manuscript

Intraday Scheduling with Patient Re-entries and Variability in Behaviours

Minglong Zhou, Gar Goei Loke, Chaithanya Bandi
Department of Analytics & Operations, NUS Business School, National University of Singapore, Singapore 119245,
minglong_zhou@u.nus.edu, gargoei@nus.edu.sg,cbandi@nus.edu.sg

Zi Qiang Glen Liau, Wilson Wang
University Orthopaedics, Hand & Reconstructive Microsurgery Cluster, National University Health System, Singapore 119228,
glen_liau@nuhs.edu.sg, wilson_wang@nuhs.edu.sg

Problem definition: We consider the intraday scheduling problem in a group of Orthopaedic clinics where the planner schedules appointment times given a sequence of appointments. We consider patient re-entry – where patients may be required to go for an X-ray examination, returning to the same doctor they have seen – and variability in patient behaviours such as walk-ins, earliness, and no-shows, which leads to inefficiency such as long patient waiting time and physician overtime.

Academic/Practical relevance: In our dataset, 25% of the patients are required to go for X-ray examination. We also found significant variability in patient behaviours. Hence, patient re-entry and variability in behaviours are common, but we found little in the literature that could handle them.

Methodology: We formulate the problem as a two-stage optimization problem, where scheduling decisions are made in the first stage. Queue dynamics in the second stage are modeled under a P-Queue paradigm (Bandi and Loke 2018), which minimizes a risk index representing the chance of violating performance targets, such as patient waiting times. The model reduces to a sequence of mixed-integer linear optimization problems.

Results: Our model achieves significant reductions, in comparative studies against a sample average approximation (SAA) model, on patient waiting times while keeping server overtime constant. Our simulations further characterize the types of uncertainties under which SAA performs poorly.

Managerial insights: We present an optimization model that is easy to implement in practice and tractable to compute. Our simulations indicate that not accounting for patient re-entry or variability in patient behaviours will lead to suboptimal policies, especially when they have specific structure that should be considered.

Key words: Optimization, Scheduling

History: February 15, 2021

1. Introduction

The appointment scheduling problem of patients in a clinic is a traditional problem in the healthcare operations management literature (*e.g.,* Ho and Lau 1992, Denton and Gupta 2003, Gupta and Denton 2008). Within this scope, the intraday scheduling problem, where patients complete their appointments within the day, is a key problem. A planner is given a known sequence of patient appointments, and makes a *here-and-now* decision on their appointment start times. In the process, she incurs costs on operations, such as resource idleness and overtime of healthcare providers. This is a standard setting in intraday scheduling literature (*e.g.,* "Theme A" in Gupta and Denton 2008) and is followed by many recent papers (*e.g.,* Mak et al. 2015, Qi 2017, Jiang et al. 2017). In this paper, we focus on the intraday scheduling problem under this standard setting. More specifically, we focus on designing better start times of appointment slots, *i.e.,* the interarrival times between patients.

This problem is challenging in practice because of patient re-entry and variability in patient behaviours, arising from no-shows, earliness, and walk-ins. We partner with the Orthopaedic clinics run by the National University Health System in Singapore. On each day, patients may not turn up, with or without informing the clinic. In our dataset, these patients account for as many as 29% of scheduled patients. Patients may also be early or late; on average, patients arrive around 7 minutes earlier than their appointment times. About 27.3% patients are late for their appointments, by 14 minutes on average, which amounts to 1.4 time slots (of 10 minutes each). Some are late by more than 60 minutes. Upon arrival, patients are routed through a registration process. Patients then meet with a doctor for their first consultation. Subsequently, they may be required to take various tests, such as X-rays, which would not be known at the point of scheduling. On average, 25% of Orthopaedic patients require X-ray examinations. For other departments, this proportion can be as large as 39%. After examination, they rejoin the queue for consultations, and are re-examined by a doctor before completing their visit. We examine this in greater detail in the simulation study.

This data illustrates that re-entry and variability in patient behaviours are very common features of a clinic's operations; and they can lead to inefficiencies if not well-controlled. No-shows and lateness create physician idleness and overtime, which are both very costly to the healthcare system. Re-entry causes heavier traffic and usually follows a different service time distribution from first consultations. As such, in scheduling patients, it is only reasonable for the planner to take into consideration all of these factors, while managing the current patients in the systems, and the expected times before they exit the system. This is a daunting challenge. In our partnering clinics, the current practice is largely to fix each appointment slot to be 10 minutes. The slots are decided at the planning stage and will not change in daily operations. When patients make appointments, they are scheduled to the first available slot. However, this equal-interval policy is generally suboptimal (Wang 1993).

Key approaches in the literature

There are three streams in the literature to approach the problem, namely stochastic programming, queueing, and robust optimization approaches. We describe the literature in each of these areas, before summarizing the present challenges associated with each of these approaches.

Stochastic Programming Approach. The two-stage stochastic programming formulation is a popular approach. Most notably, Denton and Gupta (2003) employ such a formulation where scheduling decisions are made in the first stage, before the uncertainty in patient service times materializes in the second stage. In Denton and Gupta (2003)'s model, the state and decision variables are represented by four different time durations – the interarrival time between appointments, the stochastic service time of each patient, the additional waiting time for each patient beyond their scheduled appointment time if the server is not available, and finally, the idle time incurred by the server. The dynamics are then described as a linear relationship linking up these four time durations. The objective is taken as the expected total cost, under some unit waiting, idleness and overtime costs. This model is well-studied and performs well in many intraday scheduling contexts. Consequently, this approach has been extended within a variety of applications (*e.g.,* Denton et al. 2007, Erdogan and Denton 2013, Ge et al. 2014, Berg et al. 2014).

Robust Optimization Approach. An alternative to the stochastic programming formulation is the robust optimization approach, often in the distributionally robust variant (*e.g.,* Mak et al. 2014, 2015, Padmanabhan et al. 2018, Kong et al. 2019). Mak et al. (2015) is among the first to study the distributionally robust intraday scheduling problem and they proposed a tractable formulation under a marginal moments ambiguity set. Jiang et al. (2017) model a distributionally robust single-server intraday scheduling problem with no-shows. Qi (2017) introduces a delay unpleasantness measure based on the Conditional Value-at-Risk (CVaR) to describe the delay experienced by patients, anchored on a baseline waiting time target idiosyncratic to each patient.

Queueing Approach. Broadly, there are two common approaches to evaluate queue dynamics under complex settings – fluid approximations (Braverman et al. 2017) and diffusion models (Dai and Tezcan 2011, Gurvich 2014). Specifically, the intraday scheduling problem and its variants have been studied under the assumptions of Poisson arrival processes or Erlangian service time distributions (Gurvich et al. 2010, Luo et al. 2012).

Our setting is different from those in the dynamic (or online) appointment scheduling literature (see, *e.g.,* Liu et al. 2010, Feldman et al. 2014, Wang et al. 2018). We agree with the relevance of the online approach towards the matching of patients to slots dynamically on the intraday schedule. However, in this paper, we follow the line of inquiry in Denton and Gupta (2003) and "Theme A" in Gupta and Denton (2008) to focus on the offline question of how the interarrival times between slots should be decided. This is independent of and consistent with the online matching process.

Challenges with the present approaches

There are two main difficulties with the present approaches. First, it remains challenging to incorporate all the uncertainties of patient re-entry and variability in behaviours into a single model formulation. This is because it is difficult to define how the uncertainty, in particular, the re-entry patients, interact with the decisions and other uncertainties. In the Queueing setting, this translates into difficulties in computing service times and incorporating decisions into the model. In the Stochastic programming and the Robust Optimization approach, this obscures the definition of proper sample paths for the uncertainty, either as conditional distributions or collected within an uncertainty set.

Second, it is unlikely that such a model formulation would result in a tractable form that would be computable within the intraday timeframe. Indeed, the solution methodology for many of these approaches will involve sample average approximation (SAA), wherein the number of samples required to constitute an accurate sample of the uncertainties scales rapidly with the dimension of uncertainties. In our ensuing analysis, we shall also see the challenge of using SAA to model features of the intraday scheduling problem arising from uncertainties with specific patterns.

As such, to the best of our knowledge, it remains challenging to incorporate patient re-entries, walk-ins, no-shows, and patient earliness/lateness into existing frameworks, while remaining tractable.

1.1. Our approach and contributions

Recent attempts to harmonize ideas in robust optimization with queueing theory have opened doors into tractable formulations with close fidelity to the flow dynamics (Bandi et al. 2015). Introduced by Bandi and Loke (2018), the Pipeline Queues paradigm is an alternative to modeling queues. In particular, the authors illustrated in numerical simulations that the model had relatively good performance over a range of complex networks with general service and arrival distributions. In this paper, we propose a two-stage formulation. In the first stage, the planner commits to a scheduling of appointments. In the second stage, we approximate the queueing cost as metrics of a pipeline queue. The contributions of this paper are two-pronged, in the intraday patient scheduling problem, as well as the technique of Pipeline Queues.

For the former, we illustrate that a model, which ingests high fidelity information on service times and is flexible enough to accurately replicate queue dynamics, can lead to increases in performance over existing methodologies. Specifically, we are able to handle patient re-entries, no-shows, stochastic arrival times, walk-ins, and stochastic transportation times between servers, which have traditionally been difficult to address. Our simulations illustrate this – we improve on all metrics including patient waiting times, server overtime, and maximum instantaneous waiting time (defined in Section 4), achieving reductions of as much as 18%, 10% and 26% respectively when compared against the current practice at our partnering clinics.

We also differentiate our contributions vis-á-vis Bandi and Loke (2018). At the broadest level, we extend Pipeline Queues, originally designed for the optimization and control of flows, to scheduling and demand matching problems. More specifically, we illustrate that simulating queue dynamics as a pipeline queue in the inner model remains tractable in the two-stage setting, where global parameters are controlled in the outer problem. In particular, such a second-stage formulation with Pipeline Queues is novel. In terms of theoretical contributions, we examine the situation where the class of patients is revealed *stochastically* midway through the system, as opposed to *a priori* (Proposition 6). This was not permitted by the original framework, as the modeling technique for stochastic flows relied upon critical independence assumptions to achieve reformulations. In this work, this is averted, and we show that it is possible to achieve reformulation even without further assumptions (Proposition 7). In this paper, we also illustrate the technique of using dummy queues and servers as an approximant to the true dynamics (Figure 4). Lastly, it is the first complete illustration of the Pipeline Queue paradigm on an application from an actual business environment and context. This departs from the numerical illustration in the original paper by Bandi and Loke (2018), which aims to illustrate the technique and its inner workings, as opposed to illustrating its successful application on an actual problem.

Organization of the paper

In §2, we ease the reader in by first introducing our proposed framework under a classical single-server setting, without patient re-entry. In §3, we extend the results to the fully realistic setting with patient re-entries, no-shows, uncertain arrival times, and random transportation times between stations. We conduct a simulation study in §4 using a real dataset, and conclude in §5. Though we reference the techniques in Bandi and Loke (2018), this paper is self-contained. All proofs are deferred to the Appendix A, by default, unless they can be stated succinctly.

Notation. We use boldface lowercase letters for vectors (*e.g.*, $\boldsymbol{\theta}$). We use $[N]$ to denote the running index $\{1, 2, 3, \ldots, N\}$ for a known integer N. We adopt the convention that $\inf \emptyset = +\infty$, where $\emptyset$ is the empty set. We use $\mathbb{1}$ to represent the indicator function; thus $\mathbb{1}(\mathcal{C}) = 1$ if set $\mathcal{C}$ is nonempty or $\mathbb{1}(\mathcal{C}) = 0$ if $\mathcal{C}$ is empty.

2. The Single-Server Intraday Scheduling Problem

To ease the reader into our approach, we first illustrate our model on the classical single-server intraday scheduling problem, without patient re-entry. In the succeeding section §3, only then we extend our results to an intraday scheduling setting involving patient re-entry.

The single-server intraday scheduling problem is drawn in Figure 1. The planner schedules N, which is fixed and known, appointments within T discrete periods of the clinic's operating hours. This is done before any uncertainty unfolds. We model this scheduling decision as x_t, the number of appointments scheduled to arrive at time $t \in \mathcal{T} := \{0, 1, ..., T\}$. Here, we let x_t be binary. All patients must be scheduled: $\sum_{t=0}^{T} x_t = N$. This setting closely follows that in the literature (*e.g.,* Denton and Gupta 2003), where only the timing of the slots on the schedule matters. It is not the intent to actually match patients to the slots, and operationally, the decision-maker may freely decide on the matching. In most practical situations, the number of patients to serve, N, is usually prescribed, and the number of patients far outstrip this number of slots; hence, we can always safely assume that slots will be filled. In this regard, the scheduling decisions are **static** – the scheduled times ought not change even if the patients allotted to the timings do. Moreover, in the intraday scheduling context, the time window is short enough to make it infeasible to call up new patients to arrive when there are no shows.

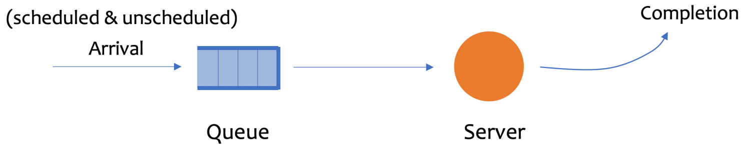

Figure 1 **Patient flow in a single-server setup**

These scheduling decisions form the first stage in our two-stage formulation. The second-stage problem is a multi-period problem. When patients arrive at their scheduled time, *e.g.,* t, they join the queue, along with walk-in patients at some known arrival rate. At this point, the uncertainty in the number and arrival times of scheduled and walk-in patients until this time period materializes. Some patients in consultation with doctors (with maximum capacity C) also complete their service and leave the system. This uncertainty is modeled by $h^{t,s}$, the hazard rate of completing service after being served for s periods at time t, and too materializes at this point. With freed up capacity, the decision-maker then makes a **recourse decision** in the form of how to route patients through the network. In a single-server network, recourse is trivial – it is simply the dispatch of patients to the server if it is idle. For more complicated networks, as we shall see later, the recourse may be non-trivial and legitimate decisions in their own right. When the next time period begins, this process repeats. The goal is to schedule and route patients in a fashion such that the waiting time target C_w is met as often as possible (possibly infeasible if $\boldsymbol{x}$ was decided poorly). Specifically, this refers to having good probabilistic guarantees against violations of the waiting time constraint. We make this clear in the exposition that will soon follow. These variables are summarized in Table 1.

A schematic of this process is illustrated in Figure 2. Specifically, time $t = 0$ is where the queue is opened, and the server starts to see patients from $t = 1$ onwards. Patients always arrive at the end of a period and can only be routed from the next period. Service completion happens at the beginning of each period, after observing which, the planner can route patients in the queue (arrived previous to this period) to see available doctors. It is important to keep in mind that

		Single-Server Setting
		Parameters and known quantities
N	:	Number of patients to schedule
T	:	Last modeling time
λ_t	:	Random arrivals (walk-ins) at time t
$h^{t,s}$	:	Known hazard rate – probability patient served for s periods at time t completes service at $t+1$
C	:	Server capacity – total number of (independent and homogeneous) doctors
C_w	:	Targeted total patient waiting time to keep under
		State and decision variables
x_t	:	Decision variable of patients scheduled to arrive at time t
$y^{t,s}$	:	Random variable of patients waiting for s periods in the **queue** at time t
$z^{t,s}$	:	Random variable of patients served for s periods in the **server** at time t
$p^{t,s}$	:	Recourse variable of patients dispatched into server after waiting for s at time t

Table 1 List of parameters and variables in the single-server setting

our primary focus is on the scheduling decision $\boldsymbol{x}$. The routing decisions only serve to approximate second-stage queueing dynamics.

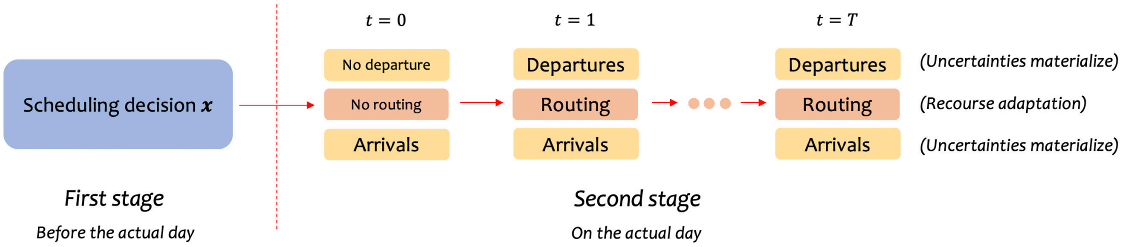

Figure 2 A rundown of events in two stages

Let us define the second-stage problem more precisely. Following the notation in Bandi and Loke (2018), consider two time dimensions – model time $t \in \mathcal{T}$ up to time horizon T and node time $s \in \mathcal{T}$ describing how long a patient has spent in a node (queue or server). Let $y^{t,s}$ and $z^{t,s}$, $t \in \mathcal{T}, s \in \mathcal{T}$, denote the number of patients that have already waited for exactly s periods but still remain in the queue and server respectively. Alternatively, $y^{t,s}$ can be thought of as the number of patients who arrived at the clinic at time $t - s$ and have yet to be served. As we start each day with the clinic empty, we would have $z^{0,s} = 0, \forall s$, and $y^{0,s} = 0$ for $s \geq 1$; $y^{0,0}$ may be positive because it corresponds to patients that arrive at the end of time 0, *i.e.*, at the beginning of operations. The definition also induces the subsequent boundary conditions: $z^{t,s} = 0$ for $\forall t \in \mathcal{T}, s \geq t$ and $y^{t,s} = 0$ for $\forall t \in \mathcal{T}, s > t$.

Next, we describe the dynamics of the queue. Inflows to the queue are made up of scheduled appointments and walk-ins. For now, we assume that the scheduled patients $\boldsymbol{x}$ arrive exactly at their appointment times. We use $\lambda_t \sim \Lambda_t$, drawn from some time non-homogeneous distribution, to represent the number of walk-in patients at time t. We assume that their moment generating functions exist, are bounded and independent across t. Thus, the total inflow to the queue at time t is $x_t + \lambda_t$. Because $y^{t,0}$ represents the number of patients in queue at time t, that have spent 0 periods in queue, it is equivocally the inflow to the queue at time t. Hence, for $\forall t \in \mathcal{T}$,

$$y^{t,0} = x_t + \lambda_t. \tag{1}$$

As the server is freed up, patients are dispatched from the queue to the server. Let $p^{t,s}$, $t, s \in \mathcal{T}$, be the second-stage recourse variable that indicates the number of patients dispatched, after waiting for s periods at time t in the queue. Necessarily, we cannot dispatch more patients than there are available, *i.e.,* $p^{t,s} \leq y^{t-1,s-1}$. We require $p^{0,s} = 0$ for $s \in \mathcal{T}$, and $p^{t,s} = 0$ for $s > t$ subsequently. These definitions lead to the dynamics:

$$y^{t,s} := y^{t-1,s-1} - p^{t,s} = y^{t-s,0} - \sum_{\tau=0}^{s-1} p^{t-\tau,s-\tau}, \quad \forall t \in [T], \forall s = 1, ..., t-1. \tag{2}$$

For the server, inflow originates from patients dispatched from the queue. As such, for $\forall t \in \mathcal{T}$,

$$z^{t,0} = \sum_{s=0}^{t} p^{t,s}, \tag{3}$$

where summing over s gives the total patients dispatched at time t.

Patients leave once their consultation ends, with probability $h^{t,s}$, after being in consultation for s periods at time t. This is the hazard rate of the service time distribution. Indeed, any general discrete-time service distribution can be modeled via this approach (Dai and Shi 2017). Moreover, these probabilities can be readily obtained from data. The distribution need not be stationary.

Assumption 1. *Service times are independent and identically distributed across patients.*

Assumption 1 is realistic because each doctor sees patients of the same ailment, one at a time. Hence, this service process is independent and identical for all patients. This induces a Binomial model on the outflow from the server. More specifically, whether a patient in consultation for s periods at time t, will complete service at time $t+1$, is a Bernoulli variable with probability $h^{t,s}$. Hence, aggregating over all patients,

$$z^{t,s} \sim \text{Bin}\left(z^{t-1,s-1}, 1 - h^{t-1,s-1}\right), \quad \forall t \in [T], s \in [T]. \tag{4}$$

Proposition 1. *The state variable $z^{t,s}$ obeys:*

$$z^{t,s} \sim Bin\left(z^{t-s,0}, \hat{h}^{t,s}\right), \quad \forall t \in [T], s \in [t],$$

where $\hat{h}^{t,s} \triangleq \prod_{\tau=1}^{s}(1 - h^{t-\tau,s-\tau})$, and $\hat{h}^{t,s} := 1$ when $s = 0$.

Proof of Proposition 1 This is most easily shown by the law of iterated expectations. □

Constant $\hat{h}^{t,s}$ can be interpreted as the cumulative survival probability for s periods amongst patients who arrived at time $t-s$. As such, this Proposition allows us to characterize the dynamics as dependent only on the *cohort* of patients entering the server.

Optimization problem

Our full model can be written formally as:

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & Q(\boldsymbol{x}) && (5)\\ \text{s.t.} \quad & \sum_{t=0}^{T} x_t = N \\ & x_t \in \{0,1\} \qquad \forall t \in \mathcal{T}, \end{aligned}$$

where $Q(\boldsymbol{x})$ represents a yet-defined second-stage queueing cost of making scheduling decisions $\boldsymbol{x}$.

The goal of the optimization is to ensure that operational targets on waiting time and capacity constraints, are attained as frequently as possible. This can be phrased as chance constraints on the upper bounds for the queue length and patient waiting time at different times, such as

$$\mathbb{P}\left[\sum_{s=0}^{t} y^{t,s}s - C_w \le 0\right] > 1-\varepsilon. \tag{6}$$

$\sum_{s=0}^{t} y^{t,s}s$ represents the total waiting time experienced by all patients currently in the queue at time t. Indeed, every patient in $y^{t,s}$ has waited in the queue for precisely s time periods. As such, they contribute $y^{t,s}s$ to the total waiting time in the queue. Summing over all s obtains the result. The decision-maker would be interested in making ε as small as possible, so as to obtain the best guarantees on the constraint being satisfied. As a result, a plausible second-stage model is:

$$\begin{aligned} Q(\boldsymbol{x}) := \min_{\boldsymbol{p},\varepsilon\ge 0} \quad & \varepsilon \\ \text{s.t.} \quad & \mathbb{P}\left[\sum_{s=0}^{t} z^{t,s} - C \le 0\right] > 1-\varepsilon && \forall t \in \mathcal{T} \\ & \mathbb{P}\left[\sum_{s=0}^{t} y^{t,s}s - C_w \le 0\right] > 1-\varepsilon && \forall t \in \mathcal{T} \\ & \mathbb{P}\left[p^{t,s} - y^{t-1,s-1} \le 0\right] > 1-\varepsilon && \forall t \in \mathcal{T}. \end{aligned}$$

The first constraint represents capacity constraints. The second constraint states that the total waiting time of all patients in the queue at any time must be bounded by C_w. The third constraint, termed ‘push constraint’, ensures patient dispatch, $\boldsymbol{p}$, does not exceed the number of patients in queue, $\boldsymbol{y}$.

However, constraints in the form of (6) are non-convex, and it is hard to derive a tractable reformulation in general.

Now, (6) can be written equivalently as $\mathbb{P}\left[\sum_{s=0}^{t} y^{t,s}s - C_w \geq 1\right] = \mathbb{P}\left[\sum_{s=0}^{t} y^{t,s}s - C_w > 0\right] \leq \varepsilon$, due to integrality requirements on $y^{t,s}$. Let us instead consider a **stronger** formulation. Suppose there is a decreasing convex function $f : \mathbb{R}^+ \to \mathbb{R}^+$, such that $f(1) \leq \varepsilon$ and $f(\delta) \to 0$ as $\delta \to \infty$. Then we want to consider the (infinite) family of chance constraints:

$$\mathbb{P}\left[\sum_{s=0}^{t} y^{t,s}s - C_w \geq \delta\right] \leq f(\delta) \quad \forall \delta > 0. \tag{7}$$

Evidently, this includes (6) by definition. Hence, if we can find some f and some policy under f that satisfies this family of chance constraints, then we are done.

Definition 1 (Entropic measure of risk). An entropic measure of risk with $k > 0$ for random variable $\tilde{\xi}$ is defined as $g_k(\tilde{\xi}) = k \log \mathbb{E}\left[\exp\left(\tilde{\xi}/k\right)\right]$. Call $g_k(\tilde{\xi}) \leq 0$ an entropic risk constraint.

The entropic measure of risk is a popular convex measure of risk (see *e.g.,* Follmer and Schied 2002, Follmer and Knispel 2011). It is convex and additive (under independence) in the uncertainty $\tilde{\xi}$, while also being convex in the risk index k. Moreover, the exponential function inside $k \log \mathbb{E}[\exp(\cdot/k)]$ operator penalizes positive values of $\tilde{\xi}$ more than proportionately to negative values, and therefore is consistent with risk-aversion. Recent works (*e.g.,* Hall et al. 2015, Jaillet et al. 2016, in portfolio management and vehicle routing respectively) are based on this measure and have been relatively successful in achieving tractable models to otherwise challenging problems.

Proposition 2. *Let $f_k(\delta) = \exp(-\delta/k)$ where $k \leq -1/\log(\varepsilon)$, and $\Gamma \in \mathbb{R}$ be a specified target that random variable $\tilde{\xi}$ ought to be kept under.*

a) f_k fulfills our requirements, i.e., *for any $k > 0$, it is a convex decreasing function with $f_k(1) \leq \varepsilon$ and $f_k(\delta) \to 0$ as $\delta \to \infty$.*

b) $g_k\left(\tilde{\xi} - \Gamma\right) \leq 0$ implies $\mathbb{P}\left[\tilde{\xi} - \Gamma \geq \delta\right] \leq f_k(\delta)$.

In particular, $g_k\left(\sum_{s=0}^{t} y^{t,s}s - C_w\right) \leq 0$ implies that the bound on probability of constraint violation (7) is satisfied.

Proof of Proposition 2. It is easy to check that f_k satisfies all properties stated in (a). As a simple consequence of the Chernoff bound, if $g_k\left(\tilde{\xi} - \Gamma\right) \leq 0$, then

$$\mathbb{P}\left(\tilde{\xi} - \Gamma \geq \delta\right) \leq \exp(-\delta/k) = f_k(\delta).$$

The last part of the proposition is a direct consequence of the above bound. □

As a result of Proposition 2, as long as we can find some particular $k \leq -1/\log(\varepsilon)$ such that the constraint $g_k\left(\sum_{s=0}^{t} y^{t,s}s - C_w\right) \leq 0$ is satisfied, then we are done. This can be done by searching for the smallest k such that $g_k\left(\sum_{s=0}^t y^{t,s}s - C_w\right) \leq 0$, then checking if $k \leq -1/\log(\varepsilon)$. The other interpretation of this is that we want to minimize the chance $f_k(\delta)$ of constraint violation by minimizing k. As a result, this leads to the following second-stage model:

$$\begin{aligned} Q(\boldsymbol{x}) = \min_{\boldsymbol{p}, k>0} \quad & k && && (8)\\ \text{s.t.} \quad & k \log \mathbb{E}\left[\exp\left(\frac{\sum_{s=0}^t z^{t,s} - C}{k}\right)\right] \leq 0 && \forall t \in [T] && (9)\\ & k \log \mathbb{E}\left[\exp\left(\frac{\sum_{s=0}^t y^{t,s}s - C_w}{k}\right)\right] \leq 0 && \forall t \in [T] && (10)\\ & k \log \mathbb{E}\left[\exp\left(\frac{p^{t,s} - y^{t-1,s-1}}{k}\right)\right] \leq 0 && \forall t \in [T], \forall s \in [T] && (11)\\ & + \textbf{overtime constraint.} \end{aligned}$$

Notice that (10) can be generalized to any affine constraint $\sum_{s=0}^t y^{t,s} r(s) \leq b$ for some constants $r(s)$, $s = 0, \ldots, t$, and b.

There are a few options on how overtime constraints can be modeled. The simplest is to require

$$k \log \mathbb{E}\left[\exp\left(\frac{\sum_{s=0}^T y^{T,s} - L}{k}\right)\right] \leq 0, \tag{12}$$

which bounds the queue length at the end of the time horizon T. It indicates that we want to clear the queue, or have no more than L patients in the queue, by clinic closure, where L is the budgeted overtime patients. Similarly, we can also impose a constraint to control the server utilization at the end of horizon, *i.e.,* we require the server to be free at T with high probability. Another approach might be to count the actual periods of overtime service, *e.g.,* as written in (28). As that would require additional machinery, we defer this discussion till the next section.

2.1. Reformulation

It turns out that these constraints can be evaluated into a form affine in decisions $\boldsymbol{x}$, and hence the optimization model (8) can be tractably solved under the following assumption:

ASSUMPTION 2. *The recourse push variable* $\boldsymbol{p}$ *is a static variable,* i.e.*, it is a function only of the decision variables* $\boldsymbol{x}$ *and distributional information* Λ_t *and* $h^{t,s}$*.*

In general, $\boldsymbol{p}$ is state-dependent; instead, $\boldsymbol{p}$ is restricted to be static. The reasoning is deferred to Remark 2 later.

PROPOSITION 3. *For any* $t \in [T]$*, capacity constraints* (9) *are affine in push variables* $\boldsymbol{p}$*:*

$$k \log \mathbb{E}\left[\exp\left(\frac{\sum_{s=0}^t z^{t,s} - C}{k}\right)\right] = \sum_{s=0}^t \beta^{t,s} \sum_{\tau=0}^{t-s} p^{t-s,\tau} - C, \tag{13}$$

for ***constants*** $\boldsymbol{\beta}$ *that can be calculated directly from primitive data:*

$$\beta^{t,s} \triangleq k \log\left(1 - \hat{h}^{t,s} + \hat{h}^{t,s} \exp\left(\frac{1}{k}\right)\right) \quad \forall t \in [T], s \in [t].$$

This Proposition states that we can reformulate the capacity constraint in a linear form in $\boldsymbol{p}$. Let us examine the representation in (13). If we had simply evaluated $\sum_{s=0}^{t} z^{t,s}$ in expectation, we would have obtained the expression $\sum_{s=0}^{t} \hat{h}^{t,s} \sum_{\tau=0}^{t-s} p^{t-s,\tau}$. As such, in moving to the entropic risk constraint, we have replaced $\hat{h}^{t,s}$ with $\beta^{t,s} = k\log\left(1 - \hat{h}^{t,s} + \hat{h}^{t,s} e^{1/k}\right)$. Figure 3 illustrates this transformation. Over $\hat{h}^{t,s} \in [0,1]$, the transformation $\beta^{t,s}$ is always larger than $\hat{h}^{t,s}$. In other words, by comparing $k\log\mathbb{E}\left[\exp\left(\frac{\sum_{s=0}^{t} z^{t,s}}{k}\right)\right]$ against the target of C_w, a buffer is allocated as opposed to the risk neutral $\mathbb{E}\left[\sum_{s=0}^{t} z^{t,s}\right]$. The index k controls how large this buffer is, approaching the risk neutral case as $k \to \infty$, and the fully robust case, *i.e.*, no violations on the constraint are permitted, as $k \to 0$. Therefore, as k decreases, the buffer grows more conservative. As such, $\beta^{t,s}$ in (13) can be interpreted as a risk averse correction to the cumulative survival probabilities $\hat{h}^{t,s}$, which yields guarantees against constraint violation in Proposition 2.

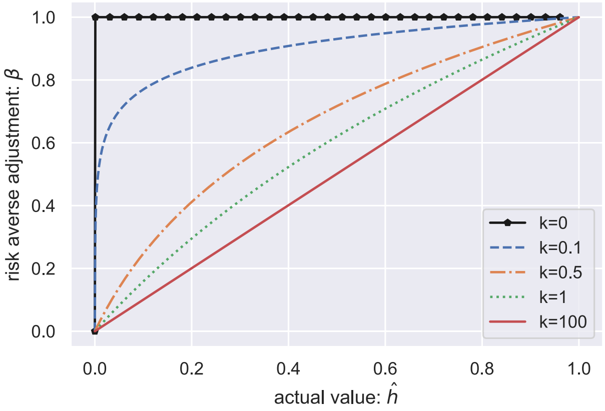

Figure 3 Illustration of how the entropic risk constraint constitutes a risk adjustment

Proposition 4. *For any* $t \in \mathcal{T}$ *and a given cost function* $r(s)$*, the expression*

$$k\log\mathbb{E}\left[\exp\left(\frac{\sum_{s=0}^{t} r(s) y^{t,s}}{k}\right)\right]$$
$$= \sum_{s=0}^{t} r(s)\left(x_{t-s} - \sum_{\tau=0}^{s-1} p^{t-\tau,s-\tau}\right) + \sum_{s=0}^{t} k\log\mathbb{E}\left[\exp\left(\lambda_{t-s} r(s)/k\right)\right] \tag{14}$$

is affine in decision variables $\boldsymbol{x}, \boldsymbol{p}$*. In particular,*

1. *Waiting cost constraints* (10) *corresponds to the case where* $r(s) = s$*:*

$$k\log\mathbb{E}\left[\exp\left(\frac{\sum_{s=0}^{t} y^{t,s} s}{k}\right)\right] = \sum_{s=1}^{t} s\left(x_{t-s} - \sum_{\tau=0}^{s-1} p^{t-\tau,s-\tau}\right) + \sum_{s=1}^{t} k\log\mathbb{E}\left[\exp\left(\lambda_{t-s} s/k\right)\right]. \tag{15}$$

2. *End-of-horizon queue length constraint* (12) *corresponds to the case where* $r(s)=1$*:*

$$k\log\mathbb{E}\left[\exp\left(\frac{\sum_{s=0}^{T}y^{T,s}}{k}\right)\right]=\sum_{s=0}^{T}\left(x_{T-s}-\sum_{\tau=0}^{s-1}p^{T-\tau,s-\tau}\right)+\sum_{s=0}^{T}k\log\mathbb{E}\left[\exp\left(\lambda_{T-s}/k\right)\right]. \qquad (16)$$

Proposition 5. *For any* $t\in\mathcal{T},s\in\mathcal{T}$*, push constraints* (11) *are affine in decision variables* $\boldsymbol{x},\boldsymbol{p}$*:*

$$k\log\mathbb{E}\left[\exp\left(\frac{p^{t,s}-y^{t-1,s-1}}{k}\right)\right]=\sum_{\tau=0}^{s-1}p^{t-\tau,s-\tau}-x_{t-s}+k\log\mathbb{E}\left[\exp\left(-\lambda_{t-s}/k\right)\right]. \qquad (17)$$

Theorem 1 **(Reformulation)**. *Under Assumptions 1 and 2, Problem* (8) *can be reformulated and solved via a bisection search where each sub-problem is a mixed-integer linear optimization problem.*

Proof of Theorem 1. By Proposition 2, our risk constraints are monotonically decreasing in k. Thus, model (8) can be solved by bisection search on k, where each sub-problem is a mixed-integer linear optimization problem by Propositions 3 – 5. □

Remark 1. At this point, we should emphasize that almost all parameters can be treated as decision variables. The model (8) is just one possible optimization approach. An alternative is to fix the risk level k and optimize some linear objective, *e.g.,* minimize the number of servers used such that all entropic risk constraints hold at a prescribed risk level k. This can be useful when the clinic wants to optimize the shift schedule of doctors; when capacity C_t at time t is treated as a decision variable, the model optimizes the (time-varying) capacity of the clinic.

Remark 2. The static nature of push decisions $\boldsymbol{p}$ is the price that we paid for tractability; to the best of our abilities, we do not know how to form a tractable model if the pushes were not static. However, the push decisions do provide us a *certificate of guarantee* for the scheduling decisions, in the sense that if we choose to execute the first-stage decisions, then there will exist at least one method of managing the queue dynamics, namely the push decisions, that guarantees desired outcomes on waiting time and overtime under the entropic measure of risk. More critically, we see, in the numerical experiments, that this, despite being a crass reduction of the solution space, is more than sufficient to arrive at a model that uniformly performs better than the baseline sample average approximation (SAA) model. This justifies the trade-off we made for tractability, because we gained in terms of flexibility to model all kinds of uncertainties, otherwise difficult in traditional formulations.

We also make one final quick comment about the multiple server setting. In the above discussion, we have used C to represent the capacity of the server. In the literature, this is often handled as a single-server system with $C=1$. Under the Pipeline Queues framework, extending from the single-server case to the multiple-server case requires nothing more than removing the requirement that $C = 1$, as long as servers are assumed to be independent and homogeneous. Indeed, one can imagine the servers (with states $z_i^{t,s}$ over index i) arranged in parallel, fanning out from the same queue. The queue decides when patients are pushed into the servers, as capacity is freed up. As such, servers do not account for any waiting time; that, is accounted by the queue. This means that all service times, no matter which servers the patients are in, are independent and identically distributed. This allows the sum $z^{t,s} := \sum_{i=1}^{C} z_i^{t,s}$ to be described by the Binomial random variable, and the difference $C - \sum_s z^{t,s}$ to be understood as the number of idle servers at time t. For further description of this, we recommend the original narrative in Bandi and Loke (2018). Hence, Theorem 1 extends to the case where $C > 1$ under the following assumption:

Assumption 3. *Service times for all patients across all doctors are assumed to be independently and identically distributed.*

3. General Setting with Re-entries

In the preceding section, we considered a single-server network to illustrate key model primitives. In this section, we extend the problem to general networks and as realistic a setting as possible.

Consider the general setting where the planner is required to schedule all N appointments by T, the last allowed slot on the schedule. As before, walk-in patients are modeled as λ_t. For scheduled patients, there is a chance that they will not show up for the appointment with probability $1 - \gamma$. If it is desired to incorporate no shows into the model, then the inflow can be modified to

$$y^{t,0} = \mathrm{Bin}\left(x_t, \gamma\right) + \lambda_t.$$

For brevity, we will consider the case where $\gamma = 1$ in the subsequent derivation. Additional variables and that are different from those in Table 1 are summarized in Table 2.

		General Setting
		Parameters and known quantities
γ	:	No show probability
$h_j^{t,s}$	:	Hazard rate—probability patient in block j served for s periods at time t completes service at $t+1$
C_2	:	Server capacity – total number of X-ray doctors
$C_{w,j}$	:	Targeted total patient waiting time to keep under in block j
		State and decision variables
x_t	:	Decision variable of patients scheduled to arrive at time t
$y_j^{t,s}$	:	Random variable of patients waiting for s periods in the queue of block j at time t
$z_j^{t,s}$	:	Random variable of patients served for s periods in the server of block j at time t
$p_j^{t,s}$	:	Recourse of patients dispatched into server of block j after waiting for s at time t

Table 2 List of additional parameters and variables in the general setting

The most important feature we want to incorporate is the element of patient re-entry. In our context, a portion of patients are required to undertake an X-ray examination, before returning to consult with the doctor again. There are three difficulties with this. First, it is not clear how to model the dynamics and uncertainty involved in patient re-entry. In our model, we shall introduce a virtual network that represents the service dynamics accurately. Second, there are now non-trivial decisions in the routing process, *e.g.,* between two patients, one who has just arrived and one returning from an X-ray examination, who should be routed to see the doctor first? In our model, we shall house these patients in two separate queues and denote separate routing decisions for each queue. Third, knowledge about whether patients require an X-ray examination only emerges after the scheduling. In our model, we use a random variable to represent this uncertainty, which can be tractably reformulated.

Consider a network with three blocks as in Figure 4. Each block consists of a queue and a server. Compared to our earlier model in Figure 1, we have two additional blocks – one for X-ray examinations and the other for re-entry. We refer to these queue-server blocks as: First consultation, X-ray examination, and Return consultation. In fact, first and return consultations feed into the same queue and server. However, patients in both queues and the decisions taken can be differentiated. We assume, for convenience, patients go for an X-ray examination at most once.

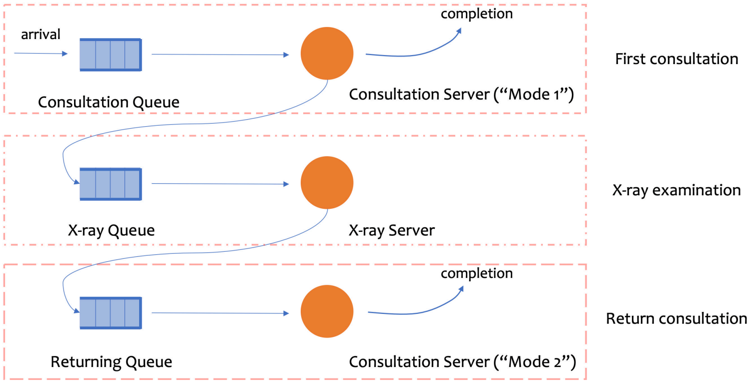

Figure 4 Patient flow network

As before, the planner makes scheduling decisions, $\boldsymbol{x}$, in the first stage, in order to obtain the best guarantees $Q(\boldsymbol{x})$ on having waiting time constraints observed.

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & Q(\boldsymbol{x}) \\ \text{s.t.} \quad & \sum_{t=0}^{T} x_t = N \\ & x_t \in \{0,1\} \qquad \forall t \in \mathcal{T}. \end{aligned} \tag{5}$$

Informally, the second-stage decision to route patients attempts to minimize the risk parameter k, so as the seek the best guarantee level ε, through a series of entropic risk constraints.

$$Q(\boldsymbol{x}) = \min_{\boldsymbol{p}\geq\boldsymbol{0},k>0} \quad k \qquad \text{s.t.} \quad \text{entropic risk constraints.} \tag{18}$$

We now proceed to describe these routing decisions and entropic risk constraints.

Variables and definition

We make the following definitions for the different blocks $j = 1, 2, 3$ representing the First consultation, X-ray examination and Return consultation: $y_j^{t,s}, z_j^{t,s}, p_j^{t,s}$ denote the number of patients in the queue, the server, and the push variables for their respective blocks with time indices t and s as before. In this model, the routing decisions are essentially the push variables. Similarly, let $h_j^{t,s}$ represent the hazard rates for patients in block j.

The complication arises in the first block, where first, we have to determine if the patient requires X-ray examination. This would be information that the planner would *not* have at the point of scheduling and only manifests at the point of the first consultation. Let us suppose that the doctor would assess each patient to require an X-ray with a known probability of q.

Second, we need to differentiate the hazard rates. In practice, doctors, with knowledge that the patient requires an X-ray examination, will likely delay their prognosis till after the X-ray results are ready. As such, patients requiring X-ray will likely end their first consultation much faster than other patients. To that end, let $h_0^{t,s}$ refer specifically only to the hazard rates of patients *not requiring X-ray examination* after first consultation (and hence leave the system thereafter) and let $h_1^{t,s}$ denote the likelihood of service completion for patients *requiring X-ray examination*, where they are routed to the X-ray queue thereafter. In general, $\boldsymbol{h}_0 \neq \boldsymbol{h}_1 \neq \boldsymbol{h}_3$.

Dynamics

The dynamics for the queues and servers in each of the blocks remain largely the same; the introduction of X-rays only affects the server in the first consultation block and the inflow to the X-ray queues. The rest are easily defined.

$$\begin{aligned}
& z_j^{t,s} \sim \mathrm{Bin}(z_j^{t-s,0}, \hat{h}_j^{t,s}) && \text{for } j = 2, 3 \\
& z_j^{t,0} = \sum_{s\in\mathcal{T}} p_j^{t,s} && \text{for } j = 1, 2, 3 \\
& y_j^{t,s} = y_j^{t-1,s-1} - p_j^{t,s} && \text{for } j = 1, 2, 3 \\
& y_1^{t,0} = x_t + \lambda_t,
\end{aligned}$$

where $\hat{h}_j^{t,s} \triangleq \prod_{\tau=1}^{s}(1-h_j^{t-\tau,s-\tau})$ for $j=2,3$, are the cumulative probabilities that extend naturally from Proposition 1.

It turns out, for the server in the First consultation block, $z_1^{t,s}$, the dynamics can be written in the same form, except where $\hat{h}_1^{t,s}$ is defined slightly differently: Consider this definition on $z_1^{t,s}$.

$$z_1^{t,s} = \sum_{\ell=1}^{z_1^{t-s,0}} \Big(\mathbb{1}(b_{\ell,t-s}=1, \text{ patient } \ell \text{ stays till time } t) + \mathbb{1}(b_{\ell,t-s}=0, \text{ patient } \ell \text{ stays till time } t)\Big), \tag{19}$$

where $b_{\ell,t} \sim \text{Bernuolli}(q)$ indicates whether the ℓth patient that is pushed into the server at time t requires X-ray examination. Let us examine this. First, patients in the server $z_1^{t,s}$ originated from the cohort $z_1^{t-s,0}$. For any patient in cohort $z_1^{t-s,0}$, there are three possibilities by the time of t:

(i) Patient is still in the server at time t, and would require an X-ray examination.

(ii) Patient is still in the server at time t, but would not require an X-ray examination.

(iii) Patient is no longer in the server at time t.

Suppose patients are labelled $\ell = 1, \dots, z_1^{t-s,0}$. Then $\mathbb{1}(b_{\ell,t-s}=1, \text{patient } \ell \text{ stays till time } t)$ denotes the first case and $\mathbb{1}(b_{\ell,t-s}=0, \text{patient } \ell \text{ stays till time } t)$ the second. The third case no longer contributes to $z_1^{t,s}$. Hence, expression (19) is obtained by summing over all the $z_1^{t-s,0}$ patients.

We are also left to define the inflows to queues at the X-ray examination and return consultation blocks, $y_j^{t,0}$ for $j=2,3$, which comprise patients who have completed service from the earlier blocks.

$$y_2^{t,0} = \sum_{s=0}^{t-1} \sum_{\ell=1}^{z_1^{t-1-s,0}} \mathbb{1}\big(b_{\ell,t-s-1}=1, u_\ell^{t-1,s}=1\big), \tag{20}$$

$$y_3^{t,0} \sim \sum_{s=0}^{t-1} \text{Bin}\left(z_2^{t-1,s}, h_2^{t-1,s}\right), \tag{21}$$

where $u_\ell^{t,s}=1$ if the ℓ^{th} patient that is pushed to first consultation server at time $t-s$ will complete his service at time $t+1$. Equation (20) is obtained from the same logic as in (19), while (21) describes that the inflow into $y_3^{t,0}$ is simply all patients who finished their X-ray examination.

Proposition 6. *For any $t \in [T], s \in [t]$,*

a) Server variables obey $z_1^{t,s} \sim Bin\Big(z_1^{t-s,0}, \hat{h}_1^{t,s}\Big)$, with cumulative survival probability after s periods for cohort $(t-s)$ given by $\hat{h}_1^{t,s} \triangleq q\prod\limits_{\tau=1}^{s}(1-h_1^{t-\tau,s-\tau}) + (1-q)\prod\limits_{\tau=1}^{s}(1-h_0^{t-\tau,s-\tau})$, $\hat{h}_1^{t,0}=1$.

Queue variables can be written as

b) $y_2^{t,0} \sim \sum\limits_{s=0}^{t-1} Bin\big(z_1^{t-s-1,0}, \bar{h}_1^{t-1,s}\big)$, where we define $\bar{h}_1^{t-1,s} \triangleq q h_1^{t-1,s}\prod\limits_{\tau=1}^{s}(1-h_1^{t-1-\tau,s-\tau})$, and

c) $y_3^{t,0} \sim \sum\limits_{s=0}^{t-1} Bin\Big(z_2^{t-s-1,0}, h_2^{t-1,s}\hat{h}_2^{t-1,s}\Big)$.

Remark 3. For fixed t, $z_1^{t,s}$ are independent across all $s \in \mathcal{T}$ because different patients are independent of each other, *i.e.*, the RHS of (19) is independent across $s \in \mathcal{T}$.

Constraints and reformulation

Consider the full problem (5) where the second-stage problem $Q(\boldsymbol{x})$ is given by

$$
\begin{aligned}
Q(\boldsymbol{x}) := & \min_{\boldsymbol{p},k>0} k && && (22)\\
\text{s.t} k\log\mathbb{E} & \left[\exp\left(\frac{\sum_{s=0}^{t}\left(z_1^{t,s}+z_3^{t,s}\right)-C}{k}\right)\right]\le 0 && \forall t\in\mathcal{T} && (23)\\
k\log\mathbb{E} & \left[\exp\left(\frac{\sum_{s=0}^{t} z_2^{t,s}-C_2}{k}\right)\right]\le 0 && \forall t\in\mathcal{T} && (24)\\
k\log\mathbb{E} & \left[\exp\left(\frac{\sum_{s=0}^{t} s y_j^{t,s}-C_{w,j}}{k}\right)\right]\le 0 && \forall t\in\mathcal{T},\forall j\in[3] && (25)\\
k\log\mathbb{E} & \left[\exp\left(\frac{p_j^{t,s}-y_j^{t-1,s-1}}{k}\right)\right]\le 0 && \forall t\in[T], s\in[T],\forall j\in[3]. && (26)
\end{aligned}
$$

The constraints we apply on the model are similar as before, where (23) and (24) are capacity constraints on the servers, (25) are the queue waiting time constraints, and (26) are the push constraints. Specifically, (23) indicates that capacity is shared between first and return consultations, because the same server (doctor) is used for them. We leave waiting time constraints separate in (25), however. This imposes different priorities among first and return consultation services.

Because of Proposition 6, we are able to express $z_1^{t,s}$ in the same form as before, so Proposition 3 applies as per usual. The only additional result we require is how to deal with the slightly different forms of $y_2^{t,0}$ and $y_3^{t,0}$. We cover the reformulation of $k\log\mathbb{E}\left[\exp\left(\sum_{s=0}^{t} r(s) y_2^{t,s}/k\right)\right]$ in the Proposition below. For all other constraints, the reformulation is written out clearly in the Appendix A.

Proposition 7. *For any $t\in[T]$, $k\log\mathbb{E}\left[\exp\left(\sum_{s=0}^{t} r(s) y_2^{t,s}/k\right)\right]$ is affine in decision variables $\boldsymbol{p}_1,\boldsymbol{p}_2$:*

$$
k\log\mathbb{E}\left[\exp\left(\sum_{s=0}^{t} r(s) y_2^{t,s}/k\right)\right] = \sum_{\bar\tau=0}^{t-1}\sum_{\tau=0}^{t-\bar\tau-1} p_1^{t-\bar\tau-1,\tau}\eta_2^{t,\bar\tau} - \sum_{s=0}^{t} r(s)\sum_{\tau=0}^{s-1} p_2^{t-\tau,s-\tau}, \qquad (27)
$$

where $\bar h_1^{t,s}$ are as defined in Proposition 6 and constants

$$
\eta_2^{t,\bar\tau} \triangleq k\log\left(1+\sum_{s=0}^{\bar\tau}\bar h_1^{t-s-1,\bar\tau-s}\left(\exp\left(\frac{r(s)}{k}\right)-1\right)\right),\ \forall t\in[T],\bar\tau=0,1,\ldots,t-1
$$

can be calculated from primitives.

Remark 4. The proof of Proposition 7 is technical but noteworthy in the sense that it features a technique in the reformulation that is not seen in the original paper by Bandi and Loke (2018). Moreover, the derivation proves that we can consider other sources of inflows to the X-ray station, *as long as they are independent*, for example, if the same X-ray station serves patients from multiple clinics, or if we also have random walk-in patients to the X-ray station.

As before, we have the result: Practically, these sub-problem, while posed as mixed-integer linear optimization problems, can be solved quickly.

Theorem 2 **(Reformulation)**. *Under Assumptions 1 – 3, Problem* (22) *can be reformulated and solved via a bisection search where each sub-problem is a mixed-integer linear optimization problem.*

Proof of Theorem 2. This follows from the proof of Theorem 1, Propositions 6 – 7, and Propositions **??** – **??** in Appendix A. □

3.1. Incorporating additional realistic features

We have set up a basic structure to obtain a scheduling with high guarantees on short wait times for the intraday scheduling problem with patient re-entry, walk-ins and no-shows. In this section, we illustrate how the model may be extended to better describe actual operations, in particular, random transportation time between service stations, and uncertain appointment arrival times.

Transportation times

To address the transportation times between servers, we can model the transportation as a service by adding new "traffic" blocks (Figure 5). Similar results can be derived for such a model.

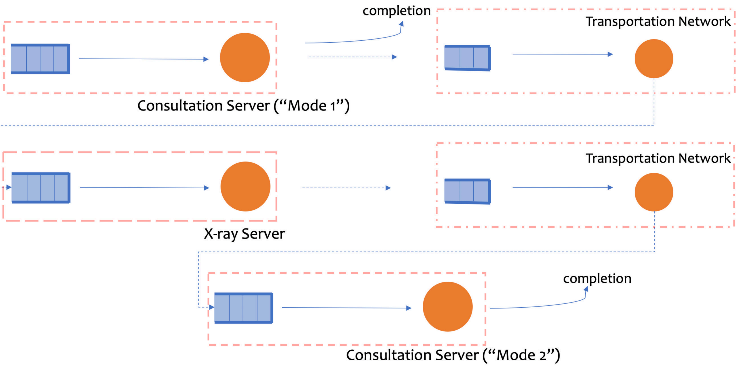

Figure 5 Patient flow network with transportations

Uncertain arrival times

In practice, scheduled patients may not show up at precisely their appointment times. Let $\tilde{A}_t$ denote the random arrival time of a patient initially scheduled to arrive at time t. Suppose we know the probability distribution of actual arrival time $\tilde{A}_t$, which are independent across $t \in \mathcal{T}$. Then we rewrite $y_1^{t,0}$ as:

$$y_1^{t,0} = \sum_{\tau=0}^{T} x_\tau \mathbb{1}\left(\tilde{A}_\tau = t\right) + \lambda_t \quad \forall t \in \mathcal{T}.$$

The entropic risk constraint (10) can still be evaluated.

Proposition 8. *The term* $k \log \mathbb{E}\left[\exp\left(\sum_{s=0}^{t}\sum_{\tau=0}^{T} x_\tau a(s) \mathbb{1}\left(\tilde{A}_\tau = t-s\right)/k\right)\right]$ *is affine in* $\boldsymbol{x}$*:*

$$k \log \mathbb{E}\left[\exp\left(\sum_{s=0}^{t}\sum_{\tau=0}^{T} x_\tau a(s) \mathbb{1}\left(\tilde{A}_\tau = t-s\right)/k\right)\right] = \sum_{\tau=0}^{T} x_\tau \alpha_{t,\tau},$$

where $\alpha_{t,\tau} \triangleq k \log \mathbb{E}\left[\exp\left(\sum_{s=0}^{t} a(s) \mathbb{1}\left(\tilde{A}_\tau = t-s\right)/k\right)\right]$ *for* $t,\tau \in \mathcal{T}$ *are constants that can be calculated from primitive data.*

Proof of Proposition 8 This follows because random variables $\tilde{A}_t$ are independent across $t \in \mathcal{T}$ and the fact that x_t for $t \in \mathcal{T}$ are binary. □

Remark 5. From data, we can estimate the probability $\mathbb{P}\left(\tilde{A}_\tau = t\right)$. Then, for all (discrete) $\tau, t \in \mathcal{T}$, the above constants $\alpha_{t,\tau}$ can be computed as:

$$\alpha_{t,\tau} = k \log\left(\sum_{s=0}^{t} \mathbb{P}\left(\tilde{A}_\tau = t-s\right) \exp\left(a(s)/k\right) + 1 - \sum_{s=0}^{t} \mathbb{P}\left(\tilde{A}_\tau = t-s\right)\right).$$

Overtime man-hours

Now we discuss how to capture the overtime man-hours, *e.g.,* the number of man-hours operated beyond the operational horizon T. We let T_c be a large constant, at which point the no patient should remain in the system. The total number of busy periods of all servers from time $T+1$ to time T_c can be written as:

$$\sum_{j=1}^{3}\sum_{t=T+1}^{T_c}\sum_{s=1}^{t} z_j^{t,s}. \tag{28}$$

Therefore, we can impose targets on overtime man-hours using entropic risk constraints. In Proposition 9, we show this can be evaluated efficiently.

Proposition 9. *The term* $k \log \mathbb{E}\left[\exp\left(\sum_{j=1}^{3}\sum_{t=T+1}^{T_c}\sum_{s=1}^{t} z_j^{t,s}/k\right)\right]$ *is affine in* $\boldsymbol{p}_1, \boldsymbol{p}_2, \boldsymbol{p}_3$*:*

$$k \log \mathbb{E}\left[\exp\left(\sum_{j=1}^{3}\sum_{t=T+1}^{T_c}\sum_{s=1}^{t} z_j^{t,s}/k\right)\right] = \sum_{j=1}^{3}\sum_{\bar{t}=0}^{T}\sum_{\tau_{\bar{t}}=1}^{\bar{t}} p_j^{\bar{t},\tau} \phi_j^{\bar{t}},$$

where $\phi_j^{\bar{t}} \triangleq k \log\left(\sum_{t=T+1}^{T_c} \hat{h}_j^{t,\bar{t}+t} \exp\left(1/k\right)\right)$ *are constants that can be calculated from primitive data.*

4. Numerical Study on Hospital Data from NUHS

In this section, we conduct a numerical study on our model (5). We illustrate that appropriately considering re-entries and variability in patient behaviours can significantly improve performance, and our model does so without compromising tractability.

4.1. The setting and data

Our data originates from clinics run by 29 Orthopaedic consultant led teams in a tertiary healthcare institution in Singapore – National University Health System (NUHS). The clinics are divided into six different sub-specializations, and the data is collected over one year for patient appointment and visits along these divisions. Our data contains over $80,000$ patient visits to over 100 doctors. Data fields include patient appointment time, arrival time (or no-show), first consultation duration, whether they are required for an X-ray examination, and return consultation duration.

Our data suggests that consultation times approximately follow a geometric distribution, with a slight skew towards shorter consultation times. We show the histogram of consultation times from one particular specialization in Figure 6. For a simple reference, the coefficient of determinant is above 97% in both fits. The consultation times from other specializations behave similarly. For convenience, we shall use geometric distributions as the underlying true service time distribution in our simulations; though our model remains tractable even if we adopted the empirical service time distributions. We observe that average service time of first consultations is slightly longer than that of return consultations. In addition, return consultations have a lighter tail.

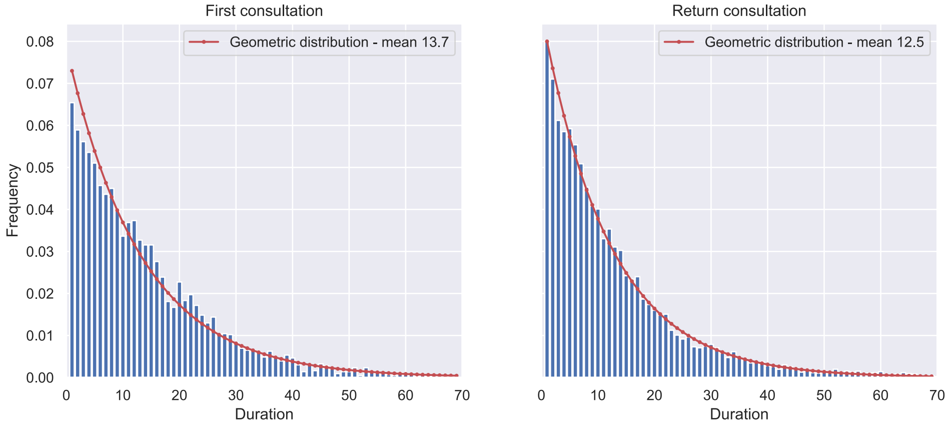

Figure 6 **Empirical distribution of consultation time**

Patient re-entry and variability in patient behaviours are apparent in our data. This is reflected in Table 3. As we can see, no-show probability can be as high as 29% and the probability that a patient is required to go for an additional X-ray examination can be as high as 39%. In addition, more than 20% of the patients arrive later than their appointment time by at least 10 minutes, and more than 47% of the patients arrive earlier by at least 10 minutes. Therefore, one cannot ignore these factors in scheduling and a high fidelity model can be helpful. It is also important to note that the patterns of patient re-entry and variability in behaviours vary significant from specialization to specialization. As such, the optimal policy for scheduling patients will expectedly vary across specializations. Moreover, some clinics are knowledgeable of the composition of the patients at the point of making the scheduling decisions. For example, some Orthopaedic clinics see a mixture of first time patients and patients on repeat consultation. The former usually are required to undergo a series of tests, whereas the latter is far less likely to require the tests. Table 3 also illustrates the proportion of first time visit. As a final note on this, the walk-ins to our partnering clinics are not systematically recorded, although they indeed exist. As such, it is not reflected in Table 3.

Table 3 Summary of no-show, X-ray probability, and proportion of first time visit

	Specialization				
	1	2	3	4	5
Average no-show probability	19%	23%	27%	29%	29%
Average X-ray probability	39%	24%	18%	15%	13%
Proportion of first time visit	27%	32%	34%	35%	41%

In these Orthopaedic clinics, the day is divided into the morning and afternoon shifts, which we can safely treat as being separate. In addition, as doctors are required to perform treatments beyond their consultation shifts, each shift is only $T = 120$ minutes long. As such, in all subsequent analysis, we will consider this as our time window of one shift. Each clinic is staffed by a single doctor in that shift. We consider a fixed number of 12 patients, which the planner needs to schedule. The current practice in the clinics is to schedule patients in equal intervals of ten minutes over these 120 minutes, filling up the slots as long as there is a backlog of patients. Equal-interval scheduling policy is common in practice due to its simplicity. It is also reasonably successful – in many cases, such equal-interval policy does not deteriorate performance severely (Stein and Cote 1994).

4.2. Simulation set-up

In this section, we examine the performance of our model (5) and attempt to understand the consequences of planning without considering patient re-entries, or variability in behaviours. To do so, we compare our model against two benchmarks, the current equal-interval scheduling policy in the clinics, and the solution obtained from the sample average approximation (SAA) model in Denton and Gupta (2003) where only service time uncertainty is considered. The rationale for the former is that equal-interval policy ignores all information regarding the stochastic nature of the random variables, hence provides a measure of the potentiality of considering them. The baseline SAA model has been extended to handle uncertainties such as sequencing decisions and no-shows (*e.g.,* Erdogan and Denton 2013, Berg et al. 2014, Jiang et al. 2017). However, to the best of our knowledge, we do not know of the means to incorporate all the uncertainties we consider into

SAA, while remaining tractable. Hence, any performance gains that our model makes over the baseline SAA model will represent both the improvement as a result of the paradigm adopted by our approach, as well as, the potential gains that can be reaped by considering the uncertainties otherwise difficult to incorporate into an SAA model.

Limitations of the SAA approach in our problem context

Specific to our problem, the difficulty arises when modeling the interaction between uncertainties and decisions. Here, SAA would be required to enumerate over all possible scenarios, leading to an exponentially large optimization problem, which is not realistic. For example, when walk-ins and no-shows are considered, the sequence of service is stochastic, and this uncertainty is decision-dependent. As such, one needs to generate samples over all possible times at which the walk-in is to occur. When re-entries are considered, the time at which the patient re-joins the consultation queue is stochastic and depends on the scheduling and routing decisions. Moreover, we can only make routing decisions after the type of patient, whether they are re-entry patient or otherwise, materializes. In all these cases, we are not aware of how the interdependence of the uncertainty and presence of counterfactual modeling could be supported by SAA, without requiring an exponential large sample of data that enumerates over all possible scenarios.

As the previous subsection illustrated, re-entries, no-shows, and walk-ins, are common not only in Orthopaedic clinics but also other clinics, in general. The setting of having different patient compositions is a particularly important feature at our partnering clinics, who observe starkly different re-entry probabilities and service times between first-time patients and repeat patients. Such settings further complicate matters.

We expect our model to outperform baseline SAA model in these realistic settings because SAA struggles to utilize all available information. We have structured the simulations later to illustrate precisely that our high fidelity model outperforms benchmarks in these practical situations at clinics. We will describe the specific experiments in detail later.

Solving the benchmark models

The solution to the SAA model is derived as follows. First, we generate 300 sample paths of the twelve patient consultation times as the input to the SAA model. Then, cost parameters (overtime cost and waiting time cost) for the SAA model are chosen such that the overtime metric in our model and the SAA model roughly matches. Finally, we solve the SAA model and get its optimal scheduling policy.

To perform the comparison, we run our optimal policy against the equal-interval policy and the baseline SAA model under $50,000$ simulations, independently and identically generated according to the information we derived from data or assumed. Under each simulation, we implement the three scheduling policies and use the same routing policy for all, which is a first-come, first-served policy. In other words, first consultation patients (including walk-ins) always arrive at the end of the first consultation queue. After X-ray examination, patients join the end of returning queue upon returning. The doctor always sees the next patient in the queues, and always clears the returning queue before first consultation patients.

We compute the metrics of total waiting time, system overtime, and maximum instantaneous waiting time for each of the policies and then average them over the $50,000$ simulations. Total waiting time is the sum of waiting times of all 12 patients in the queue, and overtime is the amount of excess time experienced by the doctor beyond $T = 120$ minutes to finish all consultations. Instantaneous waiting time at any time t is the total waiting time among patients in the queue at t, *i.e.,* it is $\sum_{i=1}^{n_t} w_{it}$, where n_t is the number of patients in the queue at time t and w_{it} is the waiting time that patient i has experienced until time t. This can be seen a proxy for the *queue length* at time t. Thus, the maximum instantaneous waiting time refers to the largest of these instantaneous waiting times amongst all times $t = 1, \dots, T$, which can be interpreted as the longest length of the queue achieved at any time point.

In the subsequent discussion, we will conduct several numerical studies. First, we study only the effect of patient re-entries, by varying the proportion of patients requiring X-ray examination. Then, upon this framework, we will now consider the effects of incorporating other modeling features. In particular, we will illustrate this for patient walk-ins and no-shows. Finally, we consider cases with heterogeneous patients. More specifically, we consider a situation where there are two types of patients (Type A and B), who may have very different consultation times, likelihoods of requiring X-ray examination, and earliness patterns, as discussed in the above section. In our clinics, the current policy is to schedule all first timers first and to fill the returning patients into later slots on the schedule. The logic behind this practice is that repeat consultations require shorter consultations and hence scheduling them later would front-load the demand and reduce server idle time. We will see in our simulations later if this is a good policy.

For each performance metric, the performance gap between the benchmark policies and our policy is calculated with our policy as the base, *i.e.*, (benchmark metric − our metric)/our metric, with positive values indicating worse performance compared to our model. *Standard deviations* of the performance metrics of our policy are also reflected, and a 1% *significance level* is assumed.

For all subsequent simulations and experiments (except experiment 2), we took care to ensure that parameters were always varied in a manner that roughly maintains the same system load. In particular, we chose to ensure that the system is always within the heavy traffic regime. This ensures that any observed differences in performance do not arise from changes in the system loads.

4.3. Varying X-ray re-entry probability

As previously motivated, the planner would not, *a priori*, know whether or not the patients require X-ray examinations. Instead, after the first consultation, each patient requires an X-ray examination with probability q. Otherwise, with probability $1-q$, the patient leaves after first consultation.

In experiment 1, we vary this probability q. We summarize the performance of our model in Table 4. Both our model and the SAA model consistently outperform the equal-interval policy, though the former two display no statistically significant difference in performance.

Table 4 Performance comparison (Experiment 1): varying X-ray probability q

		Metrics (mins)		
		Total waiting	Overtime	Max. instantaneous waiting
$q = 0.25$	Equal-interval	163.5 (**3.0%**)	28.1 (**7.3%**)	65.5 (**10.1%**)
	SAA	160.2 (**0.9%**)	26.1 (**-0.4%**)	60.0 (**0.8%**)
	Ours (± s.d.)	158.8 (± 0.61)	26.2 (± 0.10)	59.5 (± 0.25)
$q = 0.30$	Equal-interval	170.5 (**3.5%**)	29.8 (**7.2%**)	68.7 (**10.8%**)
	SAA	166.1 (**0.9%**)	27.7 (**-0.4%**)	62.6 (**1.0%**)
	Ours (± s.d.)	164.7 (± 0.62)	27.8 (± 0.11)	62.0 (± 0.27)
$q = 0.35$	Equal-interval	180.6 (**3.5%**)	32.0 (**7.0%**)	73.3 (**11%**)
	SAA	176.1 (**0.9%**)	29.8 (**-0.3%**)	66.6 (**0.9%**)
	Ours (± s.d.)	174.5 (± 0.64)	29.9 (± 0.10)	66.0 (± 0.27)

We also illustrate the structure of the optimal policies for ours and the SAA model in Figure 7. Here, the x-axis is the order of arrival of the patients, and the y-axis is the interarrival times between successive patients, that is, the gap of time before the next patient arrives. Critically, both models advocate a 'dome-shaped' structure in the interarrival times. This is a classical observation (Wang 1993). The intuition is that queues are more likely to build up in the middle of the planning horizon; therefore, a longer interarrival time in the middle balances this out. As both models advocate similar structure, their performances are similar. Nonetheless, we show in the ensuing discussion, that this classical optimality of the dome-shaped policy, can be broken.

4.4. Variability in patient behaviours and distinct patient classes

The previous discussion is only meant to serve as a basis to gain some initial intuition on the structure of the optimal policy, and what are the initial gains we can expect to reap from considering just one dimension of uncertainty. In what ensues, we conduct four groups of *independent* experiments, each of which, built upon the basic re-entry model above:

Experiments 2 and 3: In this pair of experiments, we fix the probability of re-entry at $q = 0.25$ and include exogenous uncertainties of walk-ins and no-shows. Walk-ins λ_t at time t are assumed to

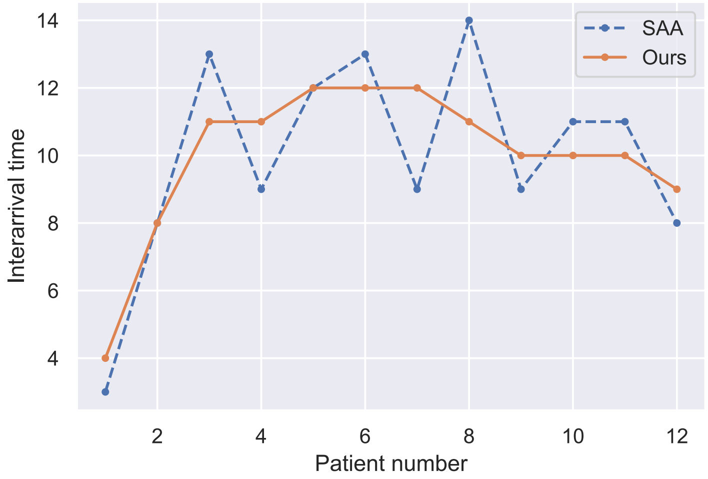

Figure 7 **Optimal interarrival times (ours and SAA) in Experiment 1**

follow a non-time-homogeneous Poisson distribution with rates α_t for $t \in [T]$. We suppose scheduled patients will *not* show up with (an independent) probability $1-\gamma$, which is allowed to vary over $\gamma = 0.8$ to $\gamma = 0.9$. This is done for two instances:

Experiment 2: Where $\alpha_{65} = 1$ and $\alpha_t = 0$ everywhere else, creating an single influx of random inflow to the system; and,

Experiment 3: Where $\alpha_{38} = \alpha_{75} = \alpha$ and $\alpha_t = 0$ everywhere else, creating two separate waves of walk-ins. α is varied for a few choices ranging from 0.7 to 0.85.

Experiments 4 and 5: In the next pair of experiments, we consider the situation where there are two distinct patient types, having different probabilities of requiring X-ray examination. Suppose there are n_A Type A patients, who are likely (with probability q) to require an X-ray examination than the $12 - n_A$ Type B patients, who do not. In this case, the sequence of service (*i.e.*, sequence of Type A and B patients to schedule) enters the decision variables. In our model, this is achieved by allowing our decision variables x_A^t and x_B^t to be indexed by the type. This is not possible, to the best of our knowledge, for the SAA model because of re-entry. Hence for SAA, like the equal-interval policy, all Type A patients are assumed to be scheduled ahead of all Type B patients. The SAA model retains the freedom to decide on the interarrival timings. Again, two instances are modeled:

Experiment 4: Where chance of re-entry q, and number of Type A patients, n_A, are varied; and,

Experiment 5: Where we fix $q = 1.0$ and $n_A = 3$, but allow Type A patients to randomly arrive earlier than their scheduled time, *uniformly* within some margin D, ranging from 2 to 5 periods.

For Experiments 2 to 5, we replicate Table 4 in appendix. These are tabulated in Tables 6 – 9 in Appendix B, summarizing the performance metrics of our model against the two benchmarks.

we summarize their differences in Figure 8. Each chart in the figure plots one of two metrics, the waiting time metric and the maximum instantaneous waiting time metric. Because we have constructed the SAA model by matching its overtime metric to coincide with ours (which can be verified in the performance tables themselves), the overtime metric has been omitted. For each chart, the horizontal and vertical axes represents the performance gap recorded by the benchmark equal-interval policy and the SAA policy in comparison to our model respectively. Each point refers to an instance in each Experiment.

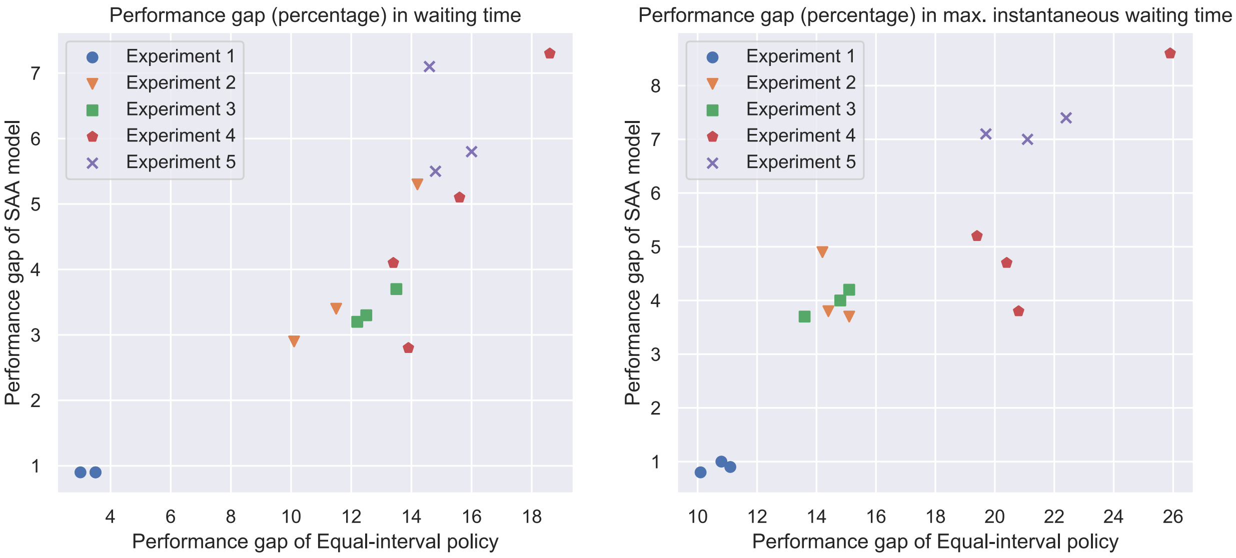

Figure 8 **Summary of performance gaps relative to our model**

Discussion of results and insights

From Figure 8, we can see that in all cases, our model is able to achieve significant improvements over the equal-interval policy, sometimes by as much as almost 20% in waiting time reductions and also in excess of 25% reductions in maximum instantaneous waiting times, and hence maximum queue lengths. More interestingly, the performance between our model and the SAA model also begins to diverge for Experiments 2 to 5. In fact, the performance of our model is superior to that of the SAA model in all these 4 experiments, showing that our model is more effective in practical settings compared to benchmarks. Where our model outperforms SAA, could broadly be due to two different reasons, both arising from the limitations of SAA to accurately model different features of the problem, whereas our model is precisely designed to do so.

Walk-ins and no-shows: Let us first examine the differences going from Experiment 1 to Experiments 2 and 3. In the former, we only had the uncertainty pertaining to whether patients required

X-ray examination or not. In the latter two, we included non-time-homogeneous walk-in patterns and no-shows, leading our model to outperform SAA when it had not before.

As we had discussed in our subsection on the ‘Limitations of the SAA approach’, once walk-ins and no-shows are incorporated, it is no longer possible to model them using the SAA approach without requiring exponential number of samples. The baseline SAA approach we considered simply ignores them, yielding the same dome-shaped policy, as before. In Figure 9, we see that our model no longer gives a dome-shaped policy. In Experiment 2, our model leaves a much longer interarrival time in the middle, anticipating the single influx of walk-ins; in Experiment 3, interarrival time is bimodal, anticipating two waves of walk-ins. In these cases, the traditional dome-shaped policy given by the baseline SAA model is suboptimal, resulting in statistically significant differences in both waiting time and maximum instantaneous waiting time. This observation is consistently replicated across different combinations of γ and α.

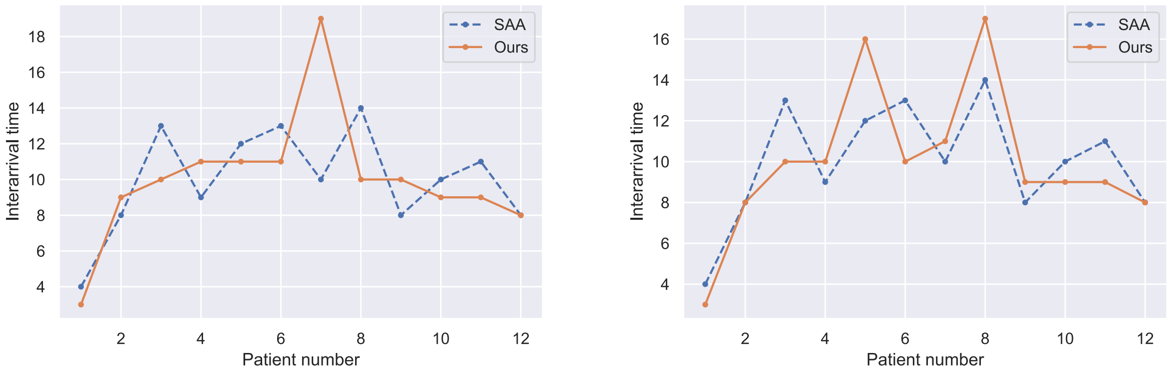

Figure 9 Optimal interarrival times (ours and SAA) in Experiments 2 (left) and 3 (right)

This behaviour arises because the walk-ins are non-time-homogeneous. Hence, the optimal policy ought to adapt to variations in walk-ins across time. In other words, failing to account for the specific structure in the underlying uncertainty, in this case, the spike in arrivals at specific times, necessarily leads to suboptimal solutions. General purpose models, such as the baseline SAA model, cannot be expected to handle such structure adequately. For instance, in Experiment 2, as show-up probability γ decreases, our model provides more significant improvements compared to the SAA model (Table 6). This is because the influx of walk-in creates a more significant relative change in traffic to the system as no-show probability increases. Indeed, capturing the structure of the uncertainty is a common theme in the literature. In many threads of Robust Optimization, the definition of the uncertainty set is critical to defining the geometry of the uncertainty, and this determines the nature of the optimal solutions (*e.g.,* Jiang et al. 2017).

Heterogeneous patients: Larger differences in performance are observed between our model and SAA in Experiments 4 and 5, showing that our model is more effective than our benchmarks in complex and practical settings. As explained in the 'Limitations of the SAA approach' subsection, differentiating two classes of patients was a planning consideration faced by our partnering clinics.

Similar to considering walk-ins and no-shows, having heterogeneous patients leads to a particular structure in the underlying uncertainties. For instance, a patient with high probability of X-ray examination creates more uncertainty in downstream waiting times, because of the greater potentiality of re-entry. Models that fail to account for this would result in suboptimal solutions. Our model reacts to such information, as we can see this in terms of the optimal policy of our model. Table 5 illustrates the case where $n_A = 3$, $q = 0.9$. It is clear from the results that the optimal sequencing of patients is not strictly arranging all the Type A patients first, even though, it does come quite close to doing so. Our model schedules two types of patients alternately. This distributes the Type A patients evenly across the earlier part of the shift and hence reduces the chance that there would be snowballing of waiting times as a result of having too many Type A returning from X-ray examinations at the same time. Nonetheless, such a change already induces a performance gap in the SAA model of as much as 7%, even if their chosen interarrival times are relatively similar. The difference is further aggravated if earliness is added to the mix.

Table 5 Scheduling policy in Experiment 4 when $n_A = 3$, $q = 0.9$

	Patient order											
	1	2	3	4	5	6	7	8	9	10	11	12
Appointment time (our model)	0	2	11	24	35	45	58	71	83	93	102	111
Patient sequence (our model)	A	B	A	B	A	B	B	B	B	B	B	B
Appointment time (SAA model)	0	4	13	25	35	47	58	70	83	93	104	114
Patient sequence (SAA model)	A	A	A	B	B	B	B	B	B	B	B	B

5. Concluding Remarks and Insights

We have considered a high fidelity intraday scheduling problem with patient re-entry, and also incorporating uncertain elements such as no-shows, walk-ins, earliness, and distinct patient types. Our model remains tractable, and simulations illustrate that the model is able to improve existing policies significantly, including a methodology that employs SAA. Most importantly, by being able to handle this myriad of uncertainties, our model surely performs better than one that is unable to handle some subset of them, hence necessarily ignoring them.

Our numerical simulations also illustrate some key insights. First, there are critical types of uncertainties that need to be handled very carefully, and that which general purpose methodologies, such as SAA, do not necessarily handle well. These usually occur when the uncertainty in question depends on either earlier uncertainties or decisions. Secondly, the classical observation that 'dome-shaped policies' are optimal in intraday scheduling problems can be broken. In general, they seem to perform well. Nonetheless, when coupled with non-time-homogeneous walk-ins, re-entries, distinct patient types, and other decision features, such as the sequencing in the types of patients, the dome-shaped structure may be insufficient to guarantee good performance.

Features we considered in our model that pertain to the optimization of scheduling are not confined to the healthcare setting. In particular, wafer fabrication, machine scheduling and ride-sharing also involve a multi-step process with re-entry and stochastic service time distributions. As such, the wider applicability of the model we introduced here to areas beyond healthcare are potentially numerous. We intend to work on these in the future.

References

Bandi, C., D. Bertsimas, N. Youssef. 2015. Robust queueing theory. *Operations Research* **63**(3) 676–700.

Bandi, C., G.G. Loke. 2018. Exploiting hidden convexity for optimal flow control in queueing networks URL `https://ssrn.com/abstract=3190874`.

Berg, B.P., B.T. Denton, Ayca E.S., T. Rohleder, T. Huschka. 2014. Optimal booking and scheduling in outpatient procedure centers. *Computers and Operations Research* **50** 24–37.

Braverman, A., J.G. Dai, X. Liu, L. Ying. 2017. Fluid-model-based car routing for modern ridesharing systems. *ACM SIGMETRICS Performance Evaluation Review* **44**(1) 11–12.

Dai, J.G., P. Shi. 2017. A two-time-scale approach to time-varying queues in hospital inpatient flow management. *Operations Research* **65**(2) 514–536.

Dai, J.G., T. Tezcan. 2011. State space collapse in many-server diffusion limits of parallel server systems. *Mathematics of Operations Research* **36**(2) 271–320.

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

Denton, B., J. Viapiano, A. Vogl. 2007. Optimization of surgery sequencing and scheduling decisions under uncertainty. *Health Care Management Science* **10**(1) 13–24.

Erdogan, A.S., B. Denton. 2013. Dynamic appointment scheduling of a stochastic server with uncertain demand. *INFORMS Journal on Computing* **25**(1) 116–132.

Feldman, J., N. Liu, H. Topaloglu, S. Ziya. 2014. Appointment scheduling under patient preference and no-show behavior. *Operations Research* **62**(4) 794–811.

Follmer, H., T. Knispel. 2011. Entropic risk measures: Coherence vs. convexity, model ambiguity, and robust large deviations. *Stochastics and Dynamics* URL `https://doi.org/10.1142/S0219493711003334`.

Follmer, H., A. Schied. 2002. Convex measures of risk and trading constraints. *Finance and Stochastics* URL `https://doi.org/10.1007/s007800200072`.

Ge, D., G. Wan, Z. Wang, J. Zhang. 2014. A note on appointment scheduling with piecewise linear cost functions. *Mathematics of Operations Research* **39**(4) 1244–1251.

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

Gurvich, I. 2014. Diffusion models and steady-state approximations for exponentially ergodic markovian queues. *The Annals of Applied Probability* **24**(6) 2527–2559.

Gurvich, I., J. Luedtke, T. Tezcan. 2010. Staffing call centers with uncertain demand forecasts: A chance-constrained optimization approach. *Management Science* **56**(7) 1093–1115.

Hall, N.G., D.Z. Long, J. Qi, M. Sim. 2015. Managing underperformance risk in project portfolio selection. *Operations Research* **63**(3) 660–675.

Ho, C.J., H.S. Lau. 1992. Minimizing total cost in scheduling outpatient appointments. *Management Science* **38**(12) 1750–1764.

Jaillet, P., J. Qi, M. Sim. 2016. Routing optimization under uncertainty. *Operations Research* **64**(1) 186–200.

Jiang, R., S. Shen, Y. Zhang. 2017. Integer programming approaches for appointment scheduling with random no-shows and service durations. *Operations Research* **65**(6) 1638–1656.

Kong, Q., S. Li, N. Liu, C.P. Teo, Z. Yan. 2019. Appointment scheduling under time-dependent patient no-show behavior. *Management Science (Forthcoming)* URL `https://ssrn.com/abstract=3359707`.

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

Luo, J., V.G. Kulkarni, S. Ziya. 2012. Appointment scheduling under patient no-shows and service interruptions. *Manufacturing & Service Operations Management* **14**(4) 670–684.

Mak, H.Y., Y. Rong, J. Zhang. 2014. Sequencing appointments for service systems using inventory approximations. *Manufacturing & Service Operations Management* **16**(2) 251–362.

Mak, H.Y., Y. Rong, J. Zhang. 2015. Appointment scheduling with limited distributional information. *Management Science* **61**(2) 316–334.

Padmanabhan, D., K. Natarajan, K. Murthy. 2018. Exploiting partial correlations in distributionally robust optimization URL `https://ssrn.com/abstract=3270706`.

Qi, J. 2017. Mitigating delays and unfairness in appointment systems. *Management Science* **63**(2).

Stein, W.E., M.J. Cote. 1994. Scheduling arrivals to a queue. *Computers and Operations Research* **21**(6) 607–614.

Wang, P.P. 1993. Static and dynamic scheduling of customer arrivals to a single-server system. *Naval Research Logistics (NRL)* **40**(3) 345–360.

Wang, X., V-A. Truong, D. Bank. 2018. Online advance admission scheduling for services, with customer preferences URL `https://arxiv.org/abs/1805.10412v1`.