Intraday Scheduling with Patient Re-entries and Variability in Behaviours

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

Chaithanya Bandi
Kellogg School of Management, Northwestern University, Evanston, IL 60208, c-bandi@kellogg.northwestern.edu

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, lateness, 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. Our model has potential wider applications, *e.g.* in machine scheduling.

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 is modelled under a P-Queue (Bandi and Loke 2018) paradigm 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: Simulations shows that our model can achieve as much as 15% reduction on various metrics including patient waiting time and server overtime over the benchmark policy.

Managerial insights: We present an optimization model that is easy to implement in practice and tractable to compute. Our simulation indicates that not accounting for patient re-entry or variability in patient behaviours will lead to sub-optimal policies, especially when X-ray rate is high and lateness has a large spread.

Key words: Optimization, Scheduling

History: September 30, 2019

1. Introduction

With the onset of the Fourth Industrial Revolution, the world has transited into an age of co-integrability between machine and machine, and between human and machine. From this paradigm, the notion of Smart Cities has emerged – megacities at the forefront of technology exploitation and integration. These technological innovations powered by the algorithms of artificial intelligence and data arising from the expansive sensor networks of the Internet of Things continually challenge the upper bounds of our standard of living.

Despite the progress, many economic sectors are still plagued by significant inefficiency. According to the study IBM Institute for Business Value (2009), the three most inefficient sectors are Healthcare, Government and Safety, and Education. In particular, the Healthcare sector suffers from the highest inefficiency – a global total of $2.5 trillion is estimated to be wasted annually. This makes up more than 40% of the total economic value of the Healthcare sector, placed at $4.3 trillion. The study further estimates that nearly 35% of this $2.5 trillion could potentially be saved, illustrating the huge room for improvement in the Healthcare sector.

However, the challenges faced by the healthcare sector are entrenched. Part of the difficulty arises from the variability in the behaviours of patients and healthcare providers alike, and the sheer complexity and interconnectedness of the sector. This contributes to the overall level of uncertainty faced in decision-making. A particularly illustrative example would be the appointment scheduling procedure. Appointment scheduling forms a major part of any hospital's daily operations. It, however, is made complicated by a variety of operational complexities and uncertainties – uncertain service time, patient re-entry, and variability in patient behaviours such as no-shows, walk-ins, and lateness. In particular, variability in patient behaviours often cannot be avoided because it is fundamentally difficult to incentivize behaviours within a public and non-profitable setting, such as the one the healthcare sector faces. Indeed, there are numerous papers written on incentivizing the right patient behaviours alone (see *e.g.* Molfenter 2013). These complexities can lead to significant inefficiencies; therefore, it is important for the planner to take these into consideration.

In this paper, we shall focus very specifically on precisely the intraday appointment scheduling problem. Even though this problem is posed in the healthcare context, we believe that such a model easily has wider applicability in other smart cities sectors, *e.g.,* machine scheduling problems in airport operations and semiconductor manufacturing.

The intraday appointment scheduling problem

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). In this paper, we are concerned about the intraday scheduling problem where patients complete their appointments within the day. A planner is given a sequence of patient appointments, and makes a *here and now* decision on when to schedule them. In the process, she incurs costs on operations, such as resource idleness and overtime of healthcare providers. The intraday scheduling problem is a key component of the broad appointment scheduling literature.

In practice, variability in patient behaviours, arising from no-shows, lateness, and walk-ins, and patient re-entry can lead to inefficiencies. In this paper, we partnered with the Orthopaedic clinics in the National University Hospital of Singapore. In these clinics, patients first request for an appointment, which the planner schedules in advance. On the appointment 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 17.5 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.

As we can see, re-entry and variability in patient behaviours are very common features of a clinic's operations. These features bring uncertainties to the system which leads to inefficiencies. No-shows and lateness create physician idleness and overtime, which are both very costly to the healthcare system. Re-entry causes a 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 for the current patients in the systems, and the expected times before they exit the system. This is a daunting challenge.

Key approaches in the literature

There are three streams in the literature to approach the problem, namely the 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) employed 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 dynamics was written as a linear sum of four different times – the inter-arrival appointment time (decisions in the first stage), the stochastic service times of each patient; from which the dynamics yields the waiting time beyond the scheduled appointment time, and the idle time incurred by the server. The objective was 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, Qi 2017).

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 propose a tractable formulation under a marginal moments ambiguity set. Jiang et al. (2017) modelled a distributionally robust single-server intraday scheduling problem with no-shows. Qi (2017) introduced 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).

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 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 approximations (SAA), which has the potential to grow with larger sample sizes.

As such, to the best of our knowledge, it remains challenging to incorporate patient re-entries, walk-ins, 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 model akin to the two-stage stochastic programming model. In the first stage, as before, 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 twofold. First, it advances the literature on intraday patient scheduling, by illustrating that incorporating a model ingesting high fidelity information on service times and queue dynamics can lead to further increases in performance over existing methodologies. Our model is 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 in previous models. Our model is shown in simulations, to provide improvement on all metrics including patient waiting times and server overtime over existing policies. We reduce patient waiting times by as much as 15% and overtime by as much as 8%. Second, it advances the stream of work in Pipeline Queues, by extending the model, which is designed for the optimization and control of flows, to scheduling and demand matching problems. Using the Pipeline Queues framework to resolve the second-stage problem is a novel application. In particular, it also adds to nascent developments in the theory of Pipeline Queues, by examining the possibility where the class of patients is determined *stochastically* midway through the system, as opposed to being *a priori* knowledge.

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.

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 N a known integer. 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 the 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 appointments within T discrete periods of the clinic's operating hours. When patients arrive at their scheduled time, they join the queue, along with walk-in patients at some arrival rate λ_t. As the availability of doctors (with maximum capacity C) is freed up, patients are dispatched to the doctor for consultation, from which they leave the system once their consultation is completed. The consultation time is random, modelled by $h^{t,s}$, the survival probability of completing service after being served for s periods at time t. The goal of the planner therefore is to ensure that total patient waiting time is controlled under some target C_w. 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.

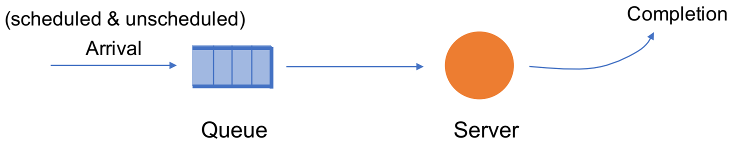

Figure 1 Patient Flow in a Single-server Setup

Single-Server Setting

		Parameters and known quantities
N	:	Number of patients to schedule
T	:	Last modelling time
λ_t	:	Random arrivals at time t
$h^{t,s}$	:	Known hazard rate – probability patient served for s periods at time t completes service
C	:	Server capacity – total number of doctors
C_w	:	Targetted 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

Variables and dynamics

Our formulation is posed as a two-stage problem. In the first stage, the planner makes a *here-and-now* decision to schedule N patients, before any uncertainty unfolds; N is fixed and known to the planner. Let x_t be a first-stage decision variable denoting the number of appointments scheduled to arrive at time $t \in \mathcal{T} := \{0, 1, ..., T\}$. In a single-server setting, x_t is binary. All patients must be scheduled: $\sum_{t=0}^{T} x_t = N$.

The second-stage problem is a multi-period problem. After the schedule $\boldsymbol{x}$ is decided, the uncertainty in the number and arrival times of walk-in patients in the first time step materializes, as well as the uncertainty in service times. 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 from queue to server. 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 route patients in a fashion such that waiting time and overtime targets are met at every time juncture (possibly infeasible if $\boldsymbol{x}$ was decided poorly). A schematic of this process is illustrated in Figure 2. It is important to keep in mind that our primary focus is on the scheduling decision $\boldsymbol{x}$. The routing decisions only serve to approximate second stage queueing costs.

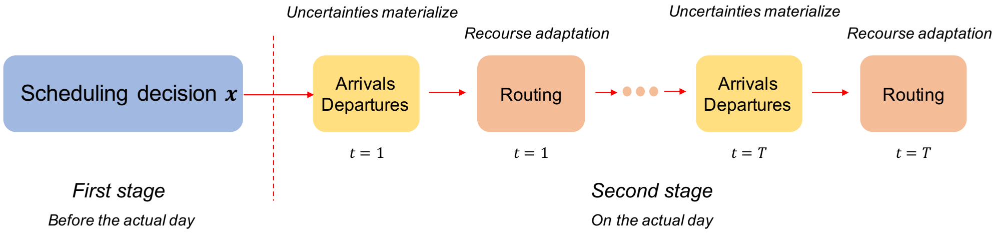

Figure 2 **A rundown of events in two stages**

Let us define the second-stage problem. 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 remain in the queue and server at time t who have already waited for exactly s periods 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. Similarly for $z^{t,s}$. 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 time 1. 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 $y_q^{t,0}$ 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.*, $p^{t,s} \le y^{t-1,s-1}$. We require $p^{0,s} = 0$ for $s \in \mathcal{T}$, and $p^{t,s} = 0$ for $s > t$ subsequently. These definitions leads to the dynamics for $\forall t \in [T]$,

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

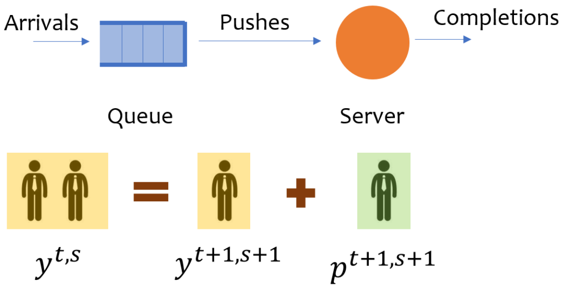

Figure 3 **Illustration of Dynamics** (2)

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 modelled via this approach (Dai and Shi 2017). Moreover, these probabilities can be readily obtained from data. The distribution need not be stationary.

This definition induces a Binomial model on the outflow from the server. More specifically, for a patient in consultation for s periods at time t, at time $t+1$, whether this patient completes his service is a Bernoulli variable with probability $h^{t,s}$. Hence, aggregating over all patients, we have

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

Proposition 1. *The state variable $z^{t,s}$ is the conditional binomial random variable:*

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

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 can be shown by induction on the law of total probability. □

Constants $\hat{h}^{t,s}$ 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.

Constraints

The goal of the second-stage is to ensure that operational targets, comprising waiting time and overtime targets, 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

$$\mathbb{P}\left[\sum_{s=0}^{t} y^{t,s}s - C_w \leq 0\right] > 1-\varepsilon. \tag{5}$$

$\sum_{s=0}^{t} y^{t,s}s$ 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. However, constraint (5) is non-convex, and it is hard to derive a tractable reformulation in general.

Now, (5) 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 $f(1) \leq \varepsilon$, $f(\delta) \to 1$ as $\delta \to 0$ and $f(\delta) \to 0$ as $\delta \to \infty$. Then we want to consider the (infinite) family of chance constraints:

$$\mathbb{P}\left[\sum_{s=0}^{t} y^{t,s}s - C_w \geq \delta\right] \leq f(\delta), \quad \forall \delta > 0. \tag{6}$$

Evidently, this includes (5) 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.,* Foellmer and Schied 2002, Foellmer 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 dis-utility function 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 the areas of 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* $k \le -1/\log(\varepsilon)$.

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

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

In particular, $g_k\left(\sum_{s=0}^{t} y^{t,s} s - C_w\right) \le 0$ *implies that the bound on probability of constraint violation (6) 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 $g_k\left(\tilde{\xi} - \Gamma\right) \le 0$, then

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

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

Remark 1. The probability measure in (6) is, in general, non-convex; therefore, the space of $\boldsymbol{y}$ satisfying (6) may not be convex. However, the measure g_k is, which ensures that the space of $\boldsymbol{y}$ satisfying $g_k\left(\sum_{s=0}^{t} y^{t,s} s - C_w\right) \le 0$ is convex.

As a result of Proposition 2, as long as we can find some particular $k \le -1/\log(\varepsilon)$ such that the constraint $g_k\left(\sum_{s=0}^{t} y^{t,s} s - C_w\right) \le 0$ is satisfied, then we are done. This can be done by searching for the smallest k such that $g_k\left(\sum_{s=0}^{t} y^{t,s} s - C_w\right) \le 0$, then checking if $k \le -1/\log(\varepsilon)$. 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:

$$\begin{aligned} Q(\boldsymbol{x}) := \min_{\boldsymbol{p}, k > 0} \quad & k \\ \text{s.t.} \quad & k \log \mathbb{E}\left[\exp\left(\left(\sum_{s=0}^{t} y^{t,s} s - C_w\right) \Big/ k\right)\right] \le 0 \qquad \forall t \in \mathcal{T} \end{aligned} \tag{7}$$

Other constraints may be addressed similarly. To this effect, let us state our full model formally.

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & Q(\boldsymbol{x}) \\ \text{s.t.} \quad & \sum_{t=0}^{T} x_t = N \\ & x_t \in \{0,1\} \qquad \forall t \in \mathcal{T} \end{aligned} \tag{8}$$

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

$$\begin{aligned}
Q(\boldsymbol{x}) = \min_{\boldsymbol{p},k>0} \quad & k && && (9)\\
\text{s.t.} \quad & k \log \mathbb{E}\left[\exp\left(\frac{\sum_{s=0}^{t} z^{t,s} - C}{k}\right)\right] \le 0 && \forall t \in [T] && (10)\\
& k \log \mathbb{E}\left[\exp\left(\frac{\sum_{s=0}^{t} y^{t,s} s - C_w}{k}\right)\right] \le 0 && \forall t \in [T] && (11)\\
& k \log \mathbb{E}\left[\exp\left(\frac{p^{t,s} - y^{t-1,s-1}}{k}\right)\right] \le 0 && \forall t \in [T], \forall s \in [T] && (12)\\
& + \textbf{overtime constraint}
\end{aligned}$$

In the first-stage, the planner seeks to minimize ε (via $Q(\boldsymbol{x})$ by Proposition 2), and best improve her chance of finding a suitable k. In the second stage problem, the first constraint (10) represents capacity constraints. The second constraint (11) states that the total waiting time of all patients in the queue at any time must be bounded by C_w. This can generalize to any affine constraint $\sum_{s=0}^{t} y^{t,s} r(s) \le b$ for some constants $r(s)$, $s = 0, \ldots, t$, and b. The third constraint (12), termed 'push constraint', ensures patient dispatch, $\boldsymbol{p}$, does not exceed the number of patients in queue, $\boldsymbol{y}$.

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

$$k \log \mathbb{E}\left[\exp\left(\frac{\sum_{s=0}^{T} y^{T,s} - L}{k}\right)\right] \le 0, \tag{13}$$

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 (33). 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 easily evaluated into a form affine in decisions $\boldsymbol{x}$, and hence the optimization model (9) can be tractably solved.

Before proceeding, we make a quick comment about the recourse variables – the push decisions $\boldsymbol{p}$. In general, $\boldsymbol{p}$ is state-dependent. Instead, we restrict ourselves to static $\boldsymbol{p}$, in other words, $\boldsymbol{p}$ is only allowed to be a function of distribution information on λ^t and the service distributions, represented by the survival rates $h^{t,s}$ (see Remark 2 for more discussion). If we make such an assumption, then Theorem 1 that reformulates model (8) holds.

Proposition 3. *For any $t \in [T]$, capacity constraints (10) are affine in push variables $\boldsymbol{p}$:*

$$k \log \mathbb{E}\left[\exp\left(\frac{\sum_{s=0}^{t} z^{t,s}}{k}\right)\right] = \sum_{s=0}^{t} \beta^{t,s} \sum_{\tau=0}^{t-s} p^{t-s,\tau}, \tag{14}$$

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 (14). 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 2.1 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 (14) 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. Let us look at the proof in detail.

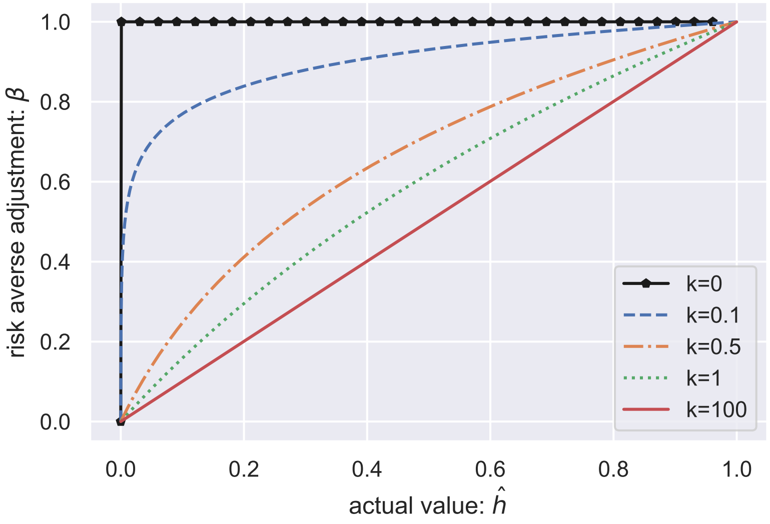

Figure 4 Illustration of how the entropic risk constraint constitute a risk adjustment

Proof of Proposition 3. Recall that by Proposition 1, the state variable $z^{t,s}$ can be written as

$$z^{t,s} \sim \text{Bin}\left(z^{t-s,0}, \hat{h}^{t,s}\right).$$

Evaluating its moment generating function and using $z^{t-s,0} = \sum_{\tau=0}^{t-s} p^{t-s,\tau}$,

$$\mathbb{E}\left[\exp\left(z^{t,s}/k\right)\right] = \exp\left(\sum_{\tau=0}^{t-s} p^{t-s,\tau} \log\left(1 - \hat{h}^{t,s} + \hat{h}^{t,s}\exp\left(\frac{1}{k}\right)\right)\right).$$

Notice that $z^{t,s} \sim \text{Bin}\left(z^{t-s,0}, \hat{h}^{t,s}\right)$, which also indicates that for a fixed time t, the state variables $z^{t,s}$ are *independent* across $s \in \mathcal{T}$. Thus, the complete reformulation can be written as

$$\begin{aligned} &k\log\mathbb{E}\left[\exp\left(\frac{\sum_{s=0}^{t} z^{t,s}}{k}\right)\right] = \sum_{s=0}^{t} k\log\mathbb{E}\left[\exp\left(\frac{z^{t,s}}{k}\right)\right] \\ =&k\log\mathbb{E}\left[\exp\left(\frac{z^{t,0}}{k}\right)\right] + \sum_{s=1}^{t} k\log\mathbb{E}\left[\exp\left(\frac{z^{t,s}}{k}\right)\Big| z^{t-s,0}\right] \\ =&\sum_{\tau=0}^{t} p^{t,\tau} + \sum_{s=1}^{t} k\log\left[\exp\left(\sum_{\tau=0}^{t-s} p^{t-s,\tau}\log\left(1-\hat{h}^{t,s}+\hat{h}^{t,s}\exp\left(\frac{1}{k}\right)\right)\right)\right] = \sum_{s=0}^{t}\beta^{t,s}\sum_{\tau=0}^{t-s}p^{t-s,\tau} \end{aligned}$$

□

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

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

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

1. *Waiting cost constraints* (11) *corresponds to the case where $r(s) = s$:*

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

2. *End-of-horizon queue length constraint* (13) *corresponds to the case where $r(s) = 1$:*

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

Proposition 5. *For any $t \in \mathcal{T}, s \in \mathcal{T}$, push constraints* (12) *are affine in decision variables $\boldsymbol{x}, \boldsymbol{p}$:*

$$k\log\mathbb{E}\left[\exp\left(\frac{p^{t,s} - y^{t-1,s-1}}{k}\right)\right] = \sum_{\tau=0}^{s-1} p^{t-\tau,s-\tau} - x_{t-s} + k\log\mathbb{E}\left[\exp\left(-\lambda_{t-s}/k\right)\right]. \tag{18}$$

As the proofs of Propositions 4 and 5 are similar to that of Proposition 3, we relegate it to Appendix A.

Theorem 1 **(Reformulation)**. *Problem* (9) *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 2a, our risk constraints are monotonically decreasing in k. Thus, model (9) can be solved by bisection search on k, where each sub-problem is a mixed integer linear optimization problem by Propositions 3 – 5. □

Remark 2. At this point, we should emphasize that almost all parameters can be treated as decision variables. The model (9) 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 schedule doctor shifts optimally; when capacity C_t at time t is treated as a decision variable, the model dynamically optimizes the capacity of the clinic.

In exchange, we have to settle for the recourse decision, which is the push variable, being static and not state dependent as it in general is. This is less consequential, since we are only interested if there exists some recourse that makes the probability of constraint violation low, given a scheduling $\boldsymbol{x}$. The static push decisions are sufficient in doing so.

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. In practice, T is often earlier than T_c, the time of the clinic closure, thereafter, any further service would construe as server overtime. The planner has a secondary goal to keep overtime capped under some target C_o. As before, walk-in patients arrive with rate λ_t. They can also have an arrival time A_t that can deviate from their stipulated time t.

No shows. For scheduled patients, there is a chance that they will not turn 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

$$y_1^{t,0} = \text{Bin}\left(x_t, \gamma\right) + \lambda_t.$$

For brevity, we will consider the case where $\gamma = 1$ in the subsequent discussion. These variables are summarized in Table 2.

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. Second, there are now non-trivial decisions in the routing process, *e.g.* between two patients, one who has just arrived and

		General Setting
		Parameters and known quantities
N	:	Number of patients to schedule
T_c	:	Time of clinic closure, after which it is considered overtime
T	:	Last modelling time
λ_t	:	Random arrivals at time t
γ	:	No show probability
A_t	:	Random arrival time of patient initially scheduled to arrive at time t
$h^{t,s}$	:	Known hazard rate – probability patient served for s periods at time t completes service
C	:	Server capacity – total number of doctors
C_w	:	Targetted total patient waiting time to keep under
		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 Parameters and Variables in the General Setting

one returning from an X-ray examination, who should be routed to see the doctor first? Third, knowledge about whether patients require an X-ray examination only emerges after the scheduling.

To deal with the first challenge, consider a network with three blocks as illustrated in Figure 5. 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 actuality, first and return consultation feed into the same queue and servers. However, patients in both queues can be differentiated, and the decisions taken can be different. We assume, for convenience, patients go for an X-ray examination at most once.

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.

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & Q(\boldsymbol{x}) \\ \text{s.t.} \quad & \sum_{t=0}^{T} x_t = N \\ & x_t \in \{0,1\} \qquad \forall t \in \mathcal{T} \end{aligned} \tag{8}$$

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.

$$Q(\boldsymbol{x}) = \min_{\text{routing decisions}, k>0} \quad k \tag{19}$$

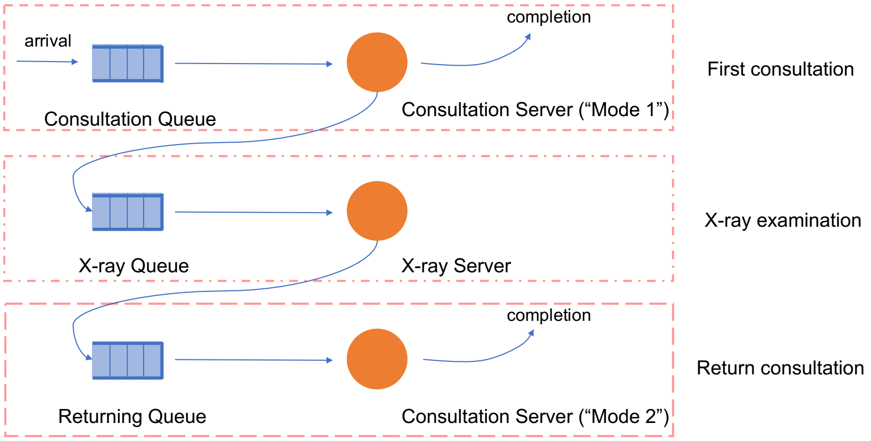

Figure 5 **Patient Flow Network**

$$\text{s.t.} \quad \text{entropic risk constraints}$$

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 survival probabilities 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 survival probabilities. 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 survival probabilities 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 remains 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.

$$z_j^{t,s} \sim \text{Bin}(z_j^{t-s,0}, \hat{h}_j^{t,s}) \qquad\qquad \text{for } j = 2, 3$$

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

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 $z_1^{t,s}$.

$$z_1^{t,s} = \sum_{\ell=1}^{z^{t-s,0}} \Big(\mathbb{1}(b_{\ell,t-s}=1, \text{ patient } \ell \text{ stays till time } t) + \mathbb{1}(b_{\ell,t-s}=0, \text{ patient } \ell \text{ stays till time } t)\Big), \tag{20}$$

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's 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 (20) 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.

$$y_2^{t,0} = \sum_{s=0}^{t-1} \sum_{\ell=1}^{z^{t-1-s,0}} \mathbb{1}\big(b_{\ell,t-s-1}=1, u_\ell^{t-1,s}=1\big), \tag{21}$$

$$y_3^{t,0} \sim \sum_{s=0}^{t-1} \text{Bin}\left(z_2^{t-1,s}, h_2^{t-1,s}\right), \tag{22}$$

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 (21) is obtained from the same logic as in (20), while (22) 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 $z_1^{t,s} \sim Bin\left(z_1^{t-s,0}, \hat{h}_1^{t,s}\right)$, with cumulative survival probability after s periods for cohort $(t-s)$ given by $\hat{h}_1^{t,s} \triangleq q\prod_{\tau=1}^{s}(1-h_1^{t-\tau,s-\tau}) + (1-q)\prod_{\tau=1}^{s}(1-h_0^{t-\tau,s-\tau})$, $\hat{h}_1^{t,0}=1$. Queue variables can be written as *b)* $y_2^{t,0} \sim \sum_{s=0}^{t-1} Bin\left(z_1^{t-s-1,0}, \bar{h}_1^{t-1,s}\right)$, *where we define* $\bar{h}_1^{t-1,s} \triangleq q h_1^{t-1,s} \prod_{\tau=1}^{s}(1-h_1^{t-1-\tau,s-\tau})$, *and*

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

Proof of Proposition 6. For any patient ℓ that has been in service since time $t-s$, by conditioning on $b_{\ell,t-s}$,

$$\mathbb{P}\Big(\mathbb{1}(\text{patient } \ell \text{ stays till time } t) = 1\Big) = \hat{h}_1^{t,s},$$

where

$$\hat{h}_1^{t,s} \triangleq q\prod_{\tau=1}^{s}(1-h_1^{t-\tau,s-\tau}) + (1-q)\prod_{\tau=1}^{s}(1-h_0^{t-\tau,s-\tau}).$$

Therefore, we can write

$$\mathbb{1}(b_{\ell,t-s}=1, \text{ patient } \ell \text{ stays till time } t) + \mathbb{1}(b_{\ell,t-s}=0, \text{ patient } \ell \text{ stays till time } t) \sim \text{Bernoulli}(\hat{h}_1^{t,s}).$$

Because random variables on different patients are independent, equation (20) is a sum of $z_1^{t-s,0}$ identical and independent Bernoulli random variables, resulting in $z_1^{t,s} \sim \text{Bin}(z_1^{t-s,0}, \hat{h}_1^{t,s})$.

To prove (b), first notice that the first consultation service times are *i.i.d.* for all patients requiring X-ray examination. In addition, by definition, we have $\mathbb{P}\Big(u_\ell^{t-1,s}=1\Big|b_{\ell,t-s-1}=1\Big) = h_1^{t-1,s}\prod_{\tau=1}^{s}(1-h_1^{t-1-\tau,s-\tau})$ for all $t\in[T], s=0,\ldots,t-1, \ell\in[z^{t-s-1}]$. Therefore,

$$\mathbb{P}\Big(b_{\ell,t-s-1}=1, u_\ell^{t-1,s}=1\Big) = \bar{h}_1^{t-1,s},$$

which indicates $\mathbb{1}\left(b_{\ell,t-s-1}=1, u_\ell^{t-1,s}=1\right) \sim \text{Bernoulli}\left(\bar{h}_1^{t-1,s}\right)$. It now follows from independence.

Part (c) easily follows from the definition of $y_3^{t,0}$ and Proposition 1. □

Remark 3. For fixed t, $z_1^{t,s}$ are independent for all $s\in\mathcal{T}$, since the RHS of (20) is independent over $s\in\mathcal{T}$ because different patients are independent of each other.

Constraints and reformulation

Consider the full problem:

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & Q(\boldsymbol{x}) \\ \text{s.t.} \quad & \sum_{t=0}^{T} x_t = N \\ & x_t \in \{0,1\} \qquad \forall t\in\mathcal{T} \end{aligned} \tag{23}$$

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

$$Q(\boldsymbol{x}) := \min_{\boldsymbol{p},k>0} \; k \tag{24}$$

$$\text{s.t. } k\log\mathbb{E}\left[\exp\left(\frac{\sum_{s=0}^{t}\left(z_1^{t,s}+z_3^{t,s}\right)-C}{k}\right)\right]\leq 0 \qquad \forall t\in\mathcal{T} \tag{25}$$

$$k\log\mathbb{E}\left[\exp\left(\frac{\sum_{s=0}^{t}z_2^{t,s}-C_2}{k}\right)\right]\leq 0 \qquad \forall t\in\mathcal{T} \tag{26}$$

$$k\log\mathbb{E}\left[\exp\left(\frac{\sum_{s=0}^{t}sy_j^{t,s}-C_{w,j}}{k}\right)\right]\leq 0 \qquad \forall t\in\mathcal{T},\forall j \tag{27}$$

$$k\log\mathbb{E}\left[\exp\left(\frac{p_j^{t,s}-y_j^{t-1,s-1}}{k}\right)\right]\leq 0 \qquad \forall t\in[T],s\in[T],\forall j \tag{28}$$

The constraints we apply on the model are similar as before, where (25) and (26) are capacity constraints on the servers, (27) are the queue waiting time constraints, and (28) are the push constraints. Specifically, (25) indicates that capacity is shared between first and return consultations, as the same server (doctor) is used for them. We leave waiting time constraints separate in (27), however. This imposes different priorities among first and return consultation services.

Because of Proposition 6a, 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.

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 $\boldsymbol{p}_1,\boldsymbol{p}_2$:*

$$k\log\mathbb{E}\left[\exp\left(\sum_{s=0}^{t}r(s)y_2^{t,s}/k\right)\right]=\sum_{\bar{\tau}=0}^{t-1}\sum_{\tau=0}^{t-\bar{\tau}-1}p_1^{t-\bar{\tau}-1,\tau}\eta_2^{t,\bar{\tau}}-\sum_{s=0}^{t}r(s)\sum_{\tau=0}^{s-1}p_2^{t-\tau,s-\tau}, \tag{29}$$

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.

Proof of Proposition 7. In the proof of Proposition 4, we have argued that for any time t, the state variables $y_1^{t,s}$ are independent for $s\in\mathcal{T}$. However, it is not true for $y_2^{t,s}$ for $s\in\mathcal{T}$; hence, we cannot use the same technique in Proposition 4. Nevertheless, we will show that $\sum_{s=0}^{t}r(s)y_2^{t,s}$ can still be evaluated efficiently. By Proposition 6,

$$y_2^{t,0}\sim\sum_{s=0}^{t-1}\mathrm{Bin}\left(z_1^{t-s-1,0},\bar{h}_1^{t-1,s}\right).$$

Then, for any time t, the distribution of $\sum_{s=0}^{t}y_2^{t-s,0}$ can be written explicitly as

$$\sum_{s=0}^{t} y_2^{t-s,0} \sim \sum_{s=0}^{t} \sum_{\tau=0}^{t-s-1} \mathrm{Bin}(z_1^{t-s-\tau-1,0}, \bar{h}_1^{t-s-1,\tau}). \tag{30}$$

We now represent the binomial random variable as a sum of Bernoulli random variables. Let $L_\ell^{t,s} \sim \mathrm{Bernoulli}(\bar{h}_1^{t,s})$ indicating whether a patient ℓ who has been in service for s periods at time t will be routed to X-ray at time $t+1$. Then, we rewrite (30) as

$$\sum_{s=0}^{t} y_2^{t-s,0} = \sum_{s=0}^{t} \sum_{\tau=0}^{t-s-1} \sum_{\ell=1}^{z_1^{t-s-\tau-1,0}} L_\ell^{t-s-1,\tau} \tag{31}$$

$$= \sum_{\bar{\tau}=0}^{t-1} \sum_{\ell_{\bar{\tau}}=1}^{z^{t-\bar{\tau}-1,0}} \left(\sum_{s=0}^{\bar{\tau}} L_{\ell_{\bar{\tau}}}^{t-s-1,\bar{\tau}-s} \right), \tag{32}$$

via a change in the order of summation and letting $\bar{\tau} = s + \tau$. In equation (32), the most inner summations $\sum_{s=0}^{\bar{\tau}} L_{\ell_{\bar{\tau}}}^{t-s-1,\bar{\tau}-s}$ are independent for all $\bar{\tau}$ and $\ell_{\bar{\tau}}$, because they correspond to some random events of different patients. As such, the following reformulation is valid,

$$\begin{aligned}
& k \log \mathbb{E}\left[\exp\left(\sum_{s=0}^{t} r(s) y_2^{t,s}/k\right)\right] \\
=& k \log \mathbb{E}\left[\exp\left(\sum_{s=0}^{t} r(s) y_2^{t-s,0}/k\right)\right] - \sum_{s=0}^{t} r(s) \sum_{\tau=0}^{s-1} p_2^{t-\tau,s-\tau} \\
=& \sum_{\bar{\tau}=0}^{t-1} \sum_{\ell_{\bar{\tau}}=1}^{z^{t-\bar{\tau}-1,0}} k \log \mathbb{E}\left[\exp\left(\sum_{s=0}^{\bar{\tau}} r(s) L_{\ell_{\bar{\tau}}}^{t-s-1,\bar{\tau}-s}/k\right)\right] - \sum_{s=0}^{t} r(s) \sum_{\tau=0}^{s-1} p_2^{t-\tau,s-\tau} \\
=& \sum_{\bar{\tau}=0}^{t-1} \sum_{\tau=0}^{t-\bar{\tau}-1} p_1^{t-\bar{\tau}-1,\tau} \eta_2^{t,\bar{\tau}} - \sum_{s=0}^{t} r(s) \sum_{\tau=0}^{s-1} p_2^{t-\tau,s-\tau},
\end{aligned}$$

where equality follows because $\boldsymbol{\eta}_2$ are constants and $z^{t-\bar{\tau}-1,0} = \sum_{\tau=0}^{t-\bar{\tau}-1} p_1^{t-\bar{\tau}-1,\tau}$. $\eta_2^{t,\bar{\tau}}$ are defined as

$$\eta_2^{t,\bar{\tau}} \triangleq k \log \mathbb{E}\left[\exp\left(\sum_{s=0}^{\bar{\tau}} r(s) L_{\ell_{\bar{\tau}}}^{t-s-1,\bar{\tau}-s}/k\right)\right], \ \forall t \in [T], \bar{\tau} = 0, 1, \ldots, t-1,$$

which evaluates to give the expression in the statement of the Proposition. □

Remark 4. 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)**. *Problem* (24) *can be reformulated and solved via a bisection search where each sub-problem is a mixed-integer linear optimization problem.*

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. This is equivalent as adding new blocks that represent the "traffic". The patient flow network is illustrated in Figure 8 (in Appendix A). Similar results can be derived for such a model.

Uncertain arrival times

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

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

The entropic risk constraint (11) 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 for $t \in \mathcal{T}$ 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 amount of man-hours operated beyond the operational horizon T. We let T_c be sufficiently large, 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:

$$\sum_{j=1}^{3}\sum_{t=T+1}^{T_c}\sum_{s=1}^{t} z_j^{t,s}. \tag{33}$$

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* $\boldsymbol{x}$*:*

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

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 (23). 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. The consultation times from other specializations behave similarly. For simplicity, we shall use geometric distributions as the underlying true service time distribution in our simulations. Note that our model remains tractable even if we adopted the empirical service time distributions. We observe that service rates differ by around 0.1 with the first consultation taking longer on average. In addition, return consultations have a lighter tail

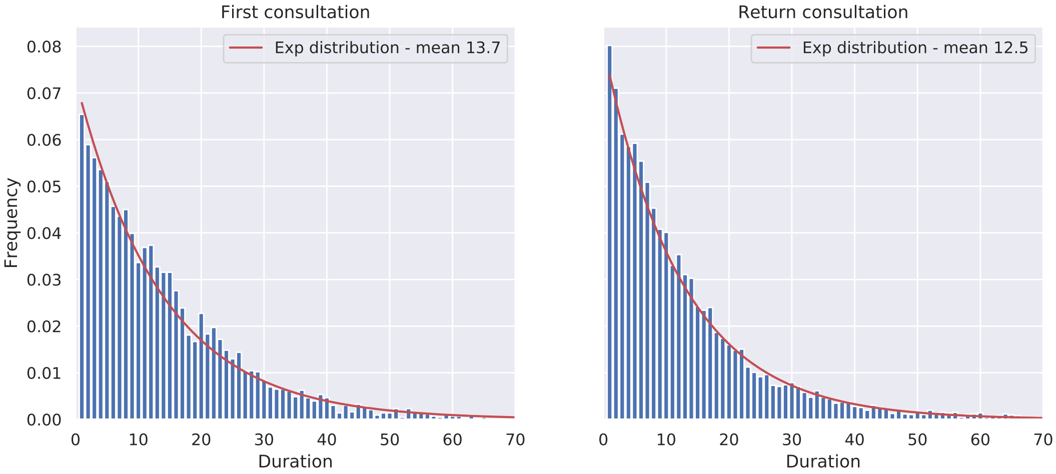

Figure 6 Empirical distribution of consultation time

than first consultations. This difference may appear small; however, we will see in simulations later that the optimal policy and performance can be significantly affected.

Patient re-entry and variability in patient behaviours are tacit 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. The histogram of arrival time difference (arrival time less scheduled time) is shown in Figure 7. Therefore, one cannot ignore these factors in making scheduling decisions. 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 be expected to vary similarly across specializations. As a final note on this, the proportion of walk-ins to our partnering clinics is very small and these data entries are not systematically recorded. As such, it is not reflected in Table 3. 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 (except specialization 1). Table 4 below illustrates the proportion of first time visit and the corresponding X-ray probability.

In these Orthopaedic clinics, the day is divided into the morning and afternoon shifts, which we can safely treat as being separate. As such, in all subsequent analysis, we will consider a time window of $T = 120$ minutes, representing one shift. Each clinic is staffed by a single doctor in that shift. The current practice in the clinics is to schedule patients in equal intervals of ten minutes

Table 3 **Summary of no-show and X-ray probability**

	Specialization				
	1	2	3	4	5
Average no-show probability	19%	23%	27%	29%	29%
Average X-ray probability	39%	24%	18%	15%	13%

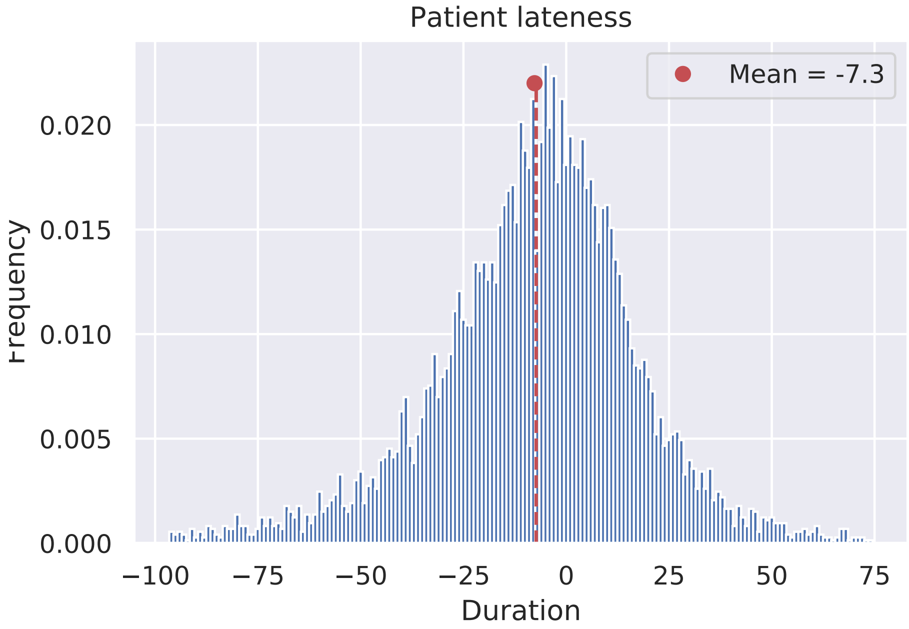

Figure 7 **Empirical distribution of arrival time difference**

Table 4 **Proportion of first time visit and its X-ray probability**

	Specialization				
	1	2	3	4	5
First time visit	26.7%	31.5%	34.0%	34.5%	41%
X-ray probability	23%	41%	25%	23%	15%

over these 120 minutes, filling up the day from the first to the twelfth slot, as long as there is a backlog of patients to be scheduled. 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 degrade the solution by much (Stein and Cote 1994).

4.2. Simulation Experiments

In this section, we hope to examine the performance of our model (23) and to understand the consequences of planning without considering patient re-entries, or variability in behaviours. To do so, we will compare the model against the current equal-interval scheduling policy in the clinics.

We find that this is a valid comparison, since the equal-interval policy can be conceived as a policy that is blind to the operating structure and characteristics of the network. The other reason for using the current operating policy as the standard of comparison is because we are unable to find any model in the existing literature that can handle all these various aspects of patient re-entry and variability in behaviours.

To perform the comparison, we shall run the optimal policy against the equal-interval policy under a simulation of a $1,000$ experiments, independently and identically generated according to the information we derived from data. Under each experiment, we implement the two scheduling policies and use the same routing policy for both, which is a first-come, first-served policy. In other words, patients always arrive at the end of the queue. We compute the metrics of total waiting time, system overtime, and maximum and average instantaneous waiting time for each of the policies and then average them over these $1,000$ 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. Thus, the maximum instantaneous waiting time refers to the largest of these instantaneous waiting times amongst all times $t = 1, \dots, T$.

In the subsequent discussion, we will conduct several separate 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 *in turn* the effects of incorporating other modeling features. In particular, we will illustrate this for patient walk-ins, no-shows, and uncertain arrival times. The choice to conduct these simulations separately, despite the ability of our model to incorporate them all at once, is so that we can examine how each element affects the performance and optimal policy. In our model, as transportation times between consultation and X-ray examinations are not logged, we are unable to model it here.

In the very last analysis, we consider a case 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 likelihoods of requiring X-ray examination, as discussed in the above section. In our clinics, the current policy is the schedule all first timers first and to fill the returning patients into later slots on the schedule. The logic 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 the best policy.

Varying X-ray probability

We consider a fixed sequence of 12 patients, which the planner needs to schedule. 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 has probability of q of requiring an

X-ray examination. Otherwise, with probability $1-q$, the patient leaves the system after first consultation.

We summarize the performance of our model in Table 5, as compared against the equal-interval benchmark. Our model consistently outperforms the benchmark. The performance of our model also seem to improve as the chance of X-ray examinations q increases. This is to be expected as our model primarily seeks to balance the risks of violation each and every constraint, and a higher q indicates a higher presence of uncertainty in the system, in terms of patient flow.

Table 5 **Performance comparison: varying X-ray probability q**

		Metrics (mins)		
		Total waiting	Overtime	Max. instantaneous waiting
$q = 0.25$	Equal-interval	125.6	21.0	51.2
	Ours	120.3	20.5	48.0
	% improvement	**4.2%**	**2.4%**	**6.3%**
$q = 0.30$	Equal-interval	138.4	23.5	57.2
	Ours	131.1	22.6	52.2
	% improvement	**5.3%**	**3.8%**	**8.7%**
$q = 0.35$	Equal-interval	147.0	25.3	60.6
	Ours	137.3	24.8	54.9
	% improvement	**6.6%**	**2.0%**	**9.4%**

Variability in behaviours: walk-ins, no-shows and lateness

In the next 3 simulations, we shall build upon the earlier model and fix the X-ray probability at $q = 0.25$. First, we further consider non-stationary walk-ins. Without loss of generality, we suppose that the number of walk-ins λ_t at time t follows a Poisson distribution with rate α_t for $t \in [T]$. To make the comparison stark, we consider a situation where walk-ins cluster around a short time interval within the 120 minutes: $\alpha_t = \alpha$, for a fixed α when $t = 50, \ldots, 59$, and $\alpha_t = 0$ elsewhere. We then vary α over a range of values. We can also compute the term relating to the walk-ins in Propositions 4 and 5 via:

$$k \log \mathbb{E}\left[\exp\left(\lambda_t/k\right)\right] = k\alpha_t(\exp(1/k) - 1).$$

We summarize the model performance in Table 6. Again, in this case, our model consistently outperforms the benchmark. Similar to the previous case, model performance tends to improve with greater α, again due to the same logic that a higher α is a greater influx of uncertainty into the model.

Table 6 Performance comparison: varying walk-in rate α

		Metrics (mins)		
		Total waiting	Overtime	Max. instantaneous waiting
$\alpha = 0.05$	Equal-interval	148.0	22.6	60.8
	Ours	142.1	22.0	56.1
	% improvement	**4.6%**	**2.7%**	**7.7%**
$\alpha = 0.075$	Equal-interval	169.3	25.4	70.4
	Ours	163.3	24.7	64.9
	% improvement	**3.5%**	**2.8%**	**7.8%**
$\alpha = 0.10$	Equal-interval	190.8	27.7	81.0
	Ours	183.1	26.3	72.2
	% improvement	**4.0%**	**5.1%**	**10.9%**

Now, we suppose scheduled patients will not show up with (an independent) probability γ. We summarize the model performance as γ is varied in Table 7. As no-show rate gets higher, the improvement over benchmark policy becomes less significant. This is because the queueing system starts to enter a low-traffic regime, hence any uncertainty arising from patient re-entry can be more easily accommodated by the system.

Table 7 Performance comparison: varying no-show probability γ

		Metrics (mins)		
		Total waiting	Overtime	Max. instantaneous waiting
$\gamma = 0.10$	Equal-interval	100.7	16.0	45.9
	Ours	95.7	15.8	43.0
	% improvement	**5.0%**	**1.2%**	**6.3%**
$\gamma = 0.15$	Equal-interval	86.8	14.1	44.0
	Ours	83.6	13.7	42.0
	% improvement	**3.7%**	**2.8%**	**4.5%**
$\gamma = 0.20$	Equal-interval	73.8	12.6	44.0
	Ours	73.7	12.0	43.7
	% improvement	**0.1%**	**4.8%**	**0.7%**

In the third simulation, we suppose scheduled patients will arrive earlier or later than their scheduled time by some margin. As we have seen earlier in our data, patient lateness is a significant issue. In this simulation, we assume that patients will arrive earlier or later than their scheduled time according to the empirical distribution (Figure 7). We, however, truncate the distribution to the window of support $[-D, D]$ in order to understand the effect of greater uncertainty on the scheduling policy.

We summarize the model performance as D is varied in Table 8. As we can see, our model provides significant improvement over benchmark policy. The average queue length is reduced by as much as 15.5%. In fact, the improvement is monotone in D. This once again validates that accounting for greater uncertainty in the model arising from patient variability in behaviours will invariably lead to policies with better performance. Notice that the patients arrive early on average in our data. We expect a larger improvement on overtime when the arrival time difference shift more towards the right.

Table 8 Performance comparison: varying arrival time difference distribution support D

		Metrics (mins)		
		Total waiting	Overtime	Max. instantaneous waiting
$D=10$	Equal-interval	141.3	21.8	83.3
	Ours	136.2	20.8	76.0
	% improvement	**3.6%**	**4.6%**	**8.8%**
$D=20$	Equal-interval	143.3	22.1	80.0
	Ours	136.6	20.6	70.2
	% improvement	**4.7%**	**6.8%**	**12.3%**
$D=30$	Equal-interval	147.0	21.6	76.6
	Ours	136.2	19.8	66.2
	% improvement	**7.3%**	**8.3%**	**13.6%**

Distinct patient classes

Finally, we study the situation of scheduling patients of two types, where one has a much higher likelihood of requiring X-ray examinations than the other. As discussed in the prelude, this is a natural setting in our case.

Suppose there are two types of patients. Type A patients are more likely (with probability q) to go through an X-ray examination while Type B patients do not need to go through one. In this simulation, we assume that there are four Type A patients and eight Type B patients. We will attempt to determine a policy that decides on both the appointment start times and the sequence of service (*i.e.*, sequence of Type A and B patients to schedule). As mentioned, the benchmark policy schedules Type A patients ahead of all Type B patients, and in equal intervals. In other words, 4 Type A patients will be scheduled at $t = 0, 10, 20, 30$, followed by Type B patients every 10 minutes.

In this simulation, we vary the chance of requiring X-ray examinations on the Type A patients. The results and their comparison against the benchmark policy are given in Table 9. We achieve consistent improvement over the benchmark. Specifically, this improvement arises from the different sequence in how our model schedules the patients of both types. Table 10 illustrates this for $q = 0.7$. On examination, our model schedules two types of patients alternately when q is high. This distributes the Type A patients evenly across the shift and hence reduces the chance that there would be snowballing of waiting times as a result of having too many Type A requiring X-ray examinations at the same time.

Table 9 **Performance comparison: varying type-A patient X-ray probability q**

		Metrics (mins)		
		Total waiting	Overtime	Max. instantaneous waiting
$q = 0.5$	Equal-interval	187.5	25.1	79.3
	Ours	169.6	23.0	67.0
	% improvement	**9.5%**	**8.4%**	**15.5%**
$q = 0.6$	Equal-interval	195.5	25.3	82.0
	Ours	177.6	23.5	71.0
	% improvement	**9.2%**	**7.1%**	**13.4%**
$q = 0.7$	Equal-interval	202.6	25.9	84.5
	Ours	187.5	25.5	72.9
	% improvement	**7.5%**	**1.5%**	**13.7%**

Table 10 **Scheduling policy when $q = 0.7$**

	Patient order											
	1	2	3	4	5	6	7	8	9	10	11	12
Appointment time (our model)	0	4	12	21	32	44	58	68	80	87	98	106
Patient sequence (our model)	A	B	A	B	B	B	B	A	B	A	B	B
Appointment time (benchmark)	0	10	20	30	40	50	60	70	80	90	100	110
Patient sequence (benchmark)	A	A	A	A	B	B	B	B	B	B	B	B

5. Concluding Remarks and Insights

We have considered the intraday scheduling problem with patient re-entry, and also incorporating uncertain elements such as no-shows, walk-ins, and lateness. Our model remains tractable, and simulations illustrate that the model is able to improve existing policies significantly in all metrics.

Our scheduling policy tends to accumulate patients towards the beginning and end of the planning horizon (*e.g.,* in Table 10). In other words, optimal inter-arrival times are longer in the middle of the planning horizon and shorter on the two ends. The intuition is that queues are more likely to build up in the middle of the planning horizon; therefore, a longer inter-arrival time balances this out. More compact scheduling at the beginning and the end reduces server idleness while keeping an acceptable level of queue length. We summarize the key insights in Table 11.

Table 11 Brief summary of key insights

Source of uncertainty	Insights
X-ray re-entry	Higher X-ray rate creates heavier traffic and more uncertainty. Improvement gets more significant as this rate increases.
Walk-ins	Higher walk-in rate brings a higher influx of uncertainty and traffic. Improvement gets more significant as this rate increases.
No-shows	Higher no-show rate leads to lighter traffic; therefore, uncertainties can be more easily accommodated.
Uncertain arrival time	The spread of distribution of arrival time difference can significantly influence the scheduling efficiency.
Distinct patients	Scheduling two types of patients alternately smooths out re-entry. Type with higher X-ray probability should be scheduled earlier.

Furthermore, such features 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

Aumann, R.J., R. Serrano. 2008. An economic index of riskiness. *Journal of Political Economy* **116**(5).

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.

Begen, M.A., R. Levi, M. Queyranne. 2012. Technical note-a sampling-based approach to appointment scheduling. *Operations Research* **60**(3) 675–681.

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.

Brown, D.B., E.D. Giorgi, M. Sim. 2012. Aspirational preferences and their representation by risk measures. *Management Science* **58**(11) 2095–2113.

Brown, D.B., M. Sim. 2009. Satisficing measures for analysis of risky positions. *Management Science* **55**(1) 71–84.

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.

Foellmer, 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`.

Foellmer, 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.

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

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

Guerriero, F., R. Guido. 2011. Operational research in the management of the operating theatre: A survey. *Health Care Management Sci* **14**(1) 89–114.

Gupta, D. 2007. Surgical suites' operations management. *Production Oper. Management* **16**(6) 689–700.

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.

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

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.

Kim, S.H., W. Whitt, W.C. Cha. 2018. A data-driven model of an appointment-generated arrival processes at an outpatient clinic. *INFORMS Journal on Computing* **30**(1) 181–199.

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`.

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.

Molfenter, T. 2013. Reducing appointment no-shows: going from theory to practice. *Substance Use & Misuse* **48**(9) 749–765.

Nazarathy, J.Y., G. Weiss. 2010. A fluid approach to large volume job shop scheduling. *Journal of Scheduling* **13**(5) 509–529.

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.

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

Appendix A: Proofs and other Technical Results

Proof of Proposition 4: Recall from (2) that $y^{t,s} = y^{t-s,0} - \sum_{\tau=0}^{s-1} p^{t-\tau,s-\tau}$. In addition, for any time t, $y^{t-s,0} = x_{t-s} + \lambda^{t-s}$ are assumed to be independent across s. As such, we can derive (15):

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

□

Proof of Proposition 5. The reformulation follows from the fact that $p^{t,s} - y^{t-1,s-1} = -y^{t,s}$.

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

□

Proof of Proposition 9 First notice that state variables $\boldsymbol{z}_j$ are independent over $j = 1, \ldots, 3$. Therefore,

$$k \log \mathbb{E}\left[\exp\left(\sum_{j=1}^{3} \sum_{t=T+1}^{\bar{T}} \sum_{s=1}^{t} z_j^{t,s}/k\right)\right] = \sum_{j=1}^{3} k \log \mathbb{E}\left[\exp\left(\sum_{t=T+1}^{\bar{T}} \sum_{s=1}^{t} \mathrm{Bin}\left(z_j^{t-s,0}, \hat{h}_j^{t,s}\right)/k\right)\right].$$

The binomial random variables in two inner summations are dependent. We will use a change of variable to identify the cohorts of different patients. We write

$$\begin{aligned}
&\sum_{j=1}^{3} k \log \mathbb{E}\left[\exp\left(\sum_{t=T+1}^{\bar{T}} \sum_{s=1}^{t} \mathrm{Bin}\left(z_j^{t-s,0}, \hat{h}_j^{t,s}\right)/k\right)\right]\\
=&\sum_{j=1}^{3} k \log \mathbb{E}\left[\exp\left(\sum_{\bar{t}=0}^{T} \sum_{t=T+1}^{\bar{T}} \mathrm{Bin}\left(z_j^{\bar{t},0}, \hat{h}_j^{t,\bar{t}+t}\right)/k\right)\right]\\
=&\sum_{j=1}^{3} \sum_{\bar{t}=0}^{T} k \log \mathbb{E}\left[\exp\left(\sum_{t=T+1}^{\bar{T}} \mathrm{Bin}\left(z_j^{\bar{t},0}, \hat{h}_j^{t,\bar{t}+t}\right)/k\right)\right],
\end{aligned}$$

where the last equation follows from the independence of patients in different cohorts. The most inner summation is a summation of dependent random variables (*e.g.,* this summation cannot be greater than $z_j^{\bar{t},0}$). Then, similar to the proof of Proposition 7, we can write

$$\sum_{t=T+1}^{\bar{T}} \text{Bin}\left(z_j^{\bar{t},0}, \hat{h}_j^{t,\bar{t}+t}\right) = \sum_{\ell_{\bar{t}}=1}^{z_j^{\bar{t},0}} \sum_{t=T+1}^{\bar{T}} \mathbb{1}(\text{patient } \ell_{\bar{t}} \text{ stays until time } t).$$

Then, by the independence of patients

$$\begin{aligned}
&\sum_{j=1}^{3} \sum_{\bar{t}=0}^{T} k \log \mathbb{E}\left[\exp\left(\sum_{t=T+1}^{\bar{T}} \text{Bin}\left(z_j^{\bar{t},0}, \hat{h}_j^{t,\bar{t}+t}\right)/k\right)\right] \\
=&\sum_{j=1}^{3} \sum_{\bar{t}=0}^{T} \sum_{\ell_{\bar{t}}=1}^{z_j^{\bar{t},0}} k \log \mathbb{E}\left[\exp\left(\sum_{t=T+1}^{\bar{T}} \mathbb{1}(\text{patient } \ell_{\bar{t}} \text{ stays till time } t)/k\right)\right] \\
=&\sum_{j=1}^{3} \sum_{\bar{t}=0}^{T} \sum_{\tau_{\bar{t}}=1}^{\bar{t}} p_j^{\bar{t},\tau} \phi_j^{\bar{t}}
\end{aligned}$$

where we define

$$\begin{aligned}
\phi_j^{\bar{t}} \triangleq & k \log \mathbb{E}\left[\exp\left(\sum_{t=T+1}^{\bar{T}} \mathbb{1}(\text{patient } \ell_{\bar{t}} \text{ stays till time } t)/k\right)\right] \\
=& k \log\left(\sum_{t=T+1}^{\bar{T}} \hat{h}_j^{t,\bar{t}+t} \exp(1/k)\right)
\end{aligned}$$

The final reformulation is affine. □

Proposition 10. *For any $t \in [T]$, the term $k \log \mathbb{E}\left[\exp\left(\sum_{s=0}^{t} z_1^{t,s}/k\right)\right]$ can be reduced to affine constraints in push variables $\boldsymbol{p}_1$. More specifically, we have:*

$$k \log \mathbb{E}\left[\exp\left(\frac{\sum_{s=0}^{t} z_1^{t,s}}{k}\right)\right] = \sum_{\tau=0}^{t} p_1^{t,\tau} + \sum_{s=1}^{t} k\delta^{t,s} \sum_{\tau=0}^{t-s} p_1^{t-s,\tau}, \tag{34}$$

where constants

$$\delta^{t,s} \triangleq \log\left(1 - \hat{h}_1^{t,s} + \hat{h}_1^{t,s} \exp\left(\frac{1}{k}\right)\right) \quad \forall t \in [T], s \in [t]$$

can be calculated from primitive data.

Proof of Proposition 10 The result follows from Proposition 6, and the proof of Proposition 3. □

Proposition 11. *For any $t \in [T], s \in \{0, \ldots, t\}$, the term $k \log \mathbb{E}\left[\exp\left(y_2^{t,s}/k\right)\right]$ can be reduced to an affine function in decision variables $\boldsymbol{p}_1, \boldsymbol{p}_2$. More specifically, we have:*

$$k \log \mathbb{E}\left[\exp\left(y_2^{t,s}/k\right)\right] = \sum_{\tau=0}^{t-s-1} \delta_2^{t-s-1,\tau} \left(\sum_{\tau_2=0}^{t-s-1-\tau} p_1^{t-s-1-\tau,\tau_2} \right) - \sum_{\tau=0}^{s-1} p_2^{t-\tau,s-\tau} \tag{35}$$

where $\delta_2^{t,s}$ are constants that can be calculated from data for all $t \in [T], s = 0, \ldots, t$:

$$\delta_2^{t,s} \triangleq k \log\left(1 - \bar{h}_1^{t,s} + \bar{h}_1^{t,s} \exp(1/k)\right).$$

Proof of Proposition 11. By Proposition 6, we have

$$y_2^{t,0} \sim \sum_{s=0}^{t-1} \operatorname{Bin}\left(z_1^{t-s-1,0}, \bar{h}_1^{t-1,s}\right),$$

where $\bar{h}_1^{t-1,s} \triangleq q h_1^{t-1,s} \prod_{\tau=1}^{s}(1 - h_1^{t-1-\tau,s-\tau})$ is as defined in Proposition 6. For simplicity, we define $\bar{h}_1^{t,0} \triangleq q h_1^{t-1,0}$ for all $t \in \mathcal{T}$. Recall that for any fixed time t, the state variables $z_!^{t,s}$ are independent for $s \in \mathcal{T}$. Then,

$$\begin{aligned}
& k \log \mathbb{E}\left[\exp\left(y_2^{t,s}/k\right)\right] \\
=& k \log \mathbb{E}\left[\exp\left(y_2^{t-s,0}/k\right)\right] - \sum_{\tau=0}^{s-1} p_2^{t-\tau,s-\tau} \\
=& \sum_{\tau=0}^{t-s-1} k \log \mathbb{E}\left[\exp\left(\operatorname{Bin}\left(z^{t-s-1-\tau,0}, \bar{h}_1^{t-s-1,\tau}\right)/k\right)\right] - \sum_{\tau=0}^{s-1} p_2^{t-\tau,s-\tau} \\
=& \sum_{\tau=0}^{t-s-1} k z^{t-s-1-\tau,0} \log\left(1 - \bar{h}_1^{t-s-1,\tau} + \bar{h}_1^{t-s-1,\tau} \exp(1/k)\right) - \sum_{\tau=0}^{s-1} p_2^{t-\tau,s-\tau} \\
=& \sum_{\tau=0}^{t-s-1} \delta_2^{t-s-1,\tau} \left(\sum_{\tau_2=0}^{t-s-1-\tau} p_1^{t-s-1-\tau,\tau_2} \right) - \sum_{\tau=0}^{s-1} p_2^{t-\tau,s-\tau},
\end{aligned}$$

where we define the constants $\delta_2^{t,s,\tau}$ for all $t \in [T], s = 0, \ldots, t$ as:

$$\delta_2^{t,s} \triangleq k \log\left(1 - \bar{h}_1^{t,s} + \bar{h}_1^{t,s} \exp(1/k)\right).$$

The last expression is the final reformulation, and it is affine. □

Proposition 12. *For any $t \in [T]$, the term $k \log \mathbb{E}\left[\exp\left(\sum_{s=0}^{t} z_2^{t,s}/k\right)\right]$ can be reduced to an affine function in decision variables $\boldsymbol{p}_x$. More specifically, we have:*

$$k \log \mathbb{E}\left[\exp\left(\sum_{s=0}^{t} z_2^{t,s}/k\right)\right] = \sum_{s=0}^{t} p_2^{t,s} + \sum_{s=1}^{t} k \beta_2^{t,s} \sum_{\tau=0}^{t-s} p_2^{t-s,\tau}, \tag{36}$$

where $\beta_2^{t,s}$ are constants that can be calculated from data for all $t \in [T], s \in [t]$:

$$\beta_2^{t,s} \triangleq \log\left(1 - \hat{h}_2^{t,s} + \hat{h}_2^{t,s} \exp\left(\frac{1}{k}\right)\right).$$

Proof of Proposition 12. The result follows from the proof of Proposition 3. □

Proposition 13. *For any $t \in [T], s \in \{0,\dots,t\}$, the term $k\log\mathbb{E}\left[\exp\left(y_3^{t,s}/k\right)\right]$ can be reduced to an affine function in decision variables $\boldsymbol{p}_2, \boldsymbol{p}_3$. More specifically, we have:*

$$k\log\mathbb{E}\left[\exp\left(y_3^{t,s}/k\right)\right] = \sum_{\tau=0}^{t-s-1} \delta_3^{t-s-1,\tau}\left(\sum_{\tau_2=0}^{t-s-1-\tau} p_2^{t-s-1-\tau,\tau_2}\right) - \sum_{\tau=0}^{s-1} p_3^{t-\tau,s-\tau}, \tag{37}$$

where $\delta_3^{t,s}$ are constants for all $t \in [T], s = 0,\dots,t$:

$$\delta_3^{t,s} \triangleq k\log\left(1 - h_2^{t,s}\hat{h}_2^{t,s} + h_2^{t,s}\hat{h}_2^{t,s}\exp(1/k)\right).$$

Proof of Proposition 13. This follows from the proof of Proposition 11.

By Proposition 6, we have

$$y_3^{t,0} \sim \sum_{s=0}^{t-1} \mathrm{Bin}(z_2^{t-s-1,0}, h_2^{t-1,s}\hat{h}_2^{t-1,s}).$$

For simplicity, we define $\hat{h}_2^{t,0} \triangleq 1$ for all $t \in \mathcal{T}$. Notice that for any fixed time t, the state variables $z_2^{t,s}$ are independent for $s \in \mathcal{T}$. Then,

$$\begin{aligned}
&k\log\mathbb{E}\left[\exp\left(y_3^{t,s}/k\right)\right] \\
=&k\log\mathbb{E}\left[\exp\left(y_3^{t-s,0}/k\right)\right] - \sum_{\tau=0}^{s-1} p_3^{t-\tau,s-\tau} \\
=&\sum_{\tau=0}^{t-s-1} k\log\mathbb{E}\left[\exp\left(\mathrm{Bin}\left(z_2^{t-s-1-\tau,0}, h_2^{t-s-1,\tau}\hat{h}_2^{t-s-1,\tau}\right)/k\right)\right] - \sum_{\tau=0}^{s-1} p_3^{t-\tau,s-\tau} \\
=&\sum_{\tau=0}^{t-s-1} \delta_3^{t-s-1,\tau}\left(\sum_{\tau_2=0}^{t-s-1-\tau} p_2^{t-s-1-\tau,\tau_2}\right) - \sum_{\tau=0}^{s-1} p_3^{t-\tau,s-\tau},
\end{aligned}$$

where we define constants for all $t \in [T], s = 0,\dots,t$ as:

$$\delta_3^{t,s} \triangleq k\log\left(1 - h_2^{t,s}\hat{h}_2^{t,s} + h_2^{t,s}\hat{h}_2^{t,s}\exp(1/k)\right).$$

The last expression is the final reformulation, and it is affine. □

Proposition 14. *For any $t \in [T]$, the term $k\log\mathbb{E}\left[\exp\left(\sum_{s=0}^{t} r(s) y_3^{t,s}/k\right)\right]$ can be reduced to an affine function in decision variables $\boldsymbol{p}, \boldsymbol{p}_2$. More specifically, we have:*

$$k\log\mathbb{E}\left[\exp\left(\sum_{s=0}^{t} r(s) y_3^{t,s}/k\right)\right] = \sum_{\bar{\tau}=0}^{t-1}\sum_{\tau=0}^{t-\bar{\tau}-1} p_2^{t-\bar{\tau}-1,\tau}\eta_3^{t,\bar{\tau}} - \sum_{s=0}^{t} r(s)\sum_{\tau=0}^{s-1} p_3^{t-\tau,s-\tau} \tag{38}$$

where $\eta_3^{t,s}$ are constants that can be calculated from data:

$$\eta_3^{t,\bar{\tau}} \triangleq k\log\left(1 + \sum_{s=0}^{\bar{\tau}} h_2^{t-s-1,\bar{\tau}-s}\hat{h}_2^{t-s-1,\bar{\tau}-s}\left(\exp\left(\frac{r(s)}{k}\right) - 1\right)\right), \; \forall t \in [T], \bar{\tau} = 0,1,\dots,t-1.$$

Proof of Proposition 14. This follows directly from the proof of Proposition 7.

For any time t, the distribution of $\sum_{s=0}^{t} y_3^{t-s,0}$ can be written as

$$\sum_{s=0}^{t} y_3^{t-s,0} \sim \sum_{s=0}^{t}\sum_{\tau=0}^{t-s-1} \mathrm{Bin}(z_2^{t-s-\tau-1,0}, h_2^{t-s-1,\tau}\hat{h}_2^{t-s-1,\tau}). \tag{39}$$

We define the random variable $M_\ell^{t,s} \sim \mathrm{Bernoulli}(h_2^{t,s}\hat{h}_2^{t,s})$, which indicates whether a patient ℓ who has been in X-ray service for s periods at time t will be routed back to the consultation doctor at time $t+1$. Then, we rewrite (39) as:

$$\sum_{s=0}^{t} y_3^{t-s,0} = \sum_{\bar{\tau}=0}^{t-1}\sum_{\ell_{\bar{\tau}}=1}^{z_2^{t-\bar{\tau}-1,0}} \left(\sum_{s=0}^{\bar{\tau}} M_{\ell_{\bar{\tau}}}^{t-s-1,\bar{\tau}-s}\right). \tag{40}$$

In equation (40), the most inner summations $\sum_{s=0}^{\bar{\tau}} M_{\ell_{\bar{\tau}}}^{t-s-1,\bar{\tau}-s}$ are independent for all $\bar{\tau}$ and $\ell_{\bar{\tau}}$, because they correspond to some random events of different patients and different patients are clearly independent. Then,

$$\begin{aligned}
&k\log\mathbb{E}\left[\exp\left(\sum_{s=0}^{t} r(s)y_3^{t,s}/k\right)\right]\\
=&k\log\mathbb{E}\left[\exp\left(\sum_{s=0}^{t} r(s)y_3^{t-s,0}/k\right)\right] - \sum_{s=0}^{t} r(s)\sum_{\tau=0}^{s-1} p_3^{t-\tau,s-\tau}\\
=&\sum_{\bar{\tau}=0}^{t-1}\sum_{\ell_{\bar{\tau}}=1}^{z_2^{t-\bar{\tau}-1,0}} k\log\mathbb{E}\left[\exp\left(\sum_{s=0}^{\bar{\tau}} r(s)M_{\ell_{\bar{\tau}}}^{t-s-1,\bar{\tau}-s}/k\right)\right] - \sum_{s=0}^{t} r(s)\sum_{\tau=0}^{s-1} p_3^{t-\tau,s-\tau}\\
=&\sum_{\bar{\tau}=0}^{t-1}\sum_{\tau=0}^{t-\bar{\tau}-1} p_2^{t-\bar{\tau}-1,\tau}\eta_3^{t,\bar{\tau}} - \sum_{s=0}^{t} r(s)\sum_{\tau=0}^{s-1} p_3^{t-\tau,s-\tau}
\end{aligned}$$

where we define the constants $\eta_3^{t,\bar{\tau}}$ as

$$\eta_3^{t,\bar{\tau}} \triangleq k\log\mathbb{E}\left[\exp\left(\sum_{s=0}^{\bar{\tau}} r(s)M_{\ell_{\bar{\tau}}}^{t-s-1,\bar{\tau}-s}/k\right)\right],\ \forall t\in[T], \bar{\tau}=0,1,\ldots,t-1,$$

which can be easily calculated and its final expression is as stated in the Proposition. The final reformulation is affine. □

Proposition 15. *For any $t\in\mathcal{T}$, the term $k\log\mathbb{E}\left[\exp\left(\sum_{s=0}^{t} z_3^{t,s}/k\right)\right]$ can be reduced to an affine function in decision variables $\boldsymbol{p}_3$. More specifically, we have:*

$$k\log\mathbb{E}\left[\exp\left(\sum_{s=0}^{t} z_3^{t,s}/k\right)\right] = \sum_{s=0}^{t} p_3^{t,s} + \sum_{s=1}^{t} k\beta_3^{t,s}\sum_{\tau=0}^{t-s} p_3^{t-s,\tau}, \tag{41}$$

where $\beta_3^{t,s}$ are constants that can be calculated from data for $t\in[T], s\in[t]$:

$$\beta_3^{t,s} \triangleq \log\left(1-\hat{h}_3^{t,s}+\hat{h}_3^{t,s}\exp\left(\frac{1}{k}\right)\right).$$

Proof of Proposition 15. The proof follows directly from the proof of Proposition 3. □

Omitted Schematics

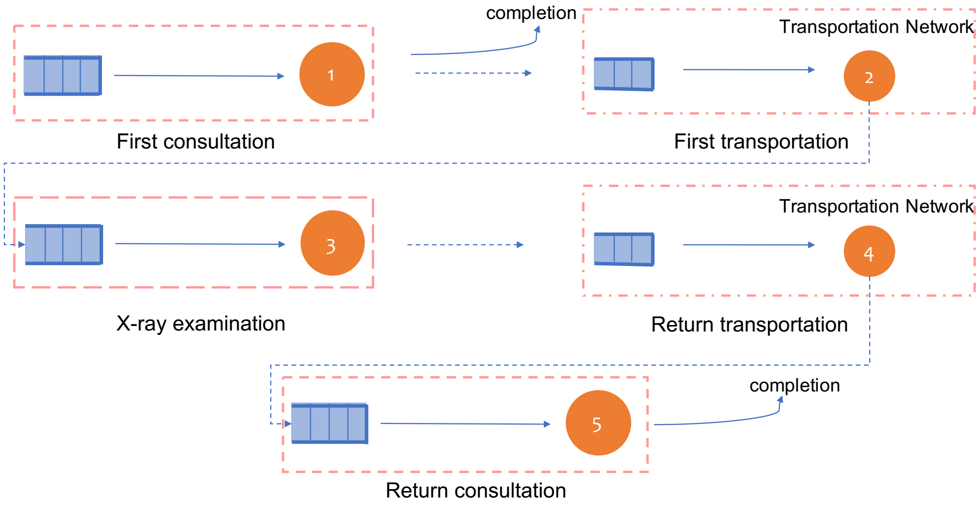

Figure 8 Patient Flow Network with Transportation