Submitted to

manuscript

Data-Driven Patient Scheduling in Emergency Departments: A Hybrid Robust–Stochastic Approach

Shuangchi He
Department of Industrial and Systems Engineering, National University of Singapore, Singapore 117576
heshuangchi@nus.edu.sg

Melvyn Sim
Department of Decision Sciences, NUS Business School, National University of Singapore, Singapore 119245
melvynsim@nus.edu.sg

Meilin Zhang
Global Asian Institute, National University of Singapore, Singapore 119077
214101@gmail.com

Emergency care necessitates adequate and timely treatment, which has unfortunately been compromised by crowding in many emergency departments (EDs). To address this issue, we study patient scheduling in EDs so that mandatory targets imposed on each patient's door-to-provider time and length of stay can be collectively met with the largest probability. Exploiting patient flow data from the ED, we propose a hybrid robust–stochastic approach to formulating the patient scheduling problem, which allows for practical features such as a time-varying patient arrival process, general consultation time distributions, and multiple heterogeneous physicians. In contrast to the conventional formulation of maximizing the joint probability of target attainment, which is computationally excruciating, the hybrid approach provides a computationally amiable formulation that yields satisfactory solutions to the patient scheduling problem. This formulation enables us to develop a dynamic scheduling algorithm for making recommendations about the next patient to be seen by each available physician. In numerical experiments, the proposed hybrid approach outperforms both the sample average approximation method and an asymptotically optimal scheduling policy.

Key words: healthcare operations, patient scheduling, robust optimization, stochastic programming, mixed integer programming, queueing network

History:

1. Introduction

Emergency department (ED) crowding and the consequential delays have been a worldwide issue and received considerable attention from governments, public media, and academic communities. ED crowding compromises the quality of and access to emergency care, putting patients at great risk of treatment errors. Numerous studies have revealed an association between crowding and increased morbidity and mortality in EDs (McHugh 2013). For hospitals, ED crowding damages their public reputation and incurs revenue loss due to ambulance diversion and patients leaving without being

seen. As pointed out by Rabin et al. (2012), widespread crowding also impedes hospitals' ability to achieve national safety and quality goals, compromises the healthcare system, and limits the regional capacity for disaster response. In some countries, the operational performance of hospitals' emergency care is closely monitored by government agencies, and some key indicators are made public on a regular basis. For example, the Centers for Medicare and Medicaid Services (CMS) in the United States publishes quality measures of timely emergency care for over 4,000 hospitals on their Hospital Compare website; some of these measures are included in the pay-for-performance program of CMS. To address the crowding issue, governments and regulatory organizations may set mandatory targets for emergency care. In 2005, England's National Health Service mandated that 98% of ED patients must be treated and either discharged home or admitted to an inpatient ward within four hours of arrival. The implementation of this "four-hour rule" greatly improved the percentage of patients spending less than four hours in EDs, from 77.3% in 2002–2003 to 97.2% in 2008–2009 (Weber et al. 2011).

Hoot and Aronsky (2008) summarized the common causes of ED crowding, including an increasing demand for emergency care, insufficient hospital bed capacity, operational inefficiencies, etc. Effective patient flow management is expected to be the solution to excessive patient delays without direct capacity expansion. To evaluate the timeliness and efficiency of emergency care, the National Quality Forum has endorsed length of stay, door-to-provider time (i.e., the time a patient spends in the ED before being seen by a healthcare provider), and leaving without being seen as quality metrics (Welch et al. 2011). Since the percentage of leaving without being seen is closely related to patients' door-to-provider times, we regard the two time metrics as major performance concerns. In general, door-to-provider times should be kept below certain safety limits according to each patient's clinical urgency. The widely used Emergency Severity Index (ESI), for example, categorizes ED patients into five groups based on their acuity levels and required medical resources; the recommended door-to-provider time targets range from "immediately" for resuscitation patients to "within one to two hours" for less urgent patients (Gilboy et al. 2011). Both clinical and operational requirements impose strict time constraints on patient flow management.

The focus of this paper is patient scheduling in EDs. From a modeling perspective, an ED can be viewed as a queueing network with medical units being the nodes, patients being the customers, and beds, medical staff, and equipments being the servers (Armony et al. 2015). Aside from prioritized customers and time-sensitive service requirements, this network is characterized by frequent returning routes of customers, i.e., after their initial consultations by a physician, most patients would undergo medical tests and return to the same physician before eventually being discharged or hospitalized. Although emergency physicians are required to provide treatment for a broad spectrum of illnesses and injuries, their expertise and work rates differ from one another. In

other words, the servers of this network are heterogeneous. When there are multiple patients waiting to be seen, their respective physicians and the sequence of their consultations must be carefully scheduled in order to meet the stringent door-to-provider and length-of-stay targets. However, the aforementioned features, including the complex network structure, server heterogeneity, highly uncertain patient arrival processes, and time-sensitive service requirements, all pose challenges in solving the patient scheduling problem.

To address this problem in a practical setting, we propose a *hybrid robust–stochastic approach* to exploiting patient flow data for real-time patient scheduling. Our intention is to maximize the percentage of patients whose door-to-provider times and lengths of stay are within the mandatory targets. Since the patient arrival pattern is highly variable, we would refrain from making assumptions, such as the arrival rate and the interarrival distribution, about patient arrivals in the future. Using the data of existing patients, the dynamic scheduling algorithm will determine the next patient to be seen whenever a physician becomes available. To make timely recommendations, the scheduling algorithm must be sufficiently efficient.

One may formulate an optimization problem to obtain the schedule that maximizes the joint probability of all waiting patients meeting the delay targets; see (6) and the discussion therein. With the joint probability of target attainment being the objective, such an optimization problem was first studied by Charnes and Cooper (1963), who termed this formulation the *P-model* in their paper. The P-model, however, is not widely used in practice, in part because evaluating the joint probability demands integration in high dimensions, which is generally computationally intractable, let alone optimizing the non-convex problem.

To tackle this issue, we incorporate features from robust optimization into our formulation by considering a *family of uncertainty sets*. Associated with a given schedule, each uncertainty set in the family consists of the feasible consultation times that patients can take without violating the mandatory delay targets. Unlike conventional robust optimization formulations where uncertainty sets are fixed, the hybrid approach searches in the family for the uncertainty set that has the largest probability of all its consultation times being feasible under the associated schedule. The schedule associated with the obtained uncertainty set is the optimal solution to the hybrid formulation. For the computational reason, we restrict the family of uncertainty sets to a collection of hyperrectangles (i.e., multi-dimensional boxes). Then, under the independence assumption of consultation times, the joint probability of all waiting patients meeting the delay targets is simply the product of the marginal probabilities of each individual patient meeting his own delay target. In this case, computing the joint probability does not involve high-dimensional integration, which would greatly improve the computational efficiency of the scheduling algorithm. In numerical experiments, the

hybrid robust–stochastic approach outperforms both the sample average approximation (SAA) method and an existing asymptotically optimal policy; see §7 for more details.

The hybrid robust–stochastic approach is of both practical and methodological importance. First, although the hybrid formulation is essentially a mixed integer program (see Theorem 2), solving this problem is practically efficient and allows real-time scheduling in EDs. As a dynamic approach driven by data, it allows for practical features such as a time-varying patient arrival process, general consultation time distributions, and heterogeneous physicians. In the literature on scheduling of queueing networks, these features are generally absent from the existing network models. As a result, the existing scheduling policies may not perform as well in practice. Second, the hybrid formulation represents an alternative perspective on solving the P-model, the objective of which is to maximize the feasibility probability of a set of randomly perturbed linear constraints. Conceivably, the hybrid formulation may produce near-optimal solutions to certain P-model problems at a far lower computational expense. As illustrated by a numerical example in §7.2, our approach may provide a highly efficient alternative to the SAA method. Besides patient scheduling, similar problems may arise from other stochastic systems with time-sensitive service requirements; see §8 for more discussion.

The remainder of this paper is organized as follows. The related literature is reviewed in §2. We introduce the queueing network model for EDs in §3. In §4, we present a tractable approach to solving the patient scheduling problem, based on a hybrid robust–stochastic formulation. This hybrid formulation is translated into a mixed integer program in §5. By introducing additional delay constraints, the hybrid formulation is incorporated into a dynamic scheduling framework in §6, which enables us to solve the patient scheduling problem sequentially according to a stochastic patient arrival process. We provide a comprehensive data-based numerical study in §7, where the hybrid approach is compared with both the SAA method and an asymptotically optimal scheduling policy. The paper is concluded in §8, where some potential applications are also discussed.

Let us close this section with frequently used notation. Scalars and vectors are denoted by lowercase and bold-face letters, respectively. Calligraphic letters are used for sets, such as $\mathcal{I}$, and we use $|\mathcal{I}|$ for the cardinality of the set. Random variables and vectors are denoted with a tilde mark, such as $\tilde{s}$ and $\tilde{\boldsymbol{s}}$. We assume that all random variables and vectors are defined on a common probability space, where $\mathbb{P}(A)$ is the probability of an event A. We reserve $\mathbb{E}(\tilde{s})$ for the expectation of a random variable $\tilde{s}$.

2. Related Literature

We sketch relevant studies so as to position our work within the previous literature. Both the literature on patient flow management and the literature on optimization of queueing networks are extensive and well established. It is not our intension to be exhaustive.

For analysis and control purposes, EDs are usually modeled as queueing networks. Although most studies are simulation-based (see Connelly and Bair 2004, Sinreich and Marmor 2005, and the references therein), several simplified queueing models are used in analytical studies. For example, to determine the staffing level of physicians, the ED occupancy process is described by a time-varying Erlang-C model in Green et al. (2006) and by a time-varying Erlang-B model in de Bruin et al. (2010). With the feature that patients may return to the same physician several times, a refined Erlang-R model for the occupancy process was proposed by Yom-Tov and Mandelbaum (2014). Saghafian et al. (2012) analyzed the practice of patient streaming (i.e., separating patients based on the predictions of whether they will be discharged or hospitalized) in EDs and proposed an improved streaming scheme. Saghafian et al. (2014) proposed a new triage system based on both clinical urgency and treatment complexity, for improving patient safety and operational efficiency.

Patient scheduling in EDs was studied by Huang et al. (2015), whose work is the most relevant to ours in the literature. In their paper, the ED is modeled as a multiclass queueing network with service deadlines and feedback routes. The authors proposed a simple yet highly effective scheduling policy that is capable of striking a balance between maintaining acceptable door-to-provider times and mitigating congestion. By means of heavy-traffic analysis, they proved that under a simplified setting, their proposed scheduling policy is asymptotically optimal for reducing the total congestion cost subject to constraints on door-to-provider times. This scheduling policy serves as an important benchmark for our hybrid approach; see §7.3 for the comparison between these two approaches.

Although the aforementioned queueing models are able to represent basic operational characteristics of an ED, they may be overly simplistic and incapable of capturing some salient features. For a queueing model to be analytically tractable, one may require probabilistic assumptions such as exponential interarrival and service time distributions, stationary arrival processes, and homogeneous servers. As pointed out by Bertsimas et al. (2011), performance analysis of queueing networks is largely unsolvable without these assumptions. However, as the ED environment is complex and changes frequently, such assumptions may not be appropriate, and conceivably, the control policies obtained under these assumptions may not necessarily work well in practice. In contrast, the proposed hybrid formulation does not rely on such assumptions. We would thus expect this data-driven approach to better fit the actual ED environment.

A robust optimization approach was studied for performance analysis of queueing networks by Bertsimas et al. (2011), Bandi and Bertsimas (2012), and Bandi et al. (2015). In their papers, randomness in arrival and service times is modeled by polyhedral uncertainty sets derived from limit laws in probability theory. More specifically, the law of the iterated logarithm was considered by Bertsimas et al. (2011) and the generalized central limit theorem was considered by Bandi and Bertsimas (2012) and Bandi et al. (2015) for constructing uncertainty sets. Using this robust

optimization approach, the authors obtained upper bounds of some performance measures for the queueing networks. Although our approach is also inspired by robust optimization, it stems from a completely different perspective. As opposed to conventional robust optimization formulations where uncertainty sets are specified as fixed constraints, our approach investigates a family of uncertainty sets and searches for the schedule that "maximizes" the uncertainty set within the family. In this sense, the obtained schedule is the most "robust" solution to the patient scheduling problem.

Finding an optimal dynamic control policy is generally difficult for queueing networks with delay or throughput time constraints. Most studies in the literature focus on simple policies that can be proved optimal in some asymptotic sense; see, e.g., Doytchinov et al. (2001), Plambeck et al. (2001), Maglaras and Van Mieghem (2005), and Huang et al. (2015). These policies are based on simplistic assumptions such as a single-server station (or a station with several identical servers), stationary customer arrival processes, and heavy-traffic conditions. Their performance may be suboptimal when these assumptions are not satisfied. In addition, the control actions of these policies depend on the service time distributions only through their first moments. In order for these policies to be near-optimal, the deadlines for delay or throughput times must be on a higher order of magnitude than service times, which may not always be a reasonable assumption. In contrast, the hybrid robust–stochastic approach allows for multiple heterogeneous servers, time-varying arrival processes, and arbitrary traffic conditions. The distributional information obtained from patient flow data is fully used in constructing uncertainty sets. In other words, the hybrid approach is able to exploit the entire service time distributions, which turns out to be a considerable advantage over the previous scheduling policies. In the numerical experiments in §7.3, the hybrid approach outperforms the asymptotically optimal scheduling policy proposed by Huang et al. (2015), even though the ED is in heavy traffic.

3. The Controlled Queueing Network Model

We use a queueing network to model the ED, which is controlled by a centralized patient scheduling system in order to meet the requirements for door-to-provider times and lengths of stay. Based on the state of current patients, the scheduling system will make sequential recommendations for the next patient to be seen for each physician.

The general flow of patients goes through the ED according to the following process: Patients arrive at the ED in a stochastic and nonstationary manner. After registration, they will be triaged by a nurse and assigned to several urgency groups based on their acuity levels and other concerns. The door-to-provider times of patients in each urgency group should be kept below a prescribed safety limit, and the safety limits of the urgency groups may differ from one anther. Leaving the

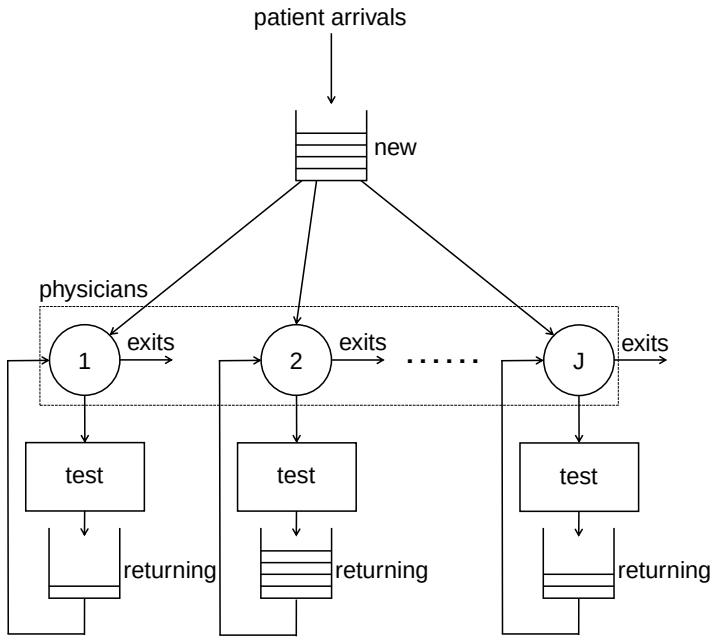

Figure 1 The queueing network model for the ED.

triage stage, patients will stay at a waiting area until they are called to be seen by a physician. These patients will be referred to as *new* patients. After the initial consultations, some patients may leave the ED, while others may be asked to undergo diagnostic tests, such as X-rays and blood tests, or to receive treatments by a nurse. When the test result is ready or the treatment is completed, the patient will return to the waiting area and become a *returning* patient, waiting to be examined by the same physician. A patient may see the same physician several times before eventually being discharged or admitted to an inpatient ward.

The scheduling system will determine the assignment of patients to each physician and the order of their consultations. A new patient can be assigned to any available physician, while a returning patient must be seen by the physician he consulted initially. We assume that the scheduling system does not manage patients who are waiting for tests or treatments by a nurse, since those patients are typically served on the first-come, first-served (FCFS) basis. A patient scheduling system is critical for the mitigation of ED crowding, because simple prioritization rules are incapable of balancing door-to-provider times and lengths of stay. If physicians give priority to new patients so as to reduce their door-to-provider times, returning patients have to spend more time waiting and form a long queue. By Little's law, the mean length of stay will be lengthened. In particular, the patients who need multiple consultations will have long total waiting times, which may create a long tail in the distribution of lengths of stay. When the ED becomes crowded, the lengths of stay of these patients will be likely to exceed the mandatory target. On the other hand, giving priority

to returning patients can effectively shorten the queue length at the waiting area, thus reducing the lengths of stay of patients. This strategy, however, will inevitably prolong the door-to-provider times of new patients, putting them at risk of treatment delays. Moreover, to maintain operational efficiency, the expertise of each physician must be considered in deciding the next patient to be seen. In general, it would be difficult to find a rule of thumb for patient scheduling under constraints on both door-to-provider times and lengths of stay.

The controlled queueing network model is depicted in Figure 1. The scheduling system makes recommendations when a physician finishes a consultation or a new patient comes to the waiting area finding at least one free physician. Let t be such an event time and consider the ED at this moment. Let $\mathcal{J}$ be the set of physicians, $\mathcal{I}^N$ the set of new patients, $\mathcal{I}^C$ the set of patients being seen, and $\mathcal{I}^R$ the set of returning patients, where the dependence on t is suppressed for notational convenience. For each $j \in \mathcal{J}$, we use $\mathcal{I}_j^C$ to denote the set of patient being seen by physician j and $\mathcal{I}_j^R$ the set of returning patients to be seen by physician j. Then,

$$\mathcal{I}^C = \bigcup_{j\in\mathcal{J}} \mathcal{I}_j^C \quad \text{and} \quad \mathcal{I}^R = \bigcup_{j\in\mathcal{J}} \mathcal{I}_j^R.$$

Moreover, $\mathcal{I}_j^C = \varnothing$ if and only if physician j is available at time t; otherwise, $\mathcal{I}_j^C$ has exactly one patient. Let $\mathcal{I}^W = \mathcal{I}^N \cup \mathcal{I}^R$ be the set of patients in the waiting area and $\mathcal{I} = \mathcal{I}^W \cup \mathcal{I}^C$ the set of patients in the ED excluding those sent to tests or treatments.

For $i \in \mathcal{I}$ and $j \in \mathcal{J}$, let $\tilde{s}_{ij}$ be the consultation time of patient i if he would be seen by physician j. For $i \in \mathcal{I}_j^C$, $\tilde{s}_{ij}$ is interpreted as the remaining consultation time of patient i, as he is being seen by physician j. We assume that $\{\tilde{s}_{ij} : i \in \mathcal{I}, j \in \mathcal{J}\}$ is a set of mutually independent random variables and use F_{ij} to denote the cumulative distribution function of $\tilde{s}_{ij}$. Since the physicians can be heterogeneous, even though $i \in \mathcal{I}$ is fixed, F_{ij} may still vary for different $j \in \mathcal{J}$. Each F_{ij} can be estimated using the historical data of physician j's consultation times, and may depend on the physician's expertise as well as the patient's status (new or returning), triage information, preliminary diagnosis, etc. In our implementation, F_{ij} is taken to be the empirical distribution function of a selected sample of consultation times. Therefore, we assume each $\tilde{s}_{ij}$ to be a discrete random variable whose values are taken from a finite set of positive numbers

$$\mathcal{S}_{ij} = \{s_{ij}(1), \ldots, s_{ij}(N_{ij})\}.$$

We use $\underline{s}_{ij}$ and $\bar{s}_{ij}$ to denote the smallest and greatest numbers in $\mathcal{S}_{ij}$. Let $\tilde{\boldsymbol{s}} = (\tilde{s}_{ij})_{i\in\mathcal{I},j\in\mathcal{J}}$ be the random vector of all these consultation times. Then, $\tilde{\boldsymbol{s}}$ takes values from the product space

$$\mathcal{S} = \prod_{i\in\mathcal{I},j\in\mathcal{J}} \mathcal{S}_{ij}.$$

The assignment of waiting patients to physicians is specified by a function $\varphi : \mathcal{I}^W \to \mathcal{J}$, where $\varphi(i)$ is the physician of patient i. Since returning patients must be seen by their initial physicians, the assignment should satisfy

$$\varphi(i) = j \quad \text{for } i \in \mathcal{I}_j^R \text{ and } j \in \mathcal{J}. \tag{1}$$

The sequencing decision for patient $i \in \mathcal{I}^W$ is specified by a correspondence $\Phi(i) : \mathcal{I}^W \to \mathcal{P}(\mathcal{I}^W)$, where $\mathcal{P}(\mathcal{I}^W)$ is the power set of $\mathcal{I}^W$ and $\Phi(i)$ is the set of patients to be seen by the same physician before patient i. The sequencing decisions for all waiting patients can thus be determined by the set of correspondences $\Phi = \{\Phi(i) : i \in \mathcal{I}^W\}$. Since $\Phi(i)$ is a set of patients to be seen by the same physician, it must satisfy

$$\varphi(k) = \varphi(i) \quad \text{for } k \in \Phi(i) \text{ and } i \in \mathcal{I}^W. \tag{2}$$

For patients to be seen by the same physician, the associated $\Phi(i)$'s form a collection of nested sets, i.e., for $i, k \in \mathcal{I}^W$ such that $\varphi(i) = \varphi(k)$, we must have

$$\Phi(i) \subset \Phi(k) \quad \text{or} \quad \Phi(k) \subset \Phi(i). \tag{3}$$

The pair of assignment and sequencing decisions (φ, Φ) is said to be an *admissible schedule* if it satisfies (1)–(3). We use $\mathcal{A}$ to denote the set of all admissible schedules. For a given schedule $(\varphi, \Phi) \in \mathcal{A}$ and a given realization of consultation times $\boldsymbol{s} = (s_{ij})_{i \in \mathcal{I}, j \in \mathcal{J}} \in \mathcal{S}$, the waiting time of patient $k \in \mathcal{I}^W$ can be obtained by

$$w_k\big(\boldsymbol{s}, (\varphi, \Phi)\big) = \sum_{\ell \in \mathcal{I}_{\varphi(k)}^C} s_{\ell\varphi(k)} + \sum_{\ell \in \Phi(k)} s_{\ell\varphi(k)}, \tag{4}$$

where the first sum on the right side is the remaining consultation time of the patient being seen by the physician (which is zero if $\mathcal{I}_{\varphi(k)}^C = \varnothing$) and the second sum is the total consultation time of the waiting patients before patient k.

We assume that each patient $i \in \mathcal{I}^W$ has a delay target τ_i. For a new patient $i \in \mathcal{I}^N$, τ_i is the duration from now until his waiting time exceeds the safety limit of his door-to-provider time. For a returning patient $i \in \mathcal{I}^R$, τ_i is specified by the scheduling system in order for his length of stay to meet the mandatory target. We will discuss how to set up delay targets for returning patients in §6. The scheduling system needs to find an admissible schedule for existing patients, under which their waiting times should not exceed the delay targets. However, since the consultation times are random, we may not be able to achieve this with complete certainty. So instead, we seek to maximize the joint probability that all patient waiting times are within the delay targets.

An *arrangement* is a function $\pi : \mathcal{S} \to \mathcal{A}$, which maps a realization of consultation times to an admissible schedule. We use $\mathcal{V}$ to denote the set of all arrangements. Under a given arrangement, we

may evaluate the joint probability that all waiting times are within the targets using (4). Therefore, the optimal arrangement can be obtained by solving the following P-model problem

$$\begin{array}{ll} \max & \mathbb{P}\big(w_i(\tilde{\boldsymbol{s}},\pi(\tilde{\boldsymbol{s}})) \le \tau_i \text{ for all } i \in \mathcal{I}^W\big) \\ \text{s.t.} & \pi \in \mathcal{V}. \end{array} \tag{5}$$

This formulation, however, is problematic because one is required to know the realization of $\tilde{\boldsymbol{s}}$ in advance to determine the admissible schedule. To fix this issue, we should confine feasible solutions to (5) within the set of *static* arrangements, i.e., the set

$$\mathcal{V}_0 = \{\pi \in \mathcal{V} : \pi(\boldsymbol{s}_1) = \pi(\boldsymbol{s}_2) \text{ for any } \boldsymbol{s}_1, \boldsymbol{s}_2 \in \mathcal{S}\}.$$

Finding the optimal static arrangement is equivalent to solving the following P-model problem

$$\begin{array}{ll} \max & \mathbb{P}\big(w_i(\tilde{\boldsymbol{s}},\mu) \le \tau_i \text{ for all } i \in \mathcal{I}^W\big) \\ \text{s.t.} & \mu \in \mathcal{A}. \end{array} \tag{6}$$

The optimal solution $\mu^\dagger = (\varphi^\dagger, \Phi^\dagger)$ to (6) specifies the assignment and sequencing decisions for all waiting patients. In particular, if there is any $i \in \mathcal{I}^W$ such that both $\Phi^\dagger(i) = \varnothing$ and $\mathcal{I}^C_{\varphi^\dagger(i)} = \varnothing$ hold, patient i will be the next patient to be seen by physician $\varphi^\dagger(i)$ and should be sent to the physician immediately. This procedure is repeated when a physician finishes a consultation or a new patient arrives at the waiting area finding at least one free physician. Each time, the scheduling system determines the next patient to be seen for available physicians.

Under a given admissible schedule, evaluating the joint probability in (6) involves multi-dimensional integration of many variables, which is computationally intractable. Nemirovski and Shapiro (2006) pointed out that computing the distribution of the sum of independent random variables is already an NP-hard problem. As a result, even finding the distribution of each patient's waiting time would be computationally prohibitive. When the number of waiting patients is large, we would be unable to obtain the optimal admissible schedule for (6) within a reasonable time that is required for dynamic patient scheduling. Therefore, we will focus on a more computationally amiable approach that has the potential for obtaining a near-optimal solution.

4. The Hybrid Robust–Stochastic Approach

Consider the function w_k given by (4) and extend its domain to $\mathbb{R}_+^{|\mathcal{I}||\mathcal{J}|} \times \mathcal{A}$. Under any $\mu \in \mathcal{A}$, the set

$$\mathcal{R}(\mu) = \big\{\boldsymbol{x} \in \mathbb{R}_+^{|\mathcal{I}||\mathcal{J}|} : w_k\big(\boldsymbol{x},\mu\big) \le \tau_k \text{ for all } k \in \mathcal{I}^W\big\}$$

is a convex polyhedron in $|\mathcal{I}||\mathcal{J}|$ dimensions. Then, we may rewrite (6) as

$$\begin{array}{ll} \max & \mathbb{P}\big(\tilde{\boldsymbol{s}} \in \mathcal{R}(\mu)\big) \\ \text{s.t.} & \mu \in \mathcal{A}, \end{array} \tag{7}$$

the optimal solution to which is the admissible schedule that maximizes the joint probability of all consultation times being within the associated convex polyhedron. Since it is difficult to evaluate $\mathbb{P}(\tilde{\boldsymbol{s}} \in \mathcal{K})$ for a general polyhedron $\mathcal{K}$ in high dimensions, it would be computationally excruciating to find the optimal solution to (7). However, if $\mathcal{K}$ happens to be hyperrectangular, e.g., $\mathcal{K} = \prod_{i\in\mathcal{I},j\in\mathcal{J}}[0,d_{ij}]$ for some $d_{ij} \geq 0$, the above joint probability can be computed by

$$\mathbb{P}(\tilde{\boldsymbol{s}} \in \mathcal{K}) = \prod_{i\in\mathcal{I},j\in\mathcal{J}} \mathbb{P}(0 \leq \tilde{s}_{ij} \leq d_{ij}) = \prod_{i\in\mathcal{I},j\in\mathcal{J}} F_{ij}(d_{ij}),$$

because the entries of $\tilde{\boldsymbol{s}}$ are mutually independent. In this case, evaluating the joint probability does not involve the tedious high-dimensional integration.

The above observation motivates us to consider an alternative formulation for the patient scheduling problem. Note that by (7), we intend to find the admissible schedule whose associated convex polyhedron has the largest probability measure induced by $\tilde{\boldsymbol{s}}$. If the probability measure of a convex polyhedron is large, we may expect the polyhedron to contain a hyperrectangular subset whose probability measure is also large. Conversely, if we can find an admissible schedule whose associated convex polyhedron contains a "large" hyperrectangle, we may also expect the polyhedron itself to be relatively "large". Hence, instead of searching for the admissible schedule that produces the "largest" convex polyhedron, we would find the admissible schedule whose associated convex polyhedron has the "largest" hyperrectangular subset. Since it is far easier to evaluate the probability measure of a hyperrectangle, the patient scheduling problem would be more computationally amiable under this formulation.

Let us consider a collection of hyperrectangular subsets of $\mathcal{S}$, given by

$$\mathcal{W} = \left\{ \mathcal{S} \cap \prod_{i\in\mathcal{I},j\in\mathcal{J}} [0,d_{ij}] : (d_{ij})_{i\in\mathcal{I},i\in\mathcal{J}} \in \mathcal{S} \right\}.$$

We would like to maximize the joint probability that the consultation times are within a set $\mathcal{Q} \in \mathcal{W}$ without exceeding the delay targets, i.e.,

$$\begin{array}{ll} \max & \mathbb{P}(\tilde{\boldsymbol{s}} \in \mathcal{Q}) \\ \text{s.t.} & w_k\big(\boldsymbol{s}, \pi(\boldsymbol{s})\big) \leq \tau_k, \quad k \in \mathcal{I}^W,\ \boldsymbol{s} \in \mathcal{Q} \\ & \mathcal{Q} \in \mathcal{W},\ \pi \in \mathcal{V} \quad . \end{array} \tag{8}$$

From a robust optimization perspective, $\mathcal{W}$ can be regarded as a family of uncertainty sets for $\tilde{\boldsymbol{s}}$, and the specific uncertainty set $\mathcal{Q}$ can be adjusted within $\mathcal{W}$ using different arrangements. The objective function in (8) involves random variables, while the constraints are based on a robust optimization formulation whose uncertainty set is adjustable. Hence, we would refer to this formulation of the patient scheduling problem as a *hybrid robust–stochastic approach*.

Comparing (8) with (5), one may raise the concern that the realization of $\tilde{\boldsymbol{s}}$ may be required in advance to determine the optimal admissible schedule. In the following theorem, we provide an equivalent form of (8), which implies that a static arrangement is optimal under the hybrid formulation.

THEOREM 1. *The hybrid optimization problem* (8) *has an equivalent form*

$$\begin{aligned} \max \ & \sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} \ln F_{ij}(d_{ij}) \\ \text{s.t.} \ & w_k(\boldsymbol{d},\mu)\le \tau_k, \quad k\in\mathcal{I}^W,\ \boldsymbol{d}=(d_{ij})_{i\in\mathcal{I},j\in\mathcal{J}} \\ & \boldsymbol{d}\in\mathcal{S},\ \mu\in\mathcal{A}. \end{aligned} \tag{9}$$

REMARK 1. Thanks to the hyperrectangular uncertainty sets, the hybrid optimization problem (8) has a computationally amiable form given by (9). With these hyperrectangular sets, computing the joint probability in (8) is reduced to the double summation in (9) without the need for high-dimensional integration. For a given uncertainty set, since the worst case occurs only when all consultation times take their largest values, we would be required to examine admissible schedules under this single scenario only. It is also worth mentioning that existing robust optimization formulations with adjustable uncertainty sets, such as the budget of uncertainty of Bertsimas and Sim (2004) and ellipsoidal uncertainty sets of Ben-Tal et al. (2004), may not lead to more tractable forms as our hybrid approach does.

Proof of Theorem 1. Let

$$\mathcal{Q}=\mathcal{S}\cap\prod_{i\in\mathcal{I},j\in\mathcal{J}}[0,d_{ij}]$$

be an element in $\mathcal{W}$, where $\boldsymbol{d}=(d_{ij})_{i\in\mathcal{I},i\in\mathcal{J}}\in\mathcal{S}$ is the boundary value. Since $\{\tilde{s}_{ij}: i\in\mathcal{I}, j\in\mathcal{J}\}$ is a set of independent random variables, maximizing $\mathbb{P}(\tilde{\boldsymbol{s}}\in\mathcal{Q})$ is equivalent to maximizing

$$\ln\mathbb{P}(\tilde{\boldsymbol{s}}\in\mathcal{Q})=\ln\prod_{i\in\mathcal{I},j\in\mathcal{J}}\mathbb{P}(\tilde{s}_{ij}\le d_{ij})=\ln\prod_{i\in\mathcal{I},j\in\mathcal{J}}F_{ij}(d_{ij})=\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}}\ln F_{ij}(d_{ij}).$$

Let $(\hat{\pi},\hat{\mathcal{Q}})$ be the optimal solution to (8) and $(\mu^\star,\boldsymbol{d}^\star)$ the optimal solution to (9), where $\hat{\pi}\in\mathcal{V}$, $\hat{\mathcal{Q}}\in\mathcal{W}$, $\mu^\star\in\mathcal{A}$, and $\boldsymbol{d}^\star=(d^\star_{ij})_{i\in\mathcal{I},i\in\mathcal{J}}\in\mathcal{S}$. If the feasible set of (8) is nonempty, we have

$$w_k\big(\boldsymbol{s},\hat{\pi}(\boldsymbol{s})\big)\le\tau_k \quad \text{for } \boldsymbol{s}\in\hat{\mathcal{Q}} \text{ and } k\in\mathcal{I}^W.$$

Let $\hat{\boldsymbol{d}}=(\hat{d}_{ij})_{i\in\mathcal{I},j\in\mathcal{J}}\in\mathcal{S}$ be the boundary value of $\hat{\mathcal{Q}}$. By taking $\hat{\mu}=\hat{\pi}(\hat{\boldsymbol{d}})$, we have

$$w_k(\hat{\boldsymbol{d}},\hat{\mu})\le\tau_k \quad \text{for } k\in\mathcal{I}^W,$$

so $(\hat{\mu},\hat{\boldsymbol{d}})$ is a feasible solution to (9). This implies that

$$\ln\mathbb{P}(\tilde{\boldsymbol{s}}\in\hat{\mathcal{Q}})=\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}}\ln F_{ij}(\hat{d}_{ij})\le\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}}\ln F_{ij}(d^\star_{ij}). \tag{10}$$

Conversely, if the feasible set of (9) is nonempty, we have

$$w_k(\boldsymbol{d}^\star, \mu^\star) \le \tau_k \quad \text{for all } k \in \mathcal{I}^W.$$

Let $\pi^\star$ be the static arrangement such that $\pi^\star(\boldsymbol{s}) = \mu^\star$ for all $\boldsymbol{s} \in \mathcal{S}$. Let $\mathcal{Q}^\star$ be the hyperrectangular subset in $\mathcal{W}$ with boundary value $\boldsymbol{d}^\star$. Since $d^\star_{ij} \ge s_{ij}$ for $\boldsymbol{s} = (s_{ij})_{i\in\mathcal{I}, i\in\mathcal{J}} \in \mathcal{Q}^\star$, it follows from (4) that

$$w_k\big(\boldsymbol{s}, \pi^\star(\boldsymbol{s})\big) \le w_k\big(\boldsymbol{d}^\star, \pi^\star(\boldsymbol{d}^\star)\big) \le \tau_k \quad \text{for all } \boldsymbol{s} \in \mathcal{Q}^\star \text{ and } k \in \mathcal{I}^W,$$

so $(\pi^\star, \mathcal{Q}^\star)$ is a feasible solution to (8). This implies that

$$\sum_{i\in\mathcal{I}} \sum_{j\in\mathcal{J}} \ln F_{ij}(d^\star_{ij}) = \ln \mathbb{P}(\tilde{\boldsymbol{s}} \in \mathcal{Q}^\star) \le \ln \mathbb{P}(\tilde{\boldsymbol{s}} \in \hat{\mathcal{Q}}). \tag{11}$$

Combining (10) and (11), we conclude that (9) is an equivalent form of (8). □

When the ED is crowded, it may happen that under any admissible schedule, there is at least one patient whose waiting time will exceed the delay target. In this case, the hybrid formulation (8) does not have a feasible solution. Since the waiting time given by (4) is increasing with each consultation time, we may determine the feasibility of (8) by examining admissible schedules when all consultation times take their smallest possible values, i.e., $\underline{\boldsymbol{s}} = (\underline{s}_{ij})_{i\in\mathcal{I}, j\in\mathcal{J}}$. To determine the feasibility of (8), we may solve the following optimization problem

$$\begin{array}{lll} \min & \alpha & \\ \text{s.t.} & w_k(\underline{\boldsymbol{s}}, \mu) \le \tau_k + \alpha, & k \in \mathcal{I}^W,\ \underline{\boldsymbol{s}} = (\underline{s}_{ij})_{i\in\mathcal{I}, j\in\mathcal{J}} \\ & \mu \in \mathcal{A} & . \end{array} \tag{12}$$

Proposition 1. *Let $\alpha^\star$ be the minimum value of α given by* (12)*. Then, the hybrid optimization problem* (8) *has a feasible solution if and only if $\alpha^\star \le 0$.*

Proof. Consider the equivalent form (9). If it has a feasible solution $(\mu, \boldsymbol{d})$, then by (4),

$$w_k(\underline{\boldsymbol{s}}, \mu) \le w_k(\boldsymbol{d}, \mu) \le \tau_k \quad \text{for } k \in \mathcal{I}^W,$$

because $\underline{s}_{ij} \le d_{ij}$ for all $i \in \mathcal{I}$ and $j \in \mathcal{J}$. Hence, $\alpha^\star \le 0$. Let $\mu^\star$ be the optimal solution to (12). Conversely, if $\alpha^\star \le 0$, we have

$$w_k(\underline{\boldsymbol{s}}, \mu^\star) \le \tau_k + \alpha^\star \le \tau_k \quad \text{for } k \in \mathcal{I}^W,$$

which implies that $(\mu^\star, \underline{\boldsymbol{s}})$ is a feasible solution to (9). □

5. A Mixed Integer Program

The equivalent form (9) can be further translated into a mixed integer program, which allows us to solve the patient scheduling problem using existing algorithms.

For $i \in \mathcal{I}$, $j \in \mathcal{J}$, and $\ell = 1, \ldots, |\mathcal{I}|$, let $x_{ij\ell}$ be the binary variable that indicates the assigned physician and consultation order of patient i, i.e.,

$$x_{ij\ell} = \begin{cases} 1 & \text{if patient } i \text{ is the } \ell\text{th patient to be seen by physician } j, \\ 0 & \text{otherwise.} \end{cases}$$

We reserve $|\mathcal{I}|$ positions for each physician so that all patients can be accommodated by any physician freely. These binary variables must jointly satisfy the following constraints: Since the first position of each queue is for the patient who is being seen or to be seen immediately by the physician, we have

$$x_{ij1} = 1 \quad \text{for } i \in \mathcal{I}_j^C \text{ and } j \in \mathcal{J}. \tag{13}$$

Furthermore, since each returning patient must be seen by their initial physicians, we have

$$\sum_{\ell=1}^{|\mathcal{I}|} x_{ij\ell} = 1 \quad \text{for } i \in \mathcal{I}_j^R \text{ and } j \in \mathcal{J}. \tag{14}$$

Under an admissible schedule, each waiting patient can be assigned to only one position, i.e.,

$$\sum_{j \in \mathcal{J}} \sum_{\ell=1}^{|\mathcal{I}|} x_{ij\ell} = 1 \quad \text{for } i \in \mathcal{I}^W, \tag{15}$$

and each position can accommodate at most one patient, i.e.,

$$\sum_{i \in \mathcal{I}_j} x_{ij\ell} \le 1 \quad \text{for } j \in \mathcal{J} \text{ and } \ell = 1, \ldots, |\mathcal{I}|, \tag{16}$$

where $\mathcal{I}_j = \mathcal{I}^N \cup \mathcal{I}_j^C \cup \mathcal{I}_j^R$ is the set of patients eligible to be seen by physician j. Beginning from the first position, we must assign patients to consecutive positions of each physician, so empty positions can appear only at the end of the queue. Since the ℓth position by physician j has a patient if and only if equality holds in (16), the above constraint is equivalent to

$$\sum_{i \in \mathcal{I}_j} x_{ij(\ell+1)} \le \sum_{i \in \mathcal{I}_j} x_{ij\ell} \quad \text{for } j \in \mathcal{J} \text{ and } \ell = 1, \ldots, |\mathcal{I}| - 1. \tag{17}$$

One can check that (13)–(17) are equivalent conditions for (1)–(3). In other words, each admissible schedule can be determined by a set of binary variables $\{x_{ij\ell} : i \in \mathcal{I}, j \in \mathcal{J}, \ell = 1, \ldots, |\mathcal{I}|\}$ that satisfies (13)–(17).

Consider the delay constraints in (9). Put

$$\bar{\tau} = \sum_{i \in \mathcal{I}, j \in \mathcal{J}} \bar{s}_{ij} \tag{18}$$

where $\bar{s}_{ij}$ is the greatest value $\tilde{s}_{ij}$ can take. Then, $\bar{\tau}$ is an upper bound of all patient waiting times. Given a set of binary variables that satisfies (13)–(17), we may write the delay constraints in (9) into

$$\sum_{\ell=1}^{m}\sum_{i\in\mathcal{I}_j} x_{ij\ell}\cdot d_{ij} \le \sum_{i\in\mathcal{I}_j^W} x_{ij(m+1)}\cdot\tau_i + \left(1-\sum_{i\in\mathcal{I}_j^W} x_{ij(m+1)}\right)\cdot\bar{\tau} \quad \text{for } j\in\mathcal{J} \text{ and } m=1,\ldots,|\mathcal{I}|-1, \tag{19}$$

where $\mathcal{I}_j^W=\mathcal{I}^N\cup\mathcal{I}_j^R$ is the set of waiting patients eligible to be seen by physician j. In this inequality, the sum on the left side is the duration physician j takes to finish the patients in the first m positions, or the waiting time until the physician begins to serve the $(m+1)$st position. If there is a patient in the $(m+1)$st position, we have

$$\sum_{i\in\mathcal{I}_j^W} x_{ij(m+1)}=1,$$

and (19) becomes

$$\sum_{\ell=1}^{m}\sum_{i\in\mathcal{I}_j} x_{ij\ell}\cdot d_{ij} \le \sum_{i\in\mathcal{I}_j^W} x_{ij(m+1)}\cdot\tau_i.$$

Since the sum on the right side is equal to the delay target, this inequality is the delay constraint for the $(m+1)$st patient. If there is no patient in the $(m+1)$st position,

$$\sum_{i\in\mathcal{I}_j^W} x_{ij(m+1)}=0$$

and the first sum on the right side of (19) becomes zero. Then, inequality (19) turns out to be

$$\sum_{\ell=1}^{m}\sum_{i\in\mathcal{I}_j} x_{ij\ell}\cdot d_{ij} \le \bar{\tau},$$

which always holds by the definition of $\bar{\tau}$. For computational convenience, let us express (19) in canonical form. By introducing a set of variables $\{u_{ij\ell}\ge 0 : i\in\mathcal{I}_j, j\in\mathcal{J}, \ell=1,\ldots,|\mathcal{I}|-1\}$, we may write (19) into two separate inequalities

$$\sum_{\ell=1}^{m}\sum_{i\in\mathcal{I}_j} u_{ij\ell} \le \sum_{i\in\mathcal{I}_j^W} x_{ij(m+1)}\cdot\tau_i + \left(1-\sum_{i\in\mathcal{I}_j^W} x_{ij(m+1)}\right)\cdot\bar{\tau} \quad \text{for } j\in\mathcal{J} \text{ and } m=1,\ldots,|\mathcal{I}|-1 \tag{20}$$

and

$$u_{ij\ell}\ge x_{ij\ell}\cdot d_{ij} \quad \text{for } i\in\mathcal{I}_j,\ j\in\mathcal{J}, \text{ and } \ell=1,\ldots,|\mathcal{I}|-1. \tag{21}$$

Since $x_{ij\ell}\in\{0,1\}$ and $d_{ij}\le\bar{s}_{ij}$, the latter inequality is equivalent to

$$u_{ij\ell}\ge d_{ij}-(1-x_{ij\ell})\cdot\bar{s}_{ij} \quad \text{for } i\in\mathcal{I}_j,\ j\in\mathcal{J}, \text{ and } \ell=1,\ldots,|\mathcal{I}|-1. \tag{22}$$

Then, inequalities (20) and (22) specify the delay constraints in canonical form.

We have obtained a set of constraints for the patient scheduling problem, given by (13)–(17), (20), and (22), which are equivalent to the constraints in (9). As suggested by the following theorem, we eventually translate the hybrid optimization problem into a mixed integer program.

THEOREM 2. *Let* $g_{ij}(n) = \ln F_{ij}(s_{ij}(n))$ *for* $i \in \mathcal{I}$, $j \in \mathcal{J}$, *and* $n = 1, \ldots, N_{ij}$. *The hybrid optimization problem* (8) *can be written as the following mixed integer program*

$$
\begin{array}{lll}
\max & \displaystyle\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}}\sum_{n=1}^{N_{ij}} y_{ij}(n)\cdot g_{ij}(n) & \\
\text{s.t.} & (13)\text{–}(17),\ (20) & \\
& \displaystyle u_{ij\ell} \geq \sum_{n=1}^{N_{ij}} y_{ij}(n)\cdot s_{ij}(n) - (1-x_{ij\ell})\cdot \bar{s}_{ij}, & i\in\mathcal{I}_j,\ j\in\mathcal{J},\ \ell = 1,\ldots,|\mathcal{I}|-1 \\
& \displaystyle\sum_{n=1}^{N_{ij}} y_{ij}(n) = 1 & , \quad i\in\mathcal{I},\ j\in\mathcal{J} \\
& x_{ij\ell}, y_{ij}(n) \in \{0,1\} & , \quad i\in\mathcal{I},\ j\in\mathcal{J},\ \ell = 1,\ldots,|\mathcal{I}|,\ n = 1,\ldots,N_{ij} \\
& u_{ij\ell} \geq 0 & , \quad i\in\mathcal{I}_j,\ j\in\mathcal{J},\ \ell = 1,\ldots,|\mathcal{I}|-1.
\end{array}
\tag{23}
$$

Proof. Since $y_{ij}(n) \in \{0,1\}$ and $\sum_{n=1}^{N_{ij}} y_{ij}(n) = 1$, there is a unique $n \in \{1,\ldots,N_{ij}\}$ such that $y_{ij}(n) = 1$ for $i \in \mathcal{I}$ and $j \in \mathcal{J}$. If $d_{ij} = s_{ij}(m)$ for some $m = 1,\ldots,N_{ij}$, by taking $y_{ij}(m) = 1$ and $y_{ij}(n) = 0$ for $n \neq m$, we have

$$d_{ij} = y_{ij}(m)\cdot s_{ij}(m) = \sum_{n=1}^{N_{ij}} y_{ij}(n)\cdot s_{ij}(n),$$

which allows us to rewrite (22) into the first inequality in (23). Since

$$\ln F(d_{ij}) = \ln F\left(\sum_{n=1}^{N_{ij}} y_{ij}(n)\cdot s_{ij}(n)\right) = \sum_{n=1}^{N_{ij}} y_{ij}(n)\cdot \ln F(s_{ij}(n)) = \sum_{n=1}^{N_{ij}} y_{ij}(n)\cdot g_{ij}(n),$$

we can also rewrite the objective function in (9) into that in (23). The rest of the proof follows from the construction of the constraints (13)–(17), (20), and (22) described above. □

6. Dynamic Scheduling Using the Hybrid Approach

Patient arrivals at the ED form a stochastic process. In order for the scheduling system to make sequential decisions accordingly, the proposed hybrid approach must be incorporated in a dynamic scheduling framework. We assume that the decision process is triggered either when a physician finishes a consultation or when a new patient comes to the waiting area finding at least one free physician. Each time, the scheduling system will recommend the next patient to be seen for the available physician.

To solve the dynamic scheduling problem by the hybrid approach, we should first determine delay targets for waiting patients when the decision process is triggered. The major concern for

new patients is their door-to-provider times, while that for returning patients is their lengths of stay. Assume that the decision process is triggered at time t. For each $i \in \mathcal{I}^W$, let a_i, D_i, and K_i be patient i's arrival time, safety limit for the door-to-provider time, and mandatory target for the length of stay, respectively. For a new patient $i \in \mathcal{I}^N$, we take the delay target as

$$\tau_i = D_i - (t - a_i),$$

which is the time until the patient's door-to-provider time reaching the safety limit. Assume that a patient can return to the same physician at most M times. To determine delay targets for returning patients, we pick a vector of M positive numbers $\boldsymbol{T}_i = (T_{i1}, \ldots, T_{iM})$ that satisfies

$$D_i < T_{i1} < \cdots < T_{iM} < K_i \quad \text{for } i \in \mathcal{I}^R,$$

where T_{im} is interpreted as a mandatory limit for the duration from patient i's arrival at the waiting area until he is seen by the physician for the $(m+1)$st time. If a returning patient $i \in \mathcal{I}^R$ is waiting to be seen for the $(m+1)$st time, we take the delay target as

$$\tau_i = T_{im} - (t - a_i).$$

Imposing additional delay constraints enables us to carry out dynamic scheduling by sequentially solving (8). With the extra delay requirements, returning patients' waiting times for individual consultations can also be maintained at a reasonable level, which may further improve patient safety and satisfaction. Clearly, the specific values of these additional mandatory limits will influence the performance of the hybrid approach in dynamic scheduling. We will discuss how to select these parameters in §7.3.

Given the delay targets for current waiting patients, the scheduling system will first determine the feasibility of the hybrid optimization problem (8) by solving (12) (which may also be converted into a mixed integer program following the procedure in §5). If the feasible set of (8) is nonempty, the scheduling system will solve the mixed integer program (23), the optimal solution to which provides the next patient to be seen for the available physician. If problem (8) turns out to be infeasible, the admissible schedule obtained by solving (12) will be used instead in our implementation. In this case, the consultation time of each waiting patient is assumed to take the minimum value. The obtained admissible schedule is the one that minimizes the longest waiting time of all waiting patients. The recommendation for the patient to be seen is made based on this admissible schedule.

7. Data-Based Validation Study

We conduct a numerical study to evaluate the performance of the hybrid robust–stochastic approach. A set of patient flow data provided by an anonymous hospital is used for validation.

This hospital adopts a four-level triage system in the ED. Levels 1 and 2 are assigned to urgent patients, who have priority over others. In this hospital, urgent patients are treated separately in a designated area with dedicated personnel and facilities. There are more than 70% of patients belonging to level 3. While their condition appears stable, these patients require timely treatment to resolve their acute symptoms. When the ED is getting crowded, this group of patients will be the most likely to suffer from prolonged waiting.[1] In fact, the crowding of level-3 patients has been the most serious problem in the ED. To address this issue, we focus on the scheduling of level-3 patients in this section. More specifically, we present some empirical findings from the data of level-3 consultation times in §7.1. The computational performance of the hybrid approach is compared with that of the SAA method in §7.2. We assess the hybrid approach for dynamic patient scheduling in §7.3, where the asymptotically optimal scheduling policy proposed by Huang et al. (2015) serves as a benchmark.

7.1. Consultation Time Categorization and Physician Heterogeneity

This set of patient flow data includes the records of around 120,000 patient visits to the ED, with over 85,000 visits made by level-3 patients. Each record contains a series of time stamps such as the start and end times of triage, consultation, and medical tests, which enable us to reconstruct the patient's entire path through the ED. Triage notes and final diagnoses can also be found from these records. In this numerical study, we divide all level-3 patients into two categories based on each one's diagnosis. The first category includes patients with the most common acute illnesses such as headache, upper respiratory tract infection, and acute gastritis, while the second category includes all other patients. According to the data set, about 40% of level-3 cases belong to the first category, and these cases are relatively simple compared with the other category. The triage nurse can easily identify cases in the first category according to patients' symptoms and vital signs, so we assume that each patient's category is known by the scheduling system when the patient arrives at the waiting area. We plot the histograms of consultation times of the two patient categories in Figure 2. The mean consultation time of the first category is 6.33 minutes and that of the second category is 6.94 minutes. In the numerical experiments, when all physicians are assumed to have the same work rates, these two empirical distributions are used in the scheduling algorithm as the distributions of unknown consultation times. Since physicians may be heterogeneous in practice,

[1] Level 4 is assigned to non-emergency patients and accounts for a negligible fraction of visits. We do not consider level-4 patients because there are no delay requirements for this group.

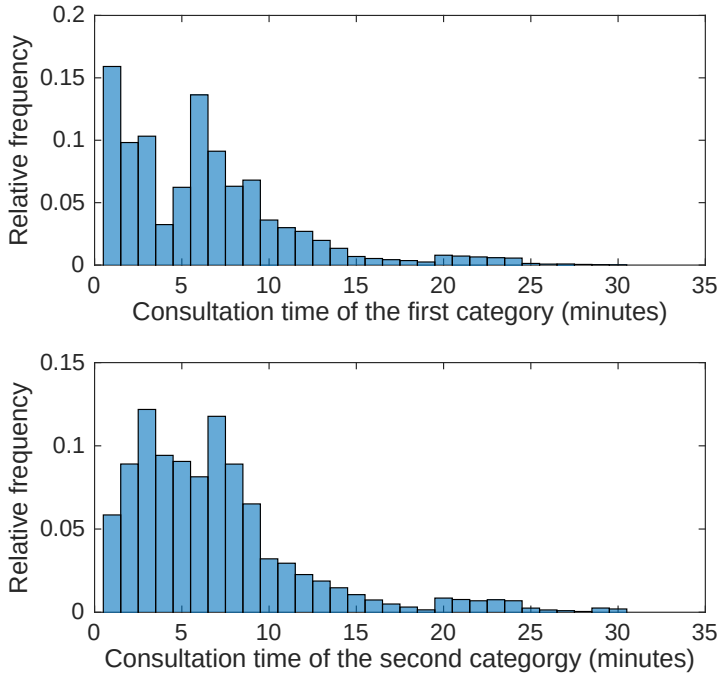

Figure 2 Histograms of consultation times of the first and second categories.

the empirical consultation time distributions of each category may differ for different physicians. In this case, the empirical distributions should be obtained using the historical consultation times by each physician.

In the ED of this hospital, there are four to six physicians working for level-3 patients in each eight-hour shift. Although emergency physicians are required to provide treatment for a wide range of illnesses and injuries, their expertise and work rates differ from one another. Among the physicians who finished more than 3,000 cases, we randomly selected five physicians and examined the records of patients seen by them. No significant differences have been found among the five patient groups. The boxplot for the consultation times by the five physicians is shown in Figure 3. We can see that both the work rates and the consultation time distributions differ a lot among these physicians. For instance, the average consultation time by physician 2 is just one half of that by physician 5, while the variability in consultation times by physician 3 is much less than that by any other physician. In order for a patient scheduling approach to be relevant to practice, the heterogeneity of physicians must be taken into account. To the best of our knowledge, the proposed hybrid robust–stochastic approach leads to the first computationally amiable scheduling algorithm that allows for heterogeneous servers in such a complex queueing network.

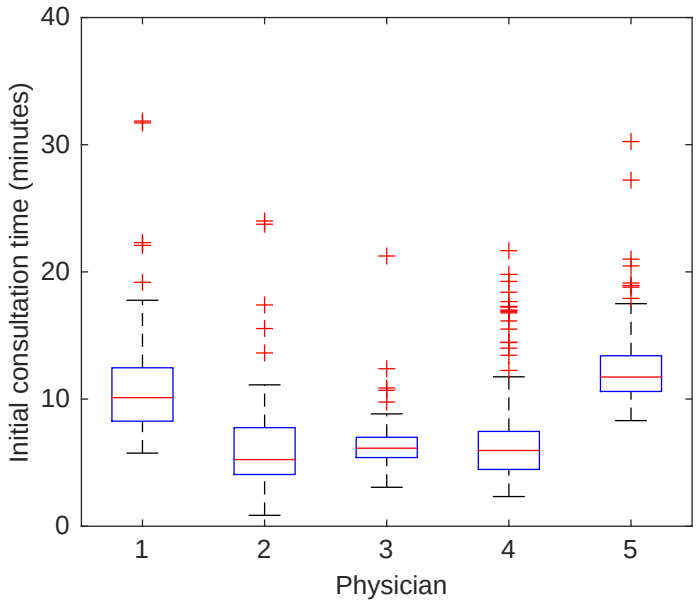

Figure 3 **Boxplot of consultation times by five different physicians.**

7.2. Comparison with Sample Average Approximations

The admissible schedule obtained by solving (8) is in general *not* the optimal solution to the P-model problem (6). Although finding the exact optimal solution to (6) is difficult, it is possible to use approximate methods such as the SAA method to obtain near-optimal solutions with reduced computational effort. Let us compare the computational performance of the hybrid approach with the SAA method.

For each $\mu \in \mathcal{A}$, let $\tilde{\chi}(\mu)$ be an indicator random variable given by

$$\tilde{\chi}(\mu) = \begin{cases} 1 & \text{if } w_k(\tilde{\boldsymbol{s}}, \mu) \le \tau_k \text{ for all } k \in \mathcal{I}^W, \\ 0 & \text{otherwise.} \end{cases}$$

Then,

$$\mathbb{E}\big(\tilde{\chi}(\mu)\big) = \mathbb{P}\big(w_k(\tilde{\boldsymbol{s}}, \mu) \le \tau_k \text{ for all } k \in \mathcal{I}^W\big).$$

Let $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N\}$ be a sample of consultation time vectors independently taken from the distribution of $\tilde{\boldsymbol{s}}$ and $\{\chi_1(\mu), \ldots, \chi_N(\mu)\}$ the corresponding realizations of $\tilde{\chi}(\mu)$. By the strong law of large numbers, the probability that all waiting patients meet the delay targets under the admissible schedule μ can be approximated by $\sum_{n=1}^{N} \chi_n(\mu)/N$. Using this fact, we formulate an approximate

problem for (6) by

$$\begin{array}{lll} \max & \sum_{n=1}^{N} z_n & \\ \text{s.t.} & w_k(\boldsymbol{s}_n, \mu) \le \tau_k + (1 - z_n) \cdot \bar{\tau}, & k \in \mathcal{I}^W,\ n = 1, \ldots, N \\ & \mu \in \mathcal{A},\ z_n \in \{0, 1\} & ,\quad n = 1, \ldots, N \end{array} \tag{24}$$

where $\bar{\tau}$ is the upper bound of patient waiting times given by (18). When $z_n = 0$, the inequality in (24) always holds for all $k \in \mathcal{I}^W$ and $\mu \in \mathcal{A}$; when $z_n = 1$, the inequality holds for all $k \in \mathcal{I}^W$ and a given $\mu \in \mathcal{A}$ if and only if $\chi_n(\mu) = 1$. Under a given $\mu \in \mathcal{A}$, the maximum value that $\sum_{n=1}^{N} z_n$ can take must be equal to $\sum_{n=1}^{N} \chi_n(\mu)$. Therefore, when the sample size N is large, the admissible schedule that maximizes the objective function in (24) should be a near-optimal solution to (6). Following the procedure in §5, one can also translate this SAA formulation into a mixed integer program.

We consider a scenario with six physicians and twenty patients in the ED. Eight and twelve patients belong to the first and second categories, respectively. There are three new patients, twelve returning patients, and five patients being seen by physicians. Each physician has two returning patients waiting to be seen. The delay targets for the three new patients are set to be 30 minutes. For the returning patients, six of them have delay targets of 20 minutes, five of them have 30 minutes, and the other one has 25 minutes. All physicians are assumed to have identical work capabilities, so the distribution of each consultation time depends only on the patient's category. All consultation times are sampled from the empirical distributions in Figure 2 according to each patient's category.

The computational performance of the SAA method is determined mainly by the sample size. With a larger sample, one may obtain a better solution to the original P-model problem at the expense of a longer computation time. In this experiment, we test the SAA formulation (24) with sample sizes $N = 20, 40, 60, 100$, and eight realizations are taken for each sample size. We obtain an admissible schedule by solving (24) for each realization of the random sample. Computation time is the major performance concern of this step. The obtained admissible schedule is then evaluated by Monte Carlo simulation, where all patients' consultation times are re-sampled from their empirical distributions. In the simulation, the patients are seen by the physicians according to the obtained admissible schedule (we assume that patients will leave the ED after they finish their current consultations). The probability of all patients meeting their delay targets is computed as the performance measure through 1,000 independent simulation runs.

We depict the performance of the SAA method in Figure 4, where SAA(N) denotes the optimal solution to (24) based on a realization of sample size N. When N is small, the solutions exhibit great

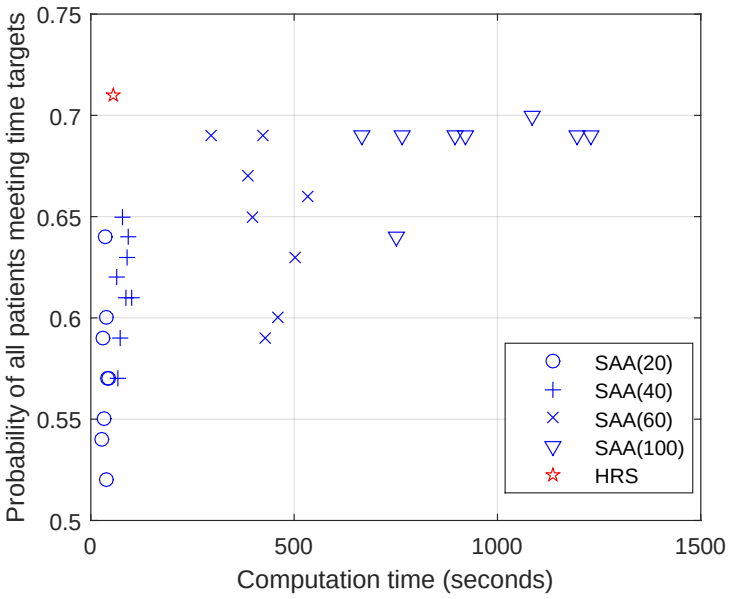

Figure 4 Computational performance of the hybrid approach and the SAA method.

variability in performance across different realizations, while most of them are not satisfactory. Increasing the sample size can stabilize and improve the performance of solutions, but at the expense of longer computation time. We also illustrate the performance of the hybrid robust–stochastic approach, which is denoted by HRS in the same figure. The solution to the hybrid formulation does not depend on specific realizations of a random sample, so no variability is present in the performance of this approach. Although the objective function in (8) is different from that of the original problem (6), the optimal solution to the hybrid formulation appears to dominate all other solutions in Figure 4. Moreover, in order to achieve comparable performance, the SAA method requires much longer computation time than the hybrid approach. This implies that with a large sample size, the SAA method cannot be used for patient scheduling on a real-time basis.

7.3. Experiments on Dynamic Patient Scheduling

Now let us evaluate the performance of the hybrid approach in dynamic patient scheduling. We focus on two performance measures: the percentage of patients whose door-to-provider times exceed the safety limits and the percentage of patients whose lengths of stay exceed the mandatory target. Since the major concern is the crowding of level-3 patients, we set the safety limit for all patients' door-to-provider times to be $D = 30$ minutes and set the mandatory target for all patients' lengths of stay to be $K = 200$ minutes. In the numerical experiments, we generate a stream of 5,000 patients arriving at the ED according to a Poisson process with rate 15.2 patients per hour. For the convenience of simulation, we assume that all medical tests and treatments by a nurse take no

time. That is to say, if a patient needs to return, he will join the queue for the same physician upon finishing the current consultation. There are four physicians in the ED. All patient consultation times are sampled from the empirical distributions in Figure 2 by each patient's category, with 40% and 60% of patients belonging to the first and second categories, respectively.

According to the patient flow data from the hospital, there are around 75% of patients returning to their initial physicians at least once before leaving the ED, while there are less than 4% of patients returning to their physicians more than three times. In the numerical experiments, we assume that a patient may return to the same physician at most three times and the probabilities of zero to three returns are 0.25, 0.40, 0.25, and 0.10, respectively. The number of returns is assumed to be independent of a patient's category. With these parameters, the ED turns out to be heavily loaded with traffic intensity $\rho = 93.57\%$. In order to solve the mixed integer program (23) sequentially, we need to specify delay targets for returning patients when the decision process is triggered. Since all patients in the experiments have identical door-to-provider time and length-of-stay requirements, we choose the same returning delay limits $\boldsymbol{T} = (T_1, T_2, T_3)$ for all returning patients.

In the numerical experiments, the performance of the hybrid approach is compared with that of the following three scheduling policies.

Global FCFS: When a physician finishes a consultation, the scheduling system will assign the patient who has the earliest registration time to him, among all patients available to the physician. Since medical tests and treatments are assumed to be instantaneous, a physician will be kept working on each patient until all consultations of this patient are completed. In this case, the Global FCFS policy is equivalent to the *returning-patients-first* policy.

New-patients-first: When a physician finishes a consultation, the scheduling system will assign the new patient who has the earliest registration time to him; if no new patients are available, the scheduling system will assign the returning patient who is eligible to be seen by the physician and has the earliest registration time.

The scheduling policy proposed by Huang et al. (2015): When a physician finishes a consultation at time t, the scheduling system will first check if there are new patients whose waiting times are about to exceed or have exceeded the door-to-provider time limits, i.e., if there exists any $i \in \mathcal{I}^N$ such that $D - (t - a_i) < \epsilon$, where $\epsilon > 0$ is a small number. The scheduling system will give priority to new patients if such a patient is found, and will give priority to returning patients otherwise. In the numerical experiments, we follow the recommendation by Huang et al. (2015), taking $\epsilon = 3$ minutes. If a new patient is to be served, the scheduling system will send the one who arrived the earliest to the available physician. (Huang et al. employed the *shortest-deadline-first* policy for new patients, which is reduced to the FCFS policy when all patients have the same door-to-provider time limit.) If a returning patient is to be served, the scheduling system will select the one with the

Table 1 Performance comparison of door-to-provider times (W) and lengths of stay (L) under different scheduling policies, with arrival rate 15.2 per hour and four homogeneous physicians.

	$\bar{W}$	$\bar{L}$	$\%(W>30)$	$\%(L>200)$
Global FCFS	37.64	52.17	45.88%	0.46%
New-patients-first	3.87	74.96	0.10%	8.50%
Huang et al. (2015)	19.63	59.91	24.34%	4.90%
HRS $(105, 140, 185)$	15.68	64.67	16.18%	1.98%
HRS $(115, 145, 190)$	14.52	66.04	13.80%	2.14%
HRS $(120, 150, 195)$	14.24	65.81	13.70%	2.20%

earliest registration time among the returning patients who have the shortest expected remaining consultation times. In other words, the returning patient who is the closest to finishing the ED visit will be sent to the available physician. (Huang et al. adopted a modified generalized $c\mu$-rule for returning patients so as to minimize the cumulative congestion cost. The scheduling policy is reduced to the *closest-to-exit-first* policy if the length of stay of each patient is regarded as the congestion cost.) Under the assumption that all physicians are homogeneous, this scheduling policy is proved to be asymptotically optimal in minimizing the mean length of stay with constraints on door-to-provider times. We use this policy as the benchmark.

In the above three scheduling policies, we assume that when a new patient arrives at the waiting area finding at least one free physician, the scheduling system will randomly select a free physician and send the patient to him immediately.

In the first numerical example, we assume that the four physicians have identical work capabilities, so the distribution of a patient's consultation time does not depend on specific physicians. We compare several performance measures in Table 1, including the mean door-to-provider time, the mean length of stay, the percentage of door-to-provider times exceeding 30 minutes, and the percentage of lengths of stay exceeding 200 minutes, under different scheduling policies. Giving priority to returning patients, the global FCFS policy yields short lengths of stay, but at the expense of long door-to-provider times. If the new-patients-first policy is used, the resulting door-to-provider times are short, whereas the lengths of stay turn out to be much longer. When the ED is crowded, neither policy can be used for patient scheduling in order to meet the stringent time constraints. The benchmark policy proposed by Huang et al. (2015) is able to strike a balance among these performance measures. As we discussed earlier, this policy depends on the consultation time distributions only through their first moments. The hybrid robust–stochastic approach is denoted by HRS (T_1, T_2, T_3) in the table, where (T_1, T_2, T_3) specifies the returning delay limits used

Table 2 Performance comparison of door-to-provider times (W) and lengths of stay (L) under different scheduling policies, with arrival rate 15.2 per hour and four heterogeneous physicians.

	$\bar{W}$	$\bar{L}$	$\%(W>30)$	$\%(L>200)$
Global FCFS	38.69	53.19	46.60%	0.36%
New-patients-first	3.88	72.19	0.12%	8.14%
Huang et al. (2015)	20.06	63.32	25.84%	6.20%
HRS $(105, 140, 185)$	13.06	56.52	12.74%	1.62%
HRS $(115, 145, 190)$	10.45	54.22	8.26%	0.72%
HRS $(120, 150, 195)$	9.97	54.91	7.28%	1.18%

in the algorithm. The hybrid approach outperforms the benchmark policy in terms of the mean door-to-provider time and the percentages of time violations. This is because the hybrid approach is able to evaluate the influence of entire consultation time distributions in patient scheduling, not just that of the first moments. Although this advantage is gained at a higher computational cost, solving the scheduling problem is still practically efficient under the hybrid formulation. In Table 1, the mean length of stay is slightly longer under our approach than under the benchmark policy, because with door-to-provider time constraints, the latter policy is asymptotically optimal in this performance measure. Our approach is designed to comply with mandatory targets for lengths of stay, and the percentage of time violations is usually regarded as a more important performance indicator.

The selection of returning delay limits may influence the performance of our scheduling policy. On the one hand, increasing these limits will accommodate more returning patients within the delay targets for their current consultations, thus allowing more new patients to meet their door-to-provider limits. On the other hand, with larger returning delay limits, the patients who have finished multiple consultations will be more likely to exceed the length-of-stay limit. To deal with these two concerns, we take T_1, the delay limit for the second consultations, to be several times longer than the door-to-provider time limit, and take T_3, the delay limit for the fourth consultations, to be sufficiently lower than the length-of-stay limit. From Table 1, we can see that the percentage of door-to-provider time violations may be reduced by increasing T_1, while the percentage of length-of-stay violations may be reduced by decreasing T_3.

The most important advantage of the hybrid approach is the capability of patient scheduling in the presence of heterogeneous physicians. When physicians have different work rates, their expertise should be taken into account in making scheduling decisions. Consider the following scenario: A physician is an expert in treating patients in category 1 but not familiar with cases in category 2.

When the physician becomes available, there is a new patient of category 2 whose door-to-provider time is about to exceed the safety limit. At the moment, should we send the category-2 patient to this "slow" physician, or keep the category-2 patient waiting and send a category-1 patient so that the physician can be working at a "fast" rate? In this case, a trade-off must be made between preventing an immediate time violation and reducing future crowding, which requires the scheduling policy to be able to evaluate the consequences of both actions. Unfortunately, the benchmark policy by Huang et al. (2015) does not allow for heterogeneous physicians. It is no longer asymptotically optimal in reducing the mean length of stay when physicians have different work rates.

The four physicians are assumed to be heterogeneous in the second example. Two of them are experts in treating cases in category 1 but not good at category 2; the other two physicians are more experienced in category 2 but not familiar with category 1. In the simulation, we generate *original* consultation times using the empirical distributions by each patient's category, while the actual consultation time of a patient depends on the specific physician. If the physician is an expert in the patient's category, the actual consultation time will be 80% of the original time; otherwise, the actual consultation time will be 120% of the original time. All other simulation settings are the same as in the previous example, and some numerical results are reported in Table 2. The global FCFS, new-patients-first, and benchmark policies do not differentiate physicians in making scheduling decisions. Under these policies, each physician's average work rate is identical to that in the previous example, so the system's performance does not show much difference. The hybrid approach can better exploit each physician's expertise by sending more patients to fast physicians as long as the situation permits. In doing so, the physicians can be working at an optimized rate, which results in considerable performance improvement. In Table 2, the hybrid approach outperforms the benchmark policy by all measures. We also plot the percentages of lengths of stay exceeding 100–300 minutes in Figure 5. In contrast to the benchmark policy that yields a long tail in the length-of-stay distribution, the hybrid approach quickly suppresses the tail distribution before lengths of stay reach the mandatory target, thus preventing extremely long stays in the ED.

8. Concluding Remarks

We proposed a data-driven approach to patient scheduling in EDs, where mandatory targets are imposed on patients' door-to-provider times and lengths of stay. The main contribution of this paper is a hybrid robust–stochastic formulation to the patient scheduling problem, by which we may obtain a near-optimal solution to the P-model problem at a significantly lower computational expense. Using this computationally efficient approach and real-time patient flow data, we developed a dynamic scheduling algorithm for making recommendations about the next patient to be

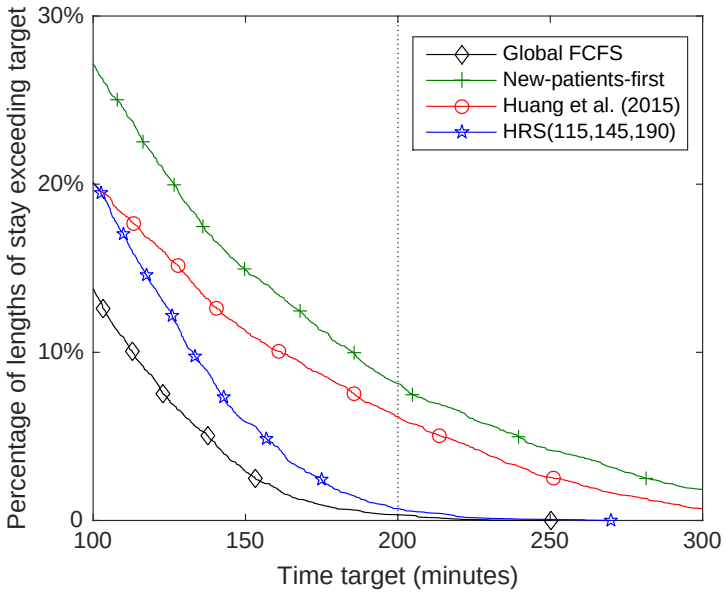

Figure 5 **Percentages of lengths of stay exceeding 100–300 minutes under different scheduling policies.**

seen by each available physician. Our hybrid robust–stochastic approach allows for practical features in the formulation and outperforms existing scheduling policies in the numerical experiments. The capability of scheduling in the presence of heterogeneous physicians, in particular, is a major advantage of this approach.

The proposed hybrid formulation may provide a computationally tractable alternative to solving optimization problems in stochastic networks with delay or throughput time constraints. Such problems can be found in healthcare systems where service requirements are time-sensitive, e.g., patient transfer from EDs to inpatient wards (Mandelbaum et al. 2012 and Shi et al. 2015), ambulance deployment (McLay and Mayorga 2013, Maxwell et al. 2014, and Chong et al. 2015), and health examinations (Baron et al. 2015). Similar problems may arise from transportation systems, e.g., taxi dispatching (Seow et al. 2010), electric vehicle charging management (Yilmaz and Krein 2013), and vehicle routing with stochastic demands and time windows (Bertsimas and van Ryzin 1993, Fisher et al. 1997, Laporte et al. 2002, and Jepsen et al. 2008).

References

Armony, M., S. Israelit, A. Mandelbaum, Y. N. Marmor, Y. Tseytlin, G. B. Yom-Tov. 2015. On patient flow in hospitals: A data-based queueing-science perspective. *Stochastic Systems* **5**(1) 1–49.

Bandi, C., D. Bertsimas. 2012. Tractable stochastic analysis in high dimensions via robust optimization. *Mathematical Programming, Series B* **134**(1) 23–70.

Bandi, C., D. Bertsimas, N. Youssef. 2015. Robust queueing theory. *Operations Research* To appear.

Baron, O., O. Berman, D. Krass, J. Wang. 2015. Strategic idling and dynamic scheduling in an open-shop service network: Case study and analysis. Preprint.

Ben-Tal, A., A. Goryashko, E. Guslitzer, A. Nemirovski. 2004. Adjustable robust solutions of uncertain linear programs. *Mathematical Programming, Series A* **99**(2) 351–376.

Bertsimas, D., D. Gamarnik, A. A. Rikun. 2011. Performance analysis of queueing networks via robust optimization. *Operations Research* **59**(2) 455–466.

Bertsimas, D., M. Sim. 2004. The price of robustness. *Operations Research* **52**(1) 35–53.

Bertsimas, D. J., G. van Ryzin. 1993. Stochastic and dynamic vehicle routing in the Euclidean plane with multiple capacitated vehicles. *Operations Research* **41**(1) 60–76.

Charnes, A., W. W. Cooper. 1963. Deterministic equivalents for optimizing and satisficing under chance constraints. *Operations Research* **11** 18–39.

Chong, K. C., S. G. Henderson, M. E. Lewis. 2015. The vehicle mix decision in emergency medical service systems. *Manufacturing & Service Operations Management* To appear.

Connelly, L. G., A. E. Bair. 2004. Discrete event simulation of emergency department activity: A platform for system-level operations research. *Academic Emergency Medicine* **11**(11) 1177–1185.

de Bruin, A. M., R. Bekker, L. van Zanten, G. M. Koole. 2010. Dimensioning hospital wards using the Erlang loss model. *Annals of Operations Research* **178**(1) 23–43.

Doytchinov, B., J. Lehoczky, S. Shreve. 2001. Real-time queues in heavy traffic with earliest-deadline-first queue discipline. *Annals of Applied Probability* **11**(2) 332–378.

Fisher, M. L., K. O. Jörnsten, O. B. G. Madsen. 1997. Vehicle routing with time windows: Two optimization algorithms. *Operations Research* **45**(3) 488–492.

Gilboy, N., P. Tanabe, D. Travers, A. M. Rosenan. 2011. *Emergency Severity Index (ESI): A Triage Tool for Emergency Department Care, Version 4. Implementation Handbook 2012 Edition*. AHRQ Publications, Rockville, MD.

Green, L. V., J. Soares, J. F. Giglio, R. A. Green. 2006. Using queueing theory to increase the effectiveness of emergency department provider staffing. *Academic Emergency Medicine* **13**(1) 61–68.

Hoot, N. R., D. Aronsky. 2008. Systematic review of emergency department crowding: Causes, effects, and solutions. *Annals of Emergency Medicine* **52**(2) 126–136.

Huang, J., B. Carmeli, A. Mandelbaum. 2015. Control of patient flow in emergency departments, or multiclass queues with deadlines and feedback. *Operations Research* To appear.

Jepsen, M., B. Petersen, S. Spoorendonk, D. Pisinger. 2008. Subset-row inequalities applied to the vehicle-routing problem with time windows. *Operations Research* **56**(2) 497–511.

Laporte, G., F. V. Louveaux, L. van Hamme. 2002. An integer *L*-shaped algorithm for the capacitated vehicle routing problem with stochastic demands. *Operations Research* **50**(3) 415–423.

Maglaras, C., J. A. Van Mieghem. 2005. Queueing systems with leadtime constraints: A fluid-model approach for admission and sequencing control. *European Journal of Operational Research* **167**(1) 179–207.

Mandelbaum, A., P. Momčilović, Y. Tseytlin. 2012. On fair routing from emergency departments to hospital wards: QED queues with heterogeneous servers. *Management Science* **58**(7) 1273–1291.

Maxwell, M. S., E. C. Ni, C. Tong, S. G. Henderson, H. Topaloglu, S. R. Hunter. 2014. A bound on the performance of an optimal ambulance redeployment policy. *Operations Research* **62**(5) 1014–1027.

McHugh, M. 2013. The conesequences of emergency department crowding and delays for patients. R. Hall, ed., *Patient Flow: Reducing Delay in Healthcare Delivery*, 2nd ed. Springer, New York, 107–127.

McLay, L. A., M. E. Mayorga. 2013. A dispatching model for server-to-customer systems that balances efficiency and equity. *Manufacturing & Service Operations Management* **15**(2) 205–220.

Nemirovski, A., A. Shapiro. 2006. Convex approximations of chance constrained programs. *SIAM Journal on Optimization* **17**(4) 969–996.

Plambeck, E., S. Kumar, J. M. Harrison. 2001. A multiclass queue in heavy traffic with throughput time constraints: Asymptotically optimal dynamic controls. *Queueing Systems* **39**(1) 23–54.

Rabin, E., K. Kocher, M. McClelland, J. Pines, U. Hwang, N. Rathlev, B. Asplin, N. S. Trueger, E. Weber. 2012. Solutions to emergency department 'boarding' and crowding are underused and may need to be legislated. *Health Affairs* **31**(8) 1757–66.

Saghafian, S., W. J. Hopp, M. P. Van Oyen, J. S. Desmond, S. L. Kronick. 2012. Patient streaming as a mechanism for improving responsiveness in emergency departments. *Operations Research* **60**(5) 1080–1097.

Saghafian, S., W. J. Hopp, M. P. Van Oyen, J. S. Desmond, S. L. Kronick. 2014. Complexity-augmented triage: A tool for improving patient safety and operational efficiency. *Manufacturing & Service Operations Management* **16**(3) 329–345.

Seow, K. T., N. H. Dang, D.-H. Lee. 2010. A collaborative multiagent taxi-dispatch system. *IEEE Transactions on Automation Science and Engineering* **7**(3) 607–616.

Shi, P., M. C. Chou, J. G. Dai, D. Ding, J. Sim. 2015. Models and insights for hospital inpatient operations: Time-dependent ED boarding time. *Management Science* To appear.

Sinreich, D., Y. Marmor. 2005. Emergency department operations: The basis for developing a simulation tool. *IIE Transactions* **37**(3) 233–245.

Weber, E. J., S. Mason, A. Carter, R. L. Hew. 2011. Emptying the corridors of shame: Organizational lessons from England's 4-hour emergency throughput target. *Annals of Emergency Medicine* **57**(2) 79–88.e1.

Welch, S. J., B. R. Asplin, S. Stone-Griffith, S. J. Davidson, J. Augustine, J. Schuur. 2011. Emergency department operational metrics, measures and definitions: Results of the second performance measures and benchmarking summit. *Annals of Emergency Medicine* **58**(1) 33–40.

Yilmaz, M., P. T. Krein. 2013. Review of battery charger topologies, charging power levels, and infrastructure for plug-in electric and hybrid vehicles. *IEEE Transactions on Power Electronics* **28**(5) 2151–2169.

Yom-Tov, G. B., A. Mandelbaum. 2014. Erlang-R: A time-varying queue with reentrant customers, in support of healthcare staffing. *Manufacturing & Service Operations Management* **16**(2) 283–299.