked wait is equal to the first available wait for 52% of appointments



Please Cite this article in press as: Osadchiy, N., D. KC. Are Patients Patient? The Role of Time to Appointment in Patient Flow. *Production and Operations Management* (2016), doi 10.1111/poms.12659

would stick with their intended physician or switch to the other provider if the wait to appointment were shorter. Of the 602 observed appointments, only 17 have booked waits that are less than the available waits for the alternative providers; 238 have exactly the same wait, and 347 have strictly greater wait (by an average of 4.5 days). In other words, patients were willing to endure significantly longer waits to schedule an appointment with a provider, even though a shorter wait time was available at another provider.

We check the *robustness* of our results to the alternative specifications of model (2). We fit the models with additional interactions, the log-transformed wait variable, the standardized wait variable, and with winsorized or dropped long waits (exceeding 90 days). The models have a similar fit to the baseline case. We continue to observe the decreasing sensitivity to wait for high-WTW patients (Table A9). Furthermore, we continue to observe this effect when we exclude the unrebooked cancellations, or include rebooked cancellations in the analysis (Table A8). As emphasized before, using the latent WTW variable is crucial for our results. Estimating a naive model based on the second stage only (arrivals and no-shows) and using deciles of actual wait does not reveal an increased sensitivity to wait for short waits (see Table A9, column 6). At the specialty level (Table A7), we observe a variability in the coefficient values, yet, directionally, the effect of increased sensitivity to wait for low-WTW deciles is robust.

To guard against uncertainty in the WTW estimates, we re-estimated model (2) using quartiles of the expected WTW. Compared to deciles, quartiles provide a coarser description of WTW, but reduce the chances of categorizing an observation into a wrong quartile. We continue to find that low-EWTW quartiles are more sensitive to wait.

We performed an additional robustness test on the simulated data to study the scenario where patients book appointments with waits that are longer than the earliest available time, but are still within their WTW. In this case, the estimated WTW distribution first-order stochastically dominates the true WTW distribution. Therefore, the estimate of the effect of wait on lost sales per section 4.1.1 can be viewed as a lower bound on the true effect of wait on lost sales.

Finally, to ascertain that our results are not driven by year-specific idiosyncracies, we obtained appointment schedule and capacity data for the same 2-month periods in the two preceding years (2010 and 2011). This data is substantively similar to the 2012 data, although it lacks the patient-level controls. Using the methods of section 4, we confirm our main findings from the year 2012. Specifically, we find that the patients' sensitivity to wait is smaller in higher deciles of WTW. We continue to observe a non-monotone counterfactual effect of offered wait on capacity utilization. Increasing wait from approximately 12 to 30 days results in a gradual increase in capacity utilization from approximately 76% to 83% in aggregate across specialties. We observe similar effects for major clinical specialties. The complete results for years 2010 and 2011 are available in the E-Companion.

8. Conclusion

In this study, we examine the effect of appointment wait times on patient flow and capacity utilization in an outpatient clinic setting. We assemble a novel and rich transaction-level appointment and patient flow dataset for over 100 distinct clinical specialties. The dataset allows us to examine the set of wait times offered to each calling patient, and the choices that they make, i.e., whether to book an appointment, and whether to show up for a scheduled appointment.

We find that the wait time significantly affects patients' decisions to book appointments, resulting in noticeable lost sales to the provider. At the same time, the wait time increases the likelihood of a no-show for already scheduled appointments. Furthermore, the sensitivity of the no-show probability to wait varies across patients: patients who choose to book appointments with long wait times are typically less sensitive to an extra day of waiting. Because of the interaction between WTW and the sensitivity to wait, the effect of wait reduction on no-shows can be non-monotone. While at long or very short waits the wait reduction reduces no-shows, the effect can be reversed at moderate waits. The intuition for this result is the following: offering shorter waits changes the mix of patients increasing the proportion of patients with a high sensitivity to wait and high probability of no-show.

In general, estimating the likelihood that a patient foregoes scheduling an appointment (or lost sales) is challenging because such patients are not observed by the econometrician, and the lost sales are usually not tracked. To impute lost sales based on the observed data, we apply a novel and generalizable non-parametric model of patient choice. Our model employs a maximum likelihood specification, which allows us to generate the distribution of patients' WTW even for those patients that we do not directly observe. This allows us to estimate lost sales as a function of wait times.

Similarly, estimating the effect of wait times on the no-show rate is challenging due to the endogenous selection of patients at the appointment-booking stage. That is, patients who choose to book an appointment may have a lower sensitivity of no-shows to wait when compared to patients who choose to balk. To generate an unbiased estimate of the effect of wait times on

Please Cite this article in press as: Osadchiy, N., D. KC. Are Patients Patient? The Role of Time to Appointment in Patient Flow. *Production and Operations Management* (2016), doi 10.1111/poms.12659

no-shows, we explicitly model the interaction between WTW and the sensitivity to wait, and correct for the attrition of observations corresponding to non-bookings by the low-WTW patients. Our analysis shows that ignoring the interaction effect with the WTW can lead to a significant underestimate of the true effect of wait on the likelihood of a no-show.

Future research should consider various alternative approaches to easing the mismatch between demand and supply. There is a growing body of literature on improving scheduling and capacity allocation decisions, including overbooking, to better match supply with demand. Our research complements this work by providing empirical estimates of the extent to which throughput and capacity utilization can be improved through more effective management of wait times.

In our analysis, we estimate a distribution of patient WTW for a specialty's entire patient population. In future work, it would be helpful to measure WTW directly and relate it to the sensitivity to wait and arrival probability. Then, one could further examine the drivers of WTW, including severity levels, the presence of outside options, insurance status, and travel distance. This could lead to other approaches to reducing the no-show rate, for example with physician calls, penalties for no-shows, or simply providing transportation to and from the clinic. Finally, even though our analyses are applied to an outpatient healthcare context, we believe that these methods could be adapted for other services with a similar multi-stage process that involves sample selection and subsequent attrition losses.

Acknowledgments

The authors are grateful to Simone Marinesi (University of Pennsylvania), Sergei Savin (University of Pennsylvania), the senior editor, and three anonymous referees for their constructive feedback.

Appendix A. Tables

Table A1 Descriptive Statistics (new patient visits, sample for analysis of no-shows)

Variable	Mean	SD	Median	Min	Max	*N*
Waiting time	20.83	31.34	11	0	363	29,089
Selected specialties						
Orthopaedics	11.86	15.28	7	0	240	4381
Spine center	18.24	29.96	12	0	355	2715
Neurology	67.91	70.17	49	0	238	1378
Otolaryngology	24.05	24.24	14	0	120	1198
Waiting time (if arrived)	18.1	26.05	9	0	363	23,544
Waiting time (if no-show, or urc)	32.42	45.89	16	0	353	5545
Among unrebooked cancellations (urc)	31	47.06	14	0	323	2757
Number of patients scheduled (per provider)	49.56	65.91	27	1	482	587
Number of patients arrived (per provider)	40.11	54.45	22	0	440	587
% arrived (per provider)	79.82	18.84	83.52	0	100	587
% no-show and urc (per provider)	20.18	18.84	16.48	0	100	587
% unrebooked cancellations	10.63	17.02	1.98	0	100	587

29,089 observations, 25,114 patients for 587 providers, 107 clinical specialties in January–February 2012. Initial data: 237,954 appointment records for 914 providers (all visits); 37,688 NPV visits for 596 providers. Less rescheduled, canceled and rebooked, pending: 157,272 appointments for 827 providers including 29,089 NPV appointments for 587 providers remain.

Table A2 Total Available Capacity for All Visits, Per Provider

Variable	Mean	SD	Median	Min	Max	*N*
Total capacity	178.68	201.53	119	2	3087	914
Mean daily capacity	5.63	5.11	4	1.02	71.79	914

163,312 slots over 914 providers in total.

Please Cite this article in press as: Osadchiy, N., D. KC. Are Patients Patient? The Role of Time to Appointment in Patient Flow. *Production and Operations Management* (2016), doi 10.1111/poms.12659

Table A3 Keywords for Determining the Urgency of a Medical Condition Based on the Clinic Visit Comment

	Medical condition	
	Urgent	Routine
Keywords	Pain, urgent, cough, complic, fever, irregular, bleeding, abnormal, nausea, swelling, blood in, injury, hurt, spasm, bad, burn, ache, flu, severe, swollen	routine, 2nd opin, consult, annual, concern, well wom, wellwom, check up, checkup, discuss, eval, physical, exam, orientation
No. of appointments	2601	9784
Median wait, days	7	13

Table A4 Demographics and Supplementary Patient, Appointment, Provider- and Specialty-Level Data

Gender	%	Ethnicity	%	Marital status	%	Payer	%
Female	59.8	Hispanic or latino	1.4	Divorced	0.1	Blue Cross	11.6
Male	40.2	Non-hispanic or latino	54.1	Life partner	6.4	HMO	41.6
		Not reported	45.9	Married	41	Medicaid	9.5
				Not reported	19.9	Medicare	21.2
				Separated	0.8	Other	1.6
				Single	27.6	Outsourced	0.1
				Widow/er	4.2	PPO	10.7
						Self Pay	3.7

	Mean	SD	Median	Min	Max	*N*
Patients						
Age, years	48.81	20.16	50.37	0.01	100.01	29,089
Distance, miles	35.63	113.8	13.1	0	4503	28,993
Day of appointment						
Average temperature, F	49.06	8.07	49	28	67.5	29,089
Average temperature departure, F	3.86	7.86	3.4	−15	18.4	29,089
Precipitation, inches	0.07	0.13	0	0	1.19	29,089
Providers						
Satisfaction rating, %	95.29	8.32	100	75	100	204
Average medicare payment, $	71.46	41.05	56.97	15.38	279.86	361
Specialties with most NPVs			*Specialties with least NPVs*			
Orhtopaedics	15.1%		Pelvic reconstructive surgery			0.03%
Spine care	9.3%		Lung transplantation			0.03%
Neurology	8.3%		Diabetes education			0.01%

Table A5 Summary of the Willingness to Wait (WTW) and Arrival Rate Estimates

	Mean	SD	Median	Min	Max	*N*
Average WTW (by specialty), days						
Population	26.98	18.64	22.38	2.00	81.96	107
Sample	39.64	35.08	28.99	1.00	231.58	107
Arrival rate (by provider)						
λ	0.23	0.21	0.17	0.00	1.00	587
NPV calls per day	4.01	5.15	2.64	$<10^{-3}$	73.00	587
Specialties with highest mean WTW			*Specialties with lowest mean WTW*			
Pediatric ophthalmology	81.9		General pathology			4.23
Neurology	75.1		GIM corporate			3.46
Glaucoma ophthalmology	73.8		Diabetes education			2

Please Cite this article in press as: Osadchiy, N., D. KC. Are Patients Patient? The Role of Time to Appointment in Patient Flow. *Production and Operations Management* (2016), doi 10.1111/poms.12659

Table A6 Wait Characteristics at a Given Level of Lost Sales, by Urgency. Summary Statistics across 107 Specialties

	Wait, days				
	All appointments			Urgent	Routine
% Lost sales	Mean	SD	Median	Mean	Mean
1	10.89	13.37	6	9.26	10.96
5	11.89	13.62	7	9.38	12.09
10	12.24	13.55	8	9.50	12.86
20	13.80	14.01	9	10.15	14.12
30	14.73	14.14	9	12.56	15.67
40	17.20	14.66	12.5	14.50	17.56
50	18.74	15.03	14	16.26	19.81
60	24.08	19.69	18.5	18.07	24.72
70	29.43	25.05	21	22.29	29.16
80	36.65	34.66	27.5	25.93	37.04
90	48.57	45.84	35.5	30.88	48.97
100	107.35	83.63	86	51.48	106.32

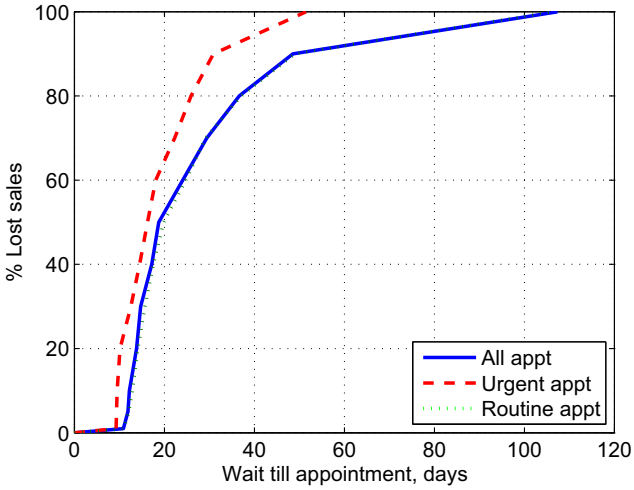


Table A7 Effect of Wait on Arrivals by Specialty

	Orthopedics (1)	Spine Center (2)	Neurology (2)	Otolaryngology (2)
Wait	−0.0507 (0.0091)***	−0.0445 (0.0066)***	0.0206 (0.0085)**	−0.0516 (0.0187)***
Wait^2	0.0002 (0.0001)***	0.0001 (0)***	0.0001 (0)***	0.0003 (0.0002)
Wait $\times \mathbb{1}\{Decile(EWTW) \leq 1\}$	−0.1303 (0.039)***	−0.0426 (0.0186)**	−0.4627 (0.1629)***	0.0497 (0.2261)
Wait $\times \mathbb{1}\{Decile(EWTW) \leq 2\}$	0.0000	0.0000	−0.1653 (0.0642)**	−0.0136 (0.0488)
Wait $\times \mathbb{1}\{Decile(EWTW) \leq 3\}$	0.0000	0.0000	−0.0613 (0.0071)***	0.0000
Wait $\times \mathbb{1}\{Decile(EWTW) \leq 4\}$	0.0000	0.0000	0.0000	0.0000
Wait $\times \mathbb{1}\{Decile(EWTW) \leq 5\}$	−0.0209 (0.04)	0.0000	0.0000	−0.0094 (0.0341)
Wait $\times \mathbb{1}\{Decile(EWTW) \leq 6\}$	−0.044 (0.0151)***	−0.0461 (0.0155)***	0.0000	0.0165 (0.0242)
Wait $\times \mathbb{1}\{Decile(EWTW) \leq 7\}$	0.0000	−0.0119 (0.0089)	−0.0578 (0.0048)***	0.0114 (0.0118)
Wait $\times \mathbb{1}\{Decile(EWTW) \leq 8\}$	−0.0129 (0.0118)	0.0000	0.0000	0.0000
Wait $\times \mathbb{1}\{Decile(EWTW) \leq 9\}$	0.0011 (0.0092)	−0.01 (0.0088)	0.0000	0.005 (0.0093)
FirstAvailable	0.2054 (0.132)	0.0071 (0.1456)	0.2139 (0.2268)	−0.3105 (0.2734)
Provider, DOW, patient & appointment controls	Yes	Yes	Yes	Yes

*10% statistical significance; **5% statistical significance; ***1% statistical significance.

Table A8 Robustness Checks with Respect to Inclusion of Cancellations (SS IPW estimator)

	No-shows and rebooked canc.	No-shows only	No-shows and all canc.
Wait	−0.0292 (0.0023)***	−0.0341 (0.0032)***	−0.0292 (0.0019)***
Wait^2	0.00009 (0.00001)***	0.0001 (0)***	0.0001 (0)***
Wait $\times \mathbb{1}\{Decile(EWTW) \leq 1\}$	−0.0086 (0.0037)**	−0.0139 (0.0051)***	−0.0106 (0.003)***
Wait $\times \mathbb{1}\{Decile(EWTW) \leq 2\}$	−0.0223 (0.0092)**	−0.029 (0.0117)**	−0.0043 (0.0038)
Wait $\times \mathbb{1}\{Decile(EWTW) \leq 3\}$	0.0026 (0.0035)	−0.0034 (0.0042)	−0.0017 (0.003)
Wait $\times \mathbb{1}\{Decile(EWTW) \leq 4\}$	−0.0054 (0.0083)	−0.0062 (0.0096)	−0.006 (0.0069)
Wait $\times \mathbb{1}\{Decile(EWTW) \leq 5\}$	−0.0047 (0.0056)	−0.0144 (0.0081)*	−0.0036 (0.0031)
Wait $\times \mathbb{1}\{Decile(EWTW) \leq 6\}$	−0.0087 (0.0022)***	−0.0093 (0.0028)***	−0.0057 (0.0018)***
Wait $\times \mathbb{1}\{Decile(EWTW) \leq 7\}$	−0.002 (0.0015)	−0.0043 (0.0022)*	−0.0036 (0.0014)**
Wait $\times \mathbb{1}\{Decile(EWTW) \leq 8\}$	−0.0032 (0.0038)	−0.006 (0.0053)	−0.006 (0.0027)**
Wait $\times \mathbb{1}\{Decile(EWTW) \leq 9\}$	−0.001 (0.0027)	−0.0032 (0.0034)	−0.0025 (0.0021)
N	27,555	24,021	35,024

All control variables as in model (2) are included.

Please Cite this article in press as: Osadchiy, N., D. KC. Are Patients Patient? The Role of Time to Appointment in Patient Flow. *Production and Operations Management* (2016), doi 10.1111/poms.12659

Table A9 Alternative Specifications of Model (2), Estimation on Subsamples with Dropped or Winsorized Waits of More Than 90 days, and Second Stage Only Estimation Based in Deciles of Wait

	Base	Log(wait+1)	Standardized wait	Dropped	Winsorized	Second stage only
Wait	−0.0292 (0.0023)***	−0.558 (0.0324)***	−0.2278 (0.0323)***	−0.0467 (0.005)***	−0.0233 (0.0019)***	−0.0293 (0.0023)***
Wait^2	0.0001 (0)***	–	–	0.0003 (0.0001)***	0.0000 (0)	0.0001 (0)***
Wait $\times \mathbb{1}\{Decile(EWTW) \leq 1\}$	−0.0086 (0.0037)**	−0.0134 (0.0497)	−0.4757 (0.0663)***	−0.0062 (0.0043)	−0.0097 (0.0037)**	−0.0248 (0.0384)
Wait $\times \mathbb{1}\{Decile(EWTW) \leq 2\}$	−0.0223 (0.0092)**	−0.1481 (0.0564)***	−0.1553 (0.117)	−0.0186 (0.0091)**	−0.0237 (0.0092)**	−0.0052 (0.0116)
Wait $\times \mathbb{1}\{Decile(EWTW) \leq 3\}$	0.0026 (0.0035)	0.014 (0.0812)	−0.2063 (0.1261)	0.0047 (0.0045)	0.0024 (0.0035)	−0.0147 (0.0075)*
Wait $\times \mathbb{1}\{Decile(EWTW) \leq 4\}$	−0.0054 (0.0083)	−0.0624 (0.0824)	0.1815 (0.2547)	−0.0054 (0.0091)	−0.006 (0.0083)	−0.0045 (0.0054)
Wait $\times \mathbb{1}\{Decile(EWTW) \leq 5\}$	−0.0047 (0.0056)	0.0891 (0.0465)*	−0.0943 (0.1878)	−0.0048 (0.0077)	−0.0052 (0.0057)	0.0047 (0.004)
Wait $\times \mathbb{1}\{Decile(EWTW) \leq 6\}$	−0.0087 (0.0022)***	−0.0542 (0.0305)*	−0.2402 (0.0883)***	−0.0061 (0.0041)	−0.0094 (0.0022)***	−0.006 (0.003)**
Wait $\times \mathbb{1}\{Decile(EWTW) \leq 7\}$	−0.002 (0.0015)	−0.0098 (0.0387)	−0.1017 (0.0922)	−0.0006 (0.0035)	−0.0034 (0.0015)**	−0.0016 (0.0026)
Wait $\times \mathbb{1}\{Decile(EWTW) \leq 8\}$	−0.0032 (0.0038)	−0.0171 (0.0394)	0.0516 (0.0976)	0.0001 (0.0042)	−0.0047 (0.0038)	−0.0038 (0.002)*
Wait $\times \mathbb{1}\{Decile(EWTW) \leq 9\}$	−0.001 (0.0027)	0.0347 (0.0338)	0.0829 (0.0685)	0.0003 (0.0035)	−0.0021 (0.0026)	−0.0026 (0.0022)
N	27,555	27,555	27,548	26,500	27,555	27,555
McFadden's pseudo R^2	0.1279	0.1304	0.1282	0.1214	0.1266	0.1286

All control variables as in model (2) are included.

References

Aksin, Z., B. Ata, S. M. Emadi, C.-L. Su. 2013. Structural estimation of callers' delay sensitivity in call centers. *Management Sci.* **59**(12): 2727–2746.

Allon, G., A. Federgruen, M. Pierson. 2011. How much is a reduction of your customers' wait worth? An empirical study of the fast-food drive-thru industry based on structural estimation methods. *Manuf. Serv. Oper. Manag.* **13**(4): 489.

Batt, R. J., C. Terwiesch. 2015. Waiting patiently: An empirical study of queue abandonment in an emergency department. *Management Sci.* **61**(1): 39–59.

Bodenheimer, T., H. Pham. 2010. Primary care: Current problems and proposed solutions. *Health Aff.* **29**(5): 799–805.

Brown, L., N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, L. Zhao. 2005. Statistical analysis of a telephone call center: A queueing-science perspective. *J. Am. Stat. Assoc.* **100** (469): 36–50.

Buell, R. W., M. I. Norton. 2011. The labor illusion: How operational transparency increases perceived value. *Management Sci.* **57**(9): 1564–1579.

Cohen, A., D. Kaplan, M. Kraus, E. Rubinshtein, D. Vardy. 2007. Nonattendance of adult otolaryngology patients for scheduled appointments. *J. Laryngol. Otol.* **121**(03): 258–261.

Commonwealth. 2011. Why not the best? Results from the national scorecard on U.S. health system performance. Technical report, The Commonwealth Fund Commission on a High Performance Health System.

Debo, L., M. Kremer. 2014. Inferring quality from wait time. *Management Sci.* **62**(10): 3023–3038.

Farias, V., S. Jagabathula, D. Shah. 2013. A new approach to modeling choice with limited data. *Management Sci.* **59**(2): 305–322.

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

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

Ghorob, A., T. Bodenheimer. 2012. Sharing the care to improve access to primary care. *N. Engl. J. Med.* **366**(21): 1955–1957.

Green, L., V. Nguyen. 2001. Strategies for cutting hospital beds: The impact on patient service. *Health Serv. Res.* **36**(2): 421.

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

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

Hall, R. 2013. *Patient Flow: Reducing Delays in Healthcare Delivery*. International Series in Operations Research & Management Science. Springer, US.

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

Heckman, J. 1979. Sample bias as a specification error. *Econometrica* **47**(1): 153–161.

Huang, X. 1995. A planning model for requirement of emergency beds. *Math. Med. Biol.* **12**(3–4): 345–353.

Kwak, N., C. Lee. 1997. A linear goal programming model for human resource allocation in a health-care organization. *J. Med. Syst.* **21**(3): 129–140.

LaGanga, L. R., S. R. Lawrence. 2012. Appointment overbooking in health care clinics to improve patient service and clinic performance. *Prod. Oper. Manag.* **21**(5): 874–888.

Leclerc, F., B. H. Schmitt, L. Dube. 1995. Waiting time and decision making: Is time like money? *J. Consum. Res.* **22**(1): 110–119.

Liu, N. 2016. Optimal choice for appointment scheduling window under patient no-show behavior. *Prod. Oper. Manag.* **25**(1): 128–142.

Liu, N., S. Ziya, V. Kulkarni. 2010. Dynamic scheduling of outpatient appointments under patient no-shows and cancellations. *Manuf. Serv. Oper. Manag.* **12**(2): 347–364.

Lu, Y., A. Musalem, M. Olivares, A. Schilkrut. 2013. Measuring the effect of queues on customer purchases. *Management Sci.* **59**(8): 1743–1763.

Luo, J., V. G. Kulkarni, S. Ziya. 2012. Appointment scheduling under patient no-shows and service interruptions. *Manuf. Serv. Oper. Manag.* **14**(4): 670–684.

Mahajan, S., G. van Ryzin. 2001. Stocking retail assortments under dynamic consumer substitution. *Oper. Res.* **49**(3): 334–351.

Maister, D. 1985. The psychology of waiting lines. *Serv. Encounter* **1**: 13–23.

Millman, M. 1993. *Access to Health Care in America*. National Academies Press, Washington DC.

Moore, C., P. Wilson-Witherspoon, J. Probst. 2001. Time and money: Effects of no-shows at a family practice residency clinic. *Fam. Med.-Kansas City* **33**(7): 522–527.

Please Cite this article in press as: Osadchiy, N., D. KC. Are Patients Patient? The Role of Time to Appointment in Patient Flow. *Production and Operations Management* (2016), doi 10.1111/poms.12659

Musalem, A., M. Olivares, E. Bradlow, C. Terwiesch, D. Corsten. 2010. Structural estimation of the effect of out-of-stocks. *Management Sci.* **56**(7): 1180–1197.

van Ryzin, G., G. Vulcano. 2014. A market discovery algorithm to estimate a general class of nonparametric choice models. *Management Sci.* **61**(2): 281–300.

Sack, K. 2008. In Massachusetts, universal coverage strains care. *New York Times*, April 5, 2008.

Sherman, M. L., D. D. Barnum, A. Buhman-Wiggs, E. Nyberg. 2009. Clinical intake of child and adolescent consumers in a rural community mental health center: does wait-time predict attendance? *Community Ment. Health J.* **45**(1): 78–84.

Von Neumann, J., O. Morgenstern, 1953. *Theory of Games and Economic Behavior*, 3rd edn. Princeton University Press, Princeton, NJ.

Wang, W.-Y., D. Gupta. 2011. Adaptive appointment systems with patient preferences. *Manuf. Serv. Oper. Manag.* **13**(3): 373–389.

Werner, R., A. Canamucio, S. Marcus, C. Terwiesch. 2012. The relationship between primary care access and ER use. Working paper, University of Pennsylvania.

Wooldridge, J. M. 2007. Inverse probability weighted estimation for general missing data problems. *J. Econom.* **141**(2): 1281–1301.

Please Cite this article in press as: Osadchiy, N., D. KC. Are Patients Patient? The Role of Time to Appointment in Patient Flow. *Production and Operations Management* (2016), doi 10.1111/poms.12659