Managing Outpatient Service with Strategic Walk-ins

Nan Liu
Carroll School of Management, Boston College, Chestnut Hill, MA 02467, USA, nan.liu@bc.edu

Willem van Jaarsveld
Department of Industrial Engineering & Innovation Sciences, Eindhoven University of Technology, 5600 MB Eindhoven, The Netherlands, W.L.v.Jaarsveld@tue.nl

Shan Wang
School of Business, Sun Yat-sen University, Guangzhou, 510275, China, wangsh337@mail.sysu.edu.cn

Guanlian Xiao
Haskayne School of Business, University of Calgary, Calgary, AB T2N 1N4, Canada, guanlian.xiao@ucalgary.ca

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 which 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: 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 tradeoff 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.

Key words: 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, e.g., 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 (abbreviated as *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 he[1] would probably

[1] For convenience, we shall refer to a patient as "he" and the provider as "she" in the rest of this article.

choose to walk in for care. However, if a patient has a less acute symptom (e.g., runny nose), then he can wait for a few days and would compare these two options. If he chooses to schedule an appointment, he endures appointment delay (i.e., waiting for some days between the appointment request and the actual appointment date), but at his scheduled visit he is likely to be served promptly. In contrast, if he chooses to walk in, he 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, i.e., appointment delay vs. 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* for short. Opening more appointment hours attracts a higher level of demand to schedule appointments, but may result in 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 (due to long appointment delay). Effective management of patient demand requires a fine balance in such capacity allocation.

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 are 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 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 could easily opt to other non-appointment alternatives after learning the appointment delay provided instantaneously. In an asynchronous system, since 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, i.e., 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 he would need to communicate with her and make an explanation (otherwise it would be considered quite rude). Second, the patient knows that it is unlikely to get a same-day appointment at time of requesting an appointment, so he could have already made up his 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 which 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 due to 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). While such a strict walk-in policy exerts a better control of walk-in hours, it may lead to lost demand due to 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 walk-in triage system—to match service capacity with patient demand. We would also like to know the impact of practice environment on the use of these levers, i.e., 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 customers[2] can choose to access via scheduled appointments or walk-in based on the trade-off between appointment delay vs. in-clinic waiting. In particular, we consider a single service provider, who needs to decide, respectively, how many appointment hours 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, where the appointment hours reserved daily is the service rate. The provider faces two independent patient demand streams. The first stream has acute symptoms and will choose to walk in without exceptions—called *exogenous walk-ins*. The second one has less acute symptoms and will make a choice strategically—called *strategic patients*.

To capture patient strategic choice when interacting with the wide spectrum of online appointment systems discussed above, we consider a general two-stage sequential decision-making process. In the first stage, a strategic patient upon his arrival thinks about whether he should engage in appointment booking or not based on his belief of the congestion of the system. He can choose to interact with the appointment booking system (e.g., open the scheduling app), or choose to walk in or balk (without engaging with the appointment system at all). If the patient engages in appointment booking at the first stage, he enters the second stage where he acquires the exact appointment delay. He 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 *disengagement cost*, because he is disengaged from his 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 could easily opt to other non-appointment alternatives. However, as discussed above, disengaging from the appointment choice after interacting with an asynchronous system is costly to the patient, because it can lead to inconvenience to himself 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 will be refunded if they hold on to the appointment option and do not deviate.) Since the disengagement cost depends on appointment system (design) and stipulates how likely patients would 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 (walk-in resp.) closely depends on the congestion during appointment hours (walk-in

[2] We use “customer” and “patient” interchangeably in this paper.

hours resp.). For the provider, she 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 due to 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 patient choice process in a wide range of online scheduling systems. Specifically, when the disengagement cost is zero, the model is equivalent to one where 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 resp.) setting and real-time (asynchronous resp.) 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 demand-capacity relationship and patient willingness to wait. When the provider's capacity falls significantly short compared to demand, then she should make delay information easily accessible by patients to attract as many of them as possible and to 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 triage system, we find that the provider's own cost tradeoff 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. Sections 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 Appendix.

2. Literature Review

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

2.1. Healthcare OM

Our work is closely related to the healthcare operations management (OM) literature on outpatient appointment scheduling; see, e.g., recent literature reviews by Gupta and Denton (2008), 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 walk-in, while 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 to strategically choose from in the first place and all walk-ins will be served. Bavafa et al. (2019) investigate a setting where 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 has studied how to manage walk-in customers in settings such as restaurants, hotels, and rental firms; see, e.g., Gans and Savin (2007), Alexandrov and Lariviere (2012), Cil and Lariviere (2013), 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, where customers first decide whether to order online or order on site, and if 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 the queueing studies with strategic customers. Starting with the seminal work by Naor (1969), extensive literature has considered 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, e.g., 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 rate in an observable queue where 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 which 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 of delay information purchase and provider decision on delay announcement receive much attention (Hassin and Roet-Green 2017, Allon et al. 2011, Yu et al. 2017, Hu 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 her 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, i.e., appointment delay and in-clinic waiting. Customers make choices in two subsequent stages, where 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 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 tradeoff 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 her 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 hours and walk-in hours; all slots are of equal length, e.g., 20 minutes. Specifically, she 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 Appendix B for several practical examples). Thus we assume that walk-in hours 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 will choose to walk in without exceptions—we call this type *exogenous* walk-ins, whose arrival behavior will not be influenced by the provider's capacity decisions. The average number of exogenous walk-ins per day is λ_E. All walk-ins will be served on the same day of visit, possibly during overtime. The second type of patients have less acute symptoms (e.g., runny nose) which could be delayed and they will 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 decision in two sequential stages. At the first stage, based on the revealed information about the system, he decides whether or not to interact with the scheduling system and if not, he chooses to walk in or balk (e.g., seek care elsewhere). If, however, he decides to interact with the scheduling system and book an appointment, then he moves onto the second stage of decision making and obtains the exact appointment delay information. At this stage, he could choose to stick with the appointment choice made in the first stage—or—to revoke that decision but to walk in or balk. If he chooses the appointment, he endures appointment delay only; at his scheduled visit he does not have to wait in clinic. If he chooses to walk in, he can see the provider on the same day, but likely has to spend time waiting in clinic. Figure 1 shows the patient decision process and our model, with additional details to be discussed.

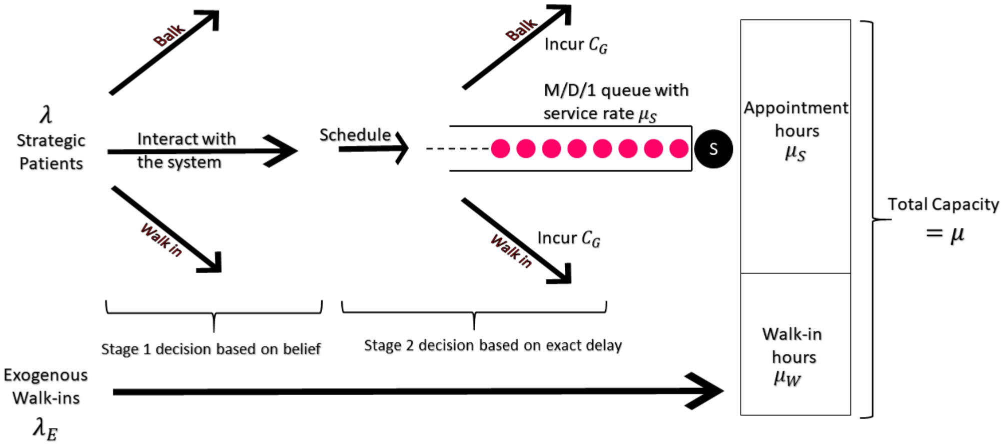


Figure 1 **Patient Decision Making and Model Schematic**

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, if a strategic patient first decides to interact with the scheduling system but then switches to walk in or balk instead (after obtaining the appointment delay information), we assume that he incurs a cost C_G which we refer to as the *disengagement* cost, because he is disengaged from his original plan. The disengagement cost can be small or large. In a real-time system like 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; he could freely opt to other non-appointment alternatives if he 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—he knows that he is likely to get an appointment sometime later in the week, so he could have already made a plan for his day, while disengaging from this plan (e.g., choosing to walk in today) most likely disrupts his life. In addition, after the provider spends efforts locating an appointment for the patient, it would be trying for him to disengage. If he indeed decides not to take the appointment, to reduce potential loss of goodwill from the provider he probably needs to reply to the confirmation email and make an explanation. He cannot "game" the system like in a real-time one by quickly observing the appointment delay and then switching to non-appointment 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 will 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 evolvement 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, he will be 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 will also accept the earliest available appointment slot. It is possible that strategic patients with more urgent needs will be provided earlier appointments if available. Our M/D/1 formulation also implicitly assumes that all patients with appointments will show up and arrive on time. The purpose of the queuing 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, while 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 emregency 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 below).

3.2. Patient Strategy and Utility

At the first stage, a strategic patient bases his decision on the *expected* utility at the second stage; and at the second stage after interacting with the scheduling system, he observes the *exact* appointment delay $\tilde{d}$, which is a random variable and influenced by the strategy adopted by all other patients. If he chooses to walk in (regardless in which stage), he does not know the exact in-clinic wait time, but he can form a blief 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 will define and identify an equilibrium strategy. Since strategic patients are *ex ante* homogeneous, we will 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 $p_S^2(\tilde{d}) + p_W^2(\tilde{d}) + p_B^2(\tilde{d}) = 1,\ \forall \tilde{d}$, since these three options are exhaustive.

We analyze patient utility backwards, 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 2^{nd} stage, denoted by u_S^2, is

$$u_S^2 = R - C_D\tilde{d}. \tag{1}$$

Note that $\tilde{d}$ is the pure delay measured by the appointment queue (in days). Since each actual service slot is of equal length (e.g., 20 minutes), the utility gained due to 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 walk-in at the 2^{nd} stage, denoted by u_W^2, is

$$u_W^2 = R - C_W\bar{w} - C_G. \tag{2}$$

The utility of balking at the 2^{nd} 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 $\tilde{d}$, he will schedule an appointment, walk in, and balk 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 he chooses to interact with the scheduling system and before observing $\tilde{d}$. The utility of interacting with the scheduling system at the 1^{st} stage, denoted by u_S^1, is the expected utility he would gain if proceeding to the 2^{nd} stage:

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

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 that the patient chooses to join the appointment queue at the 2^{nd} stage. (We will provide a more explicit form for $\mathbb{E}_{\tilde{d}}[u_S^2(\tilde{d})|\text{schedule}]$ in Section 3.3.) The utility of walk-in at the 1^{st} stage, denoted by u_W^1, is

$$u_W^1 = R - C_W\bar{w}. \tag{4}$$

Finally, the utility of balking at the 1^{st} stage, denoted by u_B^1, is normalized to be zero, i.e., $u_B^1 = 0$.

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 1^{st} stage. (Similar notations have been 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 2^{nd} 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 which represents the appointment delay observed by a strategic patient when interacting with the scheduling system. Since strategic patients arrive according to a Poisson process, $\tilde{d}$ has the same distribution as the steady-state distribution of delay in the appointment queue due to PASTA (Poisson Arrivals See Time Averages).

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 1^{st} stage is λ_W; at the 2^{nd} stage, the portion of patients who switch to walk-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 evolvement of 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 will not assume any specific form for the queueing process of walk-ins, but rather we will 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. *$w(\rho)$ 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

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

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 his 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 where 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 $(\lambda_B, \lambda_W, \lambda_B)$ we will 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 chose to interact with the appointment system as the focal patient. This patient will only schedule an appointment if scheduling has a sufficiently high utility, i.e., if he sees a sufficiently short appointment delay; otherwise he will walk in or balk. This intuition is formalized below.

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 $\tilde{d} \le \delta$, $p_S^2(\tilde{d}) = 1$; 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.*

Since 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 above) can be more explicitly expressed as follows.

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

Condition (6) ensures that when delay is no larger than δ, scheduling is no worse than walk-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 walk-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 below investigates the second-stage equilibrium when we would exogenously exclude one of the options (scheduling, walk-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 would choose if only facing two choices. For example, excluding the option of walk-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 super- and sub-scripts to represent the choices under consideration ($\underline{\text{s}}$cheduling and $\underline{\text{b}}$alking 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 walk-in is not an option for strategic patients, then there exists a unique δ_B^S such that patients $\underline{s}$chedule an appointment when the delay $\tilde{d} \le \delta_B^S$ and $\underline{b}$alk when $\tilde{d} > \delta_B^S$, where $\delta_B^S = (R + C_G)/C_D$ so that $u_S^2(\delta_B^S) = u_B^2 = -C_G$, i.e., 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 $\underline{w}$alk in with probability p_W^B and $\underline{b}$alk 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 $u_W^2(p_W^B) = u_B^2 = -C_G$, i.e., walking in with probability p_W^B makes patients feel indifferent between walk-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 $\underline{s}$chedule an appointment queue when delay $\tilde{d} \le \delta_W^S$ and $\underline{w}$alk in when $\tilde{d} > \delta_W^S$. In particular, δ_W^S solves $u_S^2(\delta_W^S) = u_W^2(\pi(\delta_W^S))$, i.e., patients feel indifferent between scheduling and walk-in when delay is δ_W^S. The proportion of strategic patients who walk in is $p_W^S = \pi(\delta_W^S)$.*

Lemmas 3.1 and 3.2 describe how patients choose between scheduling and balking, and between walk-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 walk-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 3.3.

Given these pairwise comparisons, we are ready to present the key results of this subsection in the following proposition. To avoid ambiguity, we stipulate that in the case of a tie between scheduling and balking, the patient will schedule 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* $u_W^2(0) \le -C_G$*, 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* $p_B = p_B^S - p_W^B$*; and*
3. *if* $u_W^2(p_W^S) \ge -C_G$*, 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, i.e., $u_W^2(0) \le -C_G$, implies a relatively large volume of exogenous walk-ins, which alone already make the utility of walk-in below $-C_G$ (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 walk-in can remains positive even if some strategic patients choose to walk in. The question 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 walk-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 would choose walk-in in equilibrium, and a subset would choose to balk, implying that the utilities of walk-in and balking both equal $-C_G$ in equilibrium. In other words, strategic patients who choose not to schedule would have $-C_G$ utility, and hence such patients schedule as if there were no walk-in option (see Lemma 3.1). The proportion of strategic patients walking in is such that it makes the utility of walk-in $-C_G$ (see Lemma 3.2), 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 walk-in to be no worse than balking (i.e., $u_W^2(p_W^S) \ge -C_G$ in case 3), then balking is never appealing to strategic patients. Thus, strategic patients would behave as if there were no balking option (see Lemma 3.3). We call this case *Regime SnW* (schedule and walk-in). Table 1 in Appendix D summarizes patient equilibrium in the three regimes above 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, they 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

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

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, i.e., u_S^1 defined in (3), can be rewritten as follows.

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

In (12), $\mathbb{E}_{\tilde{d}}[\tilde{d}|\tilde{d}\le\delta,\lambda_S]$ 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 $\mathbb{E}_{\tilde{d}}[\tilde{d}|\tilde{d}\le\delta,\lambda_S]$ and $\pi(\delta,\lambda_S)$ on λ_S, because λ_S is one of the patient's strategic choice at the first stage. The explicit form of (12) is helpful for our analysis of the equilibrium at the first stage.

Noticing that $u_W^1-u_B^1=u_W^2-u_B^2$, so 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 Lemma D.2 in Appendix 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 will be realized given the model parameters.

Scenario a: $u_B^2\ge u_W^2$. 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

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

We show that u_S^1 here is decreasing in λ_S, 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

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

and

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

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 LHS of (14) minus the

RHS of (14)) decreases in λ_S. Then we will 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

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

The equation above 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 their own right (see details in 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 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 $R - C_W w(\frac{\lambda_E}{\mu_W}) \le 0$, 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 $R - C_W w(\frac{\lambda_E}{\mu_W}) > 0$ 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 $R - C_W w(\frac{\lambda_E + \lambda_W + p_W \lambda_S}{\mu_W}) = 0$, $\lambda_B = \lambda - \lambda_S - \lambda_W$, $\delta = \delta^a$, $p_B = \pi(\delta^a, \lambda_S) - p_W$. In this case, λ_S and δ are unique, while 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 $R - C_W w(\frac{\lambda_E}{\mu_W}) > 0$ and $\delta^a \ge \max(\delta^b, \delta^c)$, 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 $R - C_W w(\frac{\lambda_E}{\mu_W}) \le 0$, walk-in is surely worse than balking, and thus strategic patients either schedule or balk—we draw our attention to Scenario a above. 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, i.e., $\lambda_S = \lambda$. However, if $\lambda_S^a \le \lambda$, 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 $R - C_W w(\frac{\lambda_E}{\mu_W}) > 0$, there must be some strategic patients walking in (otherwise walk-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 patients 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 above 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 walk-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 above, 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 walk-in at the first stage; then they mix again between scheduling an appointment and walk-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 her 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 below, cleaner and tractable.

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

While 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 $(\mu_S,\ \mu_W)$. 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 would 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 where 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, he chooses to join the appointment queue, otherwise he mixes between walk-in and balking. The form of the equilibrium strategy described in Theorem 1 can be simplified as (δ, p_W, p_B), which we eloborate in 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 he calculates his 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 his 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, since 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 non-negative 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 as an M/D/1/R queue where $R \in (0, +\infty)$.

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 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 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 Appendix C) provides us a way to analyzing the optimal capacity allocation problem. Second, the monotonicity of $\pi(\delta, \mu_S)$ will be 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 Table 1 in Appendix 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

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

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

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

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

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

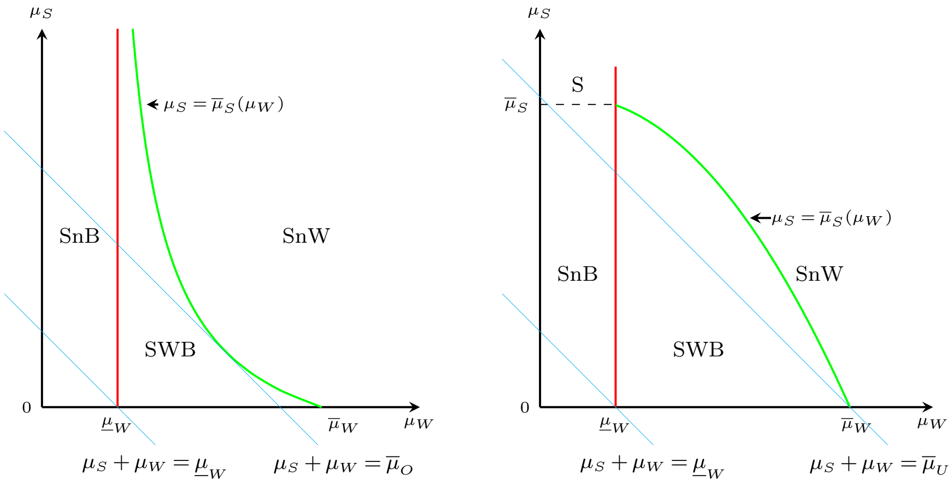
$$R - C_W w\left(\frac{\pi(\delta_B^S, \mu_S)\lambda + \lambda_E}{\mu_W}\right) = 0, \tag{17}$$

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

The intuition above 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* $\mu_W \le \underline{\mu}_W$, *the equilibrium is in Regime SnB.*
2. *If* $\underline{\mu}_W < \mu_W < \overline{\mu}_W$, *there exists a decreasing function* $\overline{\mu}_S(\mu_W)$ *such that,*
   (a) *when* $\mu_S < \overline{\mu}_S(\mu_W)$, *the equilibrium is in Regime SWB;*
   (b) *when* $\mu_S \ge \overline{\mu}_S(\mu_W)$, *the equilibrium is in Regime SnW.*
3. *If* $\mu_W \ge \overline{\mu}_W$, *the equilibrium is in Regime SnW.*



(a) Observable Setting (b) Unobservable Setting

Figure 2 Equilibrium Regimes and Capacity Allocation

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

$$\begin{array}{lll}\displaystyle\min_{\mu_S,\mu_W\ge 0} & C_L p_B \lambda + C_O o\left(\dfrac{p_W\lambda+\lambda_E}{\mu_W}\right) & \text{(Ob.P)}\\ \text{subject to:} & \mu_S+\mu_W=\mu, & \\ & (p_B,p_W)=\begin{cases}(p_B^S,0) & \text{if } \mu_W\le\underline{\mu}_W,\\ (p_B^S-p_W^B,p_W^B) & \text{if } \mu_W\ge\underline{\mu}_W,\ \mu_S\le\overline{\mu}_S(\mu_W),\\ (0,p_W^S) & \text{if } \mu_W\ge\underline{\mu}_W,\ \mu_S\ge\overline{\mu}_S(\mu_W);\end{cases} & \\ & p_B^S,p_W^B,p_W^S \text{ are defined in Lemma 3 with } \lambda_S=\lambda \text{ and } C_G=0; & \\ & \underline{\mu}_W,\overline{\mu}_S(\mu_W) \text{ are defined in (15) and (17), respectively.} & \end{array}$$

The above 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 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 $\overline{\mu}_O \geq 0$, such that SnW is optimal when $\mu \geq \overline{\mu}_O$, SnB or SWB is optimal when $\underline{\mu}_W \leq \mu < \overline{\mu}_O$, and SnB is optimal when $\mu < \underline{\mu}_W$.*

In Proposition 3, $\overline{\mu}_O$ represents the minimum total daily capacity required to attract all strategic patients to either schedule an appointment or walk in. If $\mu \geq \overline{\mu}_O$, 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 inbetween $\underline{\mu}_W$ and $\overline{\mu}_O$, which regime—SnB vs. 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 where 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, walk-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

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

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

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

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, i.e., $p_W = 0$. 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 as SnB, SWB, and SnW, respectively. More details on the equilibrium can be found in Appendix D.5.

4.2.1. Impact of μ_S and μ_W on Patient Equilibrium Behavior Recall $\underline{\mu}_W$ and $\overline{\mu}_W$ defined in (15) and (16), respectively. If $\mu_W < \underline{\mu}_W$, all strategic patients choosing not to schedule would balk. If $\mu_W > \overline{\mu}_W$, all strategic patients who choose not to schedule would walk in. Let $\overline{\mu}_S$ be such that

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

If $\mu_S > \overline{\mu}_S$, 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.

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

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

One can verify that (20) defines μ_S as a function of μ_W. With a slight abuse of notations, we write this function as $\overline{\mu}_S(\mu_W)$ as before. For a fixed μ_W, if $\mu_S > \overline{\mu}_S(\mu_W)$, then the utility of walk-in will become non-negative even if all strategic patients who choose not to schedule walk in. Thus, no strategic patients balk; they either schedule or walk in. If $\mu_S < \overline{\mu}_S(\mu_W)$, 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 walk-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 $\mu_W \le \underline{\mu}_W$, the equilibrium is in Regime SnB.*
2. *If $\underline{\mu}_W < \mu_W < \overline{\mu}_W$, there exists a decreasing function $\overline{\mu}_S(\mu_W)$ such that,*
   (a) *if $\mu_S < \overline{\mu}_S(\mu_W)$, the equilibrium is in Regime SWB;*
   (b) *if $\mu_S \ge \overline{\mu}_S(\mu_W)$, the equilibrium is in Regime SnW.*
3. *If $\mu_W \ge \overline{\mu}_W$, the equilibrium is in Regime SnW.*

REMARK 1. If $\mu_W \le \underline{\mu}_W$ and $\mu_S \ge \overline{\mu}_S$, 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, i.e., smaller than $\overline{\mu}_W$, 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, i.e., larger than $\overline{\mu}_S$, 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 to capacity—we will 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.

$$\begin{aligned} \min_{\mu_S,\mu_W \geq 0} \quad & C_L\lambda_B + C_O o\left(\frac{\lambda_W+\lambda_E}{\mu_W}\right) && \text{(Un.P)} \\ \text{subject to:} \quad & \mu_S+\mu_W=\mu, \\ & (\lambda_W,\lambda_B)=\begin{cases} (0,\lambda_B^S) & \text{if } \mu_W \leq \underline{\mu}_W, \\ (\lambda_W^B, \lambda_B^S-\lambda_W^B) & \text{if } \mu_W \geq \underline{\mu}_W,\ \mu_S \leq \overline{\mu}_S(\mu_W), \\ (\lambda_W^S, 0) & \text{if } \mu_W \geq \underline{\mu}_W,\ \mu_S \geq \overline{\mu}_S(\mu_W); \end{cases} \\ & \lambda_B^S, \lambda_W^B, \lambda_W^S \text{ are defined by Lemma D.3 in Appendix D.5;} \\ & \underline{\mu}_W, \overline{\mu}_S(\mu_W) \text{ are defined in (15) and (20), respectively.} \end{aligned}$$

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 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 walk-in and balking. Detailed analysis of problem (Un.P) is deferred to Appendix D.6. The proposition below summarizes the impact of total daily capacity on the optimal equilibrium regime.

PROPOSITION 5. *In the unobservable setting, there exists $\overline{\mu}_U > 0$ such that SnW is optimal when $\mu \geq \overline{\mu}_U$, SWB or SnB is optimal when $\underline{\mu}_W \leq \mu < \overline{\mu}_U$, 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 walk-in under optimal capacity allocation in the unobservable setting. Together with Proposition 5, this suggests that when the total capacity is sufficiently large, i.e., $\mu \geq \overline{\mu}_U$, 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 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 Setting v.s. 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 to the patient demand she is facing. Second, the provider's daily capacity is somewhat "in balance" with her patient demand. These two scenarios are of the most interest to practice because in reality healthcare 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 *costs 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 walk-in as an option for wait-sensitive strategic customers; see, e.g., Chen and Frank (2004).

When the provider's daily capacity is sufficiently large but remains in the same magnitude of her 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 while 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, where the provider has to open additional costly walk-in hours in order not to lose strategic patients from balking.

Remark 3. The analysis above 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, she 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 walk-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 Theorem E.1 in Appendix E.3.

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 shall 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 *triage model* and the original model where patients can freely choose as *strategic model*.

We focus our discussion of triage model in the context of 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.

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

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 would walk in under any feasible capacity allocation in this case; see Figure 2(a). Below we shall 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.

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

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 suppose that* $\mu > \underline{\mu}_W$,

1. *if* $\frac{C_L}{C_O} \leq \underline{\alpha}$*, the triage model costs less;*
2. *if* $\frac{C_L}{C_O} \geq \overline{\alpha}$*, 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 tradeoff. If the provider has a relatively high lost demand cost, offering patients free choice reduces balking and allows the provider to use her 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 due to 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 relative 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 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. She incurs lost demand costs due to 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 will receive detailed appointment confirmation after a relatively short amount of time, say a few hours. Such a short and yet non-trivial 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, e.g., 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 towards 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: demand-capacity relationship and patient willingness to wait. When the provider falls significantly short of capacity, she would be better off using a real-time system which provides patients with exact delay information upon their appointment requests. However, if she uses an asynchronous system in this situation, it is critical for her to boost service capacity. Otherwise many patients may choose not to come for service due to 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 has revealed heterogeneity in patient sensitivity to waiting in different medical specialties; see, e.g., 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 needs 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 patient waiting experience in clinic. One such setting could be pediatric care, where children tend to have shorter attention span and are easier to get irritated 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 multi-specialty dataset, Osadchiy and Kc (2017) find that general pathology and diabetes education are two specialties where 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 will be more appealing to patients with less urgent conditions who can tolerate longer appointment delays. While 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 triage model suggests that such an active control may either benefit or hurt a practice. In particular, it benefits a provider when she 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 tele-visit 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, 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 and 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). Last but not least, 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. Shan Wang was supported in part by the National Natural Science Foundation of China [Grant Number 72001220].

References

Ahmadi-Javid, Amir, Zahra Jalali, Kenneth J Klassen. 2017. Outpatient appointment systems in healthcare: A review of optimization studies. *European Journal of Operational Research* **258**(1) 3–34.

Alexandrov, Alexei, Martin A Lariviere. 2012. Are reservations recommended? *Manufacturing Service Oper. Management* **14**(2) 218–230.

Allon, Gad, Achal Bassamboo, Itai Gurvich. 2011. “we will be right with you”: Managing customer expectations with vague promises and cheap talk. *Operations Research* **59**(6) 1382–1394.

Anand, Krishnan S, M Fazıl Paç, Senthil Veeraraghavan. 2011. Quality–speed conundrum: Trade-offs in customer-intensive services. *Management Science* **57**(1) 40–56.

Ata, Bariş, Shiri Shneorson. 2006. Dynamic control of an m/m/1 service system with adjustable arrival and service rates. *Management Science* **52**(11) 1778–1791.

Baron, Opher, Xiaole Chen, Yang Li. 2022. Omnichannel services: the false premise and operational remedies. *Management Science* https://doi.org/10.1287/mnsc.2022.4416.

Bavafa, Hessam, Anne Canamucio, Steven C Marcus, Christian Terwiesch, Rachel M Werner. 2021. Capacity rationing in primary care: Provider availability shocks and channel diversion. *Management Science* https://doi.org/10.1287/mnsc.2021.4026.

Bavafa, Hessam, Lorin M Hitt, Christian Terwiesch. 2018. The impact of e-visits on visit frequencies and patient health: Evidence from primary care. *Management Science* **64**(12) 5461–5480.

Bavafa, Hessam, Sergei Savin, Christian Terwiesch. 2019. Managing patient panels with non-physician providers. *Production and Operations Management* **28**(6) 1577–1593.

Centre Pediatrics. 2021. Notice to our patients regarding walk-in hours. Accessed June 29, 2021: https://www.centrepediatrics.org/visits/walk-in-hours/.

Chen, Hong, Murray Frank. 2004. Monopoly pricing when customers queue. *IIE Transactions* **36**(6) 569–581.

Cil, Eren B, Martin A Lariviere. 2013. Saving seats for strategic customers. *Operations Research* **61**(6) 1321–1332.

Debo, Laurens G, L Beril Toktay, Luk N Van Wassenhove. 2008. Queuing for expert services. *Management Science* **54**(8) 1497–1512.

Dobson, Gregory, Sameer Hasija, Edieal J Pinker. 2011. Reserving capacity for urgent patients in primary care. *Production and Operations Management* **20**(3) 456–473.

Gans, Noah, Sergei Savin. 2007. Pricing and capacity rationing for rentals with uncertain durations. *Management Science* **53**(3) 390–407.

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

Guo, Pengfei, Refael Hassin. 2011. Strategic behavior and social optimization in markovian vacation queues. *Operations Research* **59**(4) 986–997.

Gupta, Diwakar, Brian Denton. 2008. Appointment scheduling in health care: Challenges and opportunities. *IIE transactions* **40**(9) 800–819.

Hassin, Refael. 2016. *Rational queueing*. CRC press.

Hassin, Refael, Moshe Haviv. 1997. Equilibrium threshold strategies: The case of queues with priorities. *Operations Research* **45**(6) 966–973.

Hassin, Refael, Ricky Roet-Green. 2017. The impact of inspection cost on equilibrium, revenue, and social welfare in a single-server queue. *Operations Research* **65**(3) 804–820.

Hu, Ming, Yang Li, Jianfu Wang. 2017. Efficient ignorance: Information heterogeneity in a queue. *Management Science* **64**(6) 2650–2671.

Ibrahim, Rouba. 2018. Sharing delay information in service systems: a literature survey. *Queueing Systems* **89**(1) 49–79.

Kc, Diwas S, Christian Terwiesch. 2009. Impact of workload on service time and patient safety: An econometric analysis of hospital operations. *Management Science* **55**(9) 1486–1498.

Liu, Jiayi, Diwas KC. 2020. Nudging patient choice: Evidence from a field experiment. Working paper.

Liu, Nan. 2016. Optimal choice for appointment scheduling window under patient no-show behavior. *Production and Operations Management* **25**(1) 128–142.

Liu, Nan, Stacey R Finkelstein, Margaret E Kruk, David Rosenthal. 2017. When waiting to see a doctor is less irritating: Understanding patient preferences and choice behavior in appointment scheduling. *Management Science* **64**(5) 1975–1996.

Liu, Nan, Serhan Ziya. 2014. Panel size and overbooking decisions for appointment-based services under patient no-shows. *Production and Operations Management* **23**(12) 2209–2223.

Naor, Pinhas. 1969. The regulation of queue size by levying tolls. *Econometrica* 15–24.

Oh, Jaelynn, Xuanming Su. 2018. Reservation policies in queues: Advance deposits, spot prices, and capacity allocation. *Production and Operations Management* **27**(4) 680–695.

Osadchiy, Nikolay, Diwas Kc. 2017. Are patients patient? the role of time to appointment in patient flow. *Production and Operations Management* **26**(3) 469–490.

Rajan, Balaraman, Tolga Tezcan, Abraham Seidmann. 2019. Service systems with heterogeneous customers: investigating the effect of telemedicine on chronic care. *Management Science* **65**(3) 1236–1267.

Thaler, Richard H, Cass R Sunstein. 2009. Nudge: Improving decisions about health, wealth, and happiness. New Haven, CT: Yale University Press.

Tunçalp, Feray, Evrim Didem Gunes, Lerzan Ormeci. 2020. Modeling strategic walk-in patients in appointment systems: Equilibrium behavior and capacity allocation. Available at SSRN 3687717.

Wang, Shan, Nan Liu, Guohua Wan. 2020. Managing appointment-based services in the presence of walk-in customers. *Management Science* **66**(2) 667–686.

Yu, Qiuping, Gad Allon, Achal Bassamboo, Seyed Iravani. 2017. Managing customer expectations and priorities in service systems. *Management Science* **64**(8) 3942–3970.

Zacharias, Christos, Mor Armony. 2016. Joint panel sizing and appointment scheduling in outpatient care. *Management Science* **63**(11) 3978–3997.

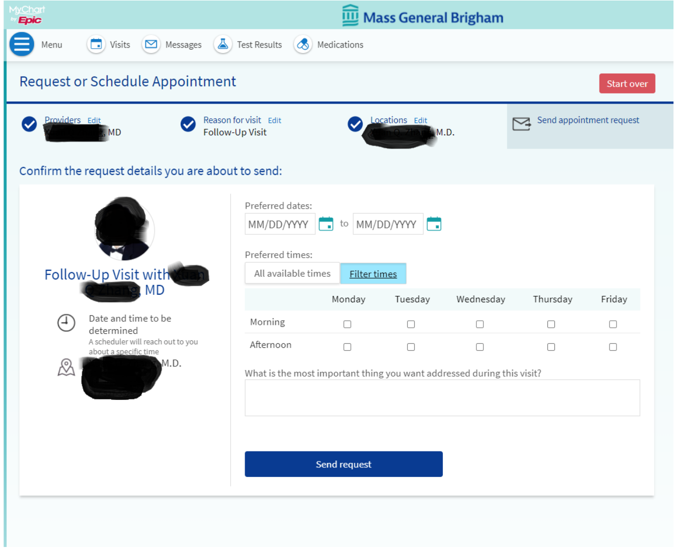
Zacharias, Christos, Tallys Yunes. 2020. Multimodularity in the stochastic appointment scheduling problem with discrete arrival epochs. *Management Science* **66**(2) 744–763.

Zhao, Peng, Illhoi Yoo, Jaie Lavoie, Beau James Lavoie, Eduardo Simoes. 2017. Web-based medical appointment systems: A systematic review. *Journal of Medical Internet Research* **19**(4) e134.

Online Appendix

This file is the electronic companion of the paper “Managing Outpatient Service with Strategic Walk-ins” by Nan Liu, Willem van Jaarsveld, Shan Wang, and Guanlian Xiao. It contains the following sections.

Appendix A: User Interface in Asynchronous Scheduling Systems



Request or Schedule Appointment

✓ Providers Edit
[redacted] MD

✓ Reason for visit Edit
Follow-Up Visit

✓ Locations Edit
[redacted] M.D.

Send appointment request

Confirm the request details you are about to send:

[redacted]

Follow-Up Visit with [redacted] MD

Date and time to be determined
A scheduler will reach out to you about a specific time

[redacted] M.D.

Preferred dates:
MM/DD/YYYY to MM/DD/YYYY

Preferred times:
All available times | Filter times

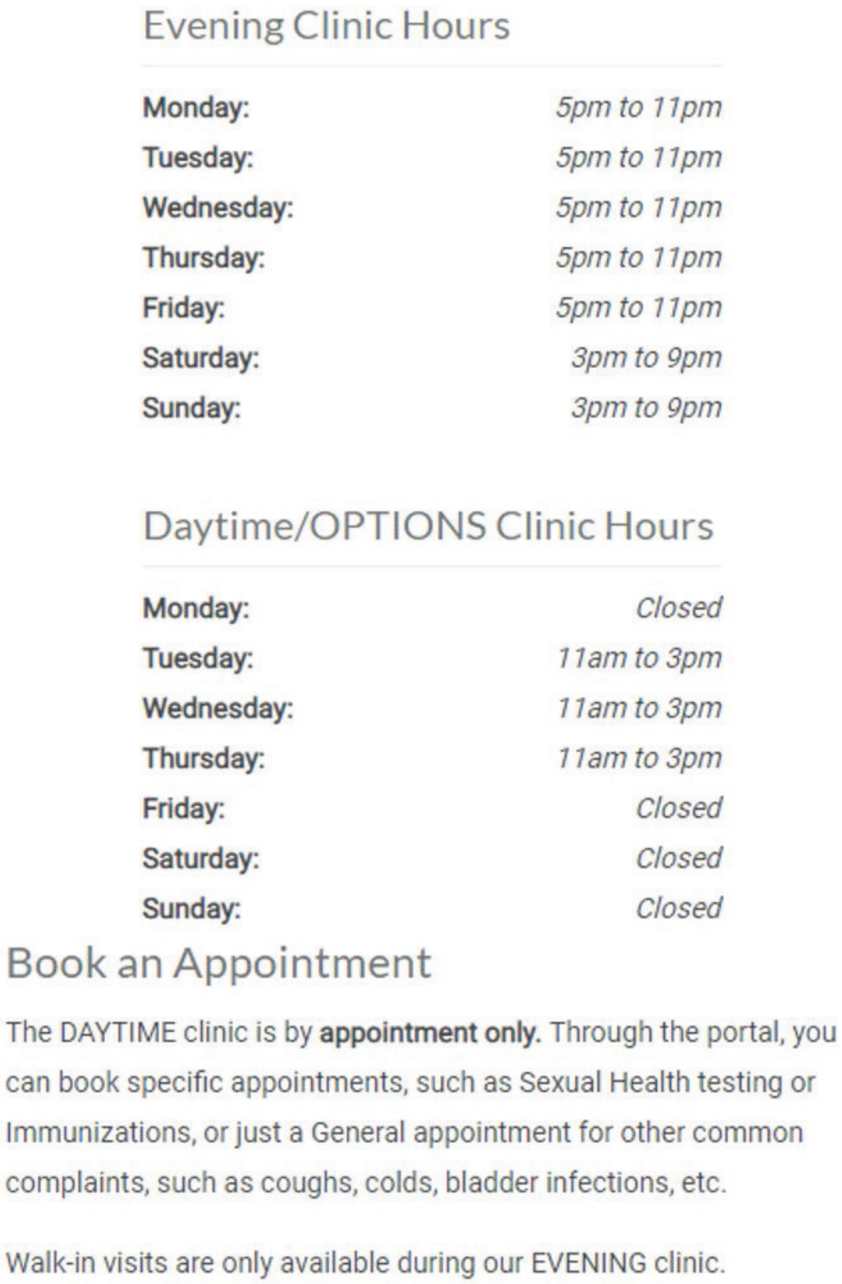
	Monday	Tuesday	Wednesday	Thursday	Friday
Morning	☐	☐	☐	☐	☐
Afternoon	☐	☐	☐	☐	☐

What is the most important thing you want addressed during this visit?

Figure A.1 Asynchronous Scheduling System at Mass General Brigham

Appendix B: Examples of Disjoint and Dedicated Appointment and Walk-in Hours

Example 1: The Onyx Medical Centre in Ontario, Canada offers evening clinic hours only for walk-ins and reserves daytime clinic hours only for appointments.[3] See Figure B.2.



Evening Clinic Hours

Monday:	*5pm to 11pm*
Tuesday:	*5pm to 11pm*
Wednesday:	*5pm to 11pm*
Thursday:	*5pm to 11pm*
Friday:	*5pm to 11pm*
Saturday:	*3pm to 9pm*
Sunday:	*3pm to 9pm*

Daytime/OPTIONS Clinic Hours

Monday:	*Closed*
Tuesday:	*11am to 3pm*
Wednesday:	*11am to 3pm*
Thursday:	*11am to 3pm*
Friday:	*Closed*
Saturday:	*Closed*
Sunday:	*Closed*

Book an Appointment

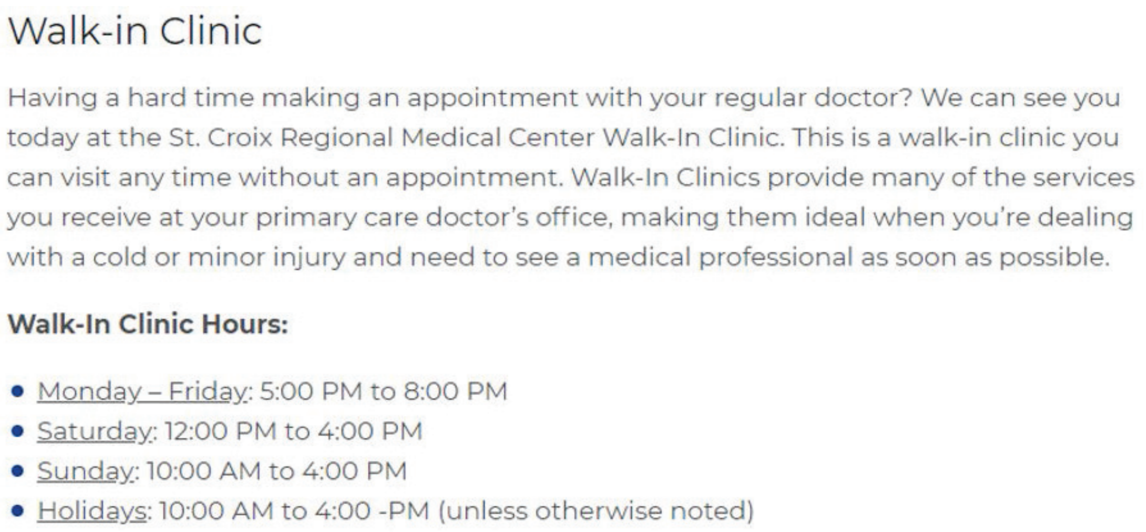
The DAYTIME clinic is by **appointment only.** Through the portal, you can book specific appointments, such as Sexual Health testing or Immunizations, or just a General appointment for other common complaints, such as coughs, colds, bladder infections, etc.

Walk-in visits are only available during our EVENING clinic.

Figure B.2 Office Hours of the Onyx Medical Centre

[3] https://onyxurgentcare.com/patient-portal/

Example 2: The St. Croix Regional Medical Center in Wisconsin, USA sets 5:00 PM to 8:00 PM daily as Walk-In Clinic, during which patients can visit without an appointment.[4] See Figure B.3.



Walk-in Clinic

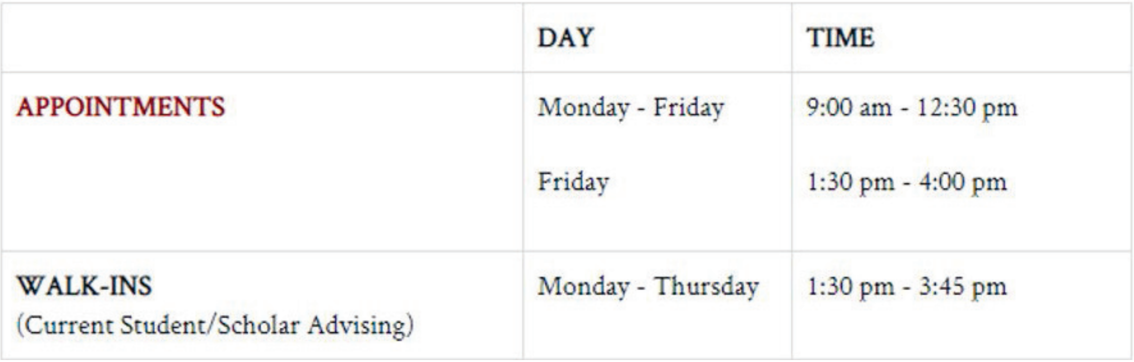
Having a hard time making an appointment with your regular doctor? We can see you today at the St. Croix Regional Medical Center Walk-In Clinic. This is a walk-in clinic you can visit any time without an appointment. Walk-In Clinics provide many of the services you receive at your primary care doctor's office, making them ideal when you're dealing with a cold or minor injury and need to see a medical professional as soon as possible.

Walk-In Clinic Hours:

- Monday – Friday: 5:00 PM to 8:00 PM
- Saturday: 12:00 PM to 4:00 PM
- Sunday: 10:00 AM to 4:00 PM
- Holidays: 10:00 AM to 4:00 -PM (unless otherwise noted)

Figure B.3 Walk-in Hours of the St. Croix Regional Medical Center

Example 3: In addition to healthcare services, other service providers also reserve walk-in hours. The University of Maryland's International Student & Scholar Services (ISSS) set the afternoon time as walk-in session, while keep the morning only for appointments.[5] See Figure B.4.



	DAY	TIME
APPOINTMENTS	Monday - Friday Friday	9:00 am - 12:30 pm 1:30 pm - 4:00 pm
WALK-INS (Current Student/Scholar Advising)	Monday - Thursday	1:30 pm - 3:45 pm

Figure B.4 Office Hours of the University of Maryland's ISSS

[4] https://www.scrmc.org/our-services/walk-in-clinic/

[5] https://globalmaryland.umd.edu/offices/international-students-scholar-services/office-hours

Appendix C: Analysis of M/D/1/R Queue with $R \in (0,\infty)$

In this section, we summarize our analysis of the appointment queue where customers use a continuous threshold-based joining strategy (as described in Section 4.1). Detailed proofs of the technical results in this section can be found in Appendix E.4. The results in this section are useful in Appendix E, where we present the proof of all technical results in the paper.

In our earlier notations, we use δ to represent this threshold such that patients join the queue when the delay is no larger than δ and balk otherwise. Recall that the arriving customer calculates his delay to be $(k+\tilde{u})/\mu_S$ where k is the number of customers waiting in the queue (excluding the one in service), $\tilde{u}$ is a standard uniform random variable, and μ_S is the service rate of the appointment queue. Let $K = \lceil \delta\mu_S \rceil$ and $p = \delta\mu_S + 1 - K$. Then $p \in [0,1]$. If an arrival sees $K > 0$ customers ahead of him (including the one in service), then his delay is $(K-1+\tilde{u})/\mu_S$. It follows that his joining probability is

$$Pr[\frac{K-1+\tilde{u}}{\mu_S} \le \delta] = Pr[\tilde{u} \le \delta\mu_S + 1 - K] = p.$$

To sum, customers join the queue if the system size is smaller than K, and if the system size is K they join the queue with probability p and balk with probability $1-p$. Following the notation introduced by Hassin and Haviv (1997), such a queue can be represented as an $M/D/1/K-1+p$ queue where $K-1+p = \delta\mu_S$ in our context. We encapsulate the discussion above in the following lemma, which states that our appointment queue with strategic customers is equivalent to a queue with a continuous buffer size.

Lemma C.1. *An $M/D/1$ queue described above where customers join based on a continuous delay threshold δ is equivalent to an $M/D/1/K-1+p$ queue with $K = \lceil \delta\mu_S \rceil$ and $p = \delta\mu_S + 1 - K$.*

Remark C.1. If $\delta\mu_S = 0$, then $K = 0$ and $p = 1$, meaning that the arriving customer only joins the queue when the system is empty.

Next we analyze our appointment queue. Let $X(t)$ be the system size at time t and $q_j(t)$ be the probability that there are j customers in the system at time t, i.e., $q_j(t) = Pr[X(t) = j]$. We are interested in the steady state distribution of system size, denoted by

$$\pi_j = \lim_{t\to\infty} q_j(t), \quad j = 0,1,\ldots,K+1.$$

We use an embedded Markov Chain approach by first analyzing the state of the system right after each customer's departure (Gross 2008). Let t_n be the time that the nth customer departs. Then $X(t_n^+)$ is the queue length right after customer n's departure. Because the arrival process is Poisson, it is clear that $\{X(t_n^+), n = 1,2,3,\ldots\}$ is a Discrete Time Markov Chain (DTMC). We first derive the stationary distribution of the queue length observed by departures. In particular, we let

$$q_j = \lim_{n\to\infty} q_j(t_n^+) = \lim_{n\to\infty} Pr[X(t_n^+) = j], \quad k = 0,1,2,\ldots,K.$$

It is important to note that $q_j \ne \pi_j$. However, given that the size of the queue can only jump up or down by at most 1 at any point of time, q_j is the same as the probability that a joining customer sees j customers ahead of him (see Theorem 7.1 in Kulkarni 1996). Specifically, we have

$$q_j = \Pr\{\text{an arrival finds } j \text{ customers ahead}|\text{he joins}\} = \frac{\pi_j}{\Pr\{\text{a customer joins}\}} \tag{21}$$

for $j = 0, 1, \ldots, K-1$; and

$$q_K = \frac{p\pi_K}{\Pr\{\text{a customer joins}\}}. \tag{22}$$

We also have

$$\lambda \Pr\{\text{a customer joins}\} = \mu(1-\pi_0), \tag{23}$$

which essentially is an inflow-outflow balance equation of the queue in steady state. Leveraging equations (21), (22), and (23), we can represent π_j's as q_j's.

Now, it is left to derive q_j's and we consider the DTMC $\{X(t_n^+), n = 1, 2, 3, \ldots\}$. Let α_j be the probability that there are j arrivals during a service time slot. Since the arrival process is Poisson with rate λ, the number of arrivals during a service time slot is a Poisson random variable with mean $\lambda/\mu_S = \rho$. So

$$\alpha_j = \frac{\rho^j e^{-\rho}}{j!}.$$

Then we have the following balance equations for q_j's.

$$\begin{aligned}
q_0 &= \alpha_0 q_1 + \alpha_0 q_0 \\
q_1 &= \alpha_0 q_2 + \alpha_1 q_1 + \alpha_1 q_0 \\
&\cdots \\
q_j &= \alpha_0 q_{j+1} + \alpha_1 q_j + \alpha_2 q_{j-1} + \cdots + \alpha_j q_1 + \alpha_j q_0 \\
&\cdots \\
q_{K-2} &= \alpha_0 q_{K-1} + \alpha_1 q_{K-2} + \alpha_2 q_{K-3} + \cdots + \alpha_{K-2} q_1 + \alpha_{K-2} q_0
\end{aligned}$$

If there are $K-1$ customers in the system right after a departure, then there are K customers before that departure, meaning that an arrival joins with probability p. This is different from the situation where there are fewer customers in the system. So we have a different equation for q_{K-1} below.

$$q_{K-1} = \alpha'_0 q_K + \alpha'_1 q_{K-1} + \alpha'_2 q_{K-2} + \cdots + \alpha'_{K-1} q_1 + \alpha'_{K-1} q_0,$$

where

$$\alpha'_j = \alpha_j + (1-p)\alpha_{j+1} + (1-p)^2 \alpha_{j+2} + \cdots + (1-p)^k \alpha_{j+k} + \ldots. \tag{24}$$

The term α'_j is the probability that exactly j customers join the queue during a service time slot that starts with $K-j$ customers in the system; note that more than j customers may arrive, but those who see K customers ahead all balk. Equation (24) gives a detailed expression for α'_j, of which the i^{th} term means that there are $j+i-1$ arrivals and the last $i-1$ arrivals choose not to join (each independently with probability $1-p$) when facing a system size of K.

C.1. Pure Threshold ($p = 1$)

If $p = 1$, i.e., $\alpha'_j = \alpha_j$, this queue is reduced to the classic $M/D/1/K$ queue where K is an integer (so customers balk if the system size is $K+1$). For completeness and a better comparison next, we summarize its analyses from the existing literature below. As shown in Brun and Garcia (2000), q_j's can be calculated recursively such that $q_j = a_j q_0$, where $a_0 = 1$, $a_1 = e^\rho - 1$, and

$$a_j = e^\rho (a_{j-1} - \sum_{i=1}^{j-1} \alpha_i a_{j-i} - \alpha_{j-1} a_0), \quad 2 \le j \le K.$$

Let $b_j = \sum_{i=0}^{j} a_i$. One can show that

$$b_j = \sum_{i=0}^{j} \frac{((i-j)\rho)^i}{i!} e^{-(i-j)\rho}. \tag{25}$$

Since $\sum_{j=0}^{j=K} q_j = 1$, we have

$$q_0 = \frac{1}{b_K} \text{ and } q_j = \frac{b_j - b_{j-1}}{b_K},\ j = 1, 2, \dots, K.$$

One can then derive that π_j, the steady state probability of j customers in the system, is

$$\pi_j = \frac{q_j}{q_0 + \rho}$$

for $j = 0, 1, \dots, K$, and

$$\pi_{K+1} = 1 - \frac{1}{q_0 + \rho}.$$

See Gross (2008) for details.

C.2. Mixed Threshold $(0 < p < 1)$

When $0 < p < 1$, we still have $q_j = a_j q_0$ for $j \le K-1$, but $q_K = a'_K q_0$ where

$$a'_K = (e^{p\rho} - 1)(1-p)^{-K+1} + (e^{p\rho} - pe^{p\rho} - 1) \sum_{i=1}^{K-1} (b_i - b_{i-1})(1-p)^{-K+i}. \tag{26}$$

The derivation of (26) appears in the Appendix E.4. Then we can write q_j's as follows.

$$q_j = \begin{cases} \dfrac{1}{b_{K-1} + a'_K} & \text{if } j = 0, \\ \dfrac{b_j - b_{j-1}}{b_{K-1} + a'_K} & \text{if } 1 \le j \le K-1, \\ \dfrac{a'_K}{b_{K-1} + a'_K} & \text{if } j = K. \end{cases} \tag{27}$$

Proposition C.1. *For an $M/D/1/K-1+p$ queue, the steady state probability that there are j customers in the system is*

$$\pi_j = \begin{cases} \dfrac{q_j}{q_0 + \rho} & \text{if } 0 \le j \le K-1, \\ \dfrac{q_K}{p(q_0 + \rho)} & \text{if } j = K, \\ 1 - \dfrac{1}{q_0 + \rho} - \dfrac{(1-p)q_K}{p(q_0 + \rho)} & \text{if } j = K+1, \end{cases} \tag{28}$$

where q_j takes the form of (27).

Recall that the buffer size $K - 1 + p$ is $\delta\mu_S$ in our appointment queue. Follow our notations earlier, let $\pi(\delta, \mu_S)$ denote the blocking probability, i.e., the probability that an arrival does not join. Then it has the following closed-form and properties.

Corollary C.1. *In the $M/D/1/\delta\mu_S$ queue,*

$$\pi(\delta, \mu_S) = \pi_{K+1} + (1-p)\pi_K = 1 - \frac{1}{q_0 + \rho},$$

where $p = \delta\mu_S + 1 - K$ and q_0 follows (27).

Proposition C.2. *In the $M/D/1/K-1+p$ queue where $K=\lceil \delta\mu_S \rceil$ and $p=\delta\mu_S+1-K$, $\pi(\delta,\mu_S)$ is continuous and strictly decreasing in δ and μ_S, respectively, when other parameters are fixed.*

Remark C.2. Proposition C.2 suggests that the blocking probability decreases in the service rate and the delay threshold. This is intuitive but with one subtlety. When μ_S increases and everything else including the delay threshold δ being fixed, the buffer size $\delta\mu_S$ also increases. Both effects naturally lead to reduction in the blocking probability. Finally, note that Lemma 5 in Section 4.1 directly follows from Proposition C.2.

Appendix D: Additional Technical Details

D.1. Illustration of Proposition 1

Table 1 summarizes patient equilibrium in the three regimes and the conditions under which these regimes take place. Equivalent conditions can be shown as interchangeable to the conditions used in Proposition 1 and provide an alternative way to categorize equilibrium into different regimes. Note that the second-stage equilibrium (δ, p_W, p_B) exists for any strategy at the first stage and that it depends on $(\lambda_S, \lambda_W, \lambda_B)$.

Table 1 Summary of Equilibrium at the Second Stage

Regime	SnB	SWB	SnW
Condition	$u_W^2(0) \le -C_G$	$u_W^2(p_W^S) < -C_G < u_W^2(0)$	$u_W^2(p_W^S) \ge -C_G$
Equivalent Condition	$p_W^B = 0$	$p_B^S > p_W^B > 0$	$p_B^S \le p_W^B$
Delay Threshold δ	$\delta_B^S = \frac{R+C_G}{C_D}$	$\delta_B^S = \frac{R+C_G}{C_D}$	δ_W^S
p_W	0	p_W^B	p_W^S
p_B	$p_B^S = \pi(\delta_B^S)$	$p_B^S - p_W^B$	0

D.2. Summary of Possible Scenarios for the Global Equilibrium

LEMMA D.2. *For any given (μ_S, μ_W), one of the following scenarios must hold in the equilibrium.*

1. $u_S^1 = u_B^1 > u_W^1$, *the second stage equilibrium strategy must be SnB, and* $\lambda_W = 0$.
2. $u_S^1 = u_W^1 = u_B^1$, *and the second stage equilibrium strategy must be SWB.*
3. $u_S^1 = u_W^1 > u_B^1$, *the second stage equilibrium strategy must be SnW, and* $\lambda_B = 0$.
4. $u_S^1 > u_W^1$ *and* $u_S^1 > u_B^1$, *the second stage equilibrium strategy can be SnB, SWB or SnW, and* $\lambda_S = \lambda$.

Lemma D.2 suggests that, 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. Note that at the second stage, the delay threshold δ in equilibrium either makes the utility of scheduling an appointment the same as that of balking or that of walk-in. Lemma D.2 simplifies our equilibrium analysis into three possible scenarios below. Each scenario is characterized by a unique pair of δ and λ_S. We use superscripts a, b, and c to differentiate these scenarios. The ultimate form of the equilibrium depends on which scenario will be realized given the model parameters.

It is clear that δ^a and δ^c are unique (see Proposition 1). The following proposition is instrumental in proving that λ_S^a, δ^b and λ_S^b are also unique. Its proof leverages a novel use of the conditional workload process observed in the system.

PROPOSITION D.3. *Considering an M/D/1 queue with λ_S as the arrival rate and arrivals who see the delay over δ will not join the queue. Recall that $\pi(\delta, \lambda_S)$ is the probability that an arrival will not join the queue and $\mathbb{E}_{\tilde{d}}[\tilde{d}|\tilde{d} \le \delta, \lambda_S]$ is the expected delay of the queue seen by a joining customer. The following results hold.*

1. *When δ is fixed, $\pi(\delta, \lambda_S)$ is strictly increasing in λ_S, and $\mathbb{E}_{\tilde{d}}[\tilde{d}|\tilde{d} \le \delta, \lambda_S]$ is increasing λ_S.*

2. When δ solves $u_S^2 = u_W^2$, i.e., δ solves (13) for a given λ_S, then δ is decreasing in λ_S, $\pi(\delta, \lambda_S)$ is strictly increasing in λ_S, and $\delta - \mathbb{E}_{\tilde{d}}[\tilde{d}|\tilde{d} \le \delta, \lambda_S]$ is decreasing λ_S.

D.3. Equilibrium in the Observable Setting

Lemma 4 indicates that, in appointment system where $C_G = 0$, patients behave as if there were no first stage decision, and the appointment system is equivalent to the system where the appointment delay is always observable to patients. In the remaining of this paper, we call it "Observable Setting".

In observable setting, a strategic patient goes to second stage directly and chooses based on the observed delay (following the threshold-based joining strategy described in Proposition 1). So we use (δ, p_W, p_B) to represent patients' strategy in this setting. Specifically, the equilibrium described in Theorem 1 can be simplified as (δ, p_W, p_B):

1. If $R - C_W w(\frac{\lambda_E}{\mu_W}) \le 0$, then $\delta = \delta_B^S$, $p_W = 0$, $p_B = p_B^S$;
2. If $R - C_W w(\frac{\lambda_E}{\mu_W}) > 0 > R - C_W w(\frac{p_W^S \lambda + \lambda_E}{\mu_W})$, then $\delta = \delta_B^S$, $p_B = p_B^S - p_W^B$;
3. If $R - C_W w(\frac{p_W^S \lambda + \lambda_E}{\mu_W}) \ge 0$, then $\delta = \delta_W^S$, $p_W = p_W^S$, $p_B = 0$;

where

- $\pi(\delta, \mu_S)$ represents the probability that the delay is over δ in an $M/D/1$ queue with μ_S as the service rate, λ as the arrival rate, and δ as the delay threshold;
- $\delta_B^S = \frac{R}{C_D}$;
- $p_B^S = \pi(\delta_B^S, \mu_S)$;
- δ_W^S solves $C_D \delta_W^S = C_W \times w(\frac{\pi(\delta_W^S, \mu_S)\lambda + \lambda_E}{\mu_W})$;
- $p_W^S = \pi(\delta_W^S, \mu_S)$;
- $p_W^B = \max\{\min\{\frac{w^{-1}(R/C_W)\mu_W - \lambda_E}{\lambda}, 1\}, 0\}$.

The interpretation of the equilibrium strategy above is same to that of the equilibrium strategy in the second stage (see Proposition 1 and Table 1).

D.4. Optimization by Regime in the Observable Setting

Let $\mu_W = \mu - \mu_S$.

Regime SnB:

$$\min_{\mu_S \in [0, \mu]} \quad C_L \pi(\frac{R}{C_D}, \mu_S)\lambda + C_O o(\frac{\lambda_E}{\mu - \mu_S}) \tag{Ob.SnB}$$
$$\mu - \mu_S \le \frac{\lambda_E}{w^{-1}(\frac{R}{C_W})}.$$

Regime SWB:

$$\min_{\mu_S \in [0, \mu]} \quad C_L \left[\lambda \pi(\frac{R}{C_D}, \mu_S) - w^{-1}(\frac{R}{C_W})(\mu - \mu_S) + \lambda_E\right] \tag{Ob.SWB}$$
$$+ C_O o\left[w^{-1}(\frac{R}{C_W})\right]$$
$$\mu - \mu_S \ge \frac{\lambda_E}{w^{-1}(\frac{R}{C_W})},$$
$$\mu - \mu_S \le \frac{\pi(\frac{R}{C_D}, \mu_S)\lambda + \lambda_E}{w^{-1}(\frac{R}{C_W})}.$$

Regime SnW:

$$\min_{\mu_S \in [0,\mu]} \quad C_O o \left[\frac{\pi(\delta_W^S, \mu_S)\lambda + \lambda_E}{\mu - \mu_S} \right] \tag{Ob.SnW}$$

$$C_D \delta_W^S = C_W w \left[\frac{\pi(\delta_W^S, \mu_S)\lambda + \lambda_E}{\mu - \mu_S} \right],$$

$$\mu - \mu_S \geq \frac{\pi(\frac{R}{C_D}, \mu_S)\lambda + \lambda_E}{w^{-1}(\frac{R}{C_W})}.$$

D.5. Equilibrium in the Unobservable Setting

Lemma D.3. *When $C_G = +\infty$, the patient equilibrium strategy described in Theorem 1 can be simplified as a triplet of $(\lambda_S, \lambda_W, \lambda_B)$ with values specified below.*

1. *If $R - C_W w(\frac{\lambda_E}{\mu_W}) \leq 0$, then $\lambda_S = \lambda_S^B$, $\lambda_W = 0$, and $\lambda_B = \lambda_B^S$;*
2. *If $R - C_W w(\frac{\lambda_W^S + \lambda_E}{\mu_W}) < 0 < R - C_W w(\frac{\lambda_E}{\mu_W})$, then $\lambda_S = \lambda_S^B$, $\lambda_W = \lambda_W^B$, and $\lambda_B = \lambda_B^S - \lambda_W^B$;*
3. *If $R - C_W w(\frac{\lambda_W^S + \lambda_E}{\mu_W}) \geq 0$, then $\lambda_S = \lambda_S^W$, $\lambda_W = \lambda_W^S$, and $\lambda_B = 0$;*

- *$\lambda_S^B = \min\{\frac{2R\mu_S^2}{C_D + 2R\mu_S}, \lambda\}$, and $\lambda_B^S = \lambda - \lambda_S^B$;*
- *$\lambda_W^B = \max\{\min\{w^{-1}(\frac{R}{C_W})\mu_W - \lambda_E, \lambda\}, 0\}$;*
- *$\lambda_S^W = \min\{\lambda_S', \lambda\}$, where λ_S' solves $u_S^1(\lambda_s') = u_W^1(\lambda - \lambda_S')$, $\lambda_W^S = \lambda - \lambda_S^W$, and $u_S^1(\cdot)$ and $u_W^1(\cdot)$ are defined in* (18) *and* (19)*, respectively.*

In Lemma D.3, λ_S^B is the equilibrium joining rate of strategy patients if they only have the options of scheduling and balking; λ_W^B is the equilibrium walk-in rate of strategy patients if they only have the options of walk-in and balking; λ_S^W (λ_W^S resp.) is the equilibrium joining (walk-in resp.) rate of strategy patients if they only have the options of scheduling and walk-in. Table 2 below summarizes the equilibrium results in the unobservable setting.

Table 2 Summary of Equilibrium In the Unobservable Setting

Regime	SnB	SWB	SnW
Condition	$u_W^1(0) \leq 0$	$u_W^1(\lambda_W^S) < 0 < u_W^1(0)$	$u_W^1(\lambda_W^s) \geq 0$
Equivalent Condition	$\lambda_W^B = 0$	$\lambda_B^S > \lambda_W^B > 0$	$\lambda_B^S \leq \lambda_W^B$
λ_S	λ_S^B	λ_S^B	λ_S^W
λ_W	0	λ_W^B	λ_W^S
λ_B	λ_B^S	$\lambda_B^S - \lambda_W^B$	0

D.6. Optimization by Regime in the Unobservable Setting

Depending on the values of (μ_S, μ_W), patient equilibrium falls into three different regimes shown in Figure 2(b). We can decompose the problem above by regime and the one with the lowest objective value is where the provider should operate.

Regime SnB:

$$\min_{\mu_S\in[0,\mu]} \ C_L\lambda_B^S + C_O o(\frac{\lambda_E}{\mu-\mu_S}) \tag{Un.SnB}$$
$$\lambda_B^S = \max\left\{\lambda, \lambda - \frac{2R(\mu_S)^2}{2R\mu_S + C_D}\right\},$$
$$\mu - \mu_S \le \frac{\lambda_E}{w^{-1}(\frac{R}{C_W})}.$$

Regime SWB:

$$\min_{\mu_S\in[0,\mu]} \ C_L\left[\lambda_B^S - w^{-1}(\frac{R}{C_W})(\mu-\mu_S) + \lambda_E\right] + C_O o\left[w^{-1}(\frac{R}{C_W})\right] \tag{Un.SWB}$$
$$\lambda_B^S = \max\left\{\lambda, \lambda - \frac{2R(\mu_S)^2}{2R\mu_S + C_D}\right\},$$
$$\mu - \mu_S \ge \frac{\lambda_E}{w^{-1}(\frac{R}{C_W})},$$
$$\mu - \mu_S \le \frac{\lambda_B^S + \lambda_E}{w^{-1}(\frac{R}{C_W})}.$$

Regime SnW:

$$\min_{\mu_S\in[0,\mu]} \ C_O o(\frac{\lambda_W^S + \lambda_E}{\mu-\mu_S}) \tag{Un.SnW}$$
$$C_D\frac{\lambda - \lambda_W^S}{2\mu_S(\mu_S - \lambda + \lambda_W^S)} = C_W w(\frac{\lambda_W^S + \lambda_E}{\mu - \mu_S}),$$
$$\mu - \mu_S \ge \frac{\lambda_W^S + \lambda_E}{w^{-1}(\frac{R}{C_W})}.$$

D.7. "Bang-Bang" Result in the Unobservable Setting

Proposition 5 offers insights similar to those in Proposition 3 for the observable setting. In particular, $\overline{\mu}_U$ is the smallest capacity required to induce all strategic patients into the SnW regime under the unobservable setting; $\underline{\mu}_W$ is defined earlier in (15). Figure 2(b) illustrates the results. One can further show that $\overline{\mu}_S(\mu_W)$ is decreasing and concave when $\mu_W \in [\underline{\mu}_W, \overline{\mu}_W]$. This allows us to obtain a closed-form expression for $\overline{\mu}_U$ and a more specific characterization of the optimal capacity allocation in the unobservable setting.

LEMMA D.4. *In the unobservable setting, the following results hold.*

1. $\overline{\mu}_U = \min\{\overline{\mu}_S + \underline{\mu}_W, \overline{\mu}_W\}$.
2. *Under the optimal capacity allocation, either* $\lambda_S = 0$ *or* $\lambda_W = 0$.

Lemma D.4 indicates that strategic patients would never mix between scheduling an appointment and walk-in under optimal capacity allocation in the unobservable setting. Together with Proposition 5, this lemma suggests that when the total capacity is sufficiently large, i.e., $\mu \ge \overline{\mu}_U$, 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.

D.8. Triage Model in the Unobservable Setting

Under the unobservable setting, the optimal capacity allocation problem of the triage model can be formulated as follows.

$$\min_{\mu_S\in[0,\mu]} \ C_L(\lambda - \lambda_S) + C_O o(\frac{\lambda_E}{\mu-\mu_S}) \tag{T.Un.P}$$
$$\lambda_S = \min\left\{\lambda, \ \frac{2R(\mu_S)^2}{2R\mu_S + C_D}\right\}.$$

Similar to the observable setting, when $\mu \le \underline{\mu}_W$, the strategic model behaves the same as the triage model; see Figure 2(b). Hence we focus on the case when $\mu > \underline{\mu}_W$. We define the following two cost ratio thresholds for the unobservable setting.

$$\overline{\beta} = \frac{o'\left(\frac{\lambda_E}{\underline{\mu}_W}\right)}{\underline{\mu}_W} \cdot \max\left\{\frac{\lambda_E}{\lambda_B^S(\mu-\underline{\mu}_W)}, 1\right\}, \quad \text{and} \quad \underline{\beta} = \frac{o'\left(\frac{\lambda_E}{\mu}\right)}{\mu}, \quad \text{where} \quad \lambda_B^S(\mu-\underline{\mu}_W) = \lambda - \frac{2R(\mu-\underline{\mu}_W)^2}{C_D + 2R(\mu-\underline{\mu}_W)}.$$

Note that $\lambda_B^S(\mu-\underline{\mu}_W)$ is the balking rate of strategic patients when they are provided with only two choices, scheduling and balking, in the unobservable setting with $\mu_S = \mu - \underline{\mu}_W$. If $\lambda_B^S(\mu-\underline{\mu}_W) < 0$, it means no one balks. Then we have the following comparison results.

PROPOSITION D.4. *Consider the same set of model parameters and suppose that* $\mu > \underline{\mu}_W$,

1. *if* $\mu \ge (\lambda+\lambda_E)\cdot\max\{\frac{\overline{\mu}_S}{\lambda}, \frac{\mu_W}{\lambda_E}\}$ *or* $\frac{C_L}{C_O} \le \underline{\beta}$, *the triage model costs less.*
2. *if* $\mu \le \overline{\mu}_S + \underline{\mu}_W$ *and* $\frac{C_L}{C_O} \ge \overline{\beta}$, *the strategic model costs less.*

Proposition D.4 reveals similar insights to those in Proposition 6 in terms of the impact due to 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. Recall that $\overline{\mu}_S$ is the minimum capacity to attract all strategic patients from balking in the unobservable setting. When the total capacity is large enough (case 1), the triage model can set $\mu_S = \overline{\mu}_S$ to attract all strategic patients to the appointment queue and then allocate the remaining capacity to serve exogenous walk-ins. In this case, strategic patients have a zero utility, while exogenous walk-ins have a positive utility. In the strategic model, however, strategic patients would choose to walk in to improve their utilities, making walk-in hours more crowded and the overtime cost higher. A sufficiently large capacity allows the triage model to benefit from "forcing" all strategic patients to stay in the appointment queue and at the same time having less congested walk-in hours and thus a lower total cost. In order for the strategic model to stand out, the total capacity needs to be small enough such that the triage model cannot take advantage of strategic patients by serving all of them at zero utility while maintaining a positive utility for exogenous walk-ins (case 2).

Appendix E: Proofs of Results

E.1. Proofs of Results in Section 3

Proof of Lemma 1: Suppose patients adopt second-stage strategy $(p_S^2(\tilde{d}), p_W^2(\tilde{d}), p_B^2(\tilde{d}))$ with $\tilde{d}_1$ and $\tilde{d}_2$ such that $\tilde{d}_1 < \tilde{d}_2$, $p_S^2(\tilde{d}_1) < 1$, and $u_S^2(\tilde{d}_2) = \max\{u_W^2, 0\}$. Then we must have $u_S^2(\tilde{d}_1) > u_S^2(\tilde{d}_2) = \max\{u_W^2, 0\}$. Letting $p_S^2(\tilde{d}_1) = 1$ yields a strictly higher utility. This leads to a contradiction and completes the proof.

□

Proof of Lemma 2 Given the capacity (μ_S, μ_W) and the first-stage strategy $(\lambda_S, \lambda_W, \lambda_B)$, by Lemma 1, patients schedule in the second stage when the observed delay is no larger than δ. Then the appointment queue becomes an M/D/1 queue with arrival rate λ_S, service rate μ_S, and delay threshold δ. Next, we examine the monotonicity of its blocking probability.

For a period from time 0 to time T, we look at a sample path of arrival process, denoted by a_1, a_2, ..., a_{N_T}, where a_i is the time interval between i^{th} arrival and $i+1^{th}$ arrival. Let $\tilde{d}_i$ denote the delay saw by $i+1^{th}$ arrival, i.e., the workload of the server just before $i+1^{th}$ arrival. Let binary variable z_i denote whether i^{th} arrival is blocked (she is blocked if there is more than δ workload just before her arrival, then $z_i = 0$). So the total number of unblocked patients during $[0, T]$ for this sample path is

$$\{\sum_{i=1}^{N_T} z_i | z_i = 1 \text{ if } \tilde{d}_{i-1} \le \delta,\ z_i = 0 \text{ o.w.};\ \tilde{d}_i = (\tilde{d}_{i-1} + \frac{z_i}{\mu_S} - a_i)^+;\ \tilde{d}_0 = 0\}.$$

Let $f(\delta)$ denote this value. It is obvious that $f(\delta)$ equals to the objective value of the following optimization problem,

$$\max_{z_i \in \{0,1\}} \sum_{i=1}^{N_T} z_i \tag{Lem.2.P1}$$

$$\tilde{d}_i = (\tilde{d}_{i-1} + \frac{z_i}{\mu_S} - a_i)^+, \text{ for } 0 \le i \le N_T$$

$$\tilde{d}_{i-1} + \frac{z_i}{\mu_S} \le \delta + \frac{1}{\mu_S}, \text{ for } 0 \le i \le N_T \tag{29}$$

$$\tilde{d}_0 = 0$$

Let $\boldsymbol{z}^*(\delta)$ be the optimal solution for the problem above. Now, consider a same sample path but a larger $\delta' > \delta$. It is obvious that if these N_T arrivals are blocked as $\boldsymbol{z}^*(\delta)$, the delay saw by an arrival is always no more than δ', indicating that $\boldsymbol{z}^*(\delta)$ is a feasible solution for (Lem.2.P1) with δ'. So $f(\delta) \le f(\delta')$. For any T and any sample path, we will have $f(\delta) \le f(\delta')$, which means $\pi(\delta) \ge \pi(\delta')$ if $\delta' > \delta$.

Next we show the strong monotonicity. While this result seems intuitive, a simple coupling argument may not work because we cannot couple the busy periods of two systems with different threshold. Our approach involves constructing renewal reward processes and sample path arguments in renewal cycles. We consider a renewal reward process (called $\underline{\text{RP}}$) which renews whenever the size of the system with threshold δ jumps from 0 to 1 and the reward is the number of customers served in a renewal cycle. Let $\underline{R}$ and τ represent the random reward and the cycle length in this process. Then, for the system with δ, the long-run average rate of customers served $\lambda \times (1 - \pi(\delta)) = \frac{E(\underline{R})}{E(\tau)}$. We next construct another renewal reward process (called $\overline{\text{RP}}$) which renews exactly at the same time when $\underline{\text{RP}}$ renews but has a strictly higher expected per-period reward than that of $\underline{\text{RP}}$. Finally, we argue that the long-run average reward of $\overline{\text{RP}}$ is no greater than the long-run average rate of customer served in the system with $\delta' > \delta$. It follows that $\pi(\delta) > \pi(\delta')$ if $\delta' > \delta$.

Without loss of generality, we suppose that a cycle starts at $t = 0$ and i is the index of the last arrival in the cycle of the process $\underline{\text{RP}}$. Construct $\overline{\text{RP}}$ in this way. Imagine a single-server queueing process with threshold $\delta' > \delta$ and it accepts customers following the principle below. Recall that z_j denotes that whether the j^{th} customer is accepted and served in process $\underline{\text{RP}}$ and we let z'_j denote that in process $\overline{\text{RP}}$. For all sample paths, set $z'_j = z_j$ for all $j < i$. Thus $\tilde{d}'_j = \tilde{d}_j$ for all $j < i$. We further claim we can find some sample path φ such that $\tilde{d}_{i-1} + \frac{1}{\mu_S} > \delta + \frac{1}{\mu_S}$, $\tilde{d}_{i-1} \le \delta + \frac{1}{\mu_S}$, $\tilde{d}'_{i-1} + \frac{1}{\mu_S} \le \delta' + \frac{1}{\mu_S}$, $\tilde{d}_{i-1} \le a_i$ and $\tilde{d}'_{i-1} + \frac{1}{\mu_S} \le a_i$. Then on this sample path, z_i must be 0, but z'_i can be 1 without violating constraint (29). For other sample paths, we have $z'_j = z_j, \forall j \le i$. Since $\tilde{d}_i = \tilde{d}'_i = 0$, for all sample paths, both queueing processes corresponding to $\underline{\text{RP}}$ and $\overline{\text{RP}}$ are empty after the service of ith customer, and the cycle ends before $i + 1^{th}$ customer arriving. By this way of construction, we ensure that the reward collected in each cycle for $\overline{\text{RP}}$ are i.i.d. random variables and the renewal reward theorem applies. As long as sample path φ has a non-zero measure, then we have $E[\sum_{j=1}^{i} z_j] < E[\sum_{j=1}^{i} z'_j]$, i.e., $E[\underline{R}] < E[\overline{R}]$. Since both processes share the same renewal cycles (the cycle starts at $t = 0$ and ends before $i + 1^{th}$ arrival), then we have the long-run average rate of customers served in process $\underline{\text{RP}}$ is strictly smaller than that in process $\overline{\text{RP}}$ by the renewal reward theorem. One can also check that $\boldsymbol{z}$ as we define in process $\overline{\text{RP}}$ is a feasible solution to problem eqn:sample.path.zi.delta with respect to δ', so the long-run average reward in $\overline{\text{RP}}$ is a lower bound for the long-run average rate of customers served in system with δ'. Combining arguments above, we have $\lambda(1 - \pi(\delta)) < \lambda(1 - \pi(\delta'))$ if $\delta' > \delta$.

Finally, it is left to show the measure of such sample path φ is strictly greater than 0 in a renewal cycle. Specifically, we need $a_i \ge \max\{\tilde{d}_{i-1}, \tilde{d}'_{i-1} + \frac{1}{\mu_S}\}$, since a_i follows exponential distribution, then the measure of such a_i is greater than 0. Thus the measure of sample path φ is strictly greater than 0.

□

Proof of Lemma 3:

1. Since $u_S^2(\tilde{d}) = R - C_D\tilde{d}$ is strictly decreasing in delay $\tilde{d}$, δ_B^S is unique.

2. Since $u_W^2(p_W) = R - C_W w(\frac{p_W\lambda_S + \lambda_W + \lambda_E}{\mu_W}) - C_G$ is strictly decreasing in p_W, δ_W^B is unique.

3. Since $\pi(\delta)$ is strictly decreasing and continuous in δ, then $u_W^2(\pi(\delta))$ is strictly increasing and continuous in δ. Notice that $u_S^2(\tilde{d})$ is strictly decreasing and continuous in $\tilde{d}$, so δ_W^S exists and is unique.

□

Proof of Proposition 1:

Existence:

It is easy to check that the equilibrium in each case satisfies all conditions (6) to (10).

Uniqueness:

Case 1: $u_W^2(0) \le -C_G$ indicates $u_W^2(p_W) < -C_G$ for any $p_W > 0$. So Walk-in option can be ignored ($p_W = 0$). By Lemma 3.1, we know there is a unique delay threshold δ_B^S such that $u_S^2(\delta_B^S) = -C_G$. Since we rule out tied cases, so patients join the appointment queue when delay $\tilde{d} \le \delta_B^S$, and balk otherwise. And $p_B = \pi(\delta_B^S) = p_B^S$ is unique.

Case 2: $u_W^2(p_W^S) < -C_G$ indicates $u_S^2(\delta_W^S) < -C_G$. By the definitions of δ_W^S and δ_B^S, we have $\delta_B^S < \delta_W^S$. So the delay threshold must be δ_B^S, and it is unique by Lemma 3.1. Since we rule out tied cases, patients join the appointment queue when delay $\tilde{d} \le \delta_B^S$. Then the probability of no scheduling is $\pi(\delta_B^S) = p_B^S$. The fact that $u_W^2(0) > -C_G$ indicates that when walk-in probability is small enough, Walk-in is more attractive than Balk. So when patients do not schedule, some of them will walk in to make the utility of Walk-in is equal to that of Balk. By Lemma 3.2, we know the corresponding walk-in proportion p_W^B is unique. Since $u_W^2(p_W^S) = u_S^2(\delta_W^S) < -C_G = u_W^2(p_W^B)$, so $p_W^B < p_W^S$. Since $\delta_B^S < \delta_W^S$, then $p_B^S > p_W^S$. So $p_B^S - p_W^B$ is positive and unique.

Case 3: $u_W^2(p_W^S) \ge -C_G$ indicates $u_S^2(\delta_W^S) \ge -C_G$. Since we rule out tied cases, then Balk option can be ignored ($p_B = 0$). By Lemma 3.3, we know there is a unique threshold δ_W^S.

□

Proof of Lemma D.2:

Notice that $u_W^1 - u_B^1 = u_W^2 - u_W^2$ and $u_W^1 - u_W^2 = C_G$.

If $u_B^1 < u_W^1$, indicating $u_B^2 < u_W^2$, then we must have SnW at the second stage.

If $u_B^1 > u_W^1$, indicating $u_B^2 > u_W^2$, then we must have SnB at the second stage.

□

Proof of Proposition D.3:

1. Let us prove that when δ is fixed, $\pi(\delta, \lambda_S)$ is strictly increasing in λ_S, and $\mathbb{E}[\tilde{d}|\tilde{d} \le \delta, \lambda_S]$ is increasing λ_S:

(a) Now we prove that $\pi(\lambda_S)$ is strictly increasing in λ_S (let us omit constant δ here):

Let us look at a sample path of arrival process with arrival rate λ_S, denoted by a_1, a_2, ..., a_N, where a_i is the time interval between i^{th} arrival and $i+1^{th}$ arrival. Let $\tilde{d}_i$ denote the delay seen by $i+1^{th}$ arrival, i.e., the workload of the server just before $i+1^{th}$ arrival. Let binary variable z_i denote whether i^{th} arrival is blocked (she is blocked if there is more than δ workload just before her arrival, then $z_i = 0$). So the total number of unblocked patients for this sample path is

$$\{\sum_{i=1}^{N} z_i | z_i = 1 \text{ if } \tilde{d}_{i-1} \le \delta,\ z_i = 0 \text{ o.w.};\ \tilde{d}_i = (\tilde{d}_{i-1} + \frac{z_i}{\mu_S} - a_i)^+;\ \tilde{d}_0 = 0\}.$$

Let $f(\lambda_S)$ denote this value. It is obvious that $f(\lambda_S)$ equals to the objective value of the following optimization problem,

$$\max_{z_i \in \{0,1\}} \sum_{i=1}^{N} z_i \qquad \text{(PropD.3.P1)}$$
$$\tilde{d}_i = (\tilde{d}_{i-1} + \frac{z_i}{\mu_S} - a_i)^+, \text{ for } 0 \le i \le N$$
$$\tilde{d}_{i-1} + \frac{z_i}{\mu_S} \le \delta + \frac{1}{\mu_S}, \text{ for } 0 \le i \le N$$
$$\tilde{d}_0 = 0$$

Let $\boldsymbol{z}^*(\lambda_S)$ be the optimal solution for the problem above.

For any sample path a_1, a_2, ..., a_N we can match it with a sample path of arrival process with arrival rate λ_S', denoted by a_1', a_2', ..., a_N', in the following way: let $Pr(\tilde{e}' < a_i') = Pr(\tilde{e} < a_i)$ where $\tilde{e}$ follows exponential distribution with $\frac{1}{\lambda_S}$ as the mean and $\tilde{e}'$ follows exponential distribution with $\frac{1}{\lambda_S'}$ as the mean. If $\lambda_S' < \lambda$, then $a_i' > a_i$. Consider such a sample path a_1', a_2', ..., a_N', it is obvious that if these N arrivals are blocked as $\boldsymbol{z}^*(\lambda_S)$, $\tilde{d}_i' \leq \tilde{d}_i$ for any i, indicating that $\boldsymbol{z}^*(\lambda_S)$ is a feasible solution for (PropD.3.P1) with λ_S'. So $f(\lambda_S) \leq f(\lambda_S')$. For any sample path, we can match it in the above way and have $f(\lambda_S) \leq f(\lambda_S')$, which means $\pi(\lambda_S) \geq \pi(\lambda_S')$ if $\lambda_S' > \lambda_S$.

As for strong monotonicity, we can follow the arguments in Proof of Lemma 2.

(b) Then we prove that $\mathbb{E}[\tilde{d}|\tilde{d} \leq \delta, \lambda_S]$ is increasing λ_S for fixed δ:

With slight abuse of notations, in this proof, we use $w(t)$ denote the stochastic process of *workload* of the server in the M/D/1 queue with δ as the delay threshold (we use w denote $w(t)$ under steady state). The workload of the server $w(t)$ represents the amount of the time needed by the server to serve all customers in the system at time t (including the remaining service time of the customer being served). Then we have

$$\mathbb{E}[\tilde{d}|\tilde{d} \leq \delta, \lambda_S] = \mathbb{E}[w|w \leq \delta, \lambda_S] \tag{30}$$

Let us only track the stochastic process of the conditional workload, i.e., $w(t)|w(t) \leq \delta$, and any time when $w(t) > \delta$ is cut off. We will use the corresponding conditional sample path (CSP) to represent such stochastic process, and use $w_c(t)$ to denote the corresponding workload at time t.

Now we show that the interarrival times in the CSP are still exponentially distributed with mean $\frac{1}{\lambda_S}$. First, observe that the arrivals to the CSP are exactly the arrivals to the original queue that occur while the workload is no larger than δ; as all other arrivals are not part of the conditional sample path. For the interarrival times, there are two cases to consider then: 1) Suppose that some arbitrary arrival to the CSP leaves the system with a workload below δ. In that case, we are sure that the next arrival is also part of the CSP. Thus, the interarrival between those two arrivals in the CSP is exponential by assumption, since it corresponds to the interarrival time between two arrivals in the original queue. 2) Now, suppose that some arrival i brings the workload in the original queue above δ. In that case, several customers may arrive in the original queue while the workload is above δ, and these arrivals are not part of the CSP. However, at some time (denote t) the workload of the original queue drops to δ again. Note that the interarrival time between arrival i and the next arrival *in the CSP* equals the time between t and the first arrival after t in the original queue. Let t' denote the last arrival in the original queue that arrives *before* t. Now consider the state of the original system at time t, and denote by X the interarrival time between the arrival at t' and the first arrival after t. X is exponentially distributed, and conditional on there being no arrival between t' and t in the original queue, the remaining time until the next arrival is distributed as $X - (t - t')|X > t - t'$, which has the same distribution as X by properties of the exponential distribution. Hence, the time between t and the first arrival after t is exponentially distributed.

In either case, the interarrival times between arrivals to the CSP are exponentially distributed, and hence the arrival process to the CSP is a Poisson process. Also, the service process is: if $w_c(t)$ is below $\delta - \frac{1}{\mu_S}$, any arrival will result in a $\frac{1}{\mu_S}$ increment in $w_c(t)$, otherwise any arrival will result in a $\delta - w_c(t)$ increment in $w_c(t)$. We let w_c be the $w_c(t)$ under steady state.

So we have

$$\mathbb{E}[\tilde{d}|\tilde{d} \leq \delta, \lambda_S] = \mathbb{E}[w|w \leq \delta, \lambda_S] = \mathbb{E}[w_c|\lambda_S] \tag{31}$$

For a period from time 0 to time T, we look at a sample path SP of Poisson arrival process, denoted by t_1, t_2, ..., t_{N_T}, where t_i is the arrival time of i^{th} arrival.

Consider an additional arrival happens at t_a, w.l.g., we assume $t_n \leq t_a < t_{n+1}$.

Let $w'_c(t)$ denote the workload at time t with the sample path SP plus the additional arrival at t_a. It is easy to check that

$$w'_c(t) = w_c(t) \text{ when } t < t_a \tag{32}$$

$$w'_c(t_a) = \min\{w_c(t_a) + \frac{1}{\mu_S}, \delta\} \tag{33}$$

$$w'_c(t) \geq w_c(t) \text{ when } t > t_a \tag{34}$$

So $w'_c(t) \geq w_c(t)$ at any t for any additional arrival and sample path SP.

Next we compare two CSPs with any different arrival rates, i.e., λ_S and $\lambda_S + \Delta\lambda_S$ respectively. We let $w_c(t|SP)$ denote the workload of CSP at time t under any sample path SP and $w_c(t|SP + t_a)$ denote the workload of CSP at time t under the sample path SP plus an additional arrival t_a. Note that $w_c(t|SP) = w_c(t)$ and $w_c(t|SP + t_a) = w'_c(t)$.

Consider a sample path of Poisson arrival process with arrival rate $\Delta\lambda_S$, denoted by $\Delta SP = (t_a^1, t_a^2, t_a^3, ...)$.

By the results above we know that $w_c(t|SP + t_a^1) \geq w_c(t|SP)$, then by induction we know $w_c(t|SP + \Delta SP) \geq \cdots \geq w_c(t|SP + t_a^1 + t_a^2 + \cdots + t_a^i) \geq \cdots \geq w_c(t|SP + t_a^1 + t_a^2) \geq w_c(t|SP + t_a^1) \geq w_c(t|SP)$.

By the memoryless property of Poisson, the arrival process with arrival rate $\lambda_S + \Delta\lambda_S$ can be decomposed to the arrival process with arrival rate λ_S and the arrival process with arrival rate $\Delta\lambda_S$. Thus we can independently construct a sample path SP for λ_S and a sample path ΔSP for $\Delta\lambda_S$. Since $w_c(t|SP + \Delta SP) \geq w_c(t|SP)$ at any t for any SP and ΔSP, thus $\mathbb{E}[w_c|\lambda_S + \Delta\lambda_S] \geq \mathbb{E}[w_c|\lambda_S]$. By (31), we know $\mathbb{E}[\tilde{d}|\tilde{d} \leq \delta, \lambda_S]$ is increasing in λ_S.

Now we have already proved that when δ is fixed, $\pi(\delta, \lambda_S)$ is strictly increasing in λ_S, and $\mathbb{E}[\tilde{d}|\tilde{d} \leq \delta, \lambda_S]$ is increasing λ_S.

2. Let δ be the solution to $u_S^2 = u_W^2$, i.e., δ solves (13) for a given λ_S, next we will give the proof in three steps:

(a) δ is decreasing in λ_S;

(b) $\pi(\delta, \lambda_S)$ is strictly increasing in λ_S;

(c) $\delta - \mathbb{E}[\tilde{d}|\tilde{d} \leq \delta, \lambda_S]$ is decreasing λ_S.

The details follow:

(a) We first show that δ is decreasing in λ_S:

i. First, by Lemma 2, we know that, given λ_S, $\pi(\delta, \lambda_S)$ is strictly decreasing in δ.

ii. Then we will show that, given δ, $[1 - \pi(\delta, \lambda_S)] \times \lambda_S$ is increasing in λ_S:

Let us omit the notation of δ since it is fixed, and let $w(t)$ denote the stochastic process of workload in the M/D/1 queue with δ as delay threshold. It is easy to check that

$$\frac{[1 - \pi(\lambda_S)]\lambda_S}{\mu_S} = \lim_{T \to +\infty} \frac{\int_{t=1}^{T} w(t)dt}{T} = E[w|\lambda_S].$$

So showing $[1-\pi(\lambda_S)]\lambda_S$ is increasing in λ_S is equivalent to showing $\lim\limits_{T\to+\infty}\frac{\int_{t=1}^{T}w(t)dt}{T}$ is increasing in λ_S.

For a period from time 0 to time T, we look at a sample path SP of Poisson arrival process, denoted by t_1, t_2, ..., t_{N_T}, where t_i is the arrival time of i^{th} arrival. Let $w(t)$ denote the workload at time t with this arrival sample path SP.

Consider an additional arrival happens at t_a, w.l.g., we assume $t_n \le t_a < t_{n+1}$. Let $w'(t)$ denote the workload at time t with the sample path SP plus the additional arrival at t_a. It is easy to check that

$$w'(t)=w(t) \text{ when } t<t_a.$$

Let t_b which is greater than t_a denote the first time such that $w'(t_b)=w(t_b)$. Thus

$$w'(t)=w(t) \text{ when } t\ge t_b.$$

If t_b does not exist, then it already shows that $\lim\limits_{T\to+\infty}\frac{\int_{t=1}^{T}w'(t)dt}{T}>\lim\limits_{T\to+\infty}\frac{\int_{t=1}^{T}w(t)dt}{T}$.

Next we assume t_b exists and it is sufficient to compare $\int_{t=t_a}^{t_b}w(t)dt$ and $\int_{t=t_a}^{t_b}w'(t)dt$. There are two situations:

- If $w(t_a)>\delta$, then the additional arrival at t_a to the sample path will not join the queue, then $w'(t_a)=w(t_a)$, thus $\int_{t=t_a}^{t_b}w(t)dt=\int_{t=t_a}^{t_b}w'(t)dt$;
- If $w(t_a)\le\delta$, then the additional arrival at t_a to the sample path will join the queue, then $w'(t_a)=w(t_a)+\frac{1}{\mu_S}$. Let $\{t_c^1,t_c^2,...,t_c^i,...\}$ denote the time points such that $w(t_c^i-\Delta t)<w'(t_c^i-\Delta t)$ and $w(t_c^i)>w'(t_c^i)$, where Δt is a sufficiently small time interval.

—If $\{t_c^i\}$ is empty, then $w'(t)\ge w(t)$ for $t_a\le t\le t_b$, thus $\int_{t=t_a}^{t_b}w(t)dt>\int_{t=t_a}^{t_b}w'(t)dt$;

—If $\{t_c^i\}$ is not empty, since $w'(t_a)=w(t_a)+\frac{1}{\mu_S}$, then before any t_c^i, there must be a time point $t_d^i\in(t_c^{i-1},t_c^i)$ such that $w(t_d^i-\Delta t)\ge w'(t_d^i-\Delta t)$ and $w(t_d^i)<w'(t_d^i)$. Specifically, $t_d^1=t_a$

$*$ We must have $w(t_c^i)-w(t_c^i-\Delta t)\to\frac{1}{\mu_S}$ and $w'(t_c^i)-w'(t_c^i-\Delta t)\to 0$, as $\Delta t\to 0$, i.e., at t_c^i, the arrival joins in the system of $w(t)$, but balks in the system $w'(t)$; while $w'(t_d^i)-w'(t_d^i-\Delta t)\to\frac{1}{\mu_S}$ and $w(t_d^i)-w(t_d^i-\Delta t)\to 0$, as $\Delta t\to 0$, i.e., at t_d^i, the arrival joins in the system of $w'(t)$, but balks in the system $w(t)$.

That is to say, at $\forall t\in(t_d^i,t_c^i)$ for any i, it is impossible that the arrival joins in the system of $w'(t)$, but balks in the system $w(t)$, because $w(t)<w'(t)$; At $\forall t\in(t_c^i,t_d^{i+1})$ for any i, it is impossible that the arrival joins in the system of $w(t)$, but balks in the system $w'(t)$, because $w(t)>w'(t)$. Thus the arrivals behave the same during (t_a,t_b) except for time points t_c^i and t_d^i.

Thus during (t_a,t_b), the number of total arrivals who join the system of $w'(t)$ is one more than or equal to the number of total arrivals who join the system of $w(t)$. So $\int_{t=t_a}^{t_b}w(t)dt\ge\int_{t=t_a}^{t_b}w'(t)dt$.
Thus we have $\int_{t=t_a}^{t_b}w(t)dt\ge\int_{t=t_a}^{t_b}w'(t)dt$, indicating $\lim\limits_{T\to+\infty}\frac{\int_{t=1}^{T}w'(t)dt}{T}\ge\lim\limits_{T\to+\infty}\frac{\int_{t=1}^{T}w(t)dt}{T}$ for any additional arrival and sample path SP.

Consider a sample path of Poisson arrival process with any arrival rate $\Delta\lambda_S$, denoted by $\Delta SP=(t_a^1,t_a^2,t_a^3,....)$. By induction we know $\lim\limits_{T\to+\infty}\frac{\int_{t=1}^{T}w(t|SP+\Delta SP)dt}{T}\ge\lim\limits_{T\to+\infty}\frac{\int_{t=1}^{T}w(t|SP)dt}{T}$ for any SP and ΔSP. By the memoryless property of Poisson, the arrival process with arrival rate $\lambda_S+\Delta\lambda_S$ can be decomposed to the arrival process with arrival rate λ_S and the arrival process with arrival rate $\Delta\lambda_S$. Thus we can independently construct a sample path SP for λ_S and a sample path ΔSP for $\Delta\lambda_S$. So $\frac{[1-\pi(\lambda_S+\Delta\lambda_S)](\lambda_S+\Delta\lambda_S)}{\mu_S} \geq \frac{[1-\pi(\lambda_S)]\lambda_S}{\mu_S}$. Then we have $[1-\pi(\lambda_S)]\lambda_S$ is increasing in λ_S.

Now we have already proved that, given δ, $[1-\pi(\delta,\lambda_S)]\lambda_S$ is increasing in λ_S.

iii. By the two results above, we can show that δ is decreasing in λ_S:

Let us rearrange the Equation (13):

$$C_D\delta = C_G + C_W w\Big(\frac{\lambda_E+\lambda-\big(1-\pi(\delta,\lambda_S)\big)\lambda_S}{\mu_W}\Big). \tag{35}$$

Now suppose δ is larger when λ_S is larger. By 2.(a).i in this proof, we know that $[1-\pi(\delta,\lambda_S)]\lambda_S$ is strictly larger when δ is larger; and by 2.(a).ii in this proof, we know that $[1-\pi(\delta,\lambda_S)]\lambda_S$ is larger when λ_S is larger. Thus if both of δ and λ_S become larger, $[1-\pi(\delta,\lambda_S)]\lambda_S$ is strictly larger, then the RHS of (35), i.e., $C_G + C_W w(\frac{\lambda_E+\lambda-[1-\pi(\delta,\lambda_S)]\lambda_S}{\mu_W})$ is strictly smaller, which contradicts with that the LHS of (35), i.e., $C_D\delta$, is larger. Then we must have δ decreases in λ_S.

(b) Then we will show that $\pi(\delta,\lambda_S)$ is strictly increasing in λ_S:

Since δ decreases in λ_S, $\pi(\delta,\lambda_S)$ strictly increases in λ_S for fixed δ (by the first point of this Proposition), and $\pi(\delta,\lambda_S)$ strictly decreases in δ for fixed λ_S (by Lemma 2). So, $\pi(\delta,\lambda_S)$ is strictly increasing in λ_S.

(c) At last, we will show that $\delta - \mathbb{E}[\tilde{d}|\tilde{d}\leq\delta,\lambda_S]$ is decreasing in λ_S:

i. First, by the first point of this Proposition, we know $\delta - \mathbb{E}[\tilde{d}|\tilde{d}\leq\delta,\lambda_S]$ is decreasing in λ_S for fixed δ.

ii. Second, we will show that $\delta - \mathbb{E}[\tilde{d}|\tilde{d}\leq\delta,\lambda_S]$ is increasing in δ for fixed λ_S:

Let us omit the notation for λ_S and look at the conditional workload $w_c(t)$ again, and we have

$$\mathbb{E}[\tilde{d}|\tilde{d}\leq\delta] = \mathbb{E}[w|w\leq\delta] = \mathbb{E}[w_c|\delta] \tag{36}$$

then

$$\delta - \mathbb{E}[\tilde{d}|\tilde{d}\leq\delta] = \mathbb{E}[\delta - w_c|\delta].$$

For a period from time 0 to time T, we look at a sample path SP of Poisson arrival process. Let $w_c(t)$ denote the workload of CSP at time t with this arrival sample path SP and workload threshold δ. Let $\{t_1,t_2,t_3,\ldots,t_i,\ldots,t_{N_T}\}$ denote the time points that $w_c(t)$ hits the threshold δ, i.e., $w_c(t_i)=\delta$ for $\forall t_i$. Consider another stochastic process $w'_c(t)$ which denote the w_c at time t with this arrival sample path SP but a larger workload threshold $\delta' > \delta$.

Since $\delta' > \delta$ and the arrivals before t_1 are exactly same in both sample paths, it is obvious that

$$\delta' - w'_c(t) > \delta - w_c(t) \text{ when } t < t_1$$

and

$$\delta' - w'_c(t_1) \geq 0 = \delta - w_c(t_1).$$

Since the arrivals during (t_1,t_2) are exactly same in both sample paths, then

$$w'_c(t) = w_c(t) + w'_c(t_1) - w_c(t_1) \leq w_c(t) + \delta' - \delta \leq \delta' \text{ for } \forall t \in (t_1,t_2).$$

So

$$\delta' - w_c'(t) \geq \delta - w_c(t) \text{ when } t \in (t_1, t_2).$$

At t_2, it is obvious that

$$\delta' - w_c'(t_2) \geq 0 = \delta - w_c(t_2).$$

And repeating the previous arguments, we can have

$$\delta' - w_c'(t) \geq \delta - w_c(t) \text{ for } \forall t \in [t_1, T].$$

Thus for any T and any sample path, we will have $\delta' - w_c'(t) \geq \delta - w_c(t)$ at any time $t \in [0, T]$, which indicates $\delta' - \mathbb{E}[\tilde{d}|\tilde{d} \leq \delta'] \geq \delta - \mathbb{E}[\tilde{d}|\tilde{d} \leq \delta]$ if $\delta' > \delta$.

iii. Finally, we can show that $\delta - \mathbb{E}[\tilde{d}|\tilde{d} \leq \delta, \lambda_S]$ is decreasing λ_S:

By the last two points, when δ is smaller and λ_S is larger, $\delta - \mathbb{E}[\tilde{d}|\tilde{d} \leq \delta, \lambda_S]$ is smaller. Since δ is decreasing in λ_S (see 2.(a) in this proof), then $\delta - \mathbb{E}[\tilde{d}|\tilde{d} \leq \delta, \lambda_S]$ is decreasing in λ_S.

□

Proof of Theorem 1:

$R - C_W w(\frac{\lambda_E}{\mu_W}) \leq 0$ indicates "Walk-in" option is always no better than "Balk". So we must have SnB at the second stage, then $\delta = \delta^a$, $p_W = 0$, $p_B = \pi(\delta^a, \lambda_S)$. Since $u_B^2 > u_W^2$, thus $u_B^1 > u_W^1$, then we know $\lambda_W = 0$. By Proposition D.3, it is easy to check $u_S^1 = R - C_D[(1 - \pi(\delta^a, \lambda_S))\mathbb{E}[\tilde{d}|\tilde{d} \leq \delta^a, \lambda_S] + \pi(\delta^a, \lambda_S)\delta^a] = R + C_D[(1 - \pi(\delta^a, \lambda_S))(\delta^a - \mathbb{E}[\tilde{d}|\tilde{d} \leq \delta^a, \lambda_S]) - \delta^a]$ is strictly decreasing in λ_S, then we know there must be a unique λ_S^a solves $u_S^1 = u_B^1 = 0$. The unique equilibrium must be $\lambda_S = \min(\lambda, \lambda_S^a)$, $\lambda_W = 0$, $\delta = \delta^a$, $p_W = 0$, $p_B = \pi(\delta^a, \lambda_S)$.

$\delta^a < \max(\delta^b, \delta^c)$ and $R - C_W w(\frac{\lambda_E}{\mu_W}) > 0$ indicate we will have SWB at the second stage and δ^a as the threshold. So we have $u_S^1 = R - C_D[(1 - \pi(\delta^a, \lambda_S^1))\mathbb{E}[\tilde{d}|\tilde{d} \leq \delta^a, \lambda_S] + \pi(\delta^a, \lambda_S)\delta^a]$ which is strictly decreasing in λ_S, and $u_W^1 = u_B^1 = 0$, so there must be a unique λ_S^a solves $u_S^1 = u_W^1 = u_B^1 = 0$. There can be multiple equilibria such that $\lambda_S = \min(\lambda_S^a, \lambda)$, $\lambda_W + p_W \lambda_S$ solves $R - C_W w(\frac{\lambda_E + \lambda_W + p_W \lambda_S}{\mu_W}) = 0$, $\delta = \delta^a$, $p_B = \pi(\delta^a, \lambda_S) - p_W$. Note that although the equilibria are multiple, the final utility of each option is 0 at the first stage; The total rate of scheduling patients is $(1 - \pi(\delta^a, \lambda_S))\lambda_S$, total rate of walk-ins is $\lambda_W + p_W \lambda_S$ and the total rate of balking patients is $\lambda - (1 - \pi(\delta^a, \lambda_S))\lambda_S - \lambda_W - p_W \lambda_S$, which are unique.

$\delta^a \geq \max(\delta^b, \delta^c)$ indicates that we must have SnW at the second stage and the threshold is $\max(\delta^b, \delta^c)$. Recall that $\delta(\lambda_S)$ solves (13) , by Proposition D.3, we know $\delta(\lambda_S)$ is decreasing in λ_S. Thus we have unique δ^b solving (13) for given λ_S, and δ^b decreases in λ_S. Under threshold δ^b, $u_S^1 = R - C_D[1 - \pi(\delta^b, \lambda_S)]\mathbb{E}[\tilde{d}|\tilde{d} \leq \delta^b, \lambda_S] - C_D \pi(\delta^b, \lambda_S)\delta^b$ and $u_W^1(\delta^b) = u_W^2(\delta^b) + C_G = R - C_D \delta^b + C_G$. Then $u_S^1(\delta^b) - u_W^1(\delta^b) = C_D[1 - \pi(\delta^b, \lambda_S)](\delta^b - \mathbb{E}[\tilde{d}|\tilde{d} \leq \delta^b, \lambda_S]) - C_G$, by Proposition D.3, we know $u_S s^1 - u_W^1$ is strictly decreasing in λ_S. Thus we must have unique λ_S^b solves $u_S^1(\delta^b) = u_W^1(\delta^b)$. Since δ^c is a constant, thus the unique equilibrium is $\lambda_S = \min(\lambda, \lambda_S^b)$, $p_W = \lambda - \lambda_S$, $\delta = \max(\delta^b, \delta^c)$, $p_W = \pi(\delta, \lambda_S)$, $p_B = 0$.

□

E.2. Proofs of Results in Section 4

Proof of Lemma 4:

When $C_G = 0$, $u_W^1 = u_W^2$ and $u_B^1 = u_B^2$. Since $u_S^1 = [1 - \pi(\delta, \lambda_S)](R - C_D \mathbb{E}[\tilde{d}|\tilde{d} \leq \delta, \lambda_S]) + \pi(\delta, \lambda_S)\min\{u_W^2, u_B^2\}$. Since $\mathbb{E}[\tilde{d}|\tilde{d} \leq \delta, \lambda_S] \leq \delta$, and $\min\{u_W^2, u_B^2\} = R - C_D\delta$, then it is obvious that $u_S^1 \leq \min\{u_W^2, u_B^2\} = \min\{u_W^1, u_B^1\}$. Thus $\lambda_S = \lambda$.

□

Proof of Lemma 5:

If follows Proposition C.2 in Appendix C.

□

Before proceeding to prove other results, we present an ancillary result first.

Proposition E.5. *Recall $\underline{\mu}_W$ and $\overline{\mu}_W$ defined in* (15) *and* (16)*, respectively. p_B^S, p_W^B, p_W^S and δ_W^S are defined in Lemma 3. Then, we have the following.*

1. p_B^S is decreasing in μ_S.

2. p_W^B is increasing and linear in μ_W when $\mu_W \in [\underline{\mu}_W, \overline{\mu}_W]$.

3. p_W^S is decreasing in μ_S and increasing in μ_W when $\mu_W \in [\underline{\mu}_W, \overline{\mu}_W]$.

4. δ_W^S is decreasing in μ_S and μ_W when $\mu_W \in [\underline{\mu}_W, \overline{\mu}_W]$.

Proof of Proposition E.5:

1. $p_B^S = \pi(\delta_B^S, \mu_S)$. As δ_B^S is a constant, so p_B^S is decreasing in μ_S by Lemma 5.

2. By definition of p_W^B.

3. p_W^S is decreasing in μ_S: By definition we have $u_W(p_W^S) = u_S(\delta_W^S)$ and $p_W^S = \pi(\delta_W^S, \mu_S)$. Suppose that p_W^S is increasing in μ_S, then $p_W^S = \pi(\delta_W^S, \mu_S)$ implies that δ_W^S must be decreasing in μ_S as $\pi(\delta, \mu_S)$ decreases in δ and μ_S. Then when μ_S is larger, p_W^S is larger and δ_W^S is smaller, leading to a contradiction that $u_W(p_W^S) < u_S(\delta_W^S)$. So p_W^S is decreasing in μ_S.

Increasing in μ_W: By definition we have $u_W(p_W^S, \mu_W) = u_S(\delta_W^S)$ and $p_W^S = \pi(\delta_W^S)$. Suppose that p_W^S is decreasing in μ_W, then $p_W^S = \pi(\delta_W^S)$ implies that δ_W^S is increasing in μ_W. Then when μ_W is larger, p_W^S is smaller and δ_W^S is larger, leading to a contradiction that $u_W(p_W^S) > u_S(\delta_W^S)$. So p_W^S is increasing in μ_W.

4. δ_W^S is decreasing in μ_S: By definition we have $u_W\big(\pi(\delta_W^S, \mu_S)\big) = u_S(\delta_W^S)$. Suppose δ_W^S is increasing in μ_S, then $\pi(\delta_W^S, \mu_S)$ is decreasing in μ_S. Then when μ_S is larger, $\pi(\delta_W^S, \mu_S)$ is smaller and δ_W^S is larger, leading to a contradiction that $u_W\big(\pi(\delta_W^S, \mu_S)\big) > u_S(\delta_W^S)$. Hence it must be that δ_W^S is decreasing in μ_S.

δ_W^S is decreasing in μ_W: By definition we have $u_W\big(\pi(\delta_W^S), \mu_W\big) = u_S(\delta_W^S)$. Suppose δ_W^S is increasing in μ_W, then $\pi(\delta_W^S)$ is decreasing in μ_W. Then when μ_W is larger, $\pi(\delta_W^S)$ is smaller and δ_W^S is larger, leading to a contradiction that $u_W\big(\pi(\delta_W^S), \mu_W\big) > u_S(\delta_W^S)$. Hence it must be that δ_W^S is decreasing in μ_W.

□

Proof of Proposition 2:

If follows from the definition of $\underline{\mu}_W$ and $\overline{\mu}_S(\mu_W)$. In addition, when $(\mu_S, \mu_W) = \big(\overline{\mu}_S(\mu_W), \mu_W\big)$, for $\underline{\mu}_W \le \mu_W \le \overline{\mu}_W$, it is evident that $p_B^S = p_W^B$ and $\delta_B^S = \delta_W^S$ thus $p_W^S = p_B^S$.

□

Proof of Proposition 3:

For any $\mu_S \ge 0$, define $\mu_W(\mu_S) = \frac{\pi(\frac{R}{C_D}, \mu_S)\lambda + \lambda_E}{w^{-1}(\frac{R}{C_W})}$, i.e., setting μ_W such that $\mu_S + \mu_W(\mu_S)$ can attract all patients from balking.

Let $\overline{\mu}_O = \min_{\mu_S \ge 0} \mu_S + \mu_W(\mu_S)$, i.e., $\overline{\mu}_O$ is the minimum total capacity which can attract all strategic patients.

Let $\mu_S^*(\mu)$ solve $\mu_S + \mu_W(\mu_S) = \mu$ where $\mu \ge \overline{\mu}_O$ (since $\overline{\mu}_O = \min_{\mu_S \ge 0} \mu_S + \mu_W(\mu_S)$, and $\mu_S + \mu_W(\mu_S)$ is continuous in μ_S and can go to infinity, so for any $\mu \ge \overline{\mu}_O$, there exist at least one solution $\mu_S^*(\mu)$), then $\Big(\mu_S^*(\mu), \mu_W\big(\mu_S^*(\mu)\big)\Big)$ is a feasible solution in SnW, which yields an overtime cost $C_O o\big(w^{-1}(\frac{R}{C_W})\big)$ which is no larger than the overtime cost of any feasible solution in Regime SWB and SnB. Thus SnW is optimal.

For $\mu < \overline{\mu}_O$, it can not attract all patients from balking, thus SnW is not feasible, and SnB or SWB is optimal.

Furthermore, if $\mu < \underline{\mu}_W$, only SnB is feasible.

□

Proof of Lemma 6:

When $C_G = +\infty$, in the second stage, $u_W^2 < u_S^2$ and $u_B^2 < u_S^2$. Thus we must have $\delta = +\infty$ and $p_W = p_B = 0$.

□

Proof of Lemma D.3:

First, by (18) and (19), it is obvious that $u_S^1(\lambda_S)$ and $u_W^1(\lambda_W)$ are strictly decreasing in λ_S and λ_W, respectively, everything else being fixed.

Case 1: Here $\lambda_S^a = \frac{2R(\mu_S)^2}{C_D + 2R\mu_S}$ which solves $u_S^1 = u_B^1$. Then Case 1 here follows Case 1 in Theorem 1.

Case 2: $u_W^1(\lambda_W^S) < 0$ indicates $u_S^1(\lambda_S^W) < 0$. By the definition of λ_W^S and λ_B^S, we have $\lambda_B^S < \lambda_W^S$. So the rate of a strategic patient choosing not to schedule is λ_B^S, which is unique since $u_S^1(\lambda_S)$ is monotone. The fact that $u_W^1(0) > 0$ indicates that when walk-in rate is small enough, Walk-in is more attractive than Balk. So when patients do not schedule, some of them will walk in to make the utility of Walk-in be equal to that of Balk. By the monotonicity of $u_W^1(\lambda_W)$, we know the corresponding walk-in rate λ_W^B is unique. Since $u_W^1(\lambda_W^S) = u_S^1(\lambda_S^W) < 0 = u_W^1(\lambda_W^B)$, so $\lambda_W^B < \lambda_W^S$. Since $\lambda_B^S > \lambda_W^S$, so $\lambda_B^S - \lambda_W^B$ is positive and unique.

Case 3: $u_W^1(\lambda_W^S) \ge 0$ indicates that $u_S^1(\lambda_S^W) \ge 0$. Since we rule out tied cases, then Balk option can be ignored ($\lambda_B = 0$). Since $u_S^1(\lambda_S)$ and $u_W^1(\lambda_W)$ are monotone, we know there is a unique walk-in rate which is λ_W^S.

□

Proof of Proposition 4: According to the definition of $\underline{\mu}_W$ and $\overline{\mu}_S(\mu_W)$, all the results are clear except for the decreasing and concave properties of $\overline{\mu}_S(\mu_W)$. Next we will complete the proof by showing these properties of $\overline{\mu}_S(\mu_W)$. When $\underline{\mu}_W \le \mu_W \le \overline{\mu}_W$, $\overline{\mu}_S(\mu_W)$ ensures $\lambda_B^S = \lambda_W^B$. Note that when $\underline{\mu}_W \le \mu_W \le \overline{\mu}_W$,

$$\lambda_B^S = \lambda - \frac{2R(\mu_S)^2}{C_D + 2R\mu_S}.$$

It is clear that λ_B^S is decreasing and concave in μ_S. Recall that λ_W^B is increasing and linear in μ_W by Proposition E.5, then μ_W as a function of μ_S, denoted by $\mu_W(\mu_S)$, is decreasing and concave in μ_S. So $\overline{\mu}_S(\mu_W)$ is decreasing and concave in μ_W, because $\overline{\mu}_S(\mu_W)$ is the inverse function of $\mu_W(\mu_S)$.

□

Proof of Proposition 5:

For any $\mu_S \ge 0$, define $\lambda_S(\mu_S) = \min\{\frac{2R(\mu_S)^2}{2R\mu_S + C_D}, \lambda\}$, and $\mu_W(\mu_S) = \frac{\lambda - \lambda_S(\mu_S) + \lambda_E}{w^{-1}(\frac{R}{C_W})}$, then $\mu_S + \mu_W(\mu_S)$ can attract all patients from balking.

Let $\overline{\mu}_U = \min_{\mu_S \ge 0} \mu_S + \mu_W(\mu_S)$. $\overline{\mu}_U$ is the minimum total capacity which can attract all strategic patients.

Let $\mu_S^*(\mu)$ solve $\mu_S + \mu_W(\mu_S) = \mu$ where $\mu \ge \overline{\mu}_U$ (since $\overline{\mu}_U = \min_{\mu_S \ge 0} \mu_S + \mu_W(\mu_S)$, and $\mu_S + \mu_W(\mu_S)$ is continuous in μ_S and can go to infinity, so for any $\mu \ge \overline{\mu}_U$, there exist at least one solution), then $\Big(\mu_S^*(\mu), \mu_W\big(\mu_S^*(\mu)\big)\Big)$ is a feasible solution in SnW, yielding an overtime cost $C_O o\big(w^{-1}(\frac{R}{C_W})\big)$ which is no larger than the overtime cost of any feasible solution in Regime *SWB* and *SnB*. Thus SnW is optimal.

For $\mu < \overline{\mu}_U$, it can not attract all patients from balking, thus SnW is not feasible, and SnB or SWB is optimal.

Furthermore, if $\mu < \underline{\mu}_W$, only SnB is feasible.

□

Proof of Lemma D.4:

1. Let us show that $\overline{\mu}_U = \min\{\overline{\mu}_S + \underline{\mu}_W, \overline{\mu}_W\}$:

By Equation (20) and the definition of λ_W^S in Lemma D.3, we have the close form of the inverse function of $\overline{\mu}_S(\mu_W)$ when $\mu_W \in [\underline{\mu}_W, \overline{\mu}_W]$:

$$\mu_W = \overline{\mu}_S^{-1}(\mu_S) = \lambda + \lambda_E - \frac{2R(\mu_S)^2}{C_D + 2R\mu_S}.$$

One can verify that $\overline{\mu}_S^{-1}(\mu_S)$ is a decreasing concave function when $\mu_S \in [0, \overline{\mu}_S]$, indicating $\overline{\mu}_S(\mu_W)$ is decreasing and concave when $\mu_W \in [\underline{\mu}_W, \overline{\mu}_W]$. By the definition of $\overline{\mu}_U$, we have $\overline{\mu}_U = \min\{\overline{\mu}_S + \underline{\mu}_W, \overline{\mu}_W\}$.

2. Now we will prove that, under optimal capacity allocation (μ_S^*, μ_W^*) in unobservable setting, either $\lambda_S^* = 0$ or $\lambda_W^* = 0$:

Suppose we have $0 < \lambda_S^* < \lambda$ and $0 < \lambda_W^* < \lambda$, then they must satisfy the following equation:

$$C_W w(\frac{\lambda_W^* + \lambda_E}{\mu_W^*}) = C_D \frac{\lambda_S^*}{2\mu_S^*(\mu_S^* - \lambda_S^*)}$$

Let $\lambda' = \lambda_S^* + \lambda_W^*$. Let $R' = C_W w(\frac{\lambda_W^* + \lambda_E}{\mu_W^*})$, then we know $R' \le R$. μ_S^* and μ_W^* can be expressed as follows:

$$\mu_W^* = \frac{\lambda_W^* + \lambda_E}{w^{-1}(\frac{R'}{C_W})}$$

$$\mu_S^* = \frac{\lambda_S^* + \sqrt{(\lambda_S^*)^2 + \frac{2C_D\lambda_S^*}{R'}}}{2}$$

Then we consider two scenarios:

(a) When $\frac{1+\sqrt{1+\frac{2C_D}{R'\lambda'}}}{2} > \frac{\mu_W^*}{\lambda_W^*+\lambda_E}$, we first show that $\frac{\mu_W^*(\lambda'+\lambda_E)}{\lambda_W^*+\lambda_E} < \mu$.

$$\begin{aligned}
&\mu - \frac{\mu_W^*(\lambda'+\lambda_E)}{\lambda_W^*+\lambda_E}\\
=&\mu_S^* + \mu_W^* - \frac{\mu_W^*(\lambda'+\lambda_E)}{\lambda_W^*+\lambda_E}\\
=&\frac{\lambda_S^* + \sqrt{(\lambda_S^*)^2 + \frac{2C_D\lambda_S^*}{R'}}}{2} - \frac{\mu_W^*\lambda_S^*}{\lambda_W^*+\lambda_E}\\
=&\lambda_S^*\left[\frac{1+\sqrt{1+\frac{2C_D}{R'\lambda_S^*}}}{2} - \frac{\mu_W^*}{\lambda_W^*+\lambda_E}\right]\\
>&\lambda_S^*\left[\frac{1+\sqrt{1+\frac{2C_D}{R'\lambda'}}}{2} - \frac{\mu_W^*}{\lambda_W^*+\lambda_E}\right]\\
>&0
\end{aligned}$$

The first inequality holds since $\lambda_S^* < \lambda'$. The second inequality holds since $\frac{1+\sqrt{1+\frac{2C_D}{R'\lambda'}}}{2} > \frac{\mu_W^*}{\lambda_W^*+\lambda_E}$.

We set $(\mu_S, \mu_W) = (0, \mu)$. Since $\frac{\mu_W^*(\lambda'+\lambda_E)}{\lambda_W^*+\lambda_E} < \mu$, then in the equilibrium, the corresponding $(\lambda_S, \lambda_W) = (0, \lambda)$. The lost demand cost is unchanged, and the overtime $o(\frac{\lambda'+\lambda_E}{\mu})$ is smaller than $o(\frac{\lambda_W^*+\lambda_E}{\mu_W^*})$. Then we have a better solution $(0, \mu)$ compared to (μ_S^*, μ_W^*), generating $\lambda_S^* = 0$.

(b) When $\frac{1+\sqrt{1+\frac{2C_D}{R'\lambda'}}}{2} \leq \frac{\mu_W^*}{\lambda_W^*+\lambda_E}$, we first show that $\frac{\lambda'+\sqrt{(\lambda')^2+\frac{2C_D\lambda'}{R'}}}{2} + \frac{\mu_W^*\lambda_E}{\lambda_W^*+\lambda_E} < \mu$.

$$\begin{aligned}
&\mu - \frac{\lambda' + \sqrt{(\lambda')^2 + \frac{2C_D\lambda'}{R'}}}{2} - \frac{\mu_W^*\lambda_E}{\lambda_W^*+\lambda_E}\\
=&\frac{\lambda_S^* + \sqrt{(\lambda_S^*)^2 + \frac{2C_D\lambda_S^*}{R'}}}{2} + \mu_W^* - \frac{\lambda' + \sqrt{(\lambda')^2 + \frac{2C_D\lambda'}{R'}}}{2} - \frac{\mu_W^*\lambda_E}{\lambda_W^*+\lambda_E}\\
=&\frac{\mu_W^*\lambda_W^*}{\lambda_W^*+\lambda_E} + \lambda_S^*\frac{1+\sqrt{1+\frac{2C_D}{R'\lambda_S^*}}}{2} - \lambda'\frac{1+\sqrt{1+\frac{2C_D}{R'\lambda'}}}{2}\\
>&\frac{\mu_W^*\lambda_W^*}{\lambda_W^*+\lambda_E} + \lambda_S^*\frac{1+\sqrt{1+\frac{2C_D}{R'\lambda'}}}{2} - \lambda'\frac{1+\sqrt{1+\frac{2C_D}{R'\lambda'}}}{2}\\
=&\frac{\mu_W^*\lambda_W^*}{\lambda_W^*+\lambda_E} - \lambda_W^*\frac{1+\sqrt{1+\frac{2C_D}{R'\lambda'}}}{2}\\
=&\lambda_W^*\left[\frac{\mu_W^*}{\lambda_W^*+\lambda_E} - \frac{1+\sqrt{1+\frac{2C_D}{R'\lambda'}}}{2}\right]\\
\geq&0
\end{aligned}$$

The first inequality holds since $\lambda_S^* < \lambda'$. The second inequality holds since $\frac{1+\sqrt{1+\frac{2C_D}{R'\lambda'}}}{2} \leq \frac{\mu_W^*}{\lambda_W^*+\lambda_E}$.

We set $(\mu_S, \mu_W) = (\mu - \frac{\mu_W^*\lambda_E}{\lambda_W^*+\lambda_E} - \Delta, \frac{\mu_W^*\lambda_E}{\lambda_W^*+\lambda_E} + \Delta)$. Since $\mu - \frac{\mu_W^*\lambda_E}{\lambda_W^*+\lambda_E} > \frac{\lambda'+\sqrt{(\lambda')^2+\frac{2C_D\lambda'}{R'}}}{2}$, then there must be a $\Delta \in (0, \mu - \frac{\mu_W^*\lambda_E}{\lambda_W^*+\lambda_E})$ such that $C_D\frac{\lambda'}{2\mu_S(\mu_S-\lambda')} = C_W w(\frac{\lambda_E}{\mu_W})$. Then in the equilibrium, the corresponding

$(\lambda_S, \lambda_W) = (\lambda', 0)$. The lost demand cost is unchanged, and the overtime $o(\frac{\lambda_E}{\mu_W})$ is smaller than $o(\frac{\lambda_W^* + \lambda_E}{\mu_W^*})$. Then we have a better solution (μ_S, μ_W) compared to (μ_S^*, μ_W^*), generating $\lambda_W^* = 0$.

Summarizing the above analysis, we have a contradiction to $0 < \lambda_S^* < \lambda$ and $0 < \lambda_W^* < \lambda$, indicating we must have either $\lambda_S^* = 0$ or $\lambda_W^* = 0$.

□

E.3. Proofs of Results in Section 5

We first present Lemma E.5, an ancillary result that helps us prove Theorem 2. In the unobservable setting, let $\bar{d}^u$ be the expected delay in equilibrium; let λ_S^u be the equilibrium arrival rate of strategic patients who schedule an appointment, i.e., joins the appointment queue. Recall that $\pi(\delta)$ is the blocking probability in the observable appointment queue with a scheduling threshold δ. Let $\lambda_S^o(\delta)$ be thearrival rate of strategic patients who schedule an appointment in the observable setting if all strategic patients adopt the delay threshold δ, i.e., $\lambda_S^o(\delta) = \lambda[1 - \pi(\delta)]$.

LEMMA E.5. *Consider the observable and unobservable settings with the same set of model parameters* $(\lambda, \lambda_E, \mu_S, \mu_W, R, C_D, C_W)$ *and in-clinic wait time function* $w(\cdot)$ *and overtime function* $o(\cdot)$.

1. *when* $\lambda_S^o(\bar{d}^u) \geq \lambda_S^u$, *the observable setting costs less;*
2. *when* $\lambda_S^o(\bar{d}^u) < \lambda_S^u$, *the unobservable setting costs less.*

Proof of Lemma E.5:

Let $\bar{d}^u$ denote the expected delay in equilibrium under the unobservable queue, and let λ_S^u, λ_B^u and λ_W^u denote the corresponding schedule, balk, and walk-in rate. Let δ^o denote the delay threshold in equilibrium under the observable queue, and let $p_B^o(\delta^o)$ and $p_W^o(\delta^o)$ denote the corresponding balk and walk-in probability. Note that $\lambda_S^o(\delta^o) = \lambda[1 - \pi(\delta^o)]$.

If $\lambda_S^u = \lambda$, i.e., all strategic patients schedule in unobservable setting, then we must have $\lambda_S^o(\delta^o) < \lambda = \lambda_S^u$. In this case, unobservable setting has zero lost demand cost, and the smallest overtime cost (there is no strategic walk-in patient). Thus unobservable setting costs less.

If $\lambda_S^u < \lambda$: there will be 3 possible regimes for the equilibrium in observable setting.

- If the equilibrium of observable queue is in SnB, and the equilibrium of unobservable queue must be in SnB, $\delta^o = \bar{d}^u = \frac{R}{C_D}$, and the overtime cost is same.
  - —If $\lambda_S^o(\delta^o) = \lambda_S^o(\bar{d}^u) \leq \lambda_S^u$, then unobservable queue has less lost demand cost and works better;
  - —If $\lambda_S^o(\delta^o) = \lambda_S^o(\bar{d}^u) \geq \lambda_S^u$, observable queue has less lost demand cost and works better.
- If the equilibrium of observable queue is in SWB, then $R - C_D\delta^o = R - C_W w(\frac{p_W^o(\delta^o)\lambda + \lambda_E}{\mu_W}) = 0$ and $1 - \pi(\delta^o) + p_W^o(\delta^o) \leq 1$.
  - —If the equilibrium of the unobservable queue is in SWB, then $\delta^o = \bar{d}^u = \frac{R}{C_D}$, $p_W^o(\delta^o)\lambda = \lambda_W^u$, and the overtime cost is same.
    - ∗ If $\lambda_S^o(\delta^o) = \lambda_S^o(\bar{d}^u) \leq \lambda_S^u$, the unobservable queue has less lost demand cost and works better;
    - ∗ If $\lambda_S^o(\delta^o) = \lambda_S^o(\bar{d}^u) \geq \lambda_S^u$, the observable queue has less lost demand cost and works better.

—If the equilibrium of unobservable queue is in SnW, then $R - C_D\bar{d}^u = R - C_W w(\frac{\lambda_W^u+\lambda_E}{\mu_W}) \geq 0$, then $\delta^o \geq \bar{d}^u$, and $\frac{\lambda_W^u}{\lambda} \leq p_W^o(\delta^o) \leq \pi(\delta^o) \leq \pi(\bar{d}^u)$.

Suppose $\lambda_S^o(\bar{d}^u) > \lambda_S^u$, then $\lambda_S^o(\delta^o) > \lambda_S^u$. Then $\lambda_S^o(\delta^o) + \lambda p_W^o(\delta^o) > \lambda_S^u + \lambda_W^u = \lambda$, which contradicts the condition $1 - \pi(\delta^o) + p_W^o(\delta^o) \leq 1$.

So we must have $\lambda_S^o(\bar{d}^u) \geq \lambda_S^u$ and $\lambda_S(\delta^o) \geq \lambda_S^u$. Since $\frac{\lambda_W^u}{\lambda} \leq p_W^o(\delta^o)$, the unobervable queue has less overtime cost and zero lost demand cost; thus it works better.

- If the equilibrium of observable queue is in SnW, then $R - C_D\delta^o = R - C_W w(\frac{p_W^o(\delta^o)\lambda+\lambda_E}{\mu_W}) \geq 0$ and $1 - \pi(\delta^o) + p_W^o(\delta^o) = 1$.

—If the equilibrium of unobservable queue is in SnW, then $R - C_D\bar{d}^u = R - C_W w(\frac{\lambda_W^u+\lambda_E}{\mu_W}) \geq 0$, $\lambda_S^u + \lambda_W^u = \lambda$, and both queues do not have lost demand cost.

$*$ When $\lambda_S^o(\bar{d}^u) \leq \lambda_S^u$:

Suppose $\delta^o < d^u$, then $R - C_W w(\frac{p_W^o(\delta^o)\lambda+\lambda_E}{\mu_W}) = R - C_D\delta^o > R - C_D\bar{d}^u = R - C_W w(\frac{\lambda_W^u+\lambda_E}{\mu_W})$ and $\lambda_S^o(\delta^o) < \lambda_S^o(\bar{d}^u) \leq \lambda_S^u$. Thus $\lambda_S^u + \lambda_W^u > \lambda p_W^o(\delta^o) + \lambda_S^o(\delta^o) = \lambda$, which contradicts $\lambda_S^u + \lambda_W^u = \lambda$.

So we must have $\lambda_S^o(\bar{d}^u) \leq \lambda_S^u$ and $\delta^o \geq \bar{d}^u$. Then $R - C_W w(\frac{\lambda_W^u+\lambda_E}{\mu_W}) = R - C_D\bar{d}^u \geq R - C_D\delta^o = R - C_W w(\frac{p_W^o(\delta^o)\lambda+\lambda_E}{\mu_W})$, and then $\lambda_W^u \leq \lambda p_W^o(\delta^o)$. Thus, the unobservable queue has less overtime cost and works better.

$*$ When $\lambda_S^o(\bar{d}^u) \geq \lambda_S^u$,:

Suppose $\delta^o > d^u$, then $R - C_W w(\frac{p_W^o(\delta^o)\lambda+\lambda_E}{\mu_W}) = R - C_D\delta^o < R - C_D\bar{d}^u = R - C_W w(\frac{\lambda_W^u+\lambda_E}{\mu_W})$ and $\lambda_S^o(\delta^o) > \lambda_S^o(\bar{d}^u) \geq \lambda_S^u$. Thus $\lambda_S^u + \lambda_W^u < \lambda p_W^o(\delta^o) + \lambda_S^o(\delta^o) = \lambda$, which contradicts $\lambda_S^u + \lambda_W^u = \lambda$.

So we must have $\lambda_S^o(\bar{d}^u) \geq \lambda_S^u$ and $\delta^o \leq \bar{d}^u$. Then $R - C_W w(\frac{\lambda_W^u+\lambda_E}{\mu_W}) = R - C_D\bar{d}^u \leq R - C_D\delta^o = R - C_W w(\frac{p_W^o(\delta^o)\lambda+\lambda_E}{\mu_W})$, and then $\lambda_W^u \geq \lambda p_W^o(\delta^o)$. Thus, the observable queue has less overtime cost and works better.

—If the equilibrium of unobservable queue is in SWB, then $R - C_W w(\frac{\lambda_W^u+\lambda_E}{\mu_W}) = R - C_D\bar{d}^u = 0 \leq R - C_D\delta^o$ and $\lambda_S^u + \lambda_W^u \leq \lambda$. Then $\delta^o \leq \bar{d}^u$, and $\frac{\lambda_W^u}{\lambda} \geq p_W^o(\delta^o) = \pi(\delta^o) \geq \pi(\bar{d}^u)$.

Suppose $\lambda_S^o(\bar{d}^u) < \lambda_S^u$, we have $\lambda_S^o(\delta^o) < \lambda_S^u$. Then $\lambda_S^o(\delta^o) + \lambda p_W^o(\delta^o) < \lambda_S^u + \lambda_W^u \leq \lambda$, which contradicts the condition $\lambda_S^o(\delta^o) + \lambda p_W^o(\delta^o) = \lambda$.

So we must have $\lambda_S^o(\bar{d}^u) \geq \lambda_S^u$, and since $\lambda_W^u \geq \lambda p_W^o(\delta^o)$, the observable queue has less overtime cost and zero lost demand cost; hence it works better.

Summarizing above, we have that if $\lambda_S^o(\bar{d}^u) \geq \lambda_S^u$, the observable queue works better; otherwise, the unobservable queue works better.

□

Proof of Theorem 2:

1. We first show that when μ is sufficiently small, observable setting costs less. Let $(\mu_S^{u*}, \mu - \mu_S^{u*})$ denote the optimal capacity allocation in unobservable setting. If we use $(\mu_S^{u*}, \mu - \mu_S^{u*})$ as the capacity allocation in observable setting, it provides an upper bound for the optimal cost. Next we show that this upper bound is no larger than the optimal cost of unobservable setting.

If $\mu_S^{u*}=0$, i.e., all capacity is allocated to walk-in session, then all the patients make choices between walking in and balking. If we do so in observable setting, patients will have same behavior, thus observable setting with $(\mu_S^{u*},\mu-\mu_S^{u*})$ will have same objective value as unobservable setting.

Then we consider $\mu_S^{u*}>0$, In unobservable setting, let λ_S^{u*} denote the corresponding λ_S in the equilibrium. We have

$$R-C_D\frac{\lambda_S^{u*}}{2\mu_S^{u*}(\mu_S^{u*}-\lambda_S^{u*})}\geq 0,$$

i.e., if there is patient scheduling, the utility of scheduling must be better than balking. Let $\bar{d}^{u*}$ denote the corresponding expected delay, i.e.,

$$\bar{d}^{u*}=\frac{\lambda_S^{u*}}{2\mu_S^{u*}(\mu_S^{u*}-\lambda_S^{u*})}\leq\frac{R}{C_D}.$$

Then we have

$$\lambda_S^{u*}=\frac{2(\mu_S^{u*})^2\bar{d}^{u*}}{1+2\mu_S^{u*}\bar{d}^{u*}}.$$

By Lemma E.5, we know that if $[1-\pi(\bar{d}^{u*})]\lambda\geq\lambda_S^{u*}$, then observable setting with $(\mu_S^{u*},\mu-\mu_S^{u*})$ costs less. Next we will show this inequality.

By Corollary C.1,

$$\pi(\bar{d}^{u*})=1-\frac{1}{q_0(\bar{d}^{u*})+\frac{\lambda}{\mu_S^{u*}}},$$

where $q_0(\bar{d}^{u*})$ is the probability that at the end of a slot, there is zero patient in the system. We have

$$\frac{\lambda}{1+\frac{\lambda}{\mu_S^{u*}}}\leq\lambda[1-\pi(\bar{d}^{u*})]\leq\lambda.$$

Suppose $\mu\leq\sqrt{\frac{\lambda C_D}{2R}}$, then

$$\lambda\geq\frac{2(\mu)^2R}{C_D}\geq\frac{2(\mu_S^{u*})^2R}{C_D},$$

then

$$\frac{\lambda}{1+\frac{\lambda}{\mu_S^{u*}}}\geq\frac{2(\mu_S^{u*})^2R}{C_D+2\mu_S^{u*}R}.$$

Thus

$$\lambda[1-\pi(\bar{d}^{u*})]\geq\frac{2(\mu_S^{u*})^2(R/C_D)}{1+2\mu_S(R/C_D)}.$$

Since $\bar{d}^{u*}\leq\frac{R}{C_D}$, then

$$\lambda[1-\pi(\bar{d}^{u*})]\geq\frac{2(\mu_S^{u*})^2\bar{d}^{u*}}{1+2\mu_S^{u*}\bar{d}^{u*}}=\lambda_S^{u*}.$$

Thus observable setting costs less when $\mu\leq\sqrt{\frac{\lambda C_D}{2R}}$.

2. Next, let us prove case 2(a), i.e., when μ is sufficiently large (specifically, $\mu\geq\lambda+\lambda_E$), observable queue works better if $C_W w(1)\leq\min\{R,\frac{C_D}{2M}\}$, where $M\geq\frac{\mu(\mu-\lambda)}{\lambda}$.

Let the optimal capacity decision to the unobservable setting be (μ_S^{u*},μ_W^{u*}). The corresponding arrival rates of patients who schedule and walk-in strategically are λ_S^{u*} and λ_W^{u*}, respectively. When $\mu\geq\lambda+\lambda_E\geq\frac{\lambda+\lambda_E}{w^{-1}(\frac{R}{C_W})}=\overline{\mu}_W$, by Proposition 5 and Lemma D.4, we know $\lambda_S^{u*}=\lambda$ or $\lambda_W^{u*}=\lambda$.

Next, we will show that when $C_W w(1) \le \min\{R, \frac{C_D}{2M}\}$, we must have $\lambda_W^{u*} = \lambda$, i.e., letting all patients walk in outperforms letting all patients schedule. Suppose that the optimal capacity allocation lets all patients schedule, then we can construct contradiction in the following way. Now, suppose we have $\lambda_S^{u*} = \lambda$ under (μ_S^{u*}, μ_W^{u*}).

- If $\mu_S^{u*} = \overline{\mu}_S$, then $\mu_W^{u*} \le \underline{\mu}_W$ (otherwise the utility of walk-in option will be larger than 0, which is the utility of scheduling). Specifically, we have

$$R - C_D \frac{\lambda}{2\mu_S^{u*}(\mu_S^{u*} - \lambda)} = 0 \ge R - C_W w(\frac{\lambda_E}{\mu_W^{u*}}),$$

$$\mu_S^{u*} = \overline{\mu}_S = \frac{\lambda}{2} + \sqrt{\frac{(\lambda)^2}{4} + \frac{C_D\lambda}{2R}}.$$

The cost under (μ_S^{u*}, μ_W^{u*}) is $C_O o(\frac{\lambda_E}{\mu_W^{u*}})$. If we we set capacity $(0, \mu)$, the cost is $C_O o(\frac{\lambda+\lambda_E}{\mu})$. It is obvious that

$$\frac{\lambda + \lambda_E}{\mu} < \frac{\lambda + \lambda_E}{\overline{\mu}_W} = \frac{\lambda_E}{\underline{\mu}_W} \le \frac{\lambda_E}{\mu_W^{u*}},$$

which contradicts that letting all patients schedule costs less. So we must have $\lambda_W^{u*} = \lambda$.

- If $\mu_S^{u*} > \overline{\mu}_S$, then the utility of strategic patients is positive, and patients feel indifference between scheduling and walking in (otherwise it is always better to move some capacity from appointment session to walk-in session). Specifically, we have

$$R - C_D \frac{\lambda}{2\mu_S^{u*}(\mu_S^{u*} - \lambda)} = R - C_W w(\frac{\lambda_E}{\mu - \mu_S^{u*}}) = R - R' > 0,$$

$$\mu_S^{u*} = \frac{\lambda}{2} + \sqrt{\frac{(\lambda)^2}{4} + \frac{C_D\lambda}{2R'}}.$$

The cost under $(\mu_S^{u*}, \mu - \mu_S^{u*})$ is $C_O o(\frac{\lambda_E}{\mu - \lambda_S^{u*}})$. If we we set capacity $(0, \mu)$, the cost is $C_O o(\frac{\lambda+\lambda_E}{\mu})$. If we can show $\frac{\lambda_E}{\mu - \mu_S^{u*}} - \frac{\lambda+\lambda_E}{\mu} > 0$, which is equivalent to show $C_W w(\frac{\lambda_E}{\mu - \mu_S^{u*}}) - C_W w(\frac{\lambda+\lambda_E}{\mu}) > 0$, we know allocating all capacity to walk-ins costs less, i.e., we must have $\lambda_W^{u*} = \lambda$.

$$\begin{aligned} & C_W w(\frac{\lambda_E}{\mu - \mu_S^{u*}}) - C_W w(\frac{\lambda + \lambda_E}{\mu}) \\ =& C_D \frac{\lambda}{2\mu_S^{u*}(\mu_S^{u*} - \lambda)} - C_W w(\frac{\lambda + \lambda_E}{\mu}) \\ >& C_D \frac{\lambda}{2\mu(\mu - \lambda)} - C_W w(1) \\ \ge& \frac{C_D}{2M} - C_W w(1) \\ \ge& 0 \end{aligned}$$

The first inequality holds since $\mu_S^{u*} < \mu$ and $\mu \ge \lambda + \lambda_E$. The second inequality holds since $\frac{\mu(\mu-\lambda)}{\lambda} \le M$. The last inequality holds since $C_W w(1) \le \frac{C_D}{2M}$.

Now we can conclude that, when $\mu \ge \lambda + \lambda_E$ and $C_W w(1) \le \min\{R, \frac{C_D}{2M}\}$, where $M \ge \frac{\mu(\mu-\lambda)}{\lambda}$, we must have $\lambda_W^{u*} = \lambda$. Since allocating all capacity to walk-in session is a feasible solution to observable setting, and yields same patients' strategic behavior as unobservable setting. So observable setting works at least as same as unobservable setting, when $\mu \ge \lambda + \lambda_E$ and $C_W w(1) \le \min\{R, \frac{C_D}{2M}\}$.

3. Next let us prove case 2(b), i.e., when μ is sufficiently large (specifically, when $\mu \geq \overline{\mu}_S + \underline{\mu}_W$) and $\mu - \lambda$ is finite, the unobservable setting is better if $C_W w(\frac{1}{2}) > R$. (Note that $\overline{\mu}_S + \underline{\mu}_W - \lambda < \frac{C_D}{2R} + \frac{\lambda_E}{w^{-1}(\frac{R}{C_W})}$, so it is possible to have finite $\frac{\mu(\mu-\lambda)}{\lambda}$.)

(a) If the optimal allocation for observable setting is SnB regime or SWB regime, the objective value is at least the minimum overtime cost in these two regimes, i.e.,

$$\mathcal{Z}^{o*} > C_O o\big(w^{-1}(\frac{R}{C_W})\big).$$

If we set $(\mu - \underline{\mu}_W, \underline{\mu}_W)$ as the allocation for the unobservable setting, which yields an upper bound for its optimal objective value $\overline{\mathcal{Z}}^u$, then all strategic patients will schedule as $\mu - \underline{\mu}_W \geq = \overline{\mu}_S$, and the total cost is

$$\overline{\mathcal{Z}}^u = C_O o\big(w^{-1}(\frac{R}{C_W})\big) < \mathcal{Z}^{o*}.$$

Thus, in this case, unobservable setting costs less than observable setting.

(b) If the optimal allocation for observable setting must be in SnW regime. Let it be μ_S^{o*} and μ_W^{o*} respectively, then we have $\mu_S^{o*} + \mu_W^{o*} = \mu$ and $w^{-1}(\frac{R}{C_W}) \geq \frac{\pi(\delta_W^S, \mu_S^{o*})\lambda + \lambda_E}{\mu_W^*}$. Let R' denote the $C_W w(\frac{\pi(\delta_W^S, \mu_S^{o*})\lambda + \lambda_E}{\mu_W^{o*}})$, then we have $R' \leq R \leq C_W w(\frac{1}{2})$, $\mu_W^{o*} = \frac{\pi(\delta_W^S, \mu_S^{o*})\lambda + \lambda_E}{w^{-1}(\frac{R'}{C_W})}$, and $\delta_W^S = \frac{R'}{C_D}$. The corresponding objective value is $C_O o\big(w^{-1}(\frac{R'}{C_W})\big)$.

In the unobservable setting, we construct a solution $\mu'_S = \frac{\lambda}{2} + \sqrt{\frac{\lambda^2}{4} + \frac{C_D \lambda}{2R'}}$ and $\mu'_W = \frac{\lambda_E}{w^{-1}(\frac{R'}{C_W})}$. Since

$$R' = C_D \frac{\lambda}{2\mu'_S(\mu'_S - \lambda)} = C_W w(\frac{\lambda_E}{\mu'_W}),$$

then all strategic patients will choose to schedule, and the corresponding objective value is $C_O o\big(w^{-1}(\frac{R'}{C_W})\big)$. Next we show that (μ'_S, μ'_W) is a feasible solution, i.e., $\mu'_S + \mu'_W \leq \mu$. Let us denote $\mu'_S + \mu'_W - \mu$ by f.

$$\begin{aligned} f =& \frac{\lambda}{2} + \sqrt{\frac{(\lambda)^2}{4} + \frac{C_D \lambda}{2R'}} + \frac{\lambda_E}{w^{-1}(\frac{R'}{C_W})} - \mu_S^{o*} - \frac{\pi(\delta_W^S, \mu_S^{o*})\lambda + \lambda_E}{w^{-1}(\frac{R'}{C_W})} \\ =& \frac{\lambda}{2} + \sqrt{\frac{(\lambda)^2}{4} + \frac{C_D \lambda}{2R'}} - \mu_S^{o*} - \frac{\pi(\delta_W^S, \mu_S^{o*})\lambda}{w^{-1}(\frac{R'}{C_W})} \end{aligned}$$

i. If $\mu \geq \mu_S^{o*} \geq \lambda + \frac{C_D}{2R'}$, then

$$f < \lambda + \frac{C_D}{2R'} - \mu_S^{o*} \leq 0.$$

ii. If $\lambda < \mu_S^{o*} < \lambda + \frac{C_D}{2R'}$, i.e., $\mu_S^{o*} = \lambda + \epsilon$, where $0 < \epsilon < \frac{C_D}{2R'}$. Since $\mu - \lambda$ is finite, then R' is finite. When μ is sufficiently large, λ is also sufficiently large and $\frac{\lambda}{\mu_S^{o*}}$ goes to 1. Then we claim that, in the observable setting, the average amount of blocked patients in $\pi(\frac{R'}{C_D}, \mu_S^{o*})\lambda$ is at least $\frac{\lambda+\epsilon}{2\frac{R'}{C_D}\lambda} - \epsilon$ (detailed proof can be found in Lemma E.6 which follows this proof). Then we have,

$$
\begin{aligned}
f &\leq \lambda + \frac{C_D}{2R'} - \lambda - \epsilon - \frac{\frac{\lambda+\epsilon}{2\frac{R'}{C_D}\lambda} - \epsilon}{w^{-1}(\frac{R'}{C_W})} \\
&= \frac{C_D}{2R'} - \epsilon - \frac{\frac{C_D}{2R'} + \frac{\epsilon}{2\frac{R'}{C_D}\lambda} - \epsilon}{w^{-1}(\frac{R'}{C_W})} \\
&= \frac{C_D}{2R'}\Big[1 - \frac{1}{w^{-1}(\frac{R'}{C_W})}\Big] - \epsilon\Big[1 - \frac{1}{w^{-1}(\frac{R'}{C_W})}\Big] - \frac{\frac{\epsilon}{2\frac{R'}{C_D}\lambda}}{w^{-1}(\frac{R'}{C_W})} \\
&= (\frac{C_D}{2R'} - \epsilon)\Big[1 - \frac{1}{w^{-1}(\frac{R'}{C_W})}\Big] - \frac{\frac{\epsilon}{2\frac{R'}{C_D}\lambda}}{w^{-1}(\frac{R'}{C_W})} \\
&< 0
\end{aligned}
\tag{37}
$$

Since $w^{-1}(\frac{R'}{C_W}) \leq w^{-1}(\frac{R}{C_W}) < 1$ and $\frac{C_D}{2R'} > \epsilon$, the last inequality is evident. So when $\frac{\lambda}{\mu_S^*}$ goes to 1, $f < 0$.

iii. When $\mu_S^{o*} \leq \lambda - \frac{w^{-1}(\frac{R'}{C_W})C_D}{2R'[1-w^{-1}(\frac{R'}{C_W})]}$, $\pi(\frac{R'}{C_W}, \mu_S^{o*}) \geq 1 - \frac{\mu_S^{o*}}{\lambda}$.

$$
\begin{aligned}
f \leq& \lambda + \frac{C_D}{2R'} - \mu_S^{o*} - \frac{\lambda}{w^{-1}(\frac{R'}{C_W})}(1 - \frac{\mu_S^{o*}}{\lambda}) \\
=& \lambda + \frac{C_D}{2R'} - \mu_S^{o*} - \frac{\lambda}{w^{-1}(\frac{R'}{C_W})} + \frac{\mu_S^{o*}}{w^{-1}(\frac{R'}{C_W})} \\
=& (\lambda - \mu_S^{o*})\Big[1 - \frac{1}{w^{-1}(\frac{R'}{C_W})}\Big] + \frac{C_D}{2R'}
\end{aligned}
$$

Since $\lambda - \mu_S^{o*} \geq \frac{w^{-1}(\frac{R'}{C_W})C_D}{2R'[1-w^{-1}(\frac{R'}{C_W})]}$, then $f \leq 0$.

iv. When $\lambda - \frac{w^{-1}(\frac{R'}{C_W})C_D}{2R'[1-w^{-1}(\frac{R'}{C_W})]} < \mu_S^{o*} < \lambda$, then by Lemma E.6 we know that $\pi(\frac{R'}{C_D}, \mu_S^{o*})\lambda \geq \pi(\frac{R'}{C_D}, \lambda + \zeta)\lambda > \frac{C_D}{2R'} - \zeta$, when $\zeta > 0$ and ζ goes to 0 (the first inequality follows from Proposition C.2). It follows that if ζ goes to 0, we have,

$$
\begin{aligned}
f \leq& \lambda + \frac{C_D}{2R'} - \mu_S^{o*} - \frac{\frac{C_D}{2R'} - \zeta}{w^{-1}(\frac{R'}{C_W})} \\
<& \frac{w^{-1}(\frac{R'}{C_W})C_D}{2R'[1 - w^{-1}(\frac{R'}{C_W})]} + \frac{C_D}{2R'} - \frac{\frac{C_D}{2R'} - \zeta}{w^{-1}(\frac{R'}{C_W})} \\
=& \frac{1}{w^{-1}(\frac{R'}{C_W})}\Big[\frac{C_D}{2R'}\frac{2w^{-1}(\frac{R'}{C_W}) - 1}{1 - w^{-1}(\frac{R'}{C_W})} + \zeta\Big].
\end{aligned}
$$

We can easily check that, if $2w^{-1}(\frac{R'}{C_W}) < 1$, and take $\zeta < \frac{\big[1-2w^{-1}(\frac{R'}{C_W})\big]C_D}{2R'\big[1-w^{-1}(\frac{R'}{C_W})\big]}$, $\frac{C_D}{2R'}\frac{2w^{-1}(\frac{R'}{C_W})-1}{1-w^{-1}(\frac{R'}{C_W})} + \zeta$ is negative. Thus $f < 0$.

Summarizing the above analysis, unobservable setting costs less if the optimal allocation for observable setting must be in SnW regime.

Now we can conclude that, when μ is sufficiently large and $\mu - \lambda$ is finite, if $C_W w(\frac{1}{2}) > R$, unobservable setting costs less.

□

Lemma E.6. *Let w be the waiting time of a random customer in the M/D/1 queue with $\lambda < \mu$ in steady state. Let $\rho = \frac{\lambda}{\mu}$. Then the steady-state probability that the server is idle in the M/D/1/$\delta\mu$ queue with the same λ and μ is bounded from below by $Pr(w = 0|w < \delta + \frac{2}{\mu})$. Furthermore, for any $x \in (0, \frac{1}{2\delta})$ and $\mu = \lambda + x$, we find*

$$\pi(\delta, \mu)\lambda > \frac{\lambda + x}{2\delta\lambda} - x$$

when ρ is sufficiently close to 1, i.e., when λ is sufficiently large.

Proof of Lemma E.6: We start by outlining the proof.

Step 1: Assuming that $\delta\mu$ is an integer, we will

(a) introduce *M/D/1/$\delta\mu$ queue with partial blocking*, that we will denote by M/D/1/$(\delta\mu)^{pb}$;

(b) show that the blocking probability of the M/D/1/$(\delta\mu)^{pb}$ queue can be expressed by a conditional probability involving the M/D/1 queue;

(c) show that the blocking probability of the M/D/1/$\delta\mu$ queue is lower bounded by the blocking probability of the M/D/1/$(\delta\mu + 1)^{pb}$ queue;

Step 2: Use the bound of step 1 to derive a bound for any $\delta\mu$;

Step 3: Use the classic heavy traffic limit to find the number of blocked patients when λ is sufficiently large, i.e., when the traffic intensity $\rho \to 1$.

Before detailing the steps, it will be convenient to define the workload in a queue at time t as the total amount of time units required to serve all customers in the system at t. That is, for any queue with deterministic service time $1/\mu$, suppose there are k customers in the queue plus a customer in service with residual service time $r \in (0, 1/\mu]$. Then the *workload* in the system is defined as $k/\mu + r$. Since the service order is first-come-first-served (FCFS), a customer that joins a system that has workload w (before the customer joins) will incur a total waiting time of w before entering service. Thus for the M/D/1 queue, the workload distribution is the same as the waiting time distribution.

Now we provide details of our proof.

1. When $\delta\mu$ is an integer.

(a) (We introduce the M/D/1/$\delta\mu$ queue with partial blocking.) Suppose the arrival process of customers is a Poisson process with arrival rate λ and the service time is a deterministic constant $\frac{1}{\mu}$. Any customer joins the queue if he observes the workload in the system to be no longer than $\frac{\delta\mu - 1}{\mu}$, and such customers leave the system after being served for $\frac{1}{\mu}$ time units based on the FCFS principle. Customers that upon arrival observe a workload in the system w^{pb} within the range $(\frac{\delta\mu - 1}{\mu}, \delta)$ will join the queue and be committed to be served for $\delta - w^{pb}$ time units before leaving the queue. (Immediately after joining, such customers bring the workload in the queue to δ.) Any customers who see a delay longer than δ will not join. This queue is called *M/D/1/$\delta\mu$ queue with partial blocking* and we denote it by $M/D/1/(\delta\mu)^{pb}$. A detailed introduction can be found in Moran (1956), where partial blocking is also called a *dam*.

(b) Consider the steady state workload w in the M/D/1 queue conditioned on it being less than δ, i.e. $w|w \leq \delta$. This workload distribution corresponds to the workload distribution of the M/D/1 queue, where all periods where the workload exceeds δ have been *cut out*. Consider the dynamics of the resulting system:

Customers join the queue just as the M/D/1 queue (each increasing the workload by $1/\mu$), except when that brings the total workload above δ. (When that happens, the workload in M/D/1 is above δ for some time, until it drops to δ again, but that part is cut out.) When a customer would bring the queue to a total workload above δ, the workload in the system instead jumps to δ. Note also that when the workload in M/D/1 drops to δ, because of the memoryless property of the arrival process, the state of the system is as if nothing happened in between. Thus the resulting process has the same dynamics as the M/D/1/$(\delta\mu)^{pb}$ queue. Therefore,

$$Pr(w^{pb}=0)=Pr(w=0|w\leq\delta) \tag{38}$$

where w is the steady state workload distribution of the M/D/1 queue, and w^{pb} is the steady state workload distribution of the $M/D/1/(\delta\mu)^{pb}$ queue.

(c) We let w^b represent the workload distribution of M/D/1/$\delta\mu$ queue. Next we compare the M/D/1/$(\delta\mu+1)^{pb}$ queue with M/D/1/$\delta\mu$ queue when $w^b\in(\delta,\delta+1/\mu)$ and $w^b=w^{pb}$. Since $\delta\mu$ is integer by assumption, $w^b>\delta$ implies that there are $k\geq 1+\lceil\delta\mu\rceil$ customers in the system (including the one in service). Any new arrivals will thus not join under the setting of M/D/1/$\delta\mu$ queue, while they will join under the setting of M/D/1/$(\delta\mu+1)^{pb}$ queue and be committed to be served for $\delta+1/\mu-w^{pb}$ time units. When the waiting time is smaller than δ, patients will join in both settings, while for a waiting time higher than $\delta+1/\mu$ patients will join in neither setting. Thus, since patients join in the M/D/1/$(\delta\mu+1)^{pb}$ whenever they would join the M/D/1/$\delta\mu$ queue, the latter queue will spend more time idling. Using (38), we find that $Pr(w^{pb}=0)=Pr(w=0|w\leq\delta+1/\mu)$ for the M/D/1/$(\delta\mu+1)^{pb}$ queue, thus we have

$$Pr(w^b=0)>Pr(w=0|w\leq\delta+1/\mu).$$

(Remark: The additional term $1/\mu$ arises because for the M/D/1/$(\delta\mu)^{pb}$ queue, the maximum workload in the system is δ, while the maximum workload in the M/D/1/$\delta\mu$ is $\delta+1/\mu$.)

2. Recall that for any (possibly fractional) value of $\delta\mu$ by Proposition C.2, the blocking probability of M/D/1/$\delta\mu$ can be lower bounded by the blocking probability of $M/D/1/\lceil\delta\mu\rceil$. We also have $Pr(w\leq\frac{\lceil\delta\mu\rceil}{\mu})<Pr(w\leq\delta+\frac{1}{\mu})$. Thus

$$Pr(w^b=0)>Pr(w=0|w\leq\frac{\lceil\delta\mu\rceil}{\mu}+\frac{1}{\mu})=Pr(w=0|w\leq\frac{\lceil(\delta+\frac{1}{\mu})\mu\rceil}{\mu})>Pr(w=0|w\leq\delta+\frac{1}{\mu}+\frac{1}{\mu}).$$

3. We proceed the analysis for $\rho\to 1$ using the classic heavy traffic limit theory, which can be described as follows: Let S and T be the service time and interarrival time, $\alpha=-E(S-T)$, and $\beta^2=var(S-T)$. Then waiting time w tends to an exponential distribution with parameter $\frac{2\alpha}{\beta^2}$ as $\rho\to 1$ (Kingman (1962)). In our setting, we have $w\to exp(2(1-\rho)\lambda)$ when $\rho\to 1$. More precisely, we fix $x\in(0,\frac{1}{2\delta})$, let $\mu=\lambda+x$, and let $\lambda\to\infty$, which leads to $\rho\to 1$. Then by Kingman (1962), for any $\epsilon>0$, there exists a $\bar{\lambda}$ such that for all $\lambda>\bar{\lambda}$ it holds that $Pr(w\leq\delta+\frac{2}{\mu})\leq 1-e^{-2(1-\rho)\lambda(\delta+\frac{2}{\mu})}+\epsilon=1-e^{-2x\rho(\delta+\frac{2}{\mu})}+\epsilon$. In particular, we let $\epsilon'=x^2\delta^2/12$, and $\bar{\lambda}'$ be the associated $\bar{\lambda}$.

Using Taylor expansion series we have $e^{-y} > 1-y+\frac{y^2}{2}-\frac{y^3}{6}$ for any $5 \geq y \geq 0$. Particularly, $e^{-y} > 1-y+\frac{y^2}{6}$ when $y < 2$, that is,

$$1-e^{-y} < y-\frac{y^2}{6}. \tag{39}$$

Note that $2x\rho(\delta+\frac{2}{\mu}) < 2$, since $x < 1/(2\delta)$ and $\mu = \lambda + x$ will be sufficiently large as discussed below. By replacing y with $2x\rho(\delta+\frac{2}{\mu})$, we have:

$$\begin{aligned}
1-e^{-2x\rho(\delta+\frac{2}{\mu})}+\epsilon' &< 2x\rho(\delta+\frac{2}{\mu})-\frac{4x^2\rho^2(\delta+\frac{2}{\mu})^2}{6}+\epsilon' \\
&= 2x\rho\delta+\Big[4x\frac{\rho}{\mu}-\frac{2x^2\rho^2(\delta+\frac{2}{\mu})^2}{6}\Big]-\frac{2x^2\rho^2(\delta+\frac{2}{\mu})^2}{6}+\epsilon' \\
&= 2x\rho\delta+4x\frac{\rho}{\mu}\Big[1-\frac{x\rho\mu(\delta+\frac{2}{\mu})^2}{12}\Big]-\Big[\frac{2x^2\rho^2(\delta+\frac{2}{\mu})^2}{6}-\epsilon'\Big] \\
&< 2x\rho\delta.
\end{aligned}$$

The last inequality holds when λ is sufficiently large for the second term and the third term to be negative, and larger than $\bar{\lambda}'$. Now, remember that $Pr(w^b = 0) = Pr(w = 0|w \leq \delta + \frac{2}{\mu})$. Then $Pr(w^b = 0) > \frac{1}{2\lambda\delta}$ for λ large enough. In other words, the capacity utilization ratio is at most $1-\frac{1}{2\delta\lambda}$, i.e.,

$$\frac{\hat{\lambda}}{\mu} < 1-\frac{1}{2\delta\lambda} \Leftrightarrow \hat{\lambda} < \mu - \frac{\mu}{2\delta\lambda},$$

where $\hat{\lambda}$ is the steady state arrival rate of customers join the queue when capacity is μ.
Finally, remembering that $\mu = \lambda + x$, we find

$$\pi(\delta,\mu)\lambda = \lambda - \hat{\lambda} > \frac{\mu}{2\lambda\delta} - x = \frac{\lambda+x}{2\delta\lambda} - x.$$

□

Supplemental Details of Remark 3: Consider the following "lost-sales" scenario. The provider will turn away additional walk-ins when the traffic intensity of walk-in hours ρ is over 100%, i.e., when the expected walk-in arrival rate reaches the capacity allocated to walk-in hours. To model this scenario, the provider's objective shown in formulation (GP) is simply to minimize demand loss. Specifically, her objective function is revised by replacing the overtime cost $C_O o(\cdot)$ with a cost due to patients turned away; the latter cost is assumed to be $C_L(\lambda_W + \lambda_E - \mu_W)^+$, where $\lambda_W + \lambda_E$ is the total expected walk-in arrival rate including strategic and exogenous walk-ins, and μ_W is the capacity allocated to walk-ins. Here we adopt a deterministic approximation to represent the amount of patients turned away for tractability. To model the utility for walk-ins, we need to take into account their potential utility loss due to being turned down. When $\rho = \frac{\lambda_W+\lambda_E}{\mu_W} > 1$, each walk-in will be turned down with probability $1-\frac{1}{\rho}$ (assuming walk-in hours are rationed randomly). Then the utility of walk-in can be modeled as $u_W = R - C_W w(\rho) - R(1-\frac{1}{\rho})^+$, where being turned down is assumed to incur a utility loss of R and the last term represents the expected utility loss due to being turned down. For convenience, we rewrite $u_W = R - C_W t(\rho)$, where $t(\rho) = w(\rho) + \frac{R}{C_W}(1-\frac{1}{\rho})^+$, so $t(\cdot)$ plays the role of $w(\cdot)$ as in our previous analysis. With these replacements, Assumption 2 on $o(\rho)$ is no longer needed, and Assumption 1 on $w(\rho)$ is relaxed (here we only assume $t(\rho)$ is a strictly increasing function of ρ). We shall note that all previous results on patient equilibrium continue to hold, but the provider's optimization problem has changed and so does the model comparison. Theorem E.1 below shows the model comparison results of this lost-sales scenario.

Theorem E.1. *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 the lost-sales scenario described above.*

1. *When μ is sufficiently small, the observable setting costs less.*
2. *When μ is sufficiently large and if there exists an $M > 1$ such that $\mu - \lambda \leq M \times \lambda_E$,*
   (a) *if $C_W \times t(1) \leq R$, both settings have the same performance;*
   (b) *if $C_W \times t(\frac{1}{M}) > R$, the unobservable setting costs less.*

Proof of Theorem E.1: Consider the provider's cost function $p_B\lambda + (p_W\lambda + \lambda_E - \mu_W)^+$ for the observable setting and $\lambda_B + (\lambda_W + \lambda_E - \mu_W)^+$ for the unobservable setting, respectively.

1. Consider $\mu < \min\{\underline{\mu}_W, \overline{\mu}_S, \lambda_E\}$, where $\underline{\mu}_W = \frac{\lambda_E}{t^{-1}(\frac{R}{C_W})}$ and $\overline{\mu}_S = \frac{\lambda}{2} + \sqrt{\frac{(\lambda)^2}{4} + \frac{C_D\lambda}{2R}}$. Since $\mu < \underline{\mu}_W$, then $p_W = 0$ in the observable setting and $\lambda_W = 0$ in the unobservable setting.

In the unobservable setting, since $\mu < \overline{\mu}_S$, $\lambda_B = \lambda - \frac{2R\mu_S^2}{2R\mu_S + C_D}$. Since $\mu < \lambda_E$, then $\lambda_E - \mu_W > 0$. It follows that the objective function becomes $\lambda - \frac{2R\mu_S^2}{2R\mu_S + C_D} + \lambda_E - \mu + \mu_S$. The optimal solution is $\mu_W^{u*} = \mu$, which is a feasible solution to the observable setting, yielding the same objective value. So the observable setting costs less than the unobservable setting.

2(a). Let $\mu > \lambda + \lambda_E$. Since $C_W \times t(1) \leq R$, then $\mu > \lambda + \lambda_E > \frac{\lambda + \lambda_E}{t^{-1}(\frac{R}{C_W})}$. Setting $\mu_W^o = \mu_W^u = \mu$ yields $p_W = 1$ in the observable setting and $\lambda_W = \lambda$ in the unobservable setting. The corresponding objective value is 0 in both settings. So they have the same performance.

2(b). Let μ be such that $\mu - \overline{\mu}_S > \lambda_E$. In the unobservable setting, let $\mu_S^u = \overline{\mu}_S$ and $\mu_W^u = \mu - \overline{\mu}_S$, then we have $\lambda_S = \lambda$ (since $C_W \times t(1/M) > R$, then $\mu - \overline{\mu}_S < \mu - \lambda \leq M \times \lambda_E < \frac{\lambda_E}{t^{-1}(\frac{R}{C_W})}$, indicating that the utility of walk-in is negative). The corresponding objective value is $(\lambda_E - \mu + \overline{\mu}_S)^+ = 0$.

In the observable setting, define $\overline{\mu}_O = \min_{\mu_S}\{\frac{\pi(\frac{R}{C_D}, \mu_S)\lambda + \lambda_E}{t^{-1}(\frac{R}{C_W})} + \mu_S\}$ which represents the minimum capacity for the observable setting to attract all strategic patients. Since $C_W \times t(\frac{1}{M}) > R$, i.e., $\frac{1}{t^{-1}(\frac{R}{C_W})} > M$, then $\overline{\mu}_O > M\lambda_E + \mu_S^* + \pi(\frac{R}{C_D}, \mu_S^*)\lambda$ where $\mu_S^* = \arg\min\{\frac{\pi(\frac{R}{C_D}, \mu_S)\lambda + \lambda_E}{t^{-1}(\frac{R}{C_W})} + \mu_S\}$. And since $(1 - \pi_0(\frac{R}{C_D}, \mu_S^*))\mu_S^* = (1 - \pi(\frac{R}{C_D}, \mu_S^*))\lambda$, then $\mu_S^* + \pi(\frac{R}{C_D}, \mu_S^*)\lambda \geq \lambda$. So we have $\overline{\mu}_O > M\lambda_E + \lambda \geq \mu$. Clearly, there must be lost demand as the observable setting cannot attract all strategic patients, indicating that the unobservable setting costs less.

□

Proof of Proposition 6:

1. First, let us prove that if $\frac{C_L}{C_O} \leq \underline{\alpha}$, the triage model costs less.

Denote the optimal capacity allocations to the Strategic Model as (μ_S^{s*}, μ_W^{s*}), and the corresponding objective value is $\mathcal{Z}_s^*$. Let λ_S^{s*} and λ_W^{s*} denote the corresponding scheduling rate and walk-in rate of strategic patients. If we use (μ_S^{s*}, μ_W^{s*}) as the capacity allocation in Triage Model, we can get an upper bound for the optimal objective value, denoted by $\overline{\mathcal{Z}}_t$.

If $\mu_W^{s*} \leq \underline{\mu}_W$, then it is SnB in Strategic Model. Thus $\mathcal{Z}_s^* = \overline{\mathcal{Z}}_t$, i.e., Triage Model works better and has smaller optimal cost than Strategic Model.

Next we consider the case where $\mu_W^{s*} > \underline{\mu}_W$, i.e., $\mu_S^{s*} < \mu - \underline{\mu}_W$, i.e., the equilibrium of Strategic Model must be SnW or SWB, then we have $\lambda_S^{s*} \leq [1 - \pi(\frac{R}{C_D}, \mu_S^*)]\lambda$.

$$\begin{aligned}
&\mathcal{Z}_s^* - \overline{\mathcal{Z}}_t \\
=&C_L(\lambda - \lambda_S^{s*} - \lambda_W^{s*}) + C_O o(\frac{\lambda_E + \lambda_W^{s*}}{\mu_W^{s*}}) - C_L\pi(\frac{R}{C_D}, \mu_S^{s*})\lambda - C_O o(\frac{\lambda_E}{\mu_W^{s*}}) \\
\geq&C_L\Big[\pi(\frac{R}{C_D}, \mu_S^{s*})\lambda - \lambda_W^{s*}\Big] + C_O o(\frac{\lambda_E + \lambda_W^{s*}}{\mu_W^{s*}}) - C_L\pi(\frac{R}{C_D}, \mu_S^{s*})\lambda - C_O o(\frac{\lambda_E}{\mu_W^{s*}}) \\
=&C_O\Big[o(\frac{\lambda_E + \lambda_W^{s*}}{\mu_W^{s*}}) - o(\frac{\lambda_E}{\mu_W^{s*}})\Big] - C_L\lambda_W^{s*} \\
\geq&C_O o'(\frac{\lambda_E}{\mu_W^{s*}})\frac{\lambda_W^{s*}}{\mu_W^{s*}} - C_L\lambda_W^{s*} \\
=&\Big[C_O o'(\frac{\lambda_E}{\mu_W^{s*}})\frac{1}{\mu_W^{s*}} - C_L\Big]\lambda_W^{s*} \\
\geq&\Big[C_O o'(\frac{\lambda_E}{\mu})\frac{1}{\mu} - C_L\Big]\lambda_W^{s*} \\
\geq&0
\end{aligned}$$

The first inequality holds since $\lambda_S^{s*} \leq [1 - \pi(\frac{R}{C_D}, \mu_S^*)]\lambda$. The second inequality holds since $o(\cdot)$ is increasing and convex. The third inequality holds since $\mu_W^{s*} \leq \mu$. The last inequality holds since $\frac{C_L}{C_O} \leq \underline{\alpha} = o'(\frac{\lambda_E}{\mu})\frac{1}{\mu}$.

Summarizing the above analysis, when $\frac{C_L}{C_O} \leq \underline{\alpha}$, Triage Model costs less.

2. Second, let us prove that if $\frac{C_L}{C_O} \geq \overline{\alpha}$, the strategic model costs less.

Denote the optimal capacity allocations to the Triage Model as (μ_S^{t*}, μ_W^{t*}), and the corresponding objective value is $\mathcal{Z}_t^*$. If we use (μ_S^{t*}, μ_W^{t*}) as the capacity allocation in Strategic Model, we can get an upper bound for the optimal objective value, denoted by $\overline{\mathcal{Z}}_s$. Let $\lambda_W^{s,t*}$ denote the corresponding walk-in rate of strategic patients.

If $\mu_W^{t*} \leq \underline{\mu}_W$, then (μ_S^{t*}, μ_W^{t*}) is a feasible but not necessarily optimal solution to Strategic Model, thus Strategic Model works better and has no bigger optimal cost than Triage Model in this case.

So we focus on the case where $\mu_W^{t*} > \underline{\mu}_W$, i.e., $\mu_S^{t*} < \mu - \underline{\mu}_W$. In the equilibrium of Strategic Model under capacity allocation (μ_S^{t*}, μ_W^{t*}), we must be in SnW regime or SWB regime.

(a) If $\mu \geq \overline{\mu}_O$, i.e., the equilibrium of Strategic Model under capacity allocation (μ_S^{t*}, μ_W^{t*}) is in SnW regime, then we have $\frac{\lambda_W^{s,t*} + \lambda_E}{\mu_W^{t*}} \leq \frac{\lambda_E}{\underline{\mu}_W}$;

$$\begin{aligned}
&\mathcal{Z}_t^* - \overline{\mathcal{Z}}_s \\
=&C_L\pi(\frac{R}{C_D}, \mu_S^{t*})\lambda + C_O o(\frac{\lambda_E}{\mu_W^{t*}}) - C_O o(\frac{\lambda_E + \lambda_W^{s,t*}}{\mu_W^{t*}}) \\
\geq&C_L\pi(\frac{R}{C_D}, \mu - \underline{\mu}_W)\lambda - C_O o'(\frac{\lambda_E + \lambda_W^{s,t*}}{\mu_W^{t*}})\frac{\lambda_W^{s,t*}}{\mu_W^{t*}} \\
\geq&C_L\pi(\frac{R}{C_D}, \mu - \underline{\mu}_W)\lambda - C_O o'(\frac{\lambda_E}{\underline{\mu}_W})\frac{\lambda_E}{\underline{\mu}_W} \\
\geq&0
\end{aligned}$$

The first inequality holds since $\mu_S^{t*} < \mu - \underline{\mu}_W$, $\pi(\frac{R}{C_D}, \mu_S)$ is decreasing in μ_S, and $o(\cdot)$ is increasing and convex. The second inequality holds since $\frac{\lambda_W^{s,t*}}{\mu_W^{t*}} < \frac{\lambda_W^{s,t*} + \lambda_E}{\mu_W^{t*}} \leq \frac{\lambda_E}{\underline{\mu}_W}$. The last inequality holds since $\frac{C_L}{C_O} \geq \overline{\alpha} \geq o'(\frac{\lambda_E}{\underline{\mu}_W})\frac{1}{\underline{\mu}_W}\frac{\lambda_E}{\pi(\frac{R}{C_D}, \mu - \underline{\mu}_W)\lambda}$.

(b) If $\mu < \overline{\mu}_O$, i.e., the equilibrium of Strategic Model under capacity allocation (μ_S^{t*}, μ_W^{t*}) is in SWB regime, then we have $\frac{\lambda_W^{s,t*}+\lambda_E}{\mu_W^{t*}} = \frac{\lambda_E}{\underline{\mu}_W}$.

$$\begin{aligned}
&\mathcal{Z}_t^* - \overline{\mathcal{Z}}_s \\
=&C_L\pi(\frac{R}{C_D}, \mu_S^{t*})\lambda + C_O o(\frac{\lambda_E}{\mu_W^{t*}}) - C_L\left[\pi(\frac{R}{C_D}, \mu_S^{t*})\lambda - \lambda_W^{s,t*}\right] - C_O o(\frac{\lambda_W^{s,t*}+\lambda_E}{\mu_W^{t*}}) \\
=&C_L\lambda_W^{s,t*} + C_O o(\frac{\lambda_E}{\mu_W^{t*}}) - C_O o(\frac{\lambda_W^{s,t*}+\lambda_E}{\mu_W^{t*}}) \\
\geq&C_L\lambda_W^{s,t*} - C_O o'(\frac{\lambda_E+\lambda_W^{s,t*}}{\mu_W^{t*}})\frac{\lambda_W^{s,t*}}{\mu_W^{t*}} \\
\geq&C_L\lambda_W^{s,t*} - C_O o'(\frac{\lambda_E}{\underline{\mu}_W})\frac{\lambda_W^{s,t*}}{\underline{\mu}_W} \\
=&\left[C_L - C_O o'(\frac{\lambda_E}{\underline{\mu}_W})\frac{1}{\underline{\mu}_W}\right]\lambda_W^{s,t*}
\end{aligned}$$

The first inequality holds since $o(\cdot)$ is increasing and convex. The second inequality holds since $\frac{\lambda_W^{s,t*}+\lambda_E}{\mu_W^{t*}} = \frac{\lambda_E}{\underline{\mu}_W}$ and $\mu_W^{t*} > \underline{\mu}_W$. The last inequality holds since $\frac{C_L}{C_O} \geq \overline{\alpha} \geq o'(\frac{\lambda_E}{\underline{\mu}_W})\frac{1}{\underline{\mu}_W}$.

Summarizing the above analysis, when $\frac{C_L}{C_O} \geq \overline{\alpha}$, Strategic Model costs less.

□

Proof of Proposition D.4:

1. We first show that when $\mu \geq (\lambda+\lambda_E)\max\{\frac{\overline{\mu}_S}{\lambda}, \frac{\mu_W}{\lambda_E}\} = \max\{(1+\frac{\lambda_E}{\lambda})\overline{\mu}_S, \overline{\mu}_W\}$, Triage Model costs less.

Let (μ_S^{s*}, μ_W^{s*}) denote the optimal allocation in Strategic Model. The corresponding objective value is $\mathcal{Z}_s^*$, the corresponding scheduling rate is λ_S^{s*} and the corresponding walk-in rate of strategic patients is λ_W^{s*}. By Proposition 5 and Lemma D.4, we know that either $\lambda_S^{s*} = \lambda$ or $\lambda_W^{s*} = \lambda$, since $\mu \geq \overline{\mu}_W$.

- If $\lambda_S^{s*} = \lambda$, then when we use (μ_S^{s*}, μ_W^{s*}) as capacity allocation in Triage Model, the equilibrium is also $\lambda_S^{t,s*} = \lambda$, yielding same objective value as Strategic Model. So in this case, Triage Model costs less.
- If $\lambda_W^{s*} = \lambda$, since $\mu \geq \overline{\mu}_W$, then $(0, \mu)$ is the best solution to Strategic Model yielding $\lambda_W^{s*} = \lambda$. So $\mathcal{Z}_s^* = C_O o(\frac{\lambda_E+\lambda}{\mu})$. Since $\mu \geq \overline{\mu}_S$, then we can set $(\overline{\mu}_S, \mu - \overline{\mu}_S)$ in Triage Model, the corresponding objective value $\overline{\mathcal{Z}}_t = C_O o(\frac{\lambda_E}{\mu-\overline{\mu}_S})$. One can verify that, when $\mu \geq (1+\frac{\lambda_E}{\lambda})\overline{\mu}_S$, $\overline{\mathcal{Z}}_t \leq \mathcal{Z}_s^*$.

Summarizing above analysis, we have that, when $\mu \geq \max\{(1+\frac{\lambda_E}{\lambda})\overline{\mu}_S, \overline{\mu}_W\}$, Triage Model costs less.

2. Then we show that when $\frac{C_L}{C_O} \leq \underline{\beta} = o'(\frac{\lambda_E}{\mu})\frac{1}{\mu}$, Triage Model works better.

Let (μ_S^{s*}, μ_W^{s*}) denote the optimal allocation in Strategic Model. The corresponding objective value is $\mathcal{Z}_s^*$, the corresponding scheduling rate is λ_S^{s*} and the corresponding walk-in rate of strategic patients is λ_W^{s*}. Let us set (μ_S^{s*}, μ_W^{s*}) as capacity allocation in Triage Model, the corresponding objective value is an upper bound, denoted by $\overline{\mathcal{Z}}_t$. Denote the corresponding scheduling rate as $\lambda_S^{t,s*}$ and the corresponding walk-in rate of strategic patients as $\lambda_W^{t,s*}$.

- If $\lambda_W^{s*} = 0$, then when we set (μ_S^{s*}, μ_W^{s*}) as capacity allocation in Triage Model, strategic patients behave as same as Strategic Model, i.e., $(\lambda_S^{t,s*}, \lambda_W^{t,s*}) = (\lambda_S^{s*}, \lambda_W^{s*})$, yielding same objective value, i.e., $\overline{\mathcal{Z}}_t \leq \mathcal{Z}_s^*$.
- If $\lambda_W^{s*} > 0$, by Lemma D.4, we have $\lambda_S^{s*} = 0$.

—If $\mu_S^* \geq \overline{\mu}_S$, then scheduling option is better than balking option in both models, we must have $\lambda_W^{s*} = \lambda$, $\lambda_S^{t,s*} = \lambda$, $\mathcal{Z}_s^* = C_O o(\frac{\lambda+\lambda_E}{\mu_W^{s*}})$ and $\overline{\mathcal{Z}}_t = C_O o(\frac{\lambda_E}{\mu_W^{s*}})$. It is easy to check $\overline{\mathcal{Z}}_t \leq \mathcal{Z}_s^*$.

—If $\mu_S^* < \overline{\mu}_S$, we have $\lambda_S^{t,s*} = \frac{2R(\mu_S^{s*})^2}{C_D + 2R\mu_S^{s*}}$. Then,

$$\begin{aligned}
&\mathcal{Z}_s^* - \overline{\mathcal{Z}}_t \\
=&C_L(\lambda - \lambda_W^{s*}) + C_O o(\frac{\lambda_W^{s*} + \lambda_E}{\mu_W^{s*}}) - C_L\Big[\lambda - \frac{2R(\mu_S^{s*})^2}{C_D + 2R\mu_S^{s*}}\Big] - C_O o(\frac{\lambda_E}{\mu_W^{s*}}) \\
\geq& - C_L\lambda_W^{s*} + C_O o(\frac{\lambda_W^{s*} + \lambda_E}{\mu_W^{s*}}) - C_O o(\frac{\lambda_E}{\mu_W^{s*}}) \\
\geq& C_O o'(\frac{\lambda_E}{\mu_W^{s*}})\frac{\lambda_W^{s*}}{\mu_W^{s*}} - C_L\lambda_W^{s*} \\
\geq& \Big[C_O o'(\frac{\lambda_E}{\mu})\frac{1}{\mu} - C_L\Big]\lambda_W^{s*} \\
\geq& 0
\end{aligned}$$

The first inequality holds since $\frac{2R(\mu_S^{s*})^2}{C_D + 2R\mu_S^{s*}}$ is positive. The second inequality holds since $o(\cdot)$ is increasing and convex. The third inequality holds since $\frac{C_L}{C_O} \leq \underline{\beta} = o'(\frac{\lambda_E}{\mu})\frac{1}{\mu}$.

Summarizing the analysis above, we have when $\frac{C_L}{C_O} \leq \underline{\beta}$, Triage Model costs less.

3. Now we show that when $\mu \leq \overline{\mu}_S + \underline{\mu}_W$ and $\frac{C_L}{C_O} \geq \overline{\beta} = o'(\frac{\lambda_E}{\underline{\mu}_W})\frac{1}{\underline{\mu}_W} \max\big\{\frac{\lambda_E}{\lambda - \frac{2R(\mu - \underline{\mu}_W)^2}{C_D + 2R(\mu - \underline{\mu}_W)}}, 1\big\}$, Strategic Model works better.

Let (μ_S^{t*}, μ_W^{t*}) denote the optimal allocation in Triage Model. The corresponding objective value is $\mathcal{Z}_t^*$, and the corresponding scheduling rate is λ_S^{t*}. Let us set (μ_S^{t*}, μ_W^{t*}) as capacity allocation in Strategic Model, the corresponding objective value is an upper bound, denoted by $\overline{\mathcal{Z}}_s$. Denote the corresponding scheduling rate as $\lambda_S^{s,t*}$ and the corresponding walk-in rate of strategic patients as $\lambda_W^{s,t*}$.

If $\mu_W^{t*} \leq \underline{\mu}_W$, then setting (μ_S^{t*}, μ_W^{t*}) in the strategic model leads to the same cost as in the triage model, but this allocation is not necessarily optimal. So the strategic model costs less. It is left to consider the situation when $\mu_W^{t*} > \underline{\mu}_W$. In this case, the equilibrium in the strategic model must be in SnW or SWB regime, depending on how large μ is (see Figure 2(b)). Also, since $\mu \leq \overline{\mu}_S + \underline{\mu}_W$, then $\mu_S^{t*} < \overline{\mu}_S$, indicating $\lambda_S^{s,t*} < \lambda$, and $\lambda_S^{t*} = \frac{2R(\mu_S^{t*})^2}{C_D + 2R\mu_S^{t*}}$

(a) If the equilibrium of Strategic Model under capacity allocation (μ_S^{t*}, μ_W^{t*}) is in SnW regime, then we have $\frac{\lambda_W^{s,t*} + \lambda_E}{\mu_W^{t*}} \leq \frac{\lambda_E}{\underline{\mu}_W}$;

$$\begin{aligned}
&\mathcal{Z}_t^* - \overline{\mathcal{Z}}_s \\
=&C_L\Big[\lambda - \frac{2R(\mu_S^{t*})^2}{C_D + 2R\mu_S^{t*}}\Big] + C_O o(\frac{\lambda_E}{\mu_W^{t*}}) - C_O o(\frac{\lambda_E + \lambda_W^{s,t*}}{\mu_W^{t*}}) \\
\geq& C_L\Big[\lambda - \frac{2R(\mu - \underline{\mu}_W)^2}{C_D + 2R(\mu - \underline{\mu}_W)}\Big] - C_O o'(\frac{\lambda_E + \lambda_W^{s,t*}}{\mu_W^{t*}})\frac{\lambda_W^{s,t*}}{\mu_W^{t*}} \\
\geq& C_L\Big[\lambda - \frac{2R(\mu - \underline{\mu}_W)^2}{C_D + 2R(\mu - \underline{\mu}_W)}\Big] - C_O o'(\frac{\lambda_E}{\underline{\mu}_W})\frac{\lambda_E}{\underline{\mu}_W} \\
\geq& 0
\end{aligned}$$

The first inequality holds since $\mu_S^{t*} = \mu - \mu_W^{t*} < \mu - \underline{\mu}_W$, and $o(\cdot)$ is increasing and convex. The second inequality holds since $\frac{\lambda_W^{s,t*}}{\mu_W^{t*}} < \frac{\lambda_W^{s,t*} + \lambda_E}{\mu_W^{t*}} \leq \frac{\lambda_E}{\underline{\mu}_W}$. The last inequality holds since $\frac{C_L}{C_O} \geq \overline{\beta} \geq o'(\frac{\lambda_E}{\underline{\mu}_W})\frac{1}{\underline{\mu}_W}\frac{\lambda_E}{\lambda - \frac{2R(\mu - \underline{\mu}_W)^2}{C_D + 2R(\mu - \underline{\mu}_W)}}$.

(b) If the equilibrium of Strategic Model under capacity allocation (μ_S^{t*}, μ_W^{t*}) is in SWB regime, then we have $\frac{\lambda_W^{s,t*}+\lambda_E}{\mu_W^{t*}} = \frac{\lambda_E}{\underline{\mu}_W}$;

$$\begin{aligned}
&\mathcal{Z}_t^* - \overline{\mathcal{Z}}_s \\
=&C_L(\lambda - \lambda_S^{t*}) + C_O o(\frac{\lambda_E}{\mu_W^{t*}}) - C_L(\lambda - \lambda_S^{t*} - \lambda_W^{s,t*}) - C_O o(\frac{\lambda_W^{s,t*} + \lambda_E}{\mu_W^{t*}}) \\
=&C_L \lambda_W^{s,t*} + C_O o(\frac{\lambda_E}{\mu_W^{t*}}) - C_O o(\frac{\lambda_W^{s,t*} + \lambda_E}{\mu_W^{t*}}) \\
\geq&C_L \lambda_W^{s,t*} - C_O o'(\frac{\lambda_E + \lambda_W^{s,t*}}{\mu_W^{t*}}) \frac{\lambda_W^{s,t*}}{\mu_W^{t*}} \\
\geq&C_L \lambda_W^{s,t*} - C_O o'(\frac{\lambda_E}{\underline{\mu}_W}) \frac{\lambda_W^{s,t*}}{\underline{\mu}_W} \\
=&\Big[C_L - C_O o'(\frac{\lambda_E}{\underline{\mu}_W}) \frac{1}{\underline{\mu}_W}\Big] \lambda_W^{s,t*} \\
\geq&0
\end{aligned}$$

The first inequality holds since $o(\cdot)$ is increasing and convex. The second inequality holds since $\frac{\lambda_W^{s,t*}+\lambda_E}{\mu_W^{t*}} = \frac{\lambda_E}{\underline{\mu}_W}$ and $\mu_W^{t*} > \underline{\mu}_W$. The last inequality holds since $\frac{C_L}{C_O} \geq \overline{\beta} \geq \frac{o'(\frac{\lambda_E}{\underline{\mu}_W})}{\underline{\mu}_W}$.

Summarizing the analysis above, we have when $\mu \leq \overline{\mu}_S + \underline{\mu}_W$ and $\frac{C_L}{C_O} \geq \overline{\beta} = \frac{o'(\frac{\lambda_E}{\underline{\mu}_W})}{\underline{\mu}_W} \max\left\{\frac{\lambda_E}{\lambda - \frac{2R(\mu-\underline{\mu}_W)^2}{C_D + 2R(\mu - \underline{\mu}_W)}}, 1\right\}$, Strategic Model costs less.

□

E.4. Proof of Results in Appendix C

Proof of Lemma C.1:

Let us look at our $M/D/1$ queue with delay threshold δ, where customers join the queue when the observed delay is no more than δ. Customers join if the observed delay d is smaller than or equals to δ. When the system size is smaller than or equal to $K-1$, then the observed delay must be smaller than or equal to $\frac{K-1}{\mu_S} = \frac{\lceil \delta\mu_S \rceil - 1}{\mu_S} < \delta$. So customers join when the system size is smaller than or equal to $K-1$. When the system size is greater than or equal to $K+1$, then the observed delay must be greater than or equal to $\frac{K+1}{\mu_S} = \frac{\lceil \delta\mu_S \rceil + 1}{\mu_S} > \delta$. So customers balk when the system size is greater than or equal to $K+1$. When the system size is K, then the observed delay must be uniformly distributed in $[\frac{K-1}{\mu_S}, \frac{K}{\mu_S}]$. In this case, the probability of observed delay is no more than δ is $p = \delta\mu_S + 1 - K$. So when customers join with probability p when the system size is K. Thus $M/D/1$ queue with delay threshold δ implies $M/D/1/K-1+p$ queue.

Now we look at M/D/1 queue with buffer size $K-1+p$, where customers join the queue when the system size is smaller than K, balk when the system size is larger than K, and join with probability p when the system size equals to K. According to Hassin and Haviv (1997), such queue can be regarded as a queue where customers' buffer size threshold is mixed between $K-1$ and K with probability p. Thus the expected delay threshold is $\frac{K-1+p}{\mu_S} = \delta$. So $M/D/1$ queue with buffer size $K-1+p$ implies $M/D/1$ queue with delay threshold δ.

□

Derivation of Equation (26)*:*

Recall that $q_j = a_j q_0$ for $j \le K-1$. And $a_0 = 1,\ a_1 = e^{\rho} - 1,\ a_j = e^{\rho}(a_{j-1} - \sum_{i=1}^{j-1}\alpha_i a_{j-i} - \alpha_{j-1}a_0)$. Let $\beta_i = \frac{((1-p)\rho)^i}{i!}e^{-(1-p)\rho}$. Note that

$$\begin{aligned}\alpha_0' &= \alpha_0 + (1-p)\alpha_1 + (1-p)^2\alpha_2 + \dots \\ &= \frac{e^{-\rho}}{e^{-(1-p)\rho}}(\beta_0 + \beta_1 + \beta_2 + \dots) \\ &= e^{-p\rho},\end{aligned}$$

then

$$q_{K-1} = e^{-p\rho}q_K + (\alpha_1 + (1-p)\alpha_2 + (1-p)^2\alpha_3 + \dots)q_{K-1} + \dots + (\alpha_{K-1} + (1-p)\alpha_K + \dots + \dots)(q_1 + q_0).$$

It follows that

$$\begin{aligned}e^{-p\rho}\frac{q_K}{q_0} =& a_{K-1} - (\alpha_1 + (1-p)\alpha_2 + (1-p)^2\alpha_3 + \dots)a_{K-1} - (\alpha_2 + (1-p)\alpha_3 + \dots)a_{K-2} - \dots \\ & - (\alpha_{K-2} + (1-p)\alpha_{K-1} + \dots)a_2 - (\alpha_{K-1} + (1-p)\alpha_K + \dots)(a_1 + a_0) \\ =& a_{K-1} - \alpha_1 a_{K-1} - \alpha_2 a_{K-2} - \dots - \alpha_{K-2}a_2 - \alpha_{K-1}(a_1 + a_0) \\ & - \frac{e^{-p\rho}}{(1-p)}(\beta_2 + \beta_3 + \dots)a_{K-1} - \dots - \frac{e^{-p\rho}}{(1-p)^{K-1}}(\beta_K + \dots)(a_1 + a_0) \\ =& e^{-\rho}a_K - \frac{e^{-p\rho}}{(1-p)}(1 - \beta_0 - \beta_1)a_{K-1} - \frac{e^{-p\rho}}{(1-p)^2}(1 - \beta_0 - \beta_1 - \beta_2)a_{K-2} - \dots \\ & - \frac{e^{-p\rho}}{(1-p)^{K-1}}(1 - \beta_0 - \beta_1 - \dots - \beta_{K-1})(a_1 + a_0) \\ =& e^{-\rho}a_K - e^{-p\rho}(\frac{a_{K-1}}{1-p} + \dots + \frac{a_1 + a_0}{(1-p)^{K-1}}) + (\frac{\alpha_0}{1-p} + \alpha_1)a_{K-1} + \dots \\ & + (\frac{\alpha_0}{(1-p)^{K-1}} + \frac{\alpha_1}{(1-p)^{K-2}} + \dots + \alpha_{K-1})(a_1 + a_0) \\ =& -e^{-p\rho}(\frac{a_{K-1}}{1-p} + \frac{a_{K-2}}{(1-p)^2} + \dots + \frac{a_2}{(1-p)^{K-2}} + \frac{a_1 + a_0}{(1-p)^{K-1}}) \\ & + a_{K-1} + \frac{a_{K-2}}{1-p} + \frac{a_{K-3}}{(1-p)^2} + \dots + \frac{a_1}{(1-p)^{K-2}} + \frac{a_0}{(1-p)^{K-1}}.\end{aligned}$$

And hence,

$$a_K' = \frac{q_K}{q_0} = (e^{p\rho} - 1)(1-p)^{-K+1} + (e^{p\rho} - pe^{p\rho} - 1)\sum_{i=1}^{K-1}a_i(1-p)^{-K+i}.$$

□

Proof of Proposition C.1:

First, we have

$$q_j = \Pr\{\text{arrival finds } j|\text{he joins}\} = \frac{\pi_j}{\Pr\{\text{Join}\}}$$

for $j = 0, 1, \dots, K-1$ and

$$q_K = \frac{p\pi_K}{\Pr\{\text{Join}\}}.$$

Then we have $q_0 = \frac{\pi_0}{\Pr\{\text{Join}\}}$. Since $\lambda\Pr\{\text{Join}\} = \mu(1 - \pi_0)$, we have

$$\pi_0 = \frac{q_0}{q_0 + \rho}$$

and

$$1 - \Pr\{\text{Join}\} = \pi_{K+1} + (1-p)\pi_K = 1 - \frac{1}{q_0 + \rho}.$$

Then $\pi_j = q_j\Pr\{\text{Join}\} = \frac{q_j}{q_0+\rho}$ for $j \le K-1$; $\pi_K = \frac{q_K\Pr\{\text{Join}\}}{p} = \frac{q_K}{p(q_0+\rho)}$; and $\pi_{K+1} = 1 - \frac{1}{q_0+\rho} - \frac{(1-p)q_K}{p(q_0+\rho)}$.

□

Proof of Corollary C.1:

It follows from Proposition C.1.

□

Proof of Proposition C.2:

1. Strictly decreasing and continuous in δ: it is equivalent to show strictly decreasing and continuous in $K-1+p$ for $p \in (0,1]$:

When $p<1$,

$$q_K = e^{p\rho}(e^{-\rho}a_K - c_K)q_0,$$

where

$$c_K = d_1 a_0 + d_1 a_1 + d_2 a_2 + \cdots + d_{K-1}a_{K-1},$$

$$d_j = (1-p)(d_{j-1} + \alpha_{K-j}),$$

$$d_1 = (1-p)\alpha_K + (1-p)^2\alpha_{K+1} + (1-p)^3\alpha_{K+2} + \cdots + (1-p)^k\alpha_{K-1+k} + \ldots.$$

c_K strictly increases and is continuous in d_j, d_j strictly decreases and is continuous in p. So c_K strictly decreases and is continuous in p, and $e^{p\rho}(e^{-\rho}a_K - c_K)$ strictly increases and is continuous in p. Specifically, if $p=1$, $c_K=0$, $q_K = a_K q_0$.

Thus we have,

$$q_j = \begin{cases} \dfrac{1}{b_{K-1} + e^{p\rho}(e^{-\rho}a_K - c_K)} & \text{if } j=0, \\ \dfrac{b_j - b_{j-1}}{b_{K-1} + e^{p\rho}(e^{-\rho}a_K - c_K)} & \text{if } 1 \le j \le K-1, \\ \dfrac{e^{p\rho}(e^{-\rho}a_K - c_K)}{b_{K-1} + e^{p\rho}(e^{-\rho}a_K - c_K)} & \text{if } j=K \end{cases}$$

where b_j follows (25). So given K, q_0 strictly decreases and is continuous in p.

Thus given K, $\pi(\delta,\mu_S) = \pi_{K+1}(K+p) + (1-p)\pi_K(K+p) = 1 - \frac{1}{q_0+\rho}$ strictly decreases and is continuous in p. Let us check the point when K becomes $K+1$. As c_K is strictly decreasing in p, and when $p \to 0$, $a'_K \to 0$; when $p=1$, $a'_K = a_K$. So $q_0(K+p) = q_0(K+1)$ when $p=1$ and $q_0(K+p) \to q_0(K)$ when $p \to 0$. So q_0 is continuous at the point when $p=0$, i.e., where K becomes $K+1$. So $\pi(\delta,\mu_S)$ is strictly decreasing and continuous in $K+p-1$.

2. Strictly decreasing and continuous in μ_S:

Continuity is evident because all component functions are continuous in μ_S.

We prove the decreasing property using a sample path argument. For a period from time 0 to time t, we look at a sample path of arrival process, denoted by a_1, a_2, ..., a_{N_t}, where a_i is the time interval between i^{th} arrival and $i+1^{th}$ arrival. Let x_i denote the workload of the server just before $i+1^{th}$ arrival, i.e., the delay saw by $i+1^{th}$ arrival. Let u_i denote a random number following uniform distribution in $[0,1]$. Let binary variable z_i denote whether i^{th} customer is blocked (he is blocked if he believes there is more than δ delay just before her arrival, then $z_i = 0$). His belief on delay is based on the system size $\lceil x_{i-1}\mu_S \rceil$ and the random variable u_i, which $\frac{\lceil x_{i-1}\mu_S\rceil - 1 + u_i}{\mu_S}$. So the total number of unblocked customers during $[0,t]$ for this sample path is

$$\{\sum_{i=1}^{N_t} z_i | \text{if } \lceil x_{i-1}\mu_S \rceil - 1 + u_i \le \delta\mu_S,\ z_i = 1;\ \text{otherwise},\ z_i = 0;\ x_i = (x_{i-1} + \frac{z_i}{\mu_S} - a_i)^+;\ x_0 = 0\}.$$

Let $f(\mu_S)$ denote this value. It is evident that $f(\mu_S)$ equals the objective value of the following optimization problem.

$$\max_{z_i \in \{0,1\}} \quad \sum_{i=1}^{N_t} z_i \tag{P1}$$

$$x_i = (x_{i-1} + \frac{z_i}{\mu_S} - a_i)^+, \text{ for } 0 \le i \le N_t, \tag{40}$$

$$\frac{\lceil x_{i-1}\mu_S \rceil - 1 + u_i}{\mu_S} + \frac{z_i}{\mu_S} \le \delta + \frac{1}{\mu_S}, \text{ for } 0 \le i \le N_t, \tag{41}$$

$$x_0 = 0.$$

Let $\boldsymbol{z}^*(\mu)$ be the optimal solution for the problem above. Now, consider a same sample path (same a_i and u_i) but a larger $\mu'_S > \mu_S$, which means shorter service time. We argue that the solution that leads to $\boldsymbol{z}^*(\mu_S)$ when $\mu = \mu_S$ is a feasible (but not necessarily optimal) solution for (P1) when μ is replaced by μ'_S. To see this, note that $\lceil x_{i-1}\mu_S \rceil$ is the system size. Since z_i is kept the same, then $\lceil x_{i-1}\mu_S \rceil \ge \lceil x'_{i-1}\mu'_S \rceil$ by (40).It follows that $\lceil x'_{i-1}\mu'_S \rceil - 1 + u_i + z_i \le \lceil x_{i-1}\mu_S \rceil - 1 + u_i + z_i \le \delta\mu_S + 1 \le \delta\mu'_S + 1$ and hence (41) holds when μ_S is replaced by μ'_S. So $f(\mu_S) \le f(\mu'_S)$. For any t and any sample path, we will have $f(\mu_S) \le f(\mu'_S)$, which means $\pi(\delta, \mu_S) \ge \pi(\delta, \mu'_S)$ if $\mu'_S > \mu_S$.

Next we show the strong monotonicity. While this result seems intuitive, a simple coupling argument may not work because we cannot couple the busy periods of two systems with different service rates. Our approach involves constructing renewal reward processes and sample path arguments in renewal cycles. We consider a renewal reward process (called $\underline{\text{RP}}$) which renews whenever the size of the system with capacity μ_S jumps from 0 to 1 and the reward is the number of customers served in a renewal cycle. Let $\underline{R}$ and τ represent the random reward and the cycle length in this process. Then, for the system with μ_S, the long-run average rate of customers served $\lambda(1 - \pi(\delta, \mu_S)) = \frac{E(\underline{R})}{E(\tau)}$. We next construct another renewal reward process (called $\overline{\text{RP}}$) which renews exactly at the same time when $\underline{\text{RP}}$ renews but has a strictly higher expected per-period reward than that of $\underline{\text{RP}}$. Finally, we argue that the long-run average reward of $\overline{\text{RP}}$ is no greater than the long-run average rate of customer served in the system with $\mu'_S > \mu_S$. It follows that $\pi(\delta, \mu_S) > \pi(\delta, \mu'_S)$ if $\mu'_S > \mu_S$.

Without loss of generality, we suppose that a cycle starts at $t = 0$ and i is the index of the last arrival in the cycle of the process $\underline{\text{RP}}$. Construct $\overline{\text{RP}}$ in this way. Imagine a single-server queueing process with service rate $\mu'_S > \mu_S$ and it accepts customers following the principle below. Recall that z_j denotes that whether the j^{th} customer is accepted and served in process $\underline{\text{RP}}$ and we let z'_j denote that in process $\overline{\text{RP}}$. For all sample paths, set $z'_j = z_j$ for all $j < i$. We further claim we can find some sample path φ such that $\frac{\lceil x_{i-1}\mu_S\rceil - 1 + u_i}{\mu_S} + \frac{1}{\mu_S} >$

$\delta + \frac{1}{\mu_S}$, $\frac{\lceil x_{i-1}\mu_S\rceil - 1 + u_i}{\mu_S} \le \delta + \frac{1}{\mu_S}$, $\frac{\lceil x'_{i-1}\mu'_S\rceil - 1 + u_i}{\mu'_S} + \frac{1}{\mu'_S} \le \delta + \frac{1}{\mu'_S}$, $x_{i-1} \le a_i$ and $x'_{i-1} + \frac{1}{\mu'_S} \le a_i$. Then on this sample path, z_i must be 0, but z'_i can be 1 without violating constraint (41). For other sample paths, we have $z'_j = z_j$, $\forall j \le i$. Since $x_i = x'_i = 0$ for all sample paths, both queueing processes corresponding to $\underline{\text{RP}}$ and $\overline{\text{RP}}$ are empty after the service of ith customer, and the cycle ends before $i+1^{th}$ customer arriving. By this way of construction, we ensure that the reward collected in each cycle for $\overline{\text{RP}}$ are i.i.d. random variables and the renewal reward theorem applies. As long as sample path φ has a non-zero measure, then we have $E[\sum_{j=1}^{i} z_j] < E[\sum_{j=1}^{i} z'_j]$, i.e., $E[\underline{R}] < E[\overline{R}]$. Since both processes share the same renewal cycles (the cycle starts at $t = 0$ and ends before $i+1^{th}$ arrival), then we have the long-run average rate of customers served in process $\underline{\text{RP}}$ is strictly smaller than that in process $\overline{\text{RP}}$ by the renewal reward theorem. One can also check that $\boldsymbol{z}$ as we define in process $\overline{\text{RP}}$ is a feasible solution to problem (P1) with respect to μ'_S, so the long-run average reward in $\overline{\text{RP}}$ is a lower bound for the long-run average rate of customers served in system with μ'_S. Combining arguments above, we have $\lambda(1 - \pi(\delta, \mu_S)) < \lambda(1 - \pi(\delta, \mu'_S))$ if $\mu'_S > \mu_S$.

Finally, it is left to show the measure of such sample path φ is strictly greater than 0 in a renewal cycle. Let us consider $\epsilon = \delta\mu'_S - \lceil x'_{i-1}\mu'_S\rceil - \delta\mu_S + \lceil x_{i-1}\mu_S\rceil$. Since $\lceil x_{i-1}\mu_S\rceil \ge \lceil x'_{i-1}\mu'_S\rceil$ and $\mu'_S > \mu_S$, we know $\epsilon > 0$. Let $\zeta = \lceil x_{i-1}\mu_S\rceil - \delta\mu_S$, then $\lceil x'_{i-1}\mu'_S\rceil - \delta\mu'_S = \zeta - \epsilon$. On sample path φ, by $\frac{\lceil x_{i-1}\mu_S\rceil - 1 + u_i}{\mu_S} + \frac{1}{\mu_S} > \delta + \frac{1}{\mu_S}$, we have $1 - \zeta < u_i \le 1$; by $\frac{\lceil x_{i-1}\mu_S\rceil - 1 + u_i}{\mu_S} \le \delta + \frac{1}{\mu_S}$, we have $1 - \zeta \ge 1 - u_i \ge 0$. So, as long as $u_i - 1 + \zeta \le \epsilon$, i.e., u_i is very close to $1 - \zeta$, then $\lceil x'_{i-1}\mu'_S\rceil - \delta\mu'_S = \zeta - \epsilon \le 1 - u_i$, i.e., $\frac{\lceil x'_{i-1}\mu'_S\rceil - 1 + u_i}{\mu'_S} + \frac{1}{\mu'_S} \le \delta + \frac{1}{\mu'_S}$. Since $0 \le 1 - \zeta < 1$ and u_i follows the uniform distribution in $[0, 1]$, then the measure of such u_i is greater than 0. As for a_i, we need $a_i \ge \max\{x_{i-1}, x'_{i-1} + \frac{1}{\mu'_S}\}$, since a_i follows exponential distribution, then the measure of such a_i is greater than 0. As u_i and a_i are independent random variables, the measure of sample path φ is strictly greater than 0.

Note that in our $M/D/1/\delta\mu_S$ queue, the queue buffer depends on μ_S. One can use a similar idea above to show that the blocking probability is decreasing in μ_S in an $M/D/1/R$ queue, where the queue buffer does not depend on μ_S.

□

Appendix References

Brun, Olivier, Jean-Marie Garcia. 2000. Analytical solution of finite capacity m/d/1 queues. *Journal of Applied Probability* **37**(4) 1092–1098.

Gross, Donald. 2008. *Fundamentals of queueing theory*. John Wiley & Sons.

Hassin, Refael, Moshe Haviv. 1997. Equilibrium threshold strategies: The case of queues with priorities. *Operations Research* **45**(6) 966–973.

Kingman, John FC. 1962. On queues in heavy traffic. *Journal of the Royal Statistical Society: Series B (Methodological)* **24**(2) 383–392.

Kulkarni, Vidyadhar G. 1996. *Modeling and analysis of stochastic systems*. Chapman and Hall/CRC.

Moran, PAP. 1956. A probability theory of a dam with a continuous release. *The Quarterly Journal of Mathematics* **7**(1) 130–137.