

Managing Outpatient Service with Strategic Walk-ins

Citation for published version (APA):

Liu, N., van Jaarsveld, W. L., Wang, S., & Xiao, G. (2023). Managing Outpatient Service with Strategic Walk-ins. *Management Science*, 69(10), 5904-5922. <https://doi.org/10.1287/mnsc.2023.4676>

Document license:

TAVERNE

DOI:

[10.1287/mnsc.2023.4676](https://doi.org/10.1287/mnsc.2023.4676)

Document status and date:

Published: 01/10/2023

Document Version:

Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.tue.nl/taverne

Take down policy

If you believe that this document breaches copyright please contact us at:

openaccess@tue.nl

providing details and we will investigate your claim.



Management Science

Publication details, including instructions for authors and subscription information:
<http://pubsonline.informs.org>

Managing Outpatient Service with Strategic Walk-ins

Nan Liu, Willem van Jaarsveld, Shan Wang, Guanlian Xiao

To cite this article:

Nan Liu, Willem van Jaarsveld, Shan Wang, Guanlian Xiao (2023) Managing Outpatient Service with Strategic Walk-ins. Management Science 69(10):5904-5922. <https://doi.org/10.1287/mnsc.2023.4676>

Full terms and conditions of use: <https://pubsonline.informs.org/Publications/Librarians-Portal/PubsOnLine-Terms-and-Conditions>

This article may be used only for the purposes of research, teaching, and/or private study. Commercial use or systematic downloading (by robots or other automatic processes) is prohibited without explicit Publisher approval, unless otherwise noted. For more information, contact permissions@informs.org.

The Publisher does not warrant or guarantee the article's accuracy, completeness, merchantability, fitness for a particular purpose, or non-infringement. Descriptions of, or references to, products or publications, or inclusion of an advertisement in this article, neither constitutes nor implies a guarantee, endorsement, or support of claims made of that product, publication, or service.

Copyright © 2023, INFORMS

Please scroll down for article—it is on subsequent pages



With 12,500 members from nearly 90 countries, INFORMS is the largest international association of operations research (O.R.) and analytics professionals and students. INFORMS provides unique networking and learning opportunities for individual professionals, and organizations of all types and sizes, to better understand and use O.R. and analytics tools and methods to transform strategic visions and achieve better outcomes.





For more information on INFORMS, its publications, membership, or meetings visit <http://www.informs.org>

Managing Outpatient Service with Strategic Walk-ins

Nan Liu,^a Willem van Jaarsveld,^b Shan Wang,^{c,*} Guanlian Xiao^d

^aCarroll School of Management, Boston College, Chestnut Hill, Massachusetts 02467; ^bDepartment of Industrial Engineering & Innovation Sciences, Eindhoven University of Technology, 5600 MB Eindhoven, Netherlands; ^cSchool of Business, Sun Yat-sen University, Guangzhou 510275, China; ^dHaskayne School of Business, University of Calgary, Calgary, Alberta T2N 1N4, Canada

*Corresponding author

Contact: nan.liu@bc.edu,  <https://orcid.org/0000-0001-7644-7341> (NL); W.L.v.Jaarsveld@tue.nl,  <https://orcid.org/0000-0003-3620-4067> (WvJ); wangsh337@mail.sysu.edu.cn,  <https://orcid.org/0000-0002-4625-7720> (SW); guanlian.xiao@ucalgary.ca,  <https://orcid.org/0000-0003-3904-0096> (GX)

Received: August 5, 2021

Revised: April 29, 2022; June 11, 2022

Accepted: June 29, 2022

Published Online in Articles in Advance:
February 6, 2023

<https://doi.org/10.1287/mnsc.2023.4676>

Copyright: © 2023 INFORMS

Abstract. Outpatient care providers usually allow patients to access service via scheduling appointments or direct walk-in. Patients choose strategically between these two access channels (and otherwise balking) based on the trade-off of appointment delay and in-clinic waiting. How to manage outpatient care with such dual access channels, taking into account patient strategic choice behavior, is a challenge faced by providers. We study three operational levers to address this management challenge: service capacity allocation between these two channels, appointment delay information revelation via the choice and design of online scheduling systems, and a walk-in triage system that restricts the use of walk-in hours only for acute care. By studying a stylized queueing model, we find that neither a real-time online scheduling system (which offers instant access to appointment delay information at time of booking) nor an asynchronous online system (which does not directly provide delay information) can be universally more efficient. Although real-time systems appear more popular in practice, asynchronous systems sometimes can result in higher operational efficiency. Under the provider's optimal capacity allocation, which scheduling system is more efficient hinges on two key factors: the patient demand–provider capacity relationship and patient willingness to wait. For the walk-in triage system, we find that it may or may not be beneficial to adopt; the provider's own cost trade-off between lost demand and overtime work is the key determinant. Our research highlights that there is no one-size-fits-all model for outpatient care management, and the best use of operational levers critically depends on the practice environment.

History: Accepted by Jayashankar Swaminathan, operations management.

Funding: S. Wang's work was supported in part by the National Natural Science Foundation of China [Grants 72001220 and 71931008].

Supplemental Material: The online appendix is available at <https://doi.org/10.1287/mnsc.2023.4676>.

Keywords: customer strategic behavior • appointment scheduling • walk-ins • queueing models

1. Introduction

Appointment scheduling is a common tool for providers to manage demand in outpatient care services, for example, health counseling, dentistry, pediatrics, and primary care. In addition to serving patients with scheduled appointments, outpatient care providers often set aside some time to see walk-in patients (*walk-ins* hereafter), who arrive without making an appointment in advance. Facing these two channels to access outpatient care (i.e., appointments and walk-ins), patients make choices based on their health conditions and the utilities of these two options (Liu et al. 2017). If a patient develops an acute symptom (e.g., high fever), then the patient would probably choose to walk in for care. However, if a patient has a less acute symptom (e.g., runny nose), then the patient can wait for a few days and would compare these two options. If the patient chooses to schedule an appointment, the patient

endures appointment delay (i.e., waiting for some days between the appointment request and the actual appointment date), but at the scheduled visit, the patient is likely to be served promptly. In contrast, if the patient chooses to walk in, the patient can see the provider on the same day, but may have to spend some time waiting in the clinic. In other words, patients with less acute conditions face a trade-off of waiting in two different time scales, that is, appointment delay versus in-clinic waiting.

To effectively manage these dual channels to access outpatient care, providers have a few operational levers at hand. The first one is capacity allocation. Given a fixed daily capacity, the provider needs to decide, respectively, the number of service slots allocated for scheduled and walk-in patients. We call them “appointment hours” and “walk-in hours.” Opening more appointment hours attracts a higher level of demand to schedule appointments

but may result in the provider working overtime to serve walk-ins. Allocating more capacity for walk-ins can reduce provider overtime but may lead to lost demand for appointments (because of long appointment delay). Effective management of patient demand requires a fine balance in such capacity allocation.

An appointment scheduling system is the portal through which patient demand fills provider capacity. In the demand filling process, appointment delay information plays a vital role in inducing patient choice. Thus, the second operational lever of the provider is to control how appointment delay information is revealed in the appointment scheduling process. When patients choose between appointments and walk-ins, they usually do not know the exact in-clinic wait time they would experience if they were to walk in. However, depending on the appointment scheduling process, patients may or may not know the exact appointment delay right away when requesting appointments.

There is a wide spectrum of appointment scheduling systems, varying in how appointment (delay) information is conveyed to patients. In this paper, we focus on online/web-based scheduling, which becomes increasingly popular nowadays given the rise of health information technology and widespread use of smart devices. There are two grand types of web-based scheduling systems: real-time (synchronous) and asynchronous systems. Real-time systems, such as *zocdoc.com*, directly provide available appointments for patients to choose from, and hence, patients know exact appointment delay when booking appointments. Asynchronous systems, however, do not provide exact delay information directly. They require patients to submit appointment requests first through emails or electronic forms and then appointments are confirmed by the provider later. One example is “Patient Gateway,” the online portal of Mass General Brigham (see Online Appendix A for its user interface). This online portal first solicits patient preferences on days of the week and times of the day, and then the provider confirms an appointment with patients via email or telephone.

How appointment delay information is revealed affects how it is perceived by patients and ultimately their choices of care options. In a real-time system, patients can easily opt to other nonappointment alternatives after learning the appointment delay provided instantaneously. In an asynchronous system, because patients do not know the exact appointment delay right away, they make choices between requesting an appointment and other care options based on expected delay, that is, their beliefs of the system congestion. Knowing that the delay information is not readily available but still reaching out to the provider and trying to make an appointment, we pose that patients tend to stick with this care option. First, if indeed the patient decides not to take the appointment, to reduce

the potential loss of goodwill from the provider the patient needs to communicate with the provider and make an explanation (otherwise, it is considered quite rude). Second, the patient knows that it is unlikely to get a same-day appointment at the time of requesting an appointment, so the patient could have already made up a plan for the day. Third, a provider may further increase the chance that patients stick with their appointment choices (after learning the appointment information) in an asynchronous system by managerial interventions that aim to improve patient adherence to provider recommendations. The “stickiness” of patients with the appointments made based on expected delay can be quite useful for providers to manage patient demand and choice. Indeed, a large number of healthcare organizations adopt scheduling systems that do not directly offer delay information to patients. According to a recent health informatics survey (Zhao et al. 2017), 9 out of the 21 web-based appointment scheduling systems being reviewed are asynchronous, including many large reputable healthcare organizations, such as MD Anderson and Geisinger.

Both capacity allocation and appointment delay revelation are useful tools for providers to “passively” manage patient demand. An active demand management approach is to limit the use of walk-in hours only for acute care, thereby eliminating (or at least mitigating) the potentially negative impact resulting from patient strategic choice. This can be done by announcing strict walk-in policies or putting a triage system in place to guide patients with less acute symptoms to schedule appointments. For instance, a large Boston-based pediatric practice makes the following notice regarding walk-in hours: “This walk-in hour should only be used if your child has an acute health problem such as sore throat, ear pain, or fever... this walk-in hour should not be used for chronic health concerns...” (Centre Pediatrics 2021). Whereas such a strict walk-in policy exerts better control of walk-in hours, it may lead to lost demand because of restrictive access.

Our research is motivated by these operations management issues faced by outpatient care providers in running practices with dual access channels. In particular, we seek to understand how a provider could best use these operational levers—capacity allocation, appointment delay information revelation (via the choice and design of online booking systems), and the use of a walk-in triage system—to match service capacity with patient demand. We also want to know the impact of the practice environment on the use of these levers, that is, when to use what and how.

To answer our research questions, we develop a model to study optimal capacity allocation in a service system that strategic customers¹ can choose to access via scheduled appointments or walk in based on the trade-off between appointment delay versus in-clinic

waiting. In particular, we consider a single service provider who needs to decide, respectively, how many appointment and walk-in hours to allocate from a fixed total daily capacity. For scheduled patients, we model the appointment book as a single-server queue, in which the appointment hours reserved daily is the service rate. The provider faces two independent patient demand streams. The first stream has acute symptoms and chooses to walk in without exceptions—called *exogenous walk-ins*. The second one has less acute symptoms and makes a choice strategically—called *strategic patients*.

To capture patient strategic choice when interacting with the wide spectrum of online appointment systems discussed, we consider a general two-stage sequential decision-making process. In the first stage, a strategic patient, upon arrival, thinks about whether to engage in appointment booking or not based on the patient's belief of the congestion of the system. The patient can choose to interact with the appointment booking system (e.g., open the scheduling app) or walk in or balk (without engaging with the appointment system at all). If the patient engages in appointment booking at the first stage, the patient enters the second stage in which the patient acquires the exact appointment delay. The patient may choose to stick with the appointment choice or switch to walk in or balk instead (after observing the exact delay information). If the latter two options are chosen, we assume that the patient incurs a cost, called the *disengagement cost*, because the patient is disengaged from the original plan. The disengagement cost can be small or large. In a synchronous system, this cost is literally zero because patients get the delay information in real time and can easily opt to other nonappointment alternatives. However, as discussed, disengaging from the appointment choice after interacting with an asynchronous system is costly to the patient because it can lead to inconvenience to the patient and/or loss of goodwill from the provider. (Alternatively, the disengagement cost may be viewed as a “refundable” appointment information acquisition cost, which patients need to pay in order to acquire the appointment information but is refunded if they hold on to the appointment option and do not deviate.) Because the disengagement cost depends on appointment system (design) and stipulates how likely patients are to stick with their appointment choices made in the first stage, it is the media through which we study how different appointment systems and their associated ways of delay information revelation affect patient choice.

Patient choice is endogenized to provider capacity allocation because the utility of making an appointment (respectively, walking in) closely depends on the congestion during appointment hours (walk-in hours). The provider incurs two types of costs: (1) lost demand costs if patients balk and (2) overtime costs

that depend on the workload during walk-in hours. The provider seeks to minimize the expected total daily costs by allocating the right amount of appointment and walk-in hours, respectively, in anticipation of patient strategic choice.

We show that, for any given disengagement cost and provider capacity allocation, there exists an equilibrium in this queueing model. It is quite challenging to establish such an equilibrium because of the two-stage patient decision process involved (see more details in Section 3). With the disengagement cost, our model provides a unified framework to capture the patient choice process in a wide range of online scheduling systems. Specifically, when the disengagement cost is zero, the model is equivalent to one in which patients know the exact appointment delay at the first stage; this is like the real-time system, and we call it the *observable* setting. When the disengagement cost is very large, the model behaves as if patients make their decisions solely based on expected appointment delay (because they would not revoke decisions made at the first stage); we call this the *unobservable* setting, which is a stylized model to resemble the asynchronous scheduling system. For convenience, we use the terms *observable* (*unobservable*) setting and *real-time* (*asynchronous*) scheduling system interchangeably.

For tractability and interpretability, we focus capacity analysis on the observable and unobservable settings. A comparison between these two settings under the optimal capacity allocation reveals that neither a real-time system nor an asynchronous one dominates in terms of operational efficiency. This finding confirms the potential value of both types of systems and, in particular, highlights that of asynchronous systems. Although real-time systems appear more popular in practice (Zhao et al. 2017), we show that asynchronous systems sometimes can be a better choice.

Furthermore, the comparison informs two key practice environmental factors that decide which type of scheduling systems performs better, namely, the demand-capacity relationship and patient willingness to wait. When the provider's capacity falls significantly short compared with demand, then the provider should make delay information easily accessible by patients to attract as many of them as possible and avoid lost demand. When the provider's capacity is at the same order of patient demand, then it depends on patient sensitivity to in-clinic waiting. Revealing exact delay works better when patients are less sensitive to in-clinic waiting; otherwise, if patients are more sensitive (i.e., running walk-in hours to attract patients is costly), not directly offering delay information can be more efficient. We do not advocate the use of asynchronous online systems to purposefully “hide” appointment delay information from patients, but rather to highlight the potential value of such systems as a scheduling approach that encourages patients to stick with

their appointment choices (and not to walk in or balk) once they decide to engage in appointment booking.

For the use of a triage system, we find that the provider's own cost trade-off plays a deciding role here. Intuitively, a triage system "protects" walk-in hours from being overly crowded but may lead to more balking. It turns out that a triage system is preferred when the provider has a relatively high overtime cost and a relatively low lost demand cost. This result is consistent under both the observable and unobservable settings.

The remainder of this paper is organized as follows. Section 2 reviews the relevant literature and summarizes our contribution. Section 3 introduces our general modeling framework and analyzes the equilibrium. Section 4 studies the provider's optimal capacity decision under the observable and unobservable settings. Section 5 compares various models, including the triage system. Section 6 discusses managerial insights from our analytical results. All proofs can be found in the online appendix.

2. Literature Review

Our research draws upon several streams of literature, and we review each stream.

2.1. Healthcare Operations Management (OM)

Our work is closely related to the healthcare OM literature on outpatient appointment scheduling; see, for example, recent literature reviews by Gupta and Denton (2008) and Ahmadi-Javid et al. (2017). A large volume of this literature studies how to time and sequence patient arrivals in order to optimize operational efficiency; some recent studies model walk-ins as exogenous random events (Wang et al. 2020, Zacharias and Yunes 2020). Concurrently, a rising stream of works use queueing models to investigate appointment system design questions, such as panel size selection and capacity decisions (Green and Savin 2008, Liu and Ziya 2014, Liu 2016, Zacharias and Armony 2016). Departing from this broad body of literature on appointment scheduling, we consider endogenized walk-in behavior and focus on strategic-level capacity decisions in a healthcare system with dual access channels.

Three recent studies are particularly relevant and noteworthy. Dobson et al. (2011) study capacity allocation in a primary care practice facing two exogenous demand streams: urgent and routine patients. Our work significantly differs from theirs in that patient demand is endogenized in our model. Similar to Dobson et al. (2011), Tunçalp et al. (2020) consider a capacity allocation problem, but they assume that two streams of patients have different delay cost rates and are strategic. Patients choose between appointments and walking in, whereas those who choose to walk in

are not guaranteed to be served and may be forced to balk. By contrast, patients in our model have balking as a third option from which to strategically choose in the first place, and all walk-ins are served. Bavafa et al. (2019) investigate a setting in which patient demand is influenced by a physician via selection of a revisit frequency consistent with patient preferences. They investigate the impact of various reimbursement schemes on patient panel size, physician earnings, and overall patient health. At a high level, both Bavafa et al. (2019) and our work study how to manage endogenized patient demand in outpatient care. The research questions, however, are fundamentally different.

2.2. Service OM

The service OM literature studies how to manage walk-in customers in settings such as restaurants, hotels, and rental firms; see, for example, Gans and Savin (2007), Alexandrov and Lariviere (2012), Cil and Lariviere (2013), and Oh and Su (2018). In these business settings, customers do not face the trade-offs of waiting in two different time scales as in our modeling context.

In our model, customers have dual channels to access services. In this sense, our work relates to the literature on omni-channel retailing. In retailing, inventory and price are often the decision variables of interest. By contrast, we investigate service processes with entirely different management levers. Departing from the traditional omni-channel retailing, Baron et al. (2022) study a service firm running omni-channel, in which customers first decide whether to order online or on-site, and if the latter, then customers arrive on site and decide whether to wait or balk. In our setting, patients also make decisions in two sequential stages, but they face three choices in both stages and we consider different system design questions.

2.3. Queueing Studies with Strategic Customers

From the methodological point of view, our work draws upon queueing studies with strategic customers. Starting with the seminal work by Naor (1969), extensive literature considers customer join and balk behavior in queues and how to optimize system efficiency/social welfare by controlling service capacity, pricing, or setting priority schemes; see, for example, Chen and Frank (2004), Ata and Shneorson (2006), Debo et al. (2008), and Anand et al. (2011). However, to the best of our knowledge, optimizing service rates in an observable queue in which customers choose between join or balk based on the exact delay upon their arrival remains largely unexplored. Our study fills this important gap in the literature.

As our work compares the observable and unobservable settings, it is important to discuss queueing studies that consider the impact of delay information

on customer strategic behavior and system performance. Hassin (2016) and Ibrahim (2018) provide excellent reviews on this topic. An important finding in this literature is that the value of delay information provision is usually context-dependent, and it can be either beneficial or detrimental to the system performance. Recently, customer behavior on delay information purchase and provider decision on delay announcement have received much attention (Allon et al. 2011, Hassin and Roet-Green 2017, Hu et al. 2017, Yu et al. 2017). These studies usually compare the use of different delay information when service capacity is fixed. However, our investigation is under the premise that the provider can optimize the capacity decision at the same time.

To summarize our contributions, we consider a service system with dual access channels (appointments and walk-ins) facing two customer streams: urgent walk-ins and those who strategically choose between access channels based on the trade-off of waiting in two different time scales, that is, appointment delay and in-clinic waiting. Customers make choices in two subsequent stages, in which exact delay is not known in the first stage but becomes available in the second stage, and disengaging from decisions made in the first stage is costly to customers. This model captures the customer strategic choice process in a variety of online appointment scheduling systems, which vary in how appointment (delay) information is provided. We prove the existence of customer equilibrium in this general model and study optimal capacity allocation in two extreme cases, namely, the observable and unobservable settings. The comparison of these two settings sheds light on the use of delay information under optimal service capacity allocation. It reveals that neither a real-time scheduling system (which provides delay information instantaneously at time of booking) nor an asynchronous one (which does not do so) is universally more efficient. Which system is better hinges on the demand–capacity relationship and customer willingness to wait. In addition, we study the use of a triage system, which reserves the walk-in channel only for those who have urgent needs. We find that the provider's own cost trade-off between lost demand and overtime work plays a critical role in determining whether a triage system is more efficient than a system that allows customers to make strategic choices freely. Our research highlights that there is no one-size-fits-all model for outpatient care management and informs how best to use different operational levers depending on the practice environment.

3. The Model

3.1. Capacity and Demand Model

We consider a single outpatient care provider who offers two channels for patients to access care: scheduled

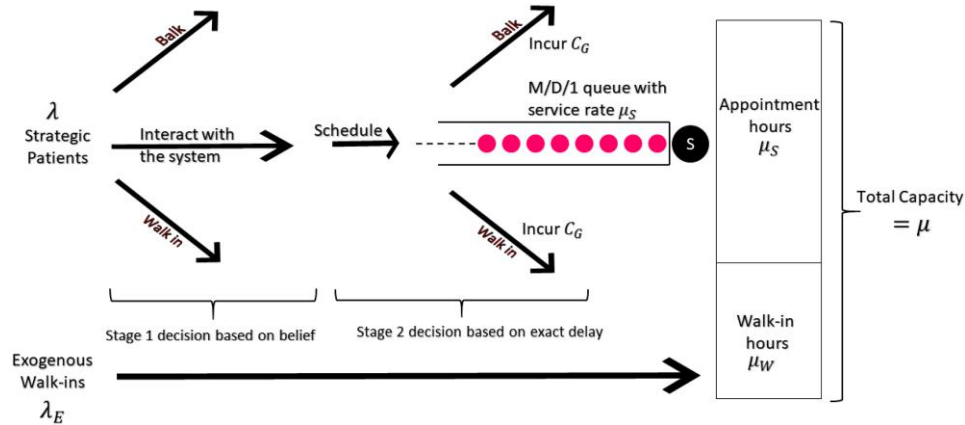
appointments and direct walk-in. The provider has a fixed total daily capacity of μ service slots to allocate between appointment and walk-in hours; all slots are of equal length, for example, 20 minutes. Specifically, the provider needs to decide the number of scheduled appointment slots (denoted by μ_S) and the number of slots reserved for walk-ins (denoted by μ_W) for each day such that $\mu_S + \mu_W = \mu$. We assume that each patient visit (scheduled appointment or walk-in) consumes one slot. It is common for practices to carve out certain time in a day (e.g., late afternoon) to serve only walk-ins (see Online Appendix B for several practical examples). Thus, we assume that walk-in and appointment hours are two disjoint blocks of time in a day.

Facing these two channels, patients make their choices based on their health conditions and the utilities of these two options. We assume that there are two types of patients based on health conditions. The first type has acute symptoms (e.g., high fever), which cannot be delayed, and they choose to walk in without exceptions; we call this type exogenous walk-ins, whose arrival behavior is not influenced by the provider's capacity decisions. The average number of exogenous walk-ins per day is λ_E . All walk-ins are served on the same day of visit, possibly during overtime. The second type of patients has less acute symptoms (e.g., runny nose), which can be delayed, and they make a choice strategically; we call them strategic patients. We assume that strategic patients arrive following a Poisson process with daily rate λ . The arrivals of exogenous and strategic patients are independent.

A strategic patient makes decisions in two sequential stages. At the first stage, based on the revealed information about the system, the patient decides whether to interact with the scheduling system and, if not, chooses to walk in or balk (e.g., seek care elsewhere). If, however, the patient decides to interact with the scheduling system and book an appointment, then the patient moves onto the second stage of decision making and obtains the exact appointment delay information. At this stage, the patient can choose to stick with the appointment choice made in the first stage or to revoke that decision and to walk in or balk. If the patient chooses the appointment, the patient endures appointment delay only; at the scheduled visit, the patient does not have to wait in the clinic. If the patient chooses to walk in, the patient can see the provider on the same day but likely has to spend time waiting in the clinic. Figure 1 shows the patient decision process and our model with additional details to be discussed.

We assume that interacting with the scheduling system and requesting an appointment incur no cost to patients because, in practice, either real-time or asynchronous systems are fairly easy to use: patients just need to make a few clicks on their devices. However,

Figure 1. (Color online) Patient Decision Making and Model Schematic



if a strategic patient first decides to interact with the scheduling system and then switches to walk in or balk instead (after obtaining the appointment delay information), we assume that the patient incurs a cost C_G , which we refer to as the disengagement cost, because the patient is disengaged from the original plan. The disengagement cost can be small or large. In a real-time system, such as ZocDoc, this cost is literally zero because patients get the appointment delay information instantaneously once interacting with the scheduling system and without causing any trouble to the provider; the patient can freely opt to other nonappointment alternatives if the patient wants. In an asynchronous online system, however, the disengagement cost is more substantial. In such systems, patients need to wait for appointment confirmation after making the request. The time gap in between is relatively short (say a few hours) but not trivial. It is this time gap that makes appointment delay information not instantaneously observable to patients and creates frictions potentially holding patients from turning down the appointment option later. In fact, deciding to request an appointment without exact delay information at hand, the patient has a strong tendency to stick with this choice; the patient knows that it is likely to get an appointment sometime later in the week, so the patient could have already made a plan for the day, whereas disengaging from this plan (e.g., choosing to walk in today) most likely disrupts the patient's life. In addition, after the provider spends effort locating an appointment for the patient, it would be trying for the patient to disengage. If the patient indeed decides not to take the appointment, to reduce potential loss of goodwill from the provider the patient probably needs to reply to the confirmation email and make an explanation. The patient cannot game the system as in a real-time one by quickly observing the appointment delay and then switching to nonappointment options without

bothering the provider. In any case, disengaging from the appointment choice made in an asynchronous system is undesirable from the perspective of patients. The disengagement cost captures such an effect. (Later, we discuss managerial interventions that can be used to influence/increase the disengagement cost.)

To model the utility of each choice, we first need to operationalize the appointment scheduling process. Inspired by the previous literature that uses stylized single-server queueing models to study strategic-level questions in appointment scheduling (Green and Savin 2008, Liu and Ziya 2014, Liu 2016, Zacharias and Armony 2016), we use an M/D/1 queue to capture the evolution of the scheduling process. Here, the queue represents the appointment queue (i.e., the virtual list of scheduled appointments yet to be served by the provider), but not the actual waiting line in clinic. Upon a patient's request of an appointment, the patient is scheduled to the end of the queue (i.e., added to the appointment backlog). Recall that the provider reserves exactly μ_S slots in a day for scheduled patients. Thus, in this stylized queue, each patient spends $1/\mu_S$ day with the server. For our purpose of modeling appointment delay, it suffices to consider deterministic service times because the provider sees a deterministic number of scheduled patients every day.

Our single-server queueing model abstracts certain operational details away from practice. Assuming a first-come, first-served order is equivalent to assuming that patients are offered and they also accept the earliest available appointment slot. It is possible that strategic patients with more urgent needs are provided earlier appointments if available. Our M/D/1 formulation also implicitly assumes that all patients with appointments show up and arrive on time. The purpose of the queueing model is to capture the overall effect of strategic demand on the appointment queue/delay. This allows us to model how a provider's capacity decision influences patient strategic choice. These stylized

assumptions render our model Markovian and tractable and still ensure that it captures the critical features in the system relevant to our research questions.

During the walk-in hours, both exogenous walk-ins and strategic patients who choose to walk in (called strategic walk-ins) come for service. Recall that the provider reserves μ_W slots for walk-ins. If too many walk-ins arrive, the provider works overtime to serve all walk-ins. Recent empirical studies lend support to our modeling assumptions that strategic patients may choose other care options when seeing a long appointment delay, whereas providers are committed to serving exogenous walk-ins even with overtime (Bavafa et al. 2021). Patients may have a range of care options other than appointments and walk-ins, such as going to urgent care centers or emergency rooms or seeing alternative providers. These other options are encapsulated in the balking option in our model, and one can adjust the service reward to reflect the utility gap between seeing the provider in person and the balking option (more on modeling details subsequently).

3.2. Patient Strategy and Utility

At the first stage, a strategic patient makes a decision based on the expected utility at the second stage, and at the second stage, after interacting with the scheduling system, the patient observes the exact appointment delay \tilde{d} , which is a random variable and influenced by the strategy adopted by all other patients. If the patient chooses to walk in (regardless in which stage), the patient does not know the exact in-clinic wait time but can form a belief on the expected in-clinic wait time, denoted by \bar{w} , which is also influenced by other patients' strategy.

We now first analyze patient utilities for any given strategy; in Section 3.3, we define and identify an equilibrium strategy. Because strategic patients are ex ante homogeneous, we identify this equilibrium in the class of mixed, symmetric strategies. Accordingly, denote any mixed strategy by $[p_S^1, p_W^1, p_B^1; p_S^2(\tilde{d}), p_W^2(\tilde{d}), p_B^2(\tilde{d}) | \forall \tilde{d}]$, where p_S^1, p_W^1 , and p_B^1 , respectively, denote the probabilities of interacting with the scheduling system, walking in, and balking at the first stage. For any patient who observes a delay \tilde{d} after choosing to interact with the scheduling system at the first stage, denote the second stage probabilities of scheduling, walking in, and balking by $p_S^2(\tilde{d}), p_W^2(\tilde{d})$, and $p_B^2(\tilde{d})$, respectively. Note that $p_S^1 + p_W^1 + p_B^1 = 1$ and $p_S^2(\tilde{d}) + p_W^2(\tilde{d}) + p_B^2(\tilde{d}) = 1, \forall \tilde{d}$ because these three options are exhaustive.

We analyze patient utility backward, starting with the second stage. Let $R > 0$ represent the reward from receiving the outpatient care service and $C_D > 0$ be the delay cost per day. After interacting with the scheduling system and observing the exact delay \tilde{d} , the utility of

scheduling an appointment at the second stage, denoted by u_S^2 , is

$$u_S^2 = R - C_D \tilde{d}. \quad (1)$$

Note that \tilde{d} is the pure delay measured by the appointment queue (in days). Because each actual service slot is of equal length (e.g., 20 minutes), the utility gained as a result of the time spent by a patient in consulting the provider is a constant (independent of μ_S) and can be conveniently included in R . Let $C_W > 0$ denote the in-clinic waiting cost per unit time and recall that C_G is the disengagement cost. Then, the utility of choosing to walk in at the second stage, denoted by u_W^2 , is

$$u_W^2 = R - C_W \bar{w} - C_G. \quad (2)$$

The utility of balking at the second stage is $u_B^2 = -C_G$. Next, we look at the first stage and consider a patient who chooses to interact with the scheduling system. After observing a delay \tilde{d} , the patient schedules an appointment, walks in, and balks with probabilities $p_S^2(\tilde{d}), p_W^2(\tilde{d})$, and $p_B^2(\tilde{d})$, respectively. Before observing \tilde{d} , the corresponding expectation can be formed as follows:

$$(p_S, p_W, p_B) = (\mathbb{E}_{\tilde{d}}[p_S^2(\tilde{d})], \mathbb{E}_{\tilde{d}}[p_W^2(\tilde{d})], \mathbb{E}_{\tilde{d}}[p_B^2(\tilde{d})]),$$

where, for instance, p_S denotes the probability that a patient chooses to schedule after choosing to interact with the scheduling system and before observing \tilde{d} . The utility of interacting with the scheduling system at the first stage, denoted by u_S^1 , is the expected utility the patient gains if proceeding to the second stage:

$$u_S^1 = p_S \mathbb{E}_{\tilde{d}}[u_S^2(\tilde{d}) | \text{schedule}] + p_W u_W^2 + p_B u_B^2, \quad (3)$$

where $u_S^2(\tilde{d})$ defined in (1) is written explicitly as a function of \tilde{d} and $\mathbb{E}_{\tilde{d}}[u_S^2(\tilde{d}) | \text{schedule}]$ is the expected utility of scheduling an appointment conditional on the patient choosing to join the appointment queue at the second stage. (We provide a more explicit form for $\mathbb{E}_{\tilde{d}}[u_S^2(\tilde{d}) | \text{schedule}]$ in Section 3.3.) The utility of walking in at the first stage, denoted by u_W^1 , is

$$u_W^1 = R - C_W \bar{w}. \quad (4)$$

Finally, the utility of balking at the first stage, denoted by u_B^1 , is normalized to be zero, that is, $u_B^1 = 0$.

Given that λ is fixed and to economize the notations, we use $(\lambda_S, \lambda_W, \lambda_B) = (p_S^1 \lambda, p_W^1 \lambda, p_B^1 \lambda)$ to represent patient strategy at the first stage. (Similar notations are used in the previous literature; see, e.g., Anand et al. 2011, Guo and Hassin 2011.) Given μ_S, μ_W , and the strategy $[\lambda_S, \lambda_W, \lambda_B; p_S^2(\tilde{d}), p_W^2(\tilde{d}), p_B^2(\tilde{d}) | \forall \tilde{d}]$ adopted by all patients, the total effective arrival rate to the appointment queue is $p_S \lambda_S$. To see this, the arrival rate of patients who interact with the scheduling system is λ_S , and at the second stage, the portion of patients who proceed to schedule appointments is $p_S = \mathbb{E}_{\tilde{d}}[p_S^2(\tilde{d})]$. Recall that \tilde{d} is

a random variable that represents the appointment delay observed by a strategic patient when interacting with the scheduling system. Because strategic patients arrive according to a Poisson process and Poisson arrivals see time averages, \tilde{d} has the same distribution as the steady-state distribution of delay in the appointment queue.

Next, consider the walk-in hours. Given μ_S , μ_W , and the strategy adopted by all patients, the average total number of walk-ins in a day is $\lambda_E + \lambda_W + \lambda_S p_W$. To see this, the arrival rate of strategic patients who choose to walk in at the first stage is λ_W ; at the second stage, the portion of patients who switch to walking in is $p_W = \mathbb{E}_{\tilde{d}}[p_W^2(\tilde{d})]$, and the exogenous walk-in rate is λ_E . It follows that the traffic intensity during the walk-in hours is $(\lambda_E + \lambda_W + p_W \lambda_S)/\mu_W$.

Patient in-clinic wait time depends on the evolution of the in-clinic wait line. Our analysis of patient choice only requires the information on expected in-clinic wait time. With the service time per customer fixed at one service slot, traffic intensity is usually sufficient to describe the expected wait time in a queueing system. So we do not assume any specific form for the queueing process of walk-ins, but rather, we use a reasonable generic function $w(\rho)$ to denote the expected in-clinic wait time, where ρ is the traffic intensity during walk-in hours.

Assumption 1. The expected in-clinic wait time $w(\rho)$ is a convex and strictly increasing function of ρ . Particularly, $w(0) = 0$.

This assumption simply states that (1) the expected in-clinic wait time increases with the congestion level during the walk-in hours and (2) the marginal increase is higher when the system is more congested. Then, \bar{w} , the expected in-clinic wait time in the system we consider is defined as

$$\bar{w} = w\left(\frac{\lambda_E + \lambda_W + p_W \lambda_S}{\mu_W}\right). \quad (5)$$

Substituting \bar{w} by (5) in (2) and (4), we obtain an explicit form for u_W^2 and u_W^1 , respectively.

3.3. Patient Equilibrium

Given μ_S and μ_W , a strategy $[\lambda_S, \lambda_W, \lambda_B; p_S^2(\tilde{d}), p_W^2(\tilde{d}), p_B^2(\tilde{d}) | \forall \tilde{d}]$ is a symmetric equilibrium strategy if it is the best response against itself. In particular, if we focus on one patient, assuming that other patients follow the symmetric equilibrium strategy, then the focal patient cannot increase expected utility by deviating from the strategy. A strategy $[\lambda_S, \lambda_W, \lambda_B; p_S^2(\tilde{d}), p_W^2(\tilde{d}), p_B^2(\tilde{d}) | \forall \tilde{d}]$ is a symmetric equilibrium if it satisfies the following conditions. Here, \tilde{d} follows the delay distribution of the appointment queue in which the arrival

rate is λ_S , service rate is μ_S , and the customers join the queue following the strategy $\{p_S^2(\tilde{d}) | \forall \tilde{d}\}$.

Condition 1. If $\lambda_S > 0$, then $u_S^1 = \max\{u_S^1, u_W^1, u_B^1\}$; if $\lambda_W > 0$, then $u_W^1 = \max\{u_S^1, u_W^1, u_B^1\}$; and if $\lambda_B > 0$, then $u_B^1 = \max\{u_S^1, u_W^1, u_B^1\}$.

Condition 2. For any observed \tilde{d} , if $p_S^2(\tilde{d}) > 0$, then $u_S^2(\tilde{d}) = \max\{u_S^2(\tilde{d}), u_W^2, u_B^2\}$; if $p_W^2(\tilde{d}) > 0$, then $u_W^2 = \max\{u_S^2(\tilde{d}), u_W^2, u_B^2\}$; and if $p_B^2(\tilde{d}) > 0$, then $u_B^2 = \max\{u_S^2(\tilde{d}), u_W^2, u_B^2\}$.

The main result of this section is the existence of a symmetric equilibrium strategy among all strategic patients. To establish this result, for any given first stage strategy $(\lambda_B, \lambda_W, \lambda_S)$, we first identify a second stage equilibrium strategy: a strategy for taking second stage decisions that satisfies Condition 2. We then identify the global equilibrium by taking into account how the second stage equilibrium strategy changes with the first stage strategy.

3.3.1. Second Stage Equilibrium. Given any first stage strategy $(\lambda_S, \lambda_W, \lambda_B)$, consider a patient who chooses to interact with the appointment system as the focal patient. This patient only schedules an appointment if scheduling has a sufficiently high utility, that is, if the patient sees a sufficiently short appointment delay; otherwise, the patient walks in or balks. This intuition is formalized as follows.

Lemma 1. For any first stage strategy $(\lambda_S, \lambda_W, \lambda_B)$ and capacity (μ_S, μ_W) , there exists a unique finite delay threshold δ such that, if $\tilde{d} \leq \delta$, $p_S^2(\tilde{d}) = 1$; otherwise, if $\tilde{d} > \delta$, $p_S^2(\tilde{d}) = 0$. Then, the triplet (δ, p_W, p_B) collectively defines the second stage equilibrium strategy.

Lemma 1 indicates that the probability of patients not scheduling an appointment is the probability that the observed delay \tilde{d} exceeds δ in the appointment queue. This result extends the classic one in Naor (1969) to the M/D/1 setting with a continuous delay threshold. The next result specifies the appointment queue when all customers join the queue following the threshold policy in Lemma 1 and some useful properties of its blocking probability.

Lemma 2. For any first stage strategy $(\lambda_S, \lambda_W, \lambda_B)$ and capacity (μ_S, μ_W) , the appointment queue is equivalent to an M/D/1 queue with λ_S as the arrival rate, μ_S as the service rate, and δ as the delay threshold. The steady-state blocking probability (i.e., the probability of delay exceeding δ), denoted by $\pi(\delta, \lambda_S, \mu_S)$, is continuous and strictly decreasing in δ .

Because the first stage strategy $(\lambda_S, \lambda_W, \lambda_B)$ and the capacity (μ_S, μ_W) are given, to economize notations, we write $\pi(\delta, \lambda_S, \mu_S)$ as $\pi(\delta)$ and $u_W^2(p_W, \lambda_S, \lambda_W, \mu_W)$ as

$u_W^2(p_W)$ whenever the context is clear. Then, the conditions for the second stage equilibrium (i.e., Condition 2) can be more explicitly expressed as follows:

$$\begin{cases} u_S^2(\tilde{d}) \geq \frac{p_W}{p_W + p_B} u_W^2(p_W) + \frac{p_B}{p_W + p_B} (-C_G), \forall \tilde{d} \leq \delta & (6) \\ u_S^2(\tilde{d}) < \frac{p_W}{p_W + p_B} u_W^2(p_W) + \frac{p_B}{p_W + p_B} (-C_G), \forall \tilde{d} > \delta & (7) \\ p_W + p_B = \pi(\delta) & (8) \\ p_W = 0, \text{ or, } u_W^2(p_W) \geq -C_G & (9) \\ p_B = 0, \text{ or, } u_W^2(p_W) \leq -C_G. & (10) \end{cases}$$

Condition (6) ensures that, when delay is no larger than δ , scheduling is no worse than walking in or balking. Condition (7) suggests that, when delay exceeds δ , scheduling is worse than walking in or balking. When delay exceeds δ , Condition (8) requires that patients either walk in or balk. Condition (9) says that either no patients walk in or walking in is no worse than balking. Finally, Condition (10) states that either no patients balk or balking is no worse than walking in.

Lemma 3 investigates the second stage equilibrium when we would exogenously exclude one of the options (scheduling, walking in, or balking) from the set of patient choices. The lemma is a stepping-stone for analyzing the general second stage equilibrium. It describes how patients choose if only facing two choices. For example, excluding the option of walking in, how would patients choose between scheduling and balking? The lemma prescribes the delay threshold beyond which a patient would balk, denoted by δ_B^S , and the corresponding balking probability, denoted by p_B^S . In these notations for patient strategic choices, we use superscripts and subscripts to represent the choices under consideration (scheduling and balking in the example).

Lemma 3. For any first stage strategy $(\lambda_S, \lambda_W, \lambda_B)$ and capacity (μ_S, μ_W) , consider the second stage problem with any given parameters (R, C_D, C_W, C_G) and in-clinic wait time function $w(\cdot)$.

1. If walking in is not an option for strategic patients, then there exists a unique δ_B^S such that patients schedule an appointment when the delay $\tilde{d} \leq \delta_B^S$ and balk when $\tilde{d} > \delta_B^S$, where $\delta_B^S = (R + C_G)/C_D$ so that $u_S^2(\delta_B^S) = u_B^2 = -C_G$; that is, patients are indifferent between scheduling an appointment and balking when the delay is δ_B^S . The proportion of strategic patients who balk is $p_B^S = \pi(\delta_B^S)$.

2. If scheduling is not an option for strategic patients, then there exists a unique $p_W^B \in [0, 1]$ such that patients walk in with probability p_W^B and balk with $1 - p_W^B$. Moreover, if $u_W^2(0) < -C_G$, $p_W^B = 0$; if $u_W^2(1) > -C_G$, $p_W^B = 1$; and otherwise, p_W^B solves $u_W^2(p_W^B) = u_B^2 = -C_G$; that is, walking in with probability p_W^B makes patients feel indifferent between walking in and balking.

3. If balking is not an option for strategic patients, then there exists a unique threshold δ_W^S such that patients schedule an appointment queue when delay $\tilde{d} \leq \delta_W^S$ and walk in when $\tilde{d} > \delta_W^S$. In particular, δ_W^S solves $u_S^2(\delta_W^S) = u_W^2(\pi(\delta_W^S))$; that is, patients feel indifferent between scheduling and walking in when delay is δ_W^S . The proportion of strategic patients who walk in is $p_W^S = \pi(\delta_W^S)$.

Lemma 3, parts 1 and 2, describe how patients choose between scheduling and balking and between walking in and balking, respectively. These two results are relatively straightforward because the utility of balking is always $-C_G$. However, the comparison between scheduling and walking in is more challenging because patient behavior influences the utilities of both options. The monotonicity and continuity of $\pi(\delta)$ established in Lemma 2 is critical for establishing Lemma 3, part 3.

Given these pairwise comparisons, we are ready to present the key results of this section in the following proposition. To avoid ambiguity, we stipulate that, in the case of a tie between scheduling and balking, the patient schedules an appointment.

Proposition 1. For any first stage strategy $(\lambda_S, \lambda_W, \lambda_B)$ and capacity (μ_S, μ_W) , consider the second stage problem with any given parameters (R, C_D, C_W, C_G) and in-clinic wait time function $w(\cdot)$. There exists a second stage equilibrium strategy (δ, p_W, p_B) that can be described as follows:

1. If $u_W^2(0) \leq -C_G$, then $\delta = \delta_B^S$, $p_W = 0$, and $p_B = p_B^S$.
2. If $u_W^2(p_W^S) < -C_G < u_W^2(0)$, then $\delta = \delta_B^S$, $p_W = p_W^B$, and $p_B = p_B^S - p_W^B$.
3. If $u_W^2(p_W^S) \geq -C_G$, then $\delta = \delta_W^S$, $p_W = p_W^S$, and $p_B = 0$, where δ_B^S , p_B^S , p_W^B , p_W^S , and δ_W^S are defined in Lemma 3.

Case 1, that is, $u_W^2(0) \leq -C_G$, implies a relatively large volume of exogenous walk-ins, which alone already makes the utility of walking in less than $-C_G$ (i.e., the utility of balking). Thus, strategic patients behave as if there were no walk-in option. We call this case *Regime SnB* (schedule and balk).

If $u_W^2(0) > -C_G$, then the utility of walking in remains positive even if some strategic patients choose to walk in. The question of how many strategic patients would walk in is then answered in the last two cases. Suppose that all strategic patients who see a delay δ_W^S walk in. If this makes walking in strictly worse than balking (i.e., $u_W^2(p_W^S) < -C_G$ in case 2), then only a subset of these patients choose walk in in equilibrium, and a subset choose to balk, implying that the utilities of walking in and balking both equal $-C_G$ in equilibrium. In other words, strategic patients who choose not to schedule have $-C_G$ utility, and hence, such patients schedule as if there were no walk-in option (see Lemma 3, part 1). The proportion of strategic patients walking in is such that it makes the utility of walking in $-C_G$ (see Lemma 3, part 2), and the rest balk. We call this case *Regime SWB* (schedule, walk in, and balk).

Finally, if admitting all strategic patients to the walk-in hours still allows walking in to be no worse than balking (i.e., $u_W^2(p_W^S) \geq -C_G$ in case 3), then balking is never appealing to strategic patients. Thus, strategic patients behave as if there were no balking option (see Lemma 3, part 3). We call this case *Regime SnW* (schedule and walk in). Table 1 in Online Appendix D summarizes patient equilibrium in these three regimes and the conditions under which these regimes take place.

3.3.2. The Global Equilibrium. In this section, we identify a first stage strategy $(\lambda_S, \lambda_W, \lambda_B)$, which, together with the second stage equilibrium analyzed in Section 3.3.1, forms a global equilibrium. The second stage equilibrium (δ, p_W, p_B) described in Proposition 1 depends on the first stage strategy $(\lambda_S, \lambda_W, \lambda_B)$ and can be summarized as follows. When the observed delay does not exceed δ , patients schedule an appointment; otherwise, they walk in or balk, and the proportions of patients who walk in and balk are p_W and p_B , respectively. Thus, we have

$$p_W u_W^2 + p_B u_B^2 = \pi(\delta) u_S^2(\delta), \quad (11)$$

which says that, at the second stage, the utility of scheduling an appointment when the observed delay is δ is indifferent compared with the expected utility of taking the other two choices in equilibrium. With (11), the expected utility of interacting with the scheduling system at the first stage, that is, u_S^1 defined in (3), can be rewritten as follows:

$$\begin{aligned} u_S^1 &= [1 - \pi(\delta, \lambda_S)] \mathbb{E}_{\tilde{d}}[u_S^2(\tilde{d}) | \tilde{d} \leq \delta, \lambda_S] + \pi(\delta, \lambda_S) u_S^2(\delta) \\ &= R - C_D [1 - \pi(\delta, \lambda_S)] \mathbb{E}_{\tilde{d}}[\tilde{d} | \tilde{d} \leq \delta, \lambda_S] - C_D \pi(\delta, \lambda_S) \delta. \end{aligned} \quad (12)$$

In (12), $\mathbb{E}_{\tilde{d}}[\tilde{d} | \tilde{d} \leq \delta, \lambda_S]$ corresponds to the term $\mathbb{E}_{\tilde{d}}[u_S^2(\tilde{d}) | \text{schedule}]$ in (3), denoting the expected delay seen by a patient who chooses to join the appointment queue based on the delay threshold δ when the arrival rate to the appointment queue is λ_S ; $\pi(\delta, \lambda_S)$ represents the probability of patients choosing not to schedule in such an appointment queue. In our notations, we highlight the dependence of $\mathbb{E}_{\tilde{d}}[\tilde{d} | \tilde{d} \leq \delta, \lambda_S]$ and $\pi(\delta, \lambda_S)$ on λ_S because λ_S is one of the patient's strategic choices at the first stage. The explicit form of (12) is helpful for our analysis of the equilibrium at the first stage.

Noticing that $u_W^1 - u_B^1 = u_W^2 - u_B^2$, if no strategic patients walk in at the second stage, then none walks in at the first stage, and if no strategic patients balk at the second stage, then none balks at the first stage (see Online Lemma D.2). This property enables us to simplify the analysis of the global equilibrium. To further illustrate, we define three possible scenarios for the equilibrium delay threshold δ at the second stage and the equilibrium scheduling rate λ_S at the first stage. We use superscripts a , b , and c to differentiate these

scenarios. The ultimate form of the global equilibrium depends on which scenario is realized given the model parameters.

Scenario a: $u_B^2 \geq u_W^2$. Define (δ^a, λ_S^a) , which jointly solve $u_S^2 = u_B^2$ and $u_S^1 = u_B^1$. That is, $\delta^a = (R + C_G)/C_D$, and $\lambda_S^a = \lambda_S$, which solves

$$\begin{aligned} u_S^1 &= R - C_D [1 - \pi(\delta^a, \lambda_S)] \mathbb{E}_{\tilde{d}}[\tilde{d} | \tilde{d} \leq \delta^a, \lambda_S] \\ &\quad - C_D \pi(\delta^a, \lambda_S) \delta^a = 0 = u_B^1. \end{aligned}$$

We show that u_S^1 here is decreasing in λ_S , and then, we can solve for a unique λ_S^a . Note that the resulting λ_S^a may be larger than λ . So, if the global equilibrium is realized in this scenario, $\lambda_S = \min\{\lambda_S^a, \lambda\}$.

Scenario b: $u_W^2 > u_B^2$ and $\lambda_S < \lambda$. Define (δ^b, λ_S^b) , which jointly solve $u_S^2 = u_W^2$ and $u_S^1 = u_W^1$. That is, $(\delta^b, \lambda_S^b) = (\delta, \lambda_S)$, which solves

$$C_D \delta = C_W w \left(\frac{\lambda_E + \lambda - \lambda_S + \pi(\delta, \lambda_S) \lambda_S}{\mu_W} \right) + C_G, \quad (13)$$

and

$$\begin{aligned} C_D [1 - \pi(\delta, \lambda_S)] \mathbb{E}_{\tilde{d}}[\tilde{d} | \tilde{d} \leq \delta, \lambda_S] + C_D \pi(\delta, \lambda_S) \delta \\ = C_W w \left(\frac{\lambda_E + \lambda - \lambda_S + \pi(\delta, \lambda_S) \lambda_S}{\mu_W} \right). \end{aligned} \quad (14)$$

We show that δ , as an implicit function of λ_S defined by (13), decreases in λ_S . Furthermore, we can show that, with δ being an implicit function of λ_S , $u_S^1 - u_W^1$ (i.e., the left-hand side of (14) minus the right-hand side of (14)) decreases in λ_S . Then, we have a unique pair of (λ_S^b, δ^b) . Note that the resulting λ_S^b may be larger than λ . If so, we need to limit λ_S to be λ and have the following scenario.

Scenario c: $u_W^2 > u_B^2$ and $\lambda_S = \lambda$. Then, we define δ^c , which solves $u_S^2 = u_W^2$ when $\lambda_S = \lambda$. That is, $\delta^c = \delta$, which solves

$$C_D \delta = C_W w \left(\frac{\lambda_E + \pi(\delta, \lambda) \lambda}{\mu_W} \right) + C_G.$$

This equation is the same as (13) with λ_S replaced by λ . Scenario c can be regarded as a special case of scenario b, and we can show that δ^c is also unique.

The uniqueness of δ^a , λ_S^a , δ^b , λ_S^b , and δ^c is crucial for establishing the existence of equilibrium. In particular, proving the uniqueness of λ_S^a , δ^b , and λ_S^b leverages a novel use of the conditional workload process observed in the system. This proof can be of theoretical interest in its own right (see details in Online Appendix D.2). Now, we are ready to present the main result of this section. Let $(\lambda_S, \lambda_W, \lambda_B; \delta, p_W, p_B)$ denote the equilibrium strategy in the system, where $(\lambda_S, \lambda_W, \lambda_B)$ and (δ, p_W, p_B) represent the patient equilibrium strategy at the first and second stages, respectively.

Theorem 1. For any given (μ_S, μ_W) , there exists an equilibrium strategy $(\lambda_S, \lambda_W, \lambda_B; \delta, p_W, p_B)$ that takes the following form in the model we consider.

1. If $R - C_W w\left(\frac{\lambda_E}{\mu_W}\right) \leq 0$, then there exists a unique equilibrium such that $\lambda_S = \min(\lambda, \lambda_S^a)$, $\lambda_W = 0$, $\lambda_B = \lambda - \lambda_S$, $\delta = \delta^a$, $p_W = 0$, $p_B = \pi(\delta^a, \lambda_S)$.
2. If $R - C_W w\left(\frac{\lambda_E}{\mu_W}\right) > 0$ and $\delta^a < \max(\delta^b, \delta^c)$, then there exist multiple equilibria such that $\lambda_S = \min(\lambda_S^a, \lambda)$, λ_W and p_W jointly solve $R - C_W w\left(\frac{\lambda_E + \lambda_W + p_W \lambda_S}{\mu_W}\right) = 0$, $\lambda_B = \lambda - \lambda_S - \lambda_W$, $\delta = \delta^a$, $p_B = \pi(\delta^a, \lambda_S) - p_W$. In this case, λ_S and δ are unique, whereas λ_W , λ_B , p_W , and p_B may take multiple values in equilibrium; however, the total rate of strategic walk-ins (i.e., $\lambda_W + p_W \lambda_S$) and the total rate of balking (i.e., $\lambda_B + p_B \lambda_S$) are unique.
3. If $R - C_W w\left(\frac{\lambda_E}{\mu_W}\right) > 0$ and $\delta^a \geq \max(\delta^b, \delta^c)$, then there exists a unique equilibrium such that $\lambda_S = \min(\lambda, \lambda_S^b)$, $\lambda_W = \lambda - \lambda_S$, $\lambda_B = 0$, $\delta = \max(\delta^b, \delta^c)$, $p_W = \pi(\delta, \lambda_S)$, $p_B = 0$.

Theorem 1 describes the global equilibrium in our model with two sequential stages of strategic decision making. When $R - C_W w\left(\frac{\lambda_E}{\mu_W}\right) \leq 0$, walking in is surely worse than balking, and thus, strategic patients either schedule or balk; we draw our attention to scenario a earlier. If $\lambda_S^a > \lambda$, then $u_S^1 > u_B^1$ even when all strategic patients choose to interact with the scheduling system (because, when δ is fixed, u_S^1 defined in (12) is shown to be decreasing in λ_S). In this case, no one balks at the first stage, and everyone chooses to interact with the scheduling system, that is, $\lambda_S = \lambda$. However, if $\lambda_S^a \leq \lambda$, then $\lambda_S = \lambda_S^a$ and the rest of the strategic patients balk at the first stage so that $u_S^1 = u_B^1 = 0$.

If $R - C_W w\left(\frac{\lambda_E}{\mu_W}\right) > 0$, there must be some strategic patients walking in (otherwise walking in would be strictly better than balking). Thus, the second stage equilibrium can be SnW or SWB. When $\delta^a < \max(\delta^b, \delta^c)$, the delay threshold at the second stage must be δ^a because a strategic patient would not walk in but balk when seeing a delay level of $\max(\delta^b, \delta^c)$. This suggests that the second stage equilibrium is SWB. Scenario a is realized such that λ_S and δ in equilibrium are $\min(\lambda_S^a, \lambda)$ and δ^a , respectively. For the walk-in option, patients can use a small walk-in rate at the first stage and a large walk-in rate at the second stage or vice versa as long as the total walk-in rate is fixed and makes no difference between the utilities of walking in and balking (at both stages).

When $\delta^a \geq \max(\delta^b, \delta^c)$, the equilibrium delay threshold at the second stage must be $\max(\delta^b, \delta^c)$ because a strategic patient who chooses to walk in when observing this delay level still has a positive utility. Thus, the second stage equilibrium is SnW, and hence, the global equilibrium occurs either in scenario b or c. As discussed, if $\delta^b < \delta^c$, then $\lambda_S^b > \lambda$, indicating that $(\delta, \lambda_S) = (\delta^c, \lambda)$ in equilibrium. In this case, scenario c is realized.

Everyone chooses to interact with the scheduling system, and those who do not choose to schedule an appointment walk in. If $\delta^b \geq \delta^c$, then $\lambda_S^b \leq \lambda$, indicating that $(\delta, \lambda_S) = (\delta^b, \lambda_S^b)$ in equilibrium. Here, scenario b is realized: strategic patients mix between interacting with the scheduling system and walking in at the first stage; then, they mix again between scheduling an appointment and walking in at the second stage; none balks in either stage.

3.4. The Provider's Problem

Facing patient strategic behavior, the provider aims to minimize the expected total daily costs by choosing μ_S and μ_W such that $\mu_S + \mu_W = \mu$, where μ is the fixed total daily capacity. The provider's daily costs include two components: a lost demand cost at rate C_L per balking patient and overtime cost incurred at rate C_O per unit time. We use a general function $o(\cdot)$ to denote the expected overtime. If the walk-in queue approaches steady state at the end of walk-in hours, then the overtime is the wait time experienced by a walk-in who would have arrived at the end of walk-in hours. Therefore, the expected overtime would share similar structural properties of the expected wait time. Following this argument, we make the following assumption.

Assumption 2. The expected overtime function $o(\rho)$ is a convex and strictly increasing function of ρ , the traffic intensity during the walk-in hours.

This assumption implies two conditions on the expected overtime: (1) it increases with the congestion of the walk-in hours, and (2) the marginal increase is higher when the system is more congested. This assumption is based on the premise that the provider keeps the same service rate. It makes our analysis of the provider's problem, which is formulated as follows, cleaner and tractable:

$$\begin{aligned} \min_{\mu_S, \mu_W \geq 0} \quad & C_L(\lambda_B + p_B \lambda_S) + C_O o\left(\frac{\lambda_W + \lambda_E + p_W \lambda_S}{\mu_W}\right) \\ & \mu_S + \mu_W = \mu, \\ & \lambda_S, \lambda_W, \lambda_B, p_W, p_B \text{ are defined in Theorem 1.} \end{aligned} \quad (\text{GP})$$

Whereas there may exist multiple equilibria, the total rates of walk-ins and balking are unique for a given set of model parameters (Theorem 1), so the optimization problem (GP) is well-defined.

3.5. Two Extreme Cases

Solving Problem (GP) requires specifying the equilibrium under any given (μ_S, μ_W) . To make our analysis more interpretable and tractable, we focus on two extreme cases: $C_G = 0$ and $C_G = +\infty$.

4. Optimal Capacity Allocation for the Provider

4.1. The Case with $C_G = 0$

When the disengagement cost $C_G = 0$, strategic patients incur no cost in revoking the decision made at

the first stage. All strategic patients choose to interact with the scheduling system and behave as if there were no first stage decisions.

Lemma 4. *If $C_G = 0$, then $\lambda_S = \lambda$ in equilibrium.*

When $C_G = 0$, the appointment system is equivalent to one in which the appointment delay is always known to patients before they make a choice. We call this setting the observable setting. In this setting, a strategic patient follows the threshold-based joining strategy described in Proposition 1. If the delay is sufficiently short, the patient chooses to join the appointment queue; otherwise, the patient mixes between walking in and balking. The form of the equilibrium strategy described in Theorem 1 can be simplified as (δ, p_W, p_B) , which we elaborate in Online Appendix D.3.

Though we can prove some structural properties of the blocking probability $\pi(\delta, \mu_S)$ in Lemma 2, it is very challenging if not impossible to analyze capacity optimization without its specific form. To get a closed-form expression of $\pi(\delta, \mu_S)$, we slightly modify the original queue in the following way. Suppose that an arriving strategic patient sees $k \geq 0$ patients waiting in the queue (excluding the one in service) and one patient is currently being served. Then, the patient calculates the appointment delay to be $(k + \tilde{u})/\mu_S$ days, where \tilde{u} is a continuous uniform random variable in $[0, 1]$. (A patient who sees nobody in the queue and nobody in service calculates the delay as zero.) We include \tilde{u} for two reasons. First, it is a stylized construct to capture the arriving patient's belief on the remaining service time of the patient currently being served. Second, as discussed next, this leads to an elegant model for the appointment queue and gives rise to a closed-form blocking probability.

Lemma 1 indicates that the probability of patients not choosing to schedule an appointment is the probability that the delay \tilde{d} exceeds δ in the appointment queue. Given that patients use this threshold-based joining strategy, the appointment queue behaves like an M/D/1 queue with a finite buffer. However, because the delay depends on the queue length plus a random term \tilde{u} , the buffer in our setting is not necessarily an integer but can be any nonnegative real number. With a slight abuse of notation, we use R to denote the buffer size. Inspired by Hassin and Haviv (1997), we call the resulting appointment queue an M/D/1/R queue in which $R \in (0, +\infty)$.

Our M/D/1/R queue behaves as follows. Let $\lceil x \rceil$ be the smallest integer that is greater than or equal to x and $p = R + 1 - \lceil R \rceil$. Then, customers join the queue if the system size (i.e., the total number of customers in the system, including those who are waiting and in service) is shorter than $\lceil R \rceil$, join with probability p and balk with probability $1 - p$ if the system size is $\lceil R \rceil$.

The next result specifies our appointment queue if all strategic patients adopt the strategy in Lemma 1 and some useful properties of its blocking probability.

Lemma 5. *In the equilibrium with δ as the delay threshold, the appointment queue is equivalent to an M/D/1/ $\delta\mu_S$ queue. The steady-state blocking probability (i.e., the probability of delay exceeding δ), denoted by $\pi(\delta, \mu_S)$, has a closed-form expression and is continuous and strictly decreasing in μ_S with a fixed δ .*

The parameter μ_S appears in the buffer size as it converts the delay threshold δ to the corresponding queue length. Analyzing this queue relies on an embedded Markov chain approach by observing the system right after each customer departure. We defer details to Online Appendix C and only present in Lemma 5 the results relevant to our discussion here. First, the closed-form expression (which can be found in Online Appendix C) provides us a way to analyze the optimal capacity allocation problem. Second, the monotonicity of $\pi(\delta, \mu_S)$ is used in the analysis that follows.

4.1.1. Impact of μ_S and μ_W on Patient Equilibrium Behavior.

Recall that there are three equilibrium regimes in the second stage (see Online Table 1). Before analyzing the provider's optimal capacity allocation, it is important to understand which regime patient equilibrium may fall into for given (μ_S, μ_W) . This section addresses this question. Let $\underline{\mu}_W$ be such that

$$R - C_W w \left(\frac{\lambda_E}{\underline{\mu}_W} \right) = 0. \quad (15)$$

Then, if $\mu_W < \underline{\mu}_W$, the utility of walking in would be strictly negative even without any strategic walk-ins, and any strategic patient who chooses not to schedule would balk. Let $\bar{\mu}_W$ be such that

$$R - C_W w \left(\frac{\lambda + \lambda_E}{\bar{\mu}_W} \right) = 0. \quad (16)$$

Then, if $\mu_W > \bar{\mu}_W$, the utility of walking in would be strictly positive even if all strategic patients chose to walk in; thus, any strategic patient who chooses not to schedule would walk in.

For $\mu_W \in [\underline{\mu}_W, \bar{\mu}_W]$, consider the following equation that involves μ_S and μ_W :

$$R - C_W w \left(\frac{\pi(\delta_B^S, \mu_S) \lambda + \lambda_E}{\mu_W} \right) = 0, \quad (17)$$

where $\delta_B^S = R/C_D$ is a constant. Thus, Equation (17) implicitly defines μ_S as a function of μ_W . We write this function as $\bar{\mu}_S(\mu_W)$. Given the walk-in hours μ_W , if the provider sets the appointment hours to be $\bar{\mu}_S(\mu_W)$ or longer, then even if all strategic patients who choose

not to schedule attend the walk-in hours, the utility of walking in is still nonnegative. Thus, strategic patients either schedule or walk in, and no one balks. However, if the provider sets the appointment hours $\mu_S < \bar{\mu}_S(\mu_W)$, then $\pi(\delta_B^S, \mu_S) > \pi(\delta_B^S, \bar{\mu}_S(\mu_W))$ (see Lemma 5). Thus, not all strategic patients who choose not to make an appointment walk in (because the utility of walking in would become negative if all of them did walk in); instead, these patients mix between walking in and balking in equilibrium.

This intuition can be formalized into the following proposition, which quantifies the impact of μ_S and μ_W on equilibrium regimes. Figure 2(a) illustrates this proposition.

Proposition 2. Consider the observable setting with any given set of parameters $(\lambda, \lambda_E, \mu_S, \mu_W, R, C_D, C_W)$ and in-clinic wait time function $w(\cdot)$:

1. If $\mu_W \leq \underline{\mu}_W$, the equilibrium is in regime SnB.
2. If $\underline{\mu}_W < \mu_W < \bar{\mu}_W$, there exists a decreasing function $\bar{\mu}_S(\mu_W)$ such that
 - a. when $\mu_S < \bar{\mu}_S(\mu_W)$, the equilibrium is in regime SWB.
 - b. when $\mu_S \geq \bar{\mu}_S(\mu_W)$, the equilibrium is in regime SnW.
3. If $\mu_W \geq \bar{\mu}_W$, the equilibrium is in regime SnW.

4.1.2. Optimal Capacity Allocation. Now, we can formulate the provider's problem (GP) in a more explicit form:

$$\begin{aligned} \min_{\mu_S, \mu_W \geq 0} \quad & C_L p_B \lambda + C_{O0} \left(\frac{p_W \lambda + \lambda_E}{\mu_W} \right) \quad (\text{Ob.P}) \\ \text{subject to:} \quad & \mu_S + \mu_W = \mu, \\ & (p_B, p_W) = \begin{cases} (p_B^S, 0) & \text{if } \mu_W \leq \underline{\mu}_W, \\ (p_B^S - p_W^B, p_W^B) & \text{if } \underline{\mu}_W < \mu_W < \bar{\mu}_W, \\ (0, p_W^S) & \text{if } \mu_W \geq \bar{\mu}_W, \end{cases} \\ & \mu_S \leq \bar{\mu}_S(\mu_W), \\ & \mu_S \geq \bar{\mu}_S(\mu_W); \\ & p_B^S, p_W^B, p_W^S \text{ are defined in Lemma 3} \\ & \text{with } \lambda_S = \lambda \text{ and } C_G = 0; \\ & \underline{\mu}_W, \bar{\mu}_S(\mu_W) \text{ are defined in (15) and (17),} \\ & \text{respectively.} \end{aligned}$$

Problem (Ob.P) can be analyzed by studying the optimization problem in each regime and comparing the outcomes of three regimes. The detailed analysis can be found in Online Appendix D.4. The next proposition discusses how patient equilibrium strategy under the optimal capacity allocation changes in the total daily capacity μ .

Proposition 3. In the observable setting, there exists $\bar{\mu}_O \geq 0$ such that SnW is optimal when $\mu \geq \bar{\mu}_O$, SnB or SWB is optimal when $\underline{\mu}_W \leq \mu < \bar{\mu}_O$, and SnB is optimal when $\mu < \underline{\mu}_W$.

In Proposition 3, $\bar{\mu}_O$ represents the minimum total daily capacity required to attract all strategic patients

to either schedule an appointment or walk in. If $\mu \geq \bar{\mu}_O$, there exists an optimal capacity allocation such that no one balks. However, if $\mu < \underline{\mu}_W$, it is impossible to attract strategic patients to walk in. Finally, if μ is in between $\underline{\mu}_W$ and $\bar{\mu}_O$, which regime—SnB versus SWB—is better depends on the trade-off between overtime cost and lost demand cost. Figure 2(a) illustrates the results.

4.2. The Case with $C_G = +\infty$

When the disengagement cost $C_G = \infty$, strategic patients behave as if there were no second stage decisions because they would join the appointment queue and not walk in or balk once they choose to interact with the scheduling system.

Lemma 6. If $C_G = +\infty$, then $\delta = +\infty$ and $p_W = p_B = 0$ in equilibrium.

When $C_G = +\infty$, the system is equivalent to one in which strategic patients make their choices solely based on expected appointment delay. We call this setting the unobservable setting. In such a setting, strategic patients mix among scheduling, walking in, and balking; the form of patient equilibrium strategy in Theorem 1 can be simplified as the triplet $(\lambda_S, \lambda_W, \lambda_B)$. Given $(\lambda_S, \lambda_W, \lambda_B)$, the appointment queue becomes an $M/D/1$ queue with λ_S as the arrival rate and μ_S as the service rate. The expected utility of choosing scheduling an appointment is

$$u_S^1(\lambda_S) = R - C_D \frac{\lambda_S}{2\mu_S(\mu_S - \lambda_S)}, \quad (18)$$

where the last fraction term is the expected delay in an $M/D/1$ queue. The expected utility of walking in is

$$u_W^1(\lambda_W) = R - C_W w \left(\frac{\lambda_W + \lambda_E}{\mu_W} \right). \quad (19)$$

We can show that the equilibrium strategy $(\lambda_S, \lambda_W, \lambda_B)$ in the unobservable setting is unique because no strategic patients walk in at the second stage; that is, $p_W = 0$. Depending on whether $\lambda_W = 0$ or $\lambda_B = 0$, we divide the equilibrium into three regimes. We adopt the same nomenclature as before and call these three regimes SnB, SWB, and SnW, respectively. More details on the equilibrium can be found in Online Appendix D.5.

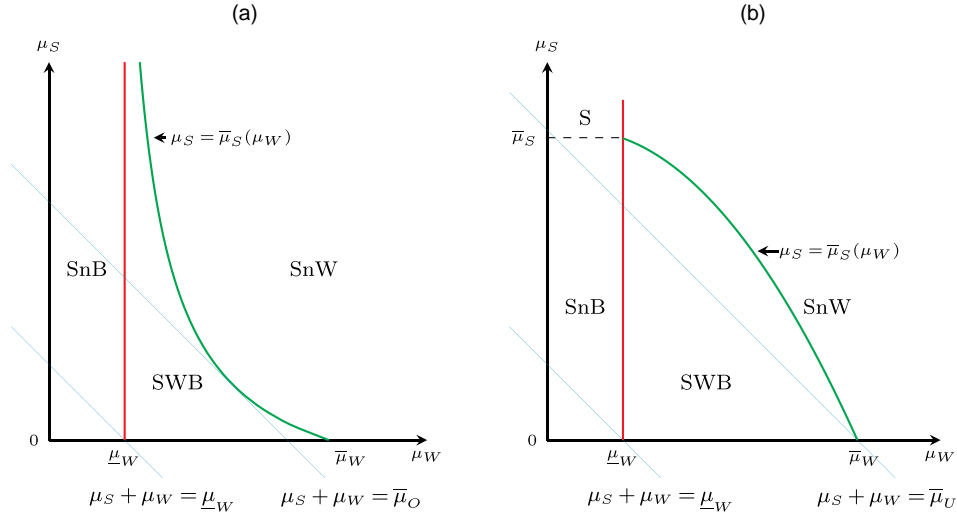
4.2.1. Impact of μ_S and μ_W on Patient Equilibrium Behavior.

Recall $\underline{\mu}_W$ and $\bar{\mu}_W$ defined in (15) and (16), respectively. If $\mu_W < \underline{\mu}_W$, all strategic patients choosing not to schedule would balk. If $\mu_W > \bar{\mu}_W$, all strategic patients who choose not to schedule would walk in. Let $\bar{\mu}_S$ be such that

$$R - C_D \frac{\lambda}{2\bar{\mu}_S(\bar{\mu}_S - \lambda)} = 0.$$

If $\mu_S > \bar{\mu}_S$, the utility of scheduling is strictly positive even if all strategic patients choose to schedule. In this case, no strategic patients would balk regardless of μ_W .

Figure 2. (Color online) Equilibrium Regimes and Capacity Allocation



Notes. (a) Observable setting. (b) Unobservable setting.

For $\mu_W \in [\underline{\mu}_W, \bar{\mu}_W]$, consider the following equation that involves μ_S and μ_W , where λ_W^S is defined in Online Lemma D.3 and is a function of both μ_S and μ_W :

$$R - C_W w \left(\frac{\lambda_W^S + \lambda_E}{\mu_W} \right) = 0. \quad (20)$$

One can verify that (20) defines μ_S as a function of μ_W . With a slight abuse of notation, we write this function as $\bar{\mu}_S(\mu_W)$ as before. For a fixed μ_W , if $\mu_S > \bar{\mu}_S(\mu_W)$, then the utility of walking in becomes nonnegative even if all strategic patients who choose not to schedule walk in. Thus, no strategic patients balk; they either schedule or walk in. If $\mu_S < \bar{\mu}_S(\mu_W)$, then some of the patients who choose not to schedule would balk because, otherwise, if they all choose to walk in then the utility of walking in would be negative. Specifically, we have the following sensitivity results on patient equilibrium with respect to changes of μ_S and μ_W . Figure 2(b) illustrates the results.

Proposition 4. Consider the unobservable setting with any given set of parameters $(\lambda, \lambda_E, \mu_S, \mu_W, R, C_D, C_W)$ and in-clinic wait time function $w(\cdot)$.

1. If $\mu_W \leq \underline{\mu}_W$, the equilibrium is in regime SnB.
2. If $\underline{\mu}_W < \mu_W < \bar{\mu}_W$, there exists a decreasing function $\bar{\mu}_S(\mu_W)$ such that
 - a. if $\mu_S < \bar{\mu}_S(\mu_W)$, the equilibrium is in regime SWB.
 - b. if $\mu_S \geq \bar{\mu}_S(\mu_W)$, the equilibrium is in regime SnW.
3. If $\mu_W \geq \bar{\mu}_W$, the equilibrium is in regime SnW.

Remark 1. If $\mu_W \leq \underline{\mu}_W$ and $\mu_S \geq \bar{\mu}_S$, then all strategic patients choose to schedule.

The unobservable setting shares some similarities with the observable setting in the equilibrium results, but a few key differences are noteworthy. In the observable setting, if μ_W is not sufficiently large, that

is, smaller than $\bar{\mu}_W$, then there are always some patients choosing to balk when they see a long appointment queue. However, in the unobservable setting, regardless of μ_W , as long as μ_S is sufficiently large, that is, larger than $\bar{\mu}_S$, no strategic patients balk because long appointment hours make the expected appointment queue length sufficiently short to induce all patients to come for service. This suggests that the unobservable setting is more “attractive” to strategic patients than the observable setting when strategic demand is relatively low compared with capacity; we come back to this point when comparing both settings in Section 5.

4.2.2. Optimal Capacity Allocation. We next analyze the problem faced by the service provider in the unobservable setting. For any given capacity decision (μ_S, μ_W) , strategic patients respond with a mixed equilibrium strategy $(\lambda_S, \lambda_W, \lambda_B)$ in the unobservable setting. Following Proposition 4, the provider’s problem (GP) can be explicitly formulated as follows:

$$\min_{\mu_S, \mu_W \geq 0} C_L \lambda_B + C_O \left(\frac{\lambda_W + \lambda_E}{\mu_W} \right) \quad (\text{Un.P})$$

subject to : $\mu_S + \mu_W = \mu$,

$$(\lambda_W, \lambda_B) = \begin{cases} (0, \lambda_B^S) & \text{if } \mu_W \leq \underline{\mu}_W, \\ (\lambda_W^B, \lambda_B^S - \lambda_W^B) & \text{if } \mu_W \geq \underline{\mu}_W, \\ & \mu_S \leq \bar{\mu}_S(\mu_W), \\ (\lambda_W^S, 0) & \text{if } \mu_W \geq \underline{\mu}_W, \\ & \mu_S \geq \bar{\mu}_S(\mu_W); \end{cases}$$

$\lambda_B^S, \lambda_W^B, \lambda_W^S$ are defined by Lemma D.3

in Appendix D.5;

$\underline{\mu}_W, \bar{\mu}_S(\mu_W)$ are defined in (15) and (20), respectively.

Note that λ_B^S , λ_W^B , and λ_W^S collectively define the equilibrium strategy for given (μ_S, μ_W) in the unobservable setting. Their specifics are given in Online Appendix D.5. Regarding interpretation, for instance, λ_W^B is the equilibrium walk-in rate of strategy patients if they only have the options of walking in and balking. Detailed analysis of Problem (Un.P) is deferred to Online Appendix D.6. The following proposition summarizes the impact of total daily capacity on the optimal equilibrium regime.

Proposition 5. *In the unobservable setting, there exists $\bar{\mu}_U > 0$ such that SnW is optimal when $\mu \geq \bar{\mu}_U$, SWB or SnB is optimal when $\underline{\mu}_W \leq \mu < \bar{\mu}_U$, and SnB is optimal when $\mu < \underline{\mu}_W$.*

Proposition 5 offers insights similar to those in Proposition 3 for the observable setting. Figure 2(b) illustrates the results. We conclude this section with the following remark.

Remark 2. We can further show that strategic patients would never mix between scheduling an appointment and walking in under optimal capacity allocation in the unobservable setting. Together with Proposition 5, this suggests that, when the total capacity is sufficiently large, that is, $\mu \geq \bar{\mu}_U$, a “bang-bang” control is optimal; it is optimal to induce a pure strategy among strategic patients so that either all of them schedule or all of them walk in. More details are shown in Online Appendix D.7.

5. Model Comparison

After analyzing optimal capacity allocation in both the observable and unobservable settings, a natural question is which system design is more operationally efficient. This section starts by shedding some light on this question. Another approach to regulate (strategic) patient demand is to institute a triage system. We also investigate when one should consider adopting such a system.

5.1. Observable vs. Unobservable Setting

We focus on two scenarios in the comparison between the observable and unobservable settings. First, the provider falls short of daily capacity compared with the patient demand the provider is facing. Second, the provider’s daily capacity is somewhat “in balance” with the patient demand. These two scenarios are of the most interest to practice because, in reality, health-care providers usually do not have abundant capacity. The comparison results are summarized in the following theorem, which characterizes conditions under which one setting costs less than the other. A setting costing less means that its total cost is no more than the total cost of its counterpart.

Theorem 2. *Consider the observable and unobservable settings with the same model parameters $(\lambda, \lambda_E, R, C_D, C_W, C_L, C_O, \mu)$, in-clinic wait time function $w(\cdot)$, and overtime function $o(\cdot)$.*

1. *When μ is sufficiently small, the observable setting costs less.*
2. *When μ is sufficiently large and if there exists an $M > 0$ such that $\frac{\mu(\mu-\lambda)}{\lambda} \leq M$,*
 - a. *If $C_W \times w(1) \leq \min\{\frac{C_D}{2M}, R\}$, the observable setting costs less.*
 - b. *If $C_W \times w(\frac{1}{2}) > R$, the unobservable setting costs less.*

When the provider has limited daily capacity, we find that the observable setting costs less. In this case, revealing exact appointment delay information makes patients more “rational” and utilizes appointment hours more efficiently (as patients in the observable setting tend to schedule as long as the current appointment queue length is short enough). In an unobservable setting, however, patients may choose to balk because the perceived appointment delay is long. Simply put, the observable appointment queue attracts more strategic patients than the unobservable one in a capacity-constrained environment, making the system more cost-efficient. This finding complements those in the previous service OM literature, which does not consider capacity optimization or walking in as an option for wait-sensitive strategic customers; see, for example, Chen and Frank (2004).

When the provider’s daily capacity is sufficiently large but remains in the same magnitude of the provider’s patient demand, we find that which system costs less depends on patient willingness to wait in both time scales. If strategic patients are less sensitive to in-clinic waiting and/or more sensitive to appointment delay (i.e., case 2a), the optimal operating regime in the unobservable setting is to induce all strategic patients to walk in because this is more cost-efficient than to attract them to schedule appointments. Thus, in this case, the observable setting costs less because inducing all strategic patients to walk in is a feasible but not necessarily optimal choice for the provider. If strategic patients are more sensitive to in-clinic waiting (i.e., case 2b), it is more costly to run walk-in hours, and attracting strategic patients to make appointments becomes easier. In the unobservable setting, the provider can open enough appointment hours so that all strategic patients choose to schedule appointments and then use the rest of the capacity for exogenous walk-ins. This turns out to be more cost-efficient than the observable setting, in which the provider has to open additional costly walk-in hours in order not to lose strategic patients from balking.

Remark 3. This analysis assumes that the provider is fully committed to serving all walk-ins, potentially

with overtime work. Another possible way to operate walk-in hours is to turn away walk-ins if the clinic is overly busy. So, instead of paying overtime costs, the provider suffers from loss of revenues. To model this “lost sales” case, one needs to replace the second term (i.e., the overtime cost) in (GP) by a term representing the revenue loss and modify patient utility of choosing walking in accordingly. With these replacements, the comparison results between the two settings remain substantively the same as those in Theorem 2, suggesting that our insights are robust under different walk-in hour practice regimes. To keep the flow of the paper, we defer all relevant technical details to Online Theorem E.1.

5.2. When to Use a Triage System

In practice, a triage system may prioritize patients with urgent needs (i.e., exogenous walk-ins in our model), and strategic patients can still access walk-in hours but face longer in-clinic wait time. Our model can capture this by increasing the expected in-clinic wait time for strategic patients. In our comparison, we concentrate on the case when strategic patients are not allowed to use walk-in hours (Centre Pediatrics 2021), i.e., their in-clinic waiting cost is in effect infinite in our model. We call such a practice model, which limits the use of walk-in hours only for acute care as a triage model and the original model in which patients can freely choose a strategic model.

We focus our discussion of the triage model in the context of the observable setting because the analysis and high-level insights in the unobservable setting are similar. Under the observable setting, the optimal capacity allocation problem of the triage model can be formulated as follows:

$$\min_{\mu_S \in [0, \mu]} C_L \pi \left(\frac{R}{C_D}, \mu_S \right) \lambda + C_{OO} \left(\frac{\lambda_E}{\mu - \mu_S} \right). \quad (\text{T.Ob.P})$$

Note that, in the triage model, walk-in hours are not accessible by strategic patients, who only choose between scheduling an appointment and balking. When $\mu \leq \underline{\mu}_W$, the strategic model behaves the same as the triage model because no strategic patients walk in under any feasible capacity allocation in this case; see Figure 2(a). We focus on the situation when $\mu > \underline{\mu}_W$. We first define two constants, each representing a threshold value for the ratio between lost demand cost rate and overtime cost rate:

$$\bar{\alpha} = \frac{o' \left(\frac{\lambda_E}{\underline{\mu}_W} \right)}{\underline{\mu}_W} \cdot \max \left\{ \frac{\lambda_E}{\lambda \pi \left(\frac{R}{C_D}, \mu - \underline{\mu}_W \right)}, 1 \right\} \quad \text{and} \\ \underline{\alpha} = \frac{o' \left(\frac{\lambda_E}{\mu} \right)}{\mu},$$

where $o'(\cdot)$ is the first order derivative of $o(\cdot)$. Then, we have the following comparison results.

Proposition 6. *Given the same set of model parameters and supposing that $\mu > \underline{\mu}_W$,*

1. *If $\frac{C_L}{C_{OO}} \leq \underline{\alpha}$, the triage model costs less.*
2. *If $\frac{C_L}{C_{OO}} \geq \bar{\alpha}$, the strategic model costs less.*

Proposition 6 reveals that patient strategic behavior may benefit or hurt the provider, depending on the provider’s own cost trade-off. If the provider has a relatively high lost demand cost, offering patients free choice reduces balking and allows the provider to use capacity in a more efficient way. However, a high overtime cost makes it important to control patient demand that goes into the walk-in hours. In this case, the triage model, which limits strategic walk-ins, is preferred.

Remark 4. Under the unobservable setting, we obtain similar insights to Proposition 6 in terms of the impact resulting from the provider’s cost structure. A key difference in the unobservable setting is that the total capacity available to allocate also plays an important role: either a relatively high lost demand cost or a sufficiently large total capacity makes the triage model more favorable; however, to make the strategic model stand out, the overtime cost has to carry a sufficiently high weight and the capacity needs to be small enough. Details can be found in Online Appendix D.8.

6. Discussion and Conclusion

In this paper, we consider a single outpatient care provider who faces two independent patient demand streams: exogenous walk-ins and strategic patients. The provider has a fixed total daily capacity to allocate between appointment hours and walk-in hours. The provider incurs lost demand costs because of patient balking and overtime/rejection costs if walk-in hours are crowded. To minimize total daily costs, the provider has three operational levers, namely, capacity allocation, appointment delay information revelation, and triage system, at hand. We develop a stylized queueing model to shed light on how best to use these levers.

One interesting and unique feature of our model is that strategic patients have dual channels to access service, and they make choices in two subsequent stages with trade-offs between waiting in two different time scales. Our model provides a general framework to capture patient choice in a wide range of online appointment scheduling systems, which vary in how delay information is revealed. We pose that delay revelation affects the disengagement cost incurred to patients, which ultimately influences their choices of care access channels. A real-time scheduling system provides instant access to appointment delay information and has a literally zero disengagement cost. In contrast, patients cannot observe exact delay when they request an appointment in an asynchronous online system and receive detailed appointment

confirmation after a relatively short amount of time, say a few hours. Such a short and yet nontrivial time gap leads to a disengagement cost, which creates information frictions potentially holding patients from turning down the appointment option after acquiring exact appointment delay.

Indeed, the disengagement cost can be adjusted via managerial interventions or scheduling system design and, thus, can be used as a management tool. In particular, to further increase disengagement costs in asynchronous online systems (i.e., to make patients more likely to stick with their appointment choice made based on expected delay and not to revoke after acquiring exact delay), the provider can engage in strategies/interventions aiming to improve patient adherence to provider recommendations, for example, building a trusting relationship between patients and the provider, improving communications with patients, and accommodating patient-stated preferences in scheduling. One particularly useful idea is to add some “nudges” when communicating with patients in appointment confirmations, either via email or personal phone contact, to make patients hold on to their appointments. Nudges are subtle changes to the design of the environment (e.g., information provided or choice of languages) meant to influence behavior in a predictable way but without restricting choices of decision makers (Thaler and Sunstein 2009). Nudges are often used to steer the decision maker toward a desired outcome and are shown to be effective in demand management for outpatient care (Liu and KC 2020). These interventions make our unobservable queueing model (i.e., the theoretical model with an infinite disengagement cost) a more accurate representation for asynchronous online systems and our comparison between the unobservable and observable settings more practically relevant and meaningful.

Our model comparison shows that neither scheduling system (real time or asynchronous online system) can be universally more efficient than its counterpart. This finding confirms the potential value of both types of systems, and in particular, highlights that of asynchronous systems. Although real-time systems appear more popular in practice (Zhao et al. 2017), we show that asynchronous systems, leveraging information frictions, sometimes can result in higher operational efficiency.

Which type of scheduling systems is more efficient depends on two key practice environmental factors: the demand–capacity relationship and patient willingness to wait. When the provider falls significantly short of capacity, the provider is better off using a real-time system that provides patients with exact delay information upon their appointment requests. However, if the provider uses an asynchronous system in this situation, it is critical for the provider to boost service capacity. Otherwise, many patients may

choose not to come for service because of perceived long appointment delays.

A perhaps more interesting and practically relevant setting is when provider capacity and patient demand are more or less in balance. Which scheduling system is better depends on patient sensitivities to in-clinic waiting/appointment delay. Previous research reveals heterogeneity in patient sensitivity to waiting in different medical specialties; see, for example, Osadchiy and KC (2017). Built upon these empirical findings, our theoretical results can inform the choice and implementation of appointment scheduling systems by clinical contexts.

In particular, if patients are more sensitive to in-clinic waiting, an asynchronous system can be more efficient because, with sufficient appointment hours, the provider can attract all strategic patients to schedule appointments and need not use costly walk-in hours to retain them. If, however, a real-time scheduling system is in place, then it is important to carefully manage the patient waiting experience in the clinic. One such setting could be pediatric care, in which children tend to have shorter attention spans and get irritated more easily than adults. Otherwise, anticipating unpleasant in-clinic waiting, those patients who do not choose to make an appointment in the first place may choose not to walk in either but opt to other care options or even skip care.

When patients are less sensitive to in-clinic waiting or, equivalently, more sensitive to appointment delay, a real-time system is better. Using a large multispecialty data set, Osadchiy and KC (2017) find that general pathology and diabetes education are two specialties for which patients are most sensitive to appointment delay. Extrapolating from this empirical finding, our analysis suggests that a real-time scheduling system appears to be a better choice than an asynchronous one in practices, such as outpatient labs and health counseling/education services.

In addition to medical specialties, patient sensitivities to waiting are likely to be influenced by urgency for care. Intuitively, asynchronous systems with information frictions are more appealing to patients with less urgent conditions who can tolerate longer appointment delays. Whereas we model strategic patients as a homogeneous population in their urgency for care, we can use their relative sensitivities to waiting in two different time scales to infer the “average” urgency level of this population. Indeed, our analytical results lend support to such intuitions that asynchronous systems work better when the patient population is less urgent in general.

Finally, our analysis of the triage model suggests that such an active control may either benefit or hurt a practice. In particular, it benefits a provider when the provider cares more about overtime but lost demand

is less of a concern. This finding has some interesting implications for the implementation of telemedicine, which has been growing tremendously in the last decade. With telemedicine, patients may televisit the provider instead of balking, and hence, the lost demand cost decreases. At the same time, offering telemedicine may compete for the provider's already-tight capacity and, thus, increase overtime cost. Both trends in cost change consistently suggest that, in the era of telemedicine, a triage system can be an effective approach to manage operations in outpatient care.

Overall, our research affirms that there is no panacea for the management of outpatient care practice and informs how best to use different operational levers depending on the practice environment. Our study also reveals several avenues for future research. First, one could incorporate additional operational details, such as patient no-shows, priority appointment offering among strategic patients who may differ in their urgency, and implementation of telemedicine (see, e.g., Bavafa et al. 2018, Rajan et al. 2019). Second, our model assumes that the provider's service rate is constant. It would be interesting to consider provider response (e.g., speeding up) in capacity planning (KC and Terwiesch 2009). Third, in our current model, strategic patients would balk if delay is sufficiently long. This partially captures the fact that patients may heal by themselves after some time, but it would be interesting to embed patient health progression dynamics explicitly in the model (see, e.g., Bavafa et al. 2019). Finally, one may model telephone scheduling as another channel to make appointments in addition to the online scheduling considered here. Such an extension results in different system dynamics and leads to new research questions.

Acknowledgments

The authors are grateful to the department editor, the associate editor, and all the referees for a thoughtful and constructive review process. The authors also thank Diwakar Gupta, Qingxia Kong, Sergei Savin, and Guohua Wan for their generous support at the early stage of this work and Ivo Adan for his inspirational comments on our analysis.

Endnote

¹ We use "customer" and "patient" interchangeably in this paper.

References

Ahmadi-Javid A, Jalali Z, Klassen KJ (2017) Outpatient appointment systems in healthcare: A review of optimization studies. *Eur. J. Oper. Res.* 258(1):3–34.
 Alexandrov A, Lariviere MA (2012) Are reservations recommended? *Manufacturing Service Oper. Management* 14(2):218–230.
 Allon G, Bassamboo A, Gurvich I (2011) "We will be right with you": Managing customer expectations with vague promises and cheap talk. *Oper. Res.* 59(6):1382–1394.
 Anand KS, Fazil Paç M, Veeraraghavan S (2011) Quality-speed conundrum: Trade-offs in customer-intensive services. *Management Sci.* 57(1):40–56.

Ata B, Shneerson S (2006) Dynamic control of an M/M/1 service system with adjustable arrival and service rates. *Management Sci.* 52(11):1778–1791.
 Baron O, Chen X, Li Y (2022) Omnichannel services: The false premise and operational remedies. *Management Sci.*, ePub ahead of print April 26, <https://doi.org/10.1287/mnsc.2017.2741>.
 Bavafa H, Hitt LM, Terwiesch C (2018) The impact of e-visits on visit frequencies and patient health: Evidence from primary care. *Management Sci.* 64(12):5461–5480.
 Bavafa H, Savin S, Terwiesch C (2019) Managing patient panels with non-physician providers. *Production Oper. Management* 28(6):1577–1593.
 Bavafa H, Canamucio A, Marcus SC, Terwiesch C, Werner RM (2021) Capacity rationing in primary care: Provider availability shocks and channel diversion. *Management Sci.* 68(4):2842–2859.
 Centre Pediatrics (2021) Notice to our patients regarding walk-in hours. Accessed June 29, 2021, <https://www.centrepediatrics.org/visits/walk-in-hours/>.
 Chen H, Frank M (2004) Monopoly pricing when customers queue. *IEE Trans.* 36(6):569–581.
 Cil EB, Lariviere MA (2013) Saving seats for strategic customers. *Oper. Res.* 61(6):1321–1332.
 Debo LG, Toktay LB, Van Wassenhove LN (2008) Queuing for expert services. *Management Sci.* 54(8):1497–1512.
 Dobson G, Hasija S, Pinker EJ (2011) Reserving capacity for urgent patients in primary care. *Production Oper. Management* 20(3):456–473.
 Gans N, Savin S (2007) Pricing and capacity rationing for rentals with uncertain durations. *Management Sci.* 53(3):390–407.
 Green LV, Savin S (2008) Reducing delays for medical appointments: A queueing approach. *Oper. Res.* 56(6):1526–1538.
 Guo P, Hassin R (2011) Strategic behavior and social optimization in Markovian vacation queues. *Oper. Res.* 59(4):986–997.
 Gupta D, Denton B (2008) Appointment scheduling in healthcare: Challenges and opportunities. *IEE Trans.* 40(9):800–819.
 Hassin R (2016) *Rational Queueing* (CRC Press, Boca Raton, FL).
 Hassin R, Haviv M (1997) Equilibrium threshold strategies: The case of queues with priorities. *Oper. Res.* 45(6):966–973.
 Hassin R, Roet-Green R (2017) The impact of inspection cost on equilibrium, revenue, and social welfare in a single-server queue. *Oper. Res.* 65(3):804–820.
 Hu M, Li Y, Wang J (2017) Efficient ignorance: Information heterogeneity in a queue. *Management Sci.* 64(6):2650–2671.
 Ibrahim R (2018) Sharing delay information in service systems: A literature survey. *Queueing Systems* 89(1):49–79.
 KC DS, Terwiesch C (2009) Impact of workload on service time and patient safety: An econometric analysis of hospital operations. *Management Sci.* 55(9):1486–1498.
 Liu J, KC D (2020) Nudging patient choice: Evidence from a field experiment. Working paper, Emory University, Atlanta.
 Liu N (2016) Optimal choice for appointment scheduling window under patient no-show behavior. *Production Oper. Management* 25(1):128–142.
 Liu N, Ziya S (2014) Panel size and overbooking decisions for appointment-based services under patient no-shows. *Production Oper. Management* 23(12):2209–2223.
 Liu N, Finkelstein SR, Kruk ME, Rosenthal D (2017) When waiting to see a doctor is less irritating: Understanding patient preferences and choice behavior in appointment scheduling. *Management Sci.* 64(5):1975–1996.
 Naor P (1969) The regulation of queue size by levying tolls. *Econometrica* 37(1):15–24.
 Oh J, Su X (2018) Reservation policies in queues: Advance deposits, spot prices, and capacity allocation. *Production Oper. Management* 27(4):680–695.
 Osadchiy N, KC D (2017) Are patients patient? The role of time to appointment in patient flow. *Production Oper. Management* 26(3):469–490.

- Rajan B, Tezcan T, Seidmann A (2019) Service systems with heterogeneous customers: Investigating the effect of telemedicine on chronic care. *Management Sci.* 65(3):1236–1267.
- Thaler RH, Sunstein CR (2009) *Nudge: Improving Decisions About Health, Wealth, and Happiness* (Yale University Press, New Haven, CT).
- Tunçalp F, Gunes ED, Ormeci L (2020) Modeling strategic walk-in patients in appointment systems: Equilibrium behavior and capacity allocation. Preprint, submitted October 26, <https://dx.doi.org/10.2139/ssrn.3687717>.
- Wang S, Liu N, Wan G (2020) Managing appointment-based services in the presence of walk-in customers. *Management Sci.* 66(2):667–686.
- Yu Q, Allon G, Bassamboo A, Iravani S (2017) Managing customer expectations and priorities in service systems. *Management Sci.* 64(8):3942–3970.
- Zacharias C, Armony M (2016) Joint panel sizing and appointment scheduling in outpatient care. *Management Sci.* 63(11):3978–3997.
- Zacharias C, Yunes T (2020) Multimodularity in the stochastic appointment scheduling problem with discrete arrival epochs. *Management Sci.* 66(2):744–763.
- Zhao P, Yoo I, Lavoie J, Lavoie BJ, Simoes E (2017) Web-based medical appointment systems: A systematic review. *J. Medical Internet Res.* 19(4):e134.