Management Science

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

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

Shuangchi He, Melvyn Sim, Meilin Zhang

To cite this article:
Shuangchi He, Melvyn Sim, Meilin Zhang (2019) Data-Driven Patient Scheduling in Emergency Departments: A Hybrid Robust-Stochastic Approach. Management Science 65(9):4123-4140. https://doi.org/10.1287/mnsc.2018.3145

http://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Vol. 65, No. 9, September 2019, pp. 4123–4140
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

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

Shuangchi He,[a] Melvyn Sim,[b] Meilin Zhang[c]

[a] Department of Industrial Systems Engineering and Management, National University of Singapore, Singapore 117576; [b] Department of Analytics and Operations, NUS Business School, National University of Singapore, Singapore 119245; [c] School of Business, Singapore University of Social Sciences, Singapore 599494

Contact: heshuangchi@nus.edu.sg, http://orcid.org/0000-0003-4107-3946 (SH); melvynsim@nus.edu.sg, http://orcid.org/0000-0001-9798-2482 (MS); zhangmeilin@suss.edu.sg, http://orcid.org/0000-0003-2880-8223 (MZ)

Received: November 22, 2015
Revised: February 19, 2017; April 16, 2018
Accepted: May 24, 2018
Published Online in Articles in Advance: May 1, 2019

https://doi.org/10.1287/mnsc.2018.3145

Abstract. 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.

History: Accepted by Yinyu Ye, optimization.
Funding: The work of S. He was supported in part by the Singapore Ministry of Education Academic Research Fund [Grant MOE2017-T2-1-012]. The work of M. Sim was supported in part by the Singapore Ministry of Education Social Science Research [Thematic Grant MOE2016-SSRTG-059].
Supplemental Material: The e-companion is available at https://doi.org/10.1287/mnsc.2018.3145.

Keywords: healthcare operations • patient scheduling • robust optimization • stochastic programming • mixed integer programming • queueing network

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 because of 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 many countries, the 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 the 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, and so forth. 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 that 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). Because 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, 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 equipment 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 with a physician, most patients 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 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. Because the patient arrival pattern is highly variable, we refrain from making assumptions, such as the arrival rate and the inter-arrival time distribution, about future patient arrivals. 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 by maximizing the joint probability of all waiting patients meeting the delay targets (please see formulations (6) and (7) and related discussion). 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*. The *P* model formulation, 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 solving the associated nonconvex optimization 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 uncertain set that has the largest probability of all of its consultation times being feasible. The schedule associated with the obtained uncertainty set is the optimal solution to the hybrid formulation. For computational reasons, we restrict the family of uncertainty sets to a collection of hyperrectangles. 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 may 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 asymptotically optimal policy; more details are in Section 7.

The hybrid robust-stochastic approach is of both practical and methodological importance. First, although the hybrid formulation is essentially a mixed integer program, solving this problem is practically efficient and allows for 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 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 problem,

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 at a far lower computational expense. As illustrated by numerical examples in Section 7.2, our approach may provide a highly efficient alternative to the SAA method. In addition to patient scheduling, similar problems arise from other stochastic systems with time-sensitive service requirements; Section 8 has more discussion.

The remainder of this paper is organized as follows. The related literature is reviewed in Section 2. We introduce the queueing network model for EDs in Section 3. In Section 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 Section 5. By introducing additional delay constraints, the hybrid formulation is incorporated into a dynamic scheduling framework in Section 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 Section 7, where the hybrid approach is compared with both the SAA method and an asymptotically optimal scheduling policy. The paper is concluded in Section 8, where some potential applications and future research are discussed. We leave the construction of nonanticipative arrangements, all proofs, and additional simulation results to the e-companion.

Let us close this section with frequently used notation. Scalars and vectors are denoted by lowercase and bold 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 that event A occurs. We reserve $\mathbb{E}(\tilde{s})$ for the expectation of a random variable $\tilde{s}$.

2. Related Literature

We sketch relevant studies to position our work within the 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 intention 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; Section 7.3 has a 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), the 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. Conceivably, 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 ED environment.

Even if the queueing network model is analytically tractable, it is still difficult to find an optimal dynamic control policy with delay or throughput time constraints. Most studies in the literature focus on simple policies that can be proved optimal in some asymptotic sense (e.g., Doytchinov et al. 2001, Plambeck et al. 2001, Maglaras and Van Mieghem 2005, and Huang et al. 2015). Some aforementioned simplistic assumptions, along with a heavy traffic condition, are necessary for asymptotic optimality to hold. The performance of these policies may be suboptimal when the simplistic assumptions are not satisfied. Moreover, the control actions of these policies depend on service time

distributions only through their first moments. For these policies to be near optimal, deadlines for delay or throughput times must be on a higher order of magnitude than service times, which may not be a reasonable assumption in the ED setting. In contrast, the hybrid robust-stochastic approach allows for multiple heterogeneous servers, time-varying arrival processes, and arbitrary traffic conditions. Distributional information obtained from patient flow data is fully used in constructing uncertainty sets. In other words, the hybrid approach can make use of entire service time distributions, which turns out to be a considerable advantage over the previous scheduling policies. In numerical experiments in Section 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.

Under the framework of robust optimization, the performance analysis of queueing networks was studied 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 using 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 performance bounds for 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.

Our hybrid robust-stochastic approach shares some features with a concurrent, independent study by Zhang et al. (2017), who considered robust optimal control of constrained linear systems with adjustable uncertainty sets. The formulation of their problem is motivated by reserve provision in electrical grids, where the reserve capacity of power is required to be periodically adjusted for cost saving without sacrificing necessary power consumption. In their formulation, a series of adjustable uncertainty sets is used to represent the reserve capacity, thus becoming decision variables as in our approach. Zhang et al. (2017) also investigated uncertainty sets of special geometric forms to render their optimization problem tractable. Despite these analogous features, our study differs significantly from their work in the following aspects. From the modeling perspective, adjustable uncertainty sets arise naturally from the formulation of reserve provision in their paper, whereas in our approach, hyperrectangular sets serve primarily as a heuristic proxy for feasible sets in solving the P model problem. Associated with a joint probability measure, these hyperrectangular sets should not be understood as uncertainty sets in the usual sense. From the methodological perspective, the main theme of their paper is how to confine both uncertainty sets and admissible policies to affine structures so that the resulting formulation becomes a convex optimization problem. In contrast, we convert the patient scheduling problem into a tractable mixed integer program, relying on the fact that, within each hyperrectangle, the worst case only occurs at a single boundary point (discussed in Theorem 1 in Section 4). It is also worth mentioning that, as decision variables, uncertainty sets in both studies are subject to joint constraints with control policies. In other words, the families of adjustable uncertainty sets may change with specific policies. This is similar to the robust optimization formulation proposed by Spacey et al. (2012), who studied software partitioning with multiple instantiation (i.e., assigning code segments of a computer program to multiple execution locations so as to minimize the overall program run time). The only decision variable in their problem is the software partition, which also determines the uncertainty set of location-aware control flows. Unlike in our study, their uncertainty set is not adjustable, and therefore, it is not a decision variable of the optimization problem.

3. The Controlled Queueing Network Model

We use a queueing network to model the ED, which is controlled by a centralized patient scheduling system to meet 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.

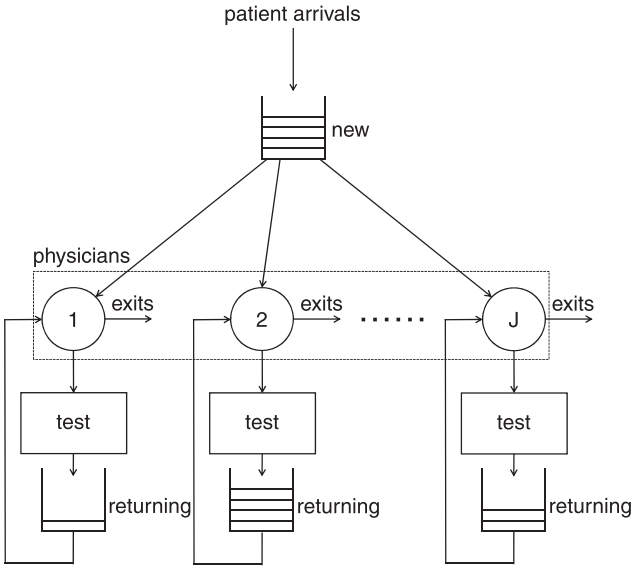
The general flow of patients goes through the ED according to the following process. Patients arrive at the ED in a stochastic and nonstationary manner. After registration, they will be triaged by a nurse and assigned to several urgency groups based on their acuity levels and other concerns. The door-to-provider 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 another. Then, patients will stay in a waiting area until they are called to be seen by a physician. These patients will be referred to as new patients. After initial consultations, some patients may leave the ED, whereas others may undergo diagnostic tests, such as x-rays and blood tests, or receive treatment from 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 hospitalized.

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, whereas a returning patient must be seen by the physician with whom he consulted initially. We assume that the scheduling system does not manage patients who are waiting for tests or treatment by a nurse, because 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 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 prolonged. In particular, 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. Giving priority to returning patients can effectively shorten the queue length in the waiting area, thus reducing lengths of stay. 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 determines the next patient to be seen when a physician completes a consultation, and determines a new patient's physician when he comes to the waiting area finding at least one free physician. Let t be such a time, and consider the ED at this moment. Let $\mathcal{J}$ be the set of physicians, $\mathcal{N}$ be the set of new patients, $\mathcal{C}$ be the set of patients being seen, and $\mathcal{R}$ be the set of returning patients, where the dependence on t is suppressed for notational convenience. For $j \in \mathcal{J}$, we use $\mathcal{C}_j$ to denote the set of patients being seen by physician j and $\mathcal{R}_j$ to denote the set of returning patients to be seen by physician j. Then, $\mathcal{C} = \bigcup_{j\in\mathcal{J}} \mathcal{C}_j$ and $\mathcal{R} = \bigcup_{j\in\mathcal{J}} \mathcal{R}_j$. Moreover, $\mathcal{C}_j = \emptyset$ if and only if physician j is available at time t; otherwise, $\mathcal{C}_j$ has exactly one patient. Let $\mathcal{W} = \mathcal{N} \cup \mathcal{R}$ be the set of patients in the waiting area and $\mathcal{I} = \mathcal{W} \cup \mathcal{C}$ be the set of patients in the ED, excluding those sent to tests or treatments.

Figure 1. Queueing Network Model of Patient Flow in an ED



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 $i \in \mathcal{C}_j$, $\tilde{s}_{ij}$ 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 $\tilde{s}_{ij}$. Because the physicians may be heterogeneous, even if $i \in \mathcal{I}$ is fixed, F_{ij} may vary for different $j \in \mathcal{J}$. Each F_{ij} can be estimated using the records 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, with values 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, respectively, 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 of 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 $\varphi : \mathcal{W} \to \mathcal{J}$, where $\varphi(i)$ is the physician of patient i. Because returning patients must be seen by their initial physicians, the assignment should satisfy

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

Sequencing decisions are specified by a correspondence $\Phi : \mathcal{W} \to \mathcal{P}(\mathcal{W})$, where $\mathcal{P}(\mathcal{W})$ is the power set of $\mathcal{W}$ and $\Phi(i)$ is the set of patients to be seen by the same physician before patient i. Then,

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

For patients to be seen by the same physician, the associated $\Phi(i)$ forms a collection of nested sets: for $i,k \in \mathcal{W}$, 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 $k \in \mathcal{W}$ can be obtained by

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

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 $\mathcal{C}_{\varphi(k)} = \emptyset$) and the second sum is the total consultation time of waiting patients before patient k.

We assume that each patient $i \in \mathcal{W}$ has a delay target τ_i. For a new patient $i \in \mathcal{N}$, τ_i is the amount of time from t until his waiting time exceeds the safety limit for his door-to-provider time. For a returning patient $i \in \mathcal{R}$, τ_i is specified by the scheduling system for his length of stay to meet the mandatory target. We will discuss how to determine delay targets for returning patients in Section 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, because consultation times are random, we may not be able to achieve this with complete certainty. 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 by (4). Therefore, an optimal arrangement can be obtained by solving the following P model problem

$$\begin{aligned} \max \quad & \mathbb{P}(w_i(\tilde{\boldsymbol{s}}, \pi(\tilde{\boldsymbol{s}})) \le \tau_i : i \in \mathcal{W}) \\ \text{s.t.} \quad & \pi \in \mathcal{V}. \end{aligned} \tag{5}$$

This formulation, however, cannot be implemented, because to determine the admissible schedule, one is required to know the realization of $\tilde{\boldsymbol{s}}$ in advance. To fix this issue, we should confine feasible solutions to (5) within the set of *nonanticipative* arrangements, which do not rely on future information to determine the patients to be seen when physicians become available. To specify a nonanticipative arrangement, we need to determine the assignment and sequencing decisions in a sequential manner. Let $w(1) \le w(2) \le \cdots$ be the times when physicians are available to start a consultation. To decide the next patient to be seen at time $w(k)$, we may exploit information on patients and physicians available before $w(k)$, including the consultation history of each physician, the identities of waiting patients, the amounts of time that the physicians have been with their current patients, etc. Using the cumulative information, we may define nonanticipative arrangements through a recursive procedure. Because the construction of nonanticipative arrangements is generally complicated, we leave the details to the e-companion. Let $\mathcal{V}_1$ be the set of all nonanticipative arrangements. We may obtain an optimal nonanticipative arrangement by solving the problem

$$\begin{aligned} \max \quad & \mathbb{P}(w_i(\tilde{\boldsymbol{s}}, \pi(\tilde{\boldsymbol{s}})) \le \tau_i : i \in \mathcal{W}) \\ \text{s.t.} \quad & \pi \in \mathcal{V}_1. \end{aligned} \tag{6}$$

The P model problem (6) turns out to be intractable owing in part to the curse of dimensionality induced by the recursive structure of nonanticipative arrangements. To simplify the computation, we may further confine feasible solutions to (5) within the set of *static* arrangements, which is given by

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

The next proposition states that static arrangements are nonanticipative.

Proposition 1. *Let $\mathcal{V}$, $\mathcal{V}_0$, and $\mathcal{V}_1$ be the sets of all arrangements, static arrangements, and nonanticipative arrangements, respectively. Then, $\mathcal{V}_0 \subset \mathcal{V}_1 \subset \mathcal{V}$.*

Because static arrangements are invariant for all realizations of consultation times, finding an optimal static arrangement would be much simpler than solving (6). It is equivalent to obtaining an optimal admissible schedule by solving the following static P model problem:

$$\begin{aligned} \max \quad & \mathbb{P}(w_i(\tilde{\boldsymbol{s}}, \mu) \le \tau_i : i \in \mathcal{W}) \\ \text{s.t.} \quad & \mu \in \mathcal{A}. \end{aligned} \tag{7}$$

The optimal solution $\mu^\dagger = (\varphi^\dagger, \Phi^\dagger)$ to (7) specifies the assignment and sequencing decisions for all waiting patients. In particular, if there is any $i \in \mathcal{W}$, such that both $\Phi^\dagger(i) = \emptyset$ and $\mathcal{C}_{\varphi^\dagger(i)} = \emptyset$ 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 completes a consultation or a new patient arrives in the waiting area finding at least one free physician. Each time, the scheduling system determines the next patient to be seen for the available physician.

Finding an optimal admissible schedule is still a considerable challenge, although solving (7) is much simpler than solving (6). Under an admissible schedule, evaluating the joint probability in (7) involves

multidimensional integration of many variables, which is computationally prohibitive. Nemirovski and Shapiro (2006) pointed out that computing the distribution of the sum of independent random variables is an NP-hard problem. As a result, even finding the distribution of a patient's waiting time would be computationally difficult. When the number of waiting patients is large, we would be unable to obtain the optimal admissible schedule for (7) within a reasonable time that is required for dynamic patient scheduling. Therefore, we would like to focus on a computationally amiable approach to 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 $\mu \in \mathcal{A}$, the set

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

is a convex polyhedron in $|\mathcal{I}| \cdot |\mathcal{J}|$ dimensions. Then, we may rewrite (7) as

$$\begin{array}{ll} \max & \mathbb{P}(\tilde{\boldsymbol{s}} \in \mathcal{X}(\mu)) \\ \text{s.t.} & \mu \in \mathcal{A}, \end{array} \tag{8}$$

the optimal solution to which is the admissible schedule that maximizes the joint probability of all consultation times being within the associated convex polyhedron. Because it is difficult to evaluate $\mathbb{P}(\tilde{\boldsymbol{s}} \in \mathcal{X})$ for a general polyhedron $\mathcal{X}$ in high dimensions, it would be computationally excruciating to find the optimal solution to (8). However, if $\mathcal{X}$ happens to be hyperrectangular (e.g., $\mathcal{X} = \prod_{i\in\mathcal{I}, j\in\mathcal{J}}[0, d_{ij}]$ for some $d_{ij} \ge 0$), the above joint probability can be computed by

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

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. Note that, by (8), we intend to find the admissible schedule with the associated convex polyhedron that 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 with a probability measure that is also large. Conversely, if we can find an admissible schedule with an associated convex polyhedron that contains a "large" hyperrectangle, we may also expect the polyhedron itself to be large. Hence, instead of searching for the admissible schedule that has the "largest" convex polyhedron, we would find the admissible schedule with an associated convex polyhedron that has the largest hyperrectangular subset. Because it is far easier to evaluate the probability measure of a hyperrectangle, the patient scheduling problem would be more computationally amiable under this formulation.

We would modify the P model problem (6) to obtain a "near-optimal" nonanticipative arrangement. To this end, let us consider a collection of hyperrectangular subsets of $\mathcal{S}$ given by

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

Our objective is to maximize the joint probability that the consultation times are within a hyperrectangular set $\mathcal{Q} \in \mathcal{H}$ without exceeding the delay targets:

$$\begin{array}{ll} \max & \mathbb{P}(\tilde{\boldsymbol{s}} \in \mathcal{Q}) \\ \text{s.t.} & w_k(\boldsymbol{s}, \pi(\boldsymbol{s})) \le \tau_k, \quad k \in \mathcal{W},\ \boldsymbol{s} \in \mathcal{Q} \\ & \mathcal{Q} \in \mathcal{H},\ \pi \in \mathcal{V}_1. \end{array} \tag{9}$$

From a robust optimization perspective, $\mathcal{H}$ can be regarded as a family of uncertainty sets for $\tilde{\boldsymbol{s}}$, and the specific uncertainty set $\mathcal{Q}$ can be adjusted within $\mathcal{H}$ using different arrangements. The objective function in (9) involves random variables, whereas the constraints are based on a robust optimization formulation with an uncertainty set that is adjustable. Hence, we refer to this formulation as a hybrid robust-stochastic approach.

Because the representation of nonanticipative arrangements is generally complicated, one may wonder if the hybrid formulation is more computationally tractable. To address this concern, let us consider a general version of formulation (9), which is not confined to scheduling problems for queueing networks. Let $\tilde{\boldsymbol{\sigma}} = (\tilde{\sigma}_1, \ldots, \tilde{\sigma}_M)$ be an M-dimensional vector of mutually independent random variables. For $k = 1, \ldots, M$, we use G_k to denote the cumulative distribution function of $\tilde{\sigma}_k$. We use $\mathcal{F}$ to denote the collection of closed hyperrectangular subsets of $\mathbb{R}^M$ that are bounded from above (i.e., $\mathcal{F} = \{\prod_{k=1}^M (-\infty, b_k] : (b_1, \ldots, b_M) \in \mathbb{R}^M\}$). Let $\mathcal{D}$ be a nonempty set and $\mathcal{G}$ be the collection of functions from $\mathbb{R}^M$ to $\mathcal{D}$. Let θ be a function from the product space $\mathbb{R}^M \times \mathcal{D}$ to $\mathbb{R}^L$, such that, for $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_M) \in \mathbb{R}^M$ and $\nu \in \mathcal{D}$, $\theta(\boldsymbol{\sigma}, \nu)$ is nondecreasing in each σ_k. Then, for a given $\boldsymbol{\eta} \in \mathbb{R}^L$, let us consider the following problem:

$$\begin{array}{ll} \max & \mathbb{P}(\tilde{\boldsymbol{\sigma}} \in \mathcal{B}) \\ \text{s.t.} & \theta(\boldsymbol{\sigma}, \psi(\boldsymbol{\sigma})) \le \boldsymbol{\eta}, \quad \boldsymbol{\sigma} \in \mathcal{B} \\ & \mathcal{B} \in \mathcal{F},\ \psi \in \mathcal{G}. \end{array} \tag{10}$$

Theorem 1 provides a simplified form of (10).

Theorem 1. *Let $(\boldsymbol{b}^\star, \nu^\star) \in \mathbb{R}^M \times \mathcal{D}$ be an optimal solution to the following problem:*

$$\begin{array}{ll} \max & \displaystyle\sum_{k=1}^M \ln G_k(b_k) \\ \text{s.t.} & \theta(\boldsymbol{b}, \nu) \le \boldsymbol{\eta}, \qquad \boldsymbol{b} = (b_1, \ldots, b_M) \\ & \boldsymbol{b} \in \mathbb{R}^M,\ \nu \in \mathcal{D}. \end{array} \tag{11}$$

Let $\mathcal{B}^\star$ be the hyperrectangular subset in $\mathcal{F}$ with boundary value $\boldsymbol{b}^\star$ and $\psi^\star$ be the constant function with $\psi^\star(\boldsymbol{\sigma}) = v^\star$ for $\boldsymbol{\sigma} \in \mathbb{R}^M$. Then, $(\mathcal{B}^\star, \psi^\star)$ is an optimal solution to (10).

This theorem allows us to obtain a simplified form of (9), for which a static arrangement is optimal.

Corollary 1. *Let $(\boldsymbol{d}^\star, \mu^\star) \in \mathcal{S} \times \mathcal{A}$ be an optimal solution to the following problem:*

$$\begin{array}{ll} \max & \displaystyle\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} \ln F_{ij}(d_{ij}) \\ \text{s.t.} & w_k(\boldsymbol{d}, \mu) \le \tau_k, \quad k \in \mathcal{W},\ \boldsymbol{d} = (d_{ij})_{i\in\mathcal{I}, j\in\mathcal{J}} \\ & \boldsymbol{d} \in \mathcal{S},\ \mu \in \mathcal{A}. \end{array} \tag{12}$$

Let $\mathcal{Q}^\star$ be the hyperrectangular uncertainty set in $\mathcal{H}$ with boundary value $\boldsymbol{d}^\star$ and $\pi^\star$ be the static arrangement with $\pi^\star(\boldsymbol{s}) = \mu^\star$ for $\boldsymbol{s} \in \mathcal{S}$. Then, $(\mathcal{Q}^\star, \pi^\star)$ is an optimal solution to (9).

Thanks to the hyperrectangular uncertainty sets, the hybrid optimization problem (9) has a computationally amiable form given by (12). With these hyperrectangular sets, computing the joint probability in (9) is reduced to the double summation in (12) without the need for high-dimensional integration. For a given uncertainty set, because 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 by Bertsimas and Sim (2004) and ellipsoidal uncertainty sets by Ben-Tal et al. (2004), may not lead to more tractable forms as our hybrid approach does.

A feasible solution $(\mathcal{Q}, \pi)$ to (9) may result in different admissible schedules for different realizations of consultation times. In this sense, it is an adjustable robust formulation analogous to the formulation proposed by Ben-Tal et al. (2004) for uncertain linear programs. In general, an adjustable robust formulation is less conservative than the nonadjustable counterpart, yielding better objective values. Corollary 1, however, implies that we may solve a nonadjustable formulation to obtain an optimal solution to (9), which is an admissible schedule invariant for all realizations of consultation times. This is because the worst cases are identical in both formulations. The hybrid optimization problem (9) may also have adjustable solutions, which are *nonstatic*, nonanticipative arrangements. As pointed out by de Ruiter et al. (2016), an adjustable solution may outperform the nonadjustable solution in terms of mean objective values, even if they are both optimal in the worst case. Hence, there could be adjustable solutions to (9) that perform better than the static solution to (12) in patient scheduling. Because the representation of nonanticipative arrangements is complicated, finding such an adjustable solution is generally difficult (Section EC.1 in the e-companion).

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 (9) does not have a feasible solution. Because the waiting time given by (4) is increasing with each consultation time, we may determine the feasibility of (9) by examining admissible schedules when all consultation times take their smallest possible values (i.e., $\underline{\boldsymbol{s}} = (\underline{s}_{ij})_{i\in\mathcal{I}, j\in\mathcal{J}}$). More specifically, we may solve the following optimization problem:

$$\begin{array}{ll} \min & \alpha \\ \text{s.t.} & w_k(\underline{\boldsymbol{s}}, \mu) \le \tau_k + \alpha, \quad k \in \mathcal{W},\ \underline{\boldsymbol{s}} = (\underline{s}_{ij})_{i\in\mathcal{I}, j\in\mathcal{J}} \\ & \mu \in \mathcal{A}, \end{array} \tag{13}$$

and we determine the feasibility of (9) by the proposition below.

Proposition 2. *Let $\alpha^\star$ be the minimum value of α given by (13). Then, the hybrid optimization problem (9) has a feasible solution if and only if $\alpha^\star \le 0$.*

5. A Mixed Integer Program

The equivalent form (12) can be 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:

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

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. Because the first position of each queue is for the patient who is being seen or to be seen immediately by the physician, we have

$$x_{ij1} = 1 \quad \text{for } i \in \mathcal{C}_j \text{ and } j \in \mathcal{J}. \tag{14}$$

Furthermore, because returning patients must be seen by their initial physicians, we have

$$\sum_{\ell=1}^{|\mathcal{I}|} x_{ij\ell} = 1 \quad \text{for } i \in \mathcal{R}_j \text{ and } j \in \mathcal{J}. \tag{15}$$

Under an admissible schedule, each waiting patient can be assigned to only one position:

$$\sum_{j\in\mathcal{J}}\sum_{\ell=1}^{|\mathcal{I}|} x_{ij\ell} = 1 \quad \text{for } i \in \mathcal{W}, \tag{16}$$

and each position can accommodate at most one patient:

$$\sum_{i\in\mathcal{I}_j} x_{ij\ell} \le 1 \quad \text{for } j\in\mathcal{J} \text{ and } \ell = 1,\dots,|\mathcal{I}|, \tag{17}$$

where $\mathcal{I}_j = \mathcal{N}\cup\mathcal{C}_j\cup\mathcal{R}_j$ 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, and therefore, empty positions can appear only at the end of the queue. Because the ℓth position by physician j has a patient if and only if equality holds in (17), the above constraint is equivalent to

$$\sum_{i\in\mathcal{I}_j} x_{ij(\ell+1)} \le \sum_{i\in\mathcal{I}_j} x_{ij\ell} \quad \text{for } j\in\mathcal{J} \text{ and } \ell = 1,\dots,|\mathcal{I}|-1. \tag{18}$$

One can check that (14)–(18) are equivalent to (1)–(3). In other words, each admissible schedule can be determined by a set of binary variables $\{x_{ij\ell} : i\in\mathcal{I}, j\in\mathcal{J}, \ell = 1,\dots,|\mathcal{I}|\}$ that satisfies (14)–(18).

Consider the delay constraints in (12). Put

$$\bar{\tau} = \max_{j\in\mathcal{J}}\left(\sum_{i\in\mathcal{I}_j}\bar{s}_{ij} - \min_{i\in\mathcal{I}_j}\bar{s}_{ij}\right), \tag{19}$$

where $\bar{s}_{ij}$ is the greatest value that $\tilde{s}_{ij}$ can take. Then, $\bar{\tau}$ is an upper bound of patient waiting times. Given a set of binary variables satisfying (14)–(18), we may write the delay constraints in (12) into

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

where $\mathcal{W}_j = \mathcal{N}\cup\mathcal{R}_j$ is the set of waiting patients eligible to be seen by physician j. In this inequality, the sum on the left side is the time that 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 $\sum_{i\in\mathcal{W}_j} x_{ij(m+1)} = 1$, and inequality (20) becomes

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

Because 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, $\sum_{i\in\mathcal{W}_j} x_{ij(m+1)} = 0$, and the first sum on the right side of (20) becomes zero. Then, inequality (20) 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 (20) in canonical form. By introducing a set of variables $\{u_{ij\ell}\ge 0 : i\in\mathcal{I}_j, j\in\mathcal{J}, \ell = 1,\dots,|\mathcal{I}|-1\}$, we may write (20) into two separate inequalities:

$$\sum_{\ell=1}^{m}\sum_{i\in\mathcal{I}_j} u_{ij\ell} \le \sum_{i\in\mathcal{W}_j} x_{ij(m+1)}\cdot\tau_i + \left(1-\sum_{i\in\mathcal{W}_j} x_{ij(m+1)}\right)\cdot\bar{\tau} \quad \text{for } j\in\mathcal{J} \text{ and } m = 1,\dots,|\mathcal{I}|-1 \tag{21}$$

and $u_{ij\ell}\ge x_{ij\ell}\cdot d_{ij}$ for $i\in\mathcal{I}_j$, $j\in\mathcal{J}$, and $\ell = 1,\dots,|\mathcal{I}|-1$. Because $x_{ij\ell}\in\{0,1\}$ and $d_{ij}\le\bar{s}_{ij}$, the latter inequality is equivalent to

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

Then, inequalities (21) and (22) specify the delay constraints in canonical form.

We have obtained the constraints for the patient scheduling problem given by (14)–(18), (21), and (22), which are equivalent to the constraints in (12). As suggested by the following theorem, we may convert 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 $n = 1,\dots,N_{ij}$. The hybrid optimization problem (9) can be written as the following mixed integer program:*

$$\begin{aligned}
\max\ & \sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}}\sum_{n=1}^{N_{ij}} y_{ij}(n)\cdot g_{ij}(n)\\
\text{s.t.}\ & \textit{constraints (14)–(18) and (21)}\\
& u_{ij\ell} \ge \sum_{n=1}^{N_{ij}} y_{ij}(n)\cdot s_{ij}(n) - (1-x_{ij\ell})\cdot\bar{s}_{ij}, \quad i\in\mathcal{I}_j, j\in\mathcal{J}, \ell = 1,\dots,|\mathcal{I}|-1\\
& \sum_{n=1}^{N_{ij}} y_{ij}(n) = 1, \quad i\in\mathcal{I}, j\in\mathcal{J}\\
& x_{ij\ell}, y_{ij}(n)\in\{0,1\}, \quad i\in\mathcal{I}, j\in\mathcal{J}, \ell = 1,\dots,|\mathcal{I}|, n = 1,\dots,N_{ij}\\
& u_{ij\ell}\ge 0, \quad i\in\mathcal{I}_j, j\in\mathcal{J}, \ell = 1,\dots,|\mathcal{I}|-1.
\end{aligned} \tag{23}$$

6. Dynamic Scheduling Using the Hybrid Approach

Patient arrivals at the ED form a stochastic process. For the scheduling system to make sequential decisions accordingly, the proposed hybrid approach must be incorporated in a dynamic scheduling framework. We assume that a decision iteration is triggered when either a physician completes a consultation or 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 need to determine delay targets for waiting patients when a decision iteration is triggered.

The major concern for new patients is their door-to-provider times, whereas that for returning patients is their lengths of stay. Assume that a decision iteration is triggered at time t. For $i \in \mathcal{W}$, let t_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 $i \in \mathcal{N}$, we take the delay target as $\tau_i = D_i - (t - t_i)$, which is the time until the patient's door-to-provider time exceeds the safety limit. Assume that a patient can return to the same physician at most B times. To determine delay targets for returning patients, we pick B positive numbers $(T_{i1}, \ldots, T_{iB})$ that satisfy $D_i < T_{i1} < \cdots < T_{iB} < K_i$ for $i \in \mathcal{R}$, where T_{im} is interpreted as the mandatory limit for the duration from patient i's arrival in the waiting area until he is seen by the physician for the $(m+1)$st time. If a returning patient $i \in \mathcal{R}$ is waiting to be seen for the $(m+1)$st time, we take the delay target as $\tau_i = T_{im} - (t - t_i)$. Imposing additional delay constraints enables us to carry out dynamic scheduling by sequentially solving (9). With the extra delay requirements, returning patients' waiting times for individual consultations can also be maintained at a reasonable level, which may further improve patients' safety and satisfaction. The specific values of these additional mandatory limits will influence the performance of dynamic scheduling. It would be desirable if delay targets for returning patients can be adjusted in each iteration according to the ED's congestion. Because finding the optimal delay target for each returning patient is generally difficult, we will use a heuristic approach to determining these parameters in our implementation; Section 7.3 has more details.

Given the delay targets for current waiting patients, the scheduling system will first determine the feasibility of the hybrid optimization problem (9) by solving (13) (which may also be converted into a mixed integer program according to the procedure in Section 5). If the feasible set of (9) is nonempty, the scheduling system will solve the mixed integer program (23), the optimal solution to which specifies the next patient to be seen for the available physician. If problem (9) turns out to be infeasible, the admissible schedule obtained by solving (13) 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.

We would like to point out that this iterative scheduling procedure is *myopic* in nature. Therefore, admissible schedules obtained from consecutive iterations could be *inconsistent* (i.e., successive assignment decisions may not satisfy Bellman's principle of optimality). Because a dynamic programming formulation would be intractable, time inconsistency is practically inevitable in solving the patient scheduling problem. Delage and Iancu (2015) discussed time consistency issues arising from robust multistage decision making.

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 ED, 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. Although their conditions appear 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 Section 7.1. The computational performance of the hybrid approach is compared with that of the SAA method in Section 7.2. We assess the hybrid approach for dynamic patient scheduling in Section 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.

We divide level 3 patients into two categories. The first category includes patients with the most common acute illnesses, such as headache, upper respiratory tract infection, and acute gastritis, whereas the second category includes all other patients. In practice, the categorization of patients should be done at the triage stage according to each patient's symptoms and vital signs. Generally speaking, patients in the first category can be easily identified by the triage nurse, and the diagnosis and treatment of these cases are relatively simple. Nowadays, more and more EDs have implemented a program known as the *fast track*, in which patients having minor illnesses and injuries are identified by the triage nurse and sent to a dedicated area for medical care (Sanchez et al. 2006). Hence, patient

categorization can be readily implemented through an ED's triage process. We assume that each patient's category is known by the scheduling system when the patient arrives in the waiting area.

In the numerical study, we categorize all level 3 cases in the dataset into the two groups according to each one's diagnosis. There are about 40% of level 3 cases belonging to the first category. 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 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 consultation times. Because physicians may be heterogeneous, the empirical consultation time distributions of each category may differ for different physicians. In this case, the empirical distributions should be generated using the records of consultation times by each physician.

In the ED of this hospital, there are five 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 physicians who completed more than 3,000 cases, we selected five physicians and examined the records of patients seen by them. No significant differences have been found among the five patient groups. We also examined the physicians' consultation times when they were on day and night shifts. The distribution of a physician's consultation times does not seem to change with time greatly. The boxplot of consultation times by the five physicians is shown in Figure 3. We can see that the distributions of consultation times differ a lot across these physicians. For instance, the median consultation time by physician 2 is just one-half of that by physician 5, whereas the variability in consultation times by physician 3 is much less than that by any other physician. For a patient scheduling approach to be relevant to practice, the heterogeneity of physicians must be taken into account.

Figure 2. (Color online) Histograms of Consultation Times of the Two Categories

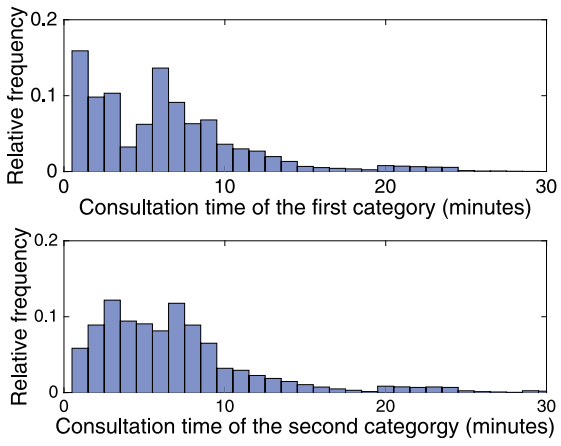
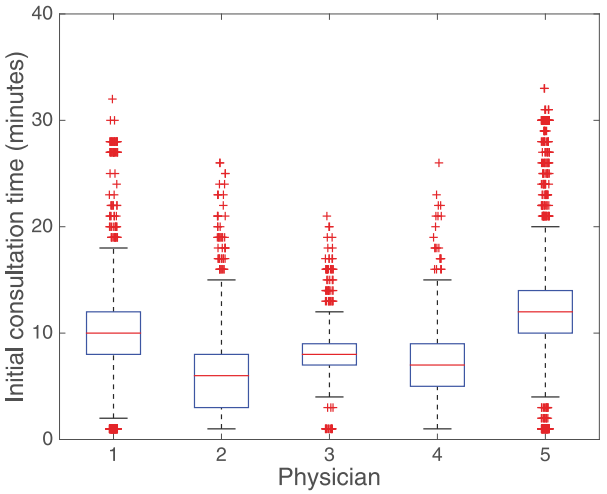


Figure 3. (Color online) Boxplot of Consultation Times by Five Physicians



7.2. Comparison with the SAA Method

The admissible schedule obtained by solving (9) is in general not the optimal solution to the static P model problem (7). Although finding the exact optimal solution to (7) is difficult, it is possible to use approximate approaches, 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 that of the SAA method.

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

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

Then, $\mathbb{E}(\tilde{\chi}(\mu)) = \mathbb{P}(w_k(\tilde{\mathbf{s}}, \mu) \le \tau_k : k \in \mathcal{W})$. Let $\{\mathbf{s}_1, \ldots, \mathbf{s}_N\}$, where $\mathbf{s}_n = (s_{ij}^n)_{i \in \mathcal{I}, j \in \mathcal{J}}$ for $n = 1, \ldots, N$, be a sample of consultation time vectors independently taken from the distribution of $\tilde{\mathbf{s}}$ and $\{\chi_1(\mu), \ldots, \chi_N(\mu)\}$ be the corresponding realizations of $\tilde{\chi}(\mu)$. By the strong law of large numbers, the probability of all waiting patients meeting 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 (7) by

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

where

$$\bar{\tau}_n = \max_{j \in \mathcal{J}} \left(\sum_{i \in \mathcal{I}_j} s_{ij}^n - \min_{i \in \mathcal{I}_j} s_{ij}^n \right)$$

is an upper bound of patient waiting times. When $z_n = 0$, the inequality in (24) always holds for all $k \in \mathcal{W}$ and $\mu \in \mathcal{A}$; when $z_n = 1$, the inequality holds for all $k \in \mathcal{W}$ 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 N is large, the admissible schedule that maximizes the objective function in (24) should be a near-optimal solution to (7). Following the procedure in Section 5, one can also convert the SAA formulation into a mixed integer program.

We consider a scenario with 6 physicians and 20 patients in the ED. Eight and 12 patients belong to the first and second categories, respectively. There are 3 new patients, 12 returning patients, and 5 patients being seen by physicians. Each physician has 2 returning patients waiting to be seen. We assume that all new patients have the same delay target τ_N and that all returning patients have the same delay target τ_R. All patients will leave the ED when they complete their current consultations. The six physicians have identical work capabilities, and therefore, the distribution of each consultation time depends on the patient's category only. All consultation times are sampled from the empirical distributions in Figure 2 according to patients' categories.

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 static P model problem (7) at the expense of longer computation time. We evaluate the SAA formulation (24) with different delay targets and different sample sizes. With each set of parameters, we take eight realizations of the random sample and then solve (24) using each realization. Computation time is the major concern of this step. The obtained admissible schedule is then evaluated by Monte Carlo simulation, where consultation times are resampled from the empirical distributions. In the simulation, the patients are seen by the physicians according to the obtained admissible schedule. The probability of all patients meeting their delay targets is computed as the performance measure through 1,000 independent simulation runs.

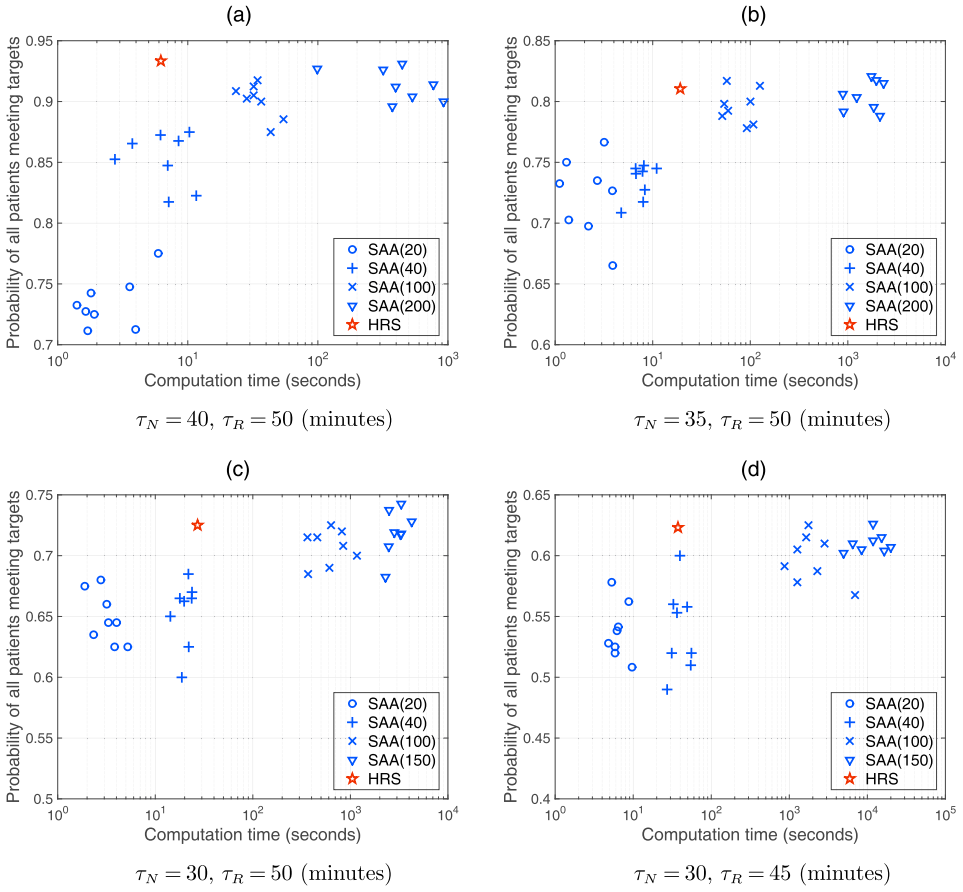
We depict the performance of the SAA method in Figure 4, where SAA(N) denotes the optimal solution to (24) based on a realization of sample size N. Because the computation time of the SAA method seems to increase exponentially with the sample size, we use a logarithmic scale for computation time in Figure 4. When the pair of delay targets is taken to be $(\tau_N, \tau_R) = (40, 50)$ and $(35, 50)$ minutes, we test the SAA formulation with sample sizes $N = 20, 40, 100$, and 200. If the sample size is small, the solutions exhibit great variability in performance across different realizations, and most of them are not satisfactory. Increasing the sample size can stabilize and improve the performance of solutions at the expense of longer computation time. In solving the mixed integer program for the SAA formulation, the computation time tends to increase when the delay targets are shorter. When the pair of delay targets is $(\tau_N, \tau_R) = (30, 50)$ and $(30, 45)$ minutes, it becomes difficult to obtain the optimal solution to (24) within hours for $N = 200$. Instead, we test these two cases with $N = 20$, 40, 100, and 150. When the sample size is $N = 150$, it took up to six hours to obtain the optimal solution for $(\tau_N, \tau_R) = (30, 45)$ minutes.

We also illustrate the performance of the hybrid robust-stochastic approach, which is denoted by HRS in Figure 4. The solution to the hybrid formulation does not depend on specific realizations of a random sample, and therefore, no variability is present in the performance of this approach. Although the objective function in (9) is different from that of the static P model problem (7), the optimal solution to the hybrid formulation outperforms most of the solutions from the SAA realizations. In Figure 4, the probability of target attainment by an SAA solution is comparable with that by the HRS solution only when the sample size is large. Although there are several SAA realizations producing better solutions than the hybrid approach, there is no SAA solution outperforming the corresponding HRS solution by more than 2% in terms of the probability of target attainment. The SAA method requires much longer computation time. To achieve comparable performance, it may take hours to obtain a solution from an SAA realization, whereas it only takes tens of seconds to obtain the HRS solution. When the sample size is large, the SAA method cannot be used for patient scheduling on a real-time basis. Hence, we will use the scheduling policy proposed by Huang et al. (2015), which is much more computationally efficient, as the benchmark policy for dynamic patient scheduling.

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 targets. Because 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 below, we generate a stream of 5,000 patients arriving at the ED according to a Poisson process with a rate of 15.2 patients per hour. For the convenience of simulation, we

Figure 4. (Color online) Computational Performance of the Hybrid Approach and the SAA Method with Different Delay Targets



assume that all medical tests and treatments by a nurse take no time. Hence, if a patient returns, he will join the queue for the same physician immediately after 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 physicians at least once before leaving the ED, whereas there are less than 4% of patients returning more than three times. In the numerical experiments, we assume that a patient may return at most three times and that 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 of 93.57%.

To solve the mixed integer program (23) sequentially, we need to specify mandatory limits for returning patients when a decision iteration is triggered. Because all patients have identical requirements for door-to-provider times and lengths of stay, we identify three time limits (T_1, T_2, T_3) for the durations from a patient's arrival until he starts the second, third, and fourth consultations, respectively. The selection of these limits will influence the performance of the scheduling algorithm. On the one hand, increasing these limits will accommodate more returning patients within delay targets for their current consultations, thus allowing more new patients to meet the safety limit for door-to-provider times. On the other hand, with larger mandatory limits, patients having multiple consultations will be more likely to exceed their length of stay target. To deal with these two concerns, we use a simple heuristic approach to determining mandatory limits for returning patients. Because most patients will return at least once, the balance of consultations between new and returning patients can be controlled by dynamically adjusting T_1, the mandatory limit for patients returning for the first time. More specifically, when a decision iteration is triggered, we take

$$T_1 = T_L + (T_U - T_L)\frac{|\mathcal{N}|}{|\mathcal{N}| + |\mathcal{R}^{(1)}|}, \tag{25}$$

where T_L and T_U are two given positive numbers with $T_L < T_U$, $|\mathcal{N}|$ is the number of new patients waiting in

the ED, and $|\mathcal{R}^{(1)}|$ is the number of patients waiting for their second consultations. If $|\mathcal{N}| = |\mathcal{R}^{(1)}| = 0$, we set $T_1 = T_L$. Following (25), when there are more new patients than patients returning for the first time, we will increase the mandatory limit for patients' second consultations, which allows more new patients to be served within their door-to-provider time limit. When there are more patients returning for the first time, we will decrease T_1, allowing more returning patients to be served within their length of stay target. In practice, we take both T_L and T_U to be several times longer than the safety limit for door-to-provider times, so that most new patients can be seen quickly. After T_1 is determined, we specify two positive numbers Δ_1 and Δ_2 and set $T_2 = T_1 + \Delta_1$ and $T_3 = T_2 + \Delta_2$. For the convenience of implementation, we assume that Δ_1 and Δ_2 are invariant in all iterations. We require T_3, the mandatory limit for the fourth consultations, to be sufficiently lower than the limit for lengths of stay.

In the first numerical example, we assume that the four physicians have identical work capabilities, and therefore, the distribution of a patient's consultation time does not depend on specific physicians. We evaluate the hybrid approach with $\Delta_1 = \Delta_2 = 30$ minutes, reporting the performance for different pairs of (T_L, T_U) in Table 1. As one may expect, increasing T_L and T_U will generally reduce door-to-provider violations but result in more length of stay violations. We compare the performance of the hybrid approach with that of the following scheduling policies.

Global FCFS. When a physician finishes a consultation, the scheduling system will send the patient who has the earliest registration time among those eligible to be seen by this physician. Because 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 the physician; 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 Benchmark Policy. This policy was 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 limit (i.e., if there exists $i \in \mathcal{N}$, such that $D - (t - t_i) < \epsilon$, where $\epsilon > 0$ is a given excess time). The scheduling system will give priority to new patients if such a patient is found, and it will give priority to returning patients otherwise. 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. (2015) used the *shortest deadline first* policy for new patients, which is reduced to the FCFS policy when all patients have the same door-to-provider limit.] If a returning patient is to be served, the scheduling system will select the one with the 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 completion will be sent to the available physician. [Huang et al. (2015) adopted a generalized $c\mu$ rule for returning patients to minimize the cumulative congestion cost. Their policy is reduced to the *closest to exit first* policy if a patient's length of stay is regarded as his congestion cost.] Under the assumption that physicians are homogeneous, this scheduling policy is proved asymptotically optimal in minimizing the mean length of stay with constraints on door-to-provider times.

In the above three scheduling policies, we assume that, when a new patient arrives in the waiting area finding one or more physicians available, the scheduling

Table 1. Performance Comparison of Door-to-Provider Times ($\bar{V}$) and Lengths of Stay ($\bar{L}$) Under Various Scheduling Policies, with Arrival Rate of 15.2 Patients Per Hour and Four Homogeneous Physicians

Scheduling policy	$\bar{V}$	$\bar{L}$	$\%\,\bar{V} > 30$	$\%\,\bar{L} > 200$	% Violations
Global FCFS	37.64	52.17	45.88	0.46	45.88
New patients first	3.87	74.96	0.10	8.50	8.50
Benchmark, $\epsilon = 1$	20.59	59.64	36.52	4.46	38.28
Benchmark, $\epsilon = 3$	19.63	59.91	24.34	4.90	25.76
Benchmark, $\epsilon = 6$	18.39	61.63	13.66	6.48	15.92
Benchmark, $\epsilon = 10$	16.39	62.08	5.80	6.64	8.35
SAA (40)	18.66	61.27	22.12	4.96	24.08
HRS, $T_L = 90$, $T_U = 120$	13.31	63.67	12.80	2.28	13.15
HRS, $T_L = 100$, $T_U = 130$	13.59	63.55	12.10	2.48	13.09
HRS, $T_L = 105$, $T_U = 130$	13.40	64.31	12.54	2.44	13.26

system will randomly select a free physician, to whom the patient will be sent immediately.

The SAA Method. When a physician completes a consultation or when a new patient arrives in the waiting area finding at least one free physician, the scheduling system will solve (24) to determine the next patient to be seen for the available physician.

We compare several performance measures in Table 1, including the mean door-to-provider time (denoted by $\bar{V}$), the mean length of stay (denoted by $\bar{L}$), the percentage of door-to-provider times exceeding 30 minutes, the percentage of lengths of stay exceeding 200 minutes, and the percentage of patients who experience time violations (i.e., either their door-to-provider times exceed 30 minutes or their lengths of stay exceed 200 minutes). 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 the ED to meet the stringent time constraints. Although the percentage of patients who experience time violations is relatively low under the new patient first policy, the lengths of stay of such patients are extensively long, resulting in severe congestion in the ED. The benchmark policy is capable of striking a balance for these performance measures. As we discussed earlier, this policy depends on consultation time distributions only through their first moments. We report the performance of this policy when the excess time is taken to be $\epsilon = 1$, 3, 6, or 10 minutes. A larger excess time allows physicians to see more new patients within the door-to-provider time limit while bringing on more length of stay violations. As the excess time increases, the performance of the benchmark policy becomes more and more analogous to that of the new patients first policy. Huang et al. (2015) recommended that ϵ should be one order of magnitude smaller than the safety limit for door-to-provider times. For example, with $D = 30$ minutes, we may take $\epsilon = 3$ minutes as a practical option. The SAA method is tested with sample size being 40 for the overall computation time to be comparable with that of the hybrid approach. Although the SAA method requires longer computation times than the benchmark policy, its performance is merely comparable with that of the benchmark policy with $\epsilon = 3$ minutes. Because no clear advantages are shown, the SAA method will not be included in subsequent numerical experiments.

Denoted by HRS in Table 1, the hybrid robust-stochastic approach outperforms the benchmark policy in terms of mean door-to-provider time and the percentage of length of stay violations in all cases. It also leads to a greater percentage of patients meeting the door-to-provider requirement, when the benchmark policy takes $\epsilon = 1$, 3, or 6 minutes. Note that, when the excess time is large (e.g., $\epsilon = 6$ or 10 minutes), the percentage of patients meeting the length of stay target is not satisfactory under the benchmark policy.

The hybrid approach achieves a better balance between door-to-provider times and lengths of stay, because it can evaluate the influence of entire consultation time distribution, 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. This is because, with door-to-provider time constraints, the benchmark policy is asymptotically optimal in terms of mean length of stay. Our approach is aimed at complying with mandatory targets for lengths of stay, and the percentage of patients meeting the targets is usually regarded as a more important performance indicator out of safety concerns.

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 in category 2 whose door-to-provider time is about to exceed the safety limit. At that 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 tradeoff must be made between preventing an immediate time violation and reducing future crowding. The scheduling policy is required to evaluate the consequences of both actions. Unfortunately, the benchmark policy does not allow for heterogeneous physicians. It is no longer asymptotically optimal in reducing the mean length of stay when physicians have different expertise.

In the second example, the four physicians are assumed to be heterogeneous. Two of them are experts in treating cases in category 1 but are not good at category 2; the other two physicians are more experienced in category 2 but are not familiar with category 1. In the simulation, we generate *original* consultation times using the empirical distributions according to patients' categories, whereas 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. In addition to

the scheduling policies mentioned earlier, we also consider the following scheduling policy that is modified from the benchmark policy.[2]

The Modified Benchmark Policy. The scheduling system follows the benchmark policy in deciding whether to serve a new patient. That is, when a physician finishes a consultation at time t, the scheduling system will check if there exists $i \in \mathcal{N}$, such that $D - (t - t_i) < \epsilon$, where $\epsilon > 0$ is the given excess time. Priority will be given to new patients if such a patient is found and given to returning patients otherwise. If a returning patient will be served, the scheduling system follows the benchmark policy, selecting the one who is the closest to completion. If a new patient will be served, the physician's expertise should be taken into account. Assume that the physician is an expert in category 1. If the new patient who has the earliest registration time also belongs to category 1, the scheduling system will send this patient to the physician directly. If this patient belongs to category 2, the urgency of serving this patient should be evaluated against the loss of efficiency caused by prolonged consultations. For $k = 1, 2$, let C_k be the expected total consultation time (including possible future returns) of a patient in category k provided by this physician. Then, $C_2 - C_1$ is the additional time if the physician would serve a patient in category 2 instead of category 1. Let $\tau^{(k)}$ be the time until the earliest new patient in category k exceeds the door-to-provider limit (with $\tau^{(1)} = D$ if there are no new patients in category 1 waiting to be seen). Then, $\tau^{(1)} - \tau^{(2)}$, the difference between the two deadlines, can be used to measure the relative urgency of serving category 2 instead of category 1. Because the new patient who has the earliest registration time belongs to category 2, we must have $\tau^{(1)} \geq \tau^{(2)}$. Clearly, it would be more urgent to serve category 2 if $\tau^{(1)} - \tau^{(2)}$ is large. The scheduling system will determine the category to be served by comparing $\tau^{(1)} - \tau^{(2)}$ with a threshold value, which is proportional to the additional time that the physician needs for serving a patient in category 2. More specifically, if $\tau^{(1)} - \tau^{(2)} \leq \delta(C_2 - C_1)$ for some $\delta > 0$, the earliest new patient in category 1 should be sent to the physician; otherwise, the earliest new patient in category 2 should be seen. Note that we may have $C_2 - C_1 \leq 0$, which implies that the slow physician can also serve a patient in category 2 quickly. In this case, the physician will always serve the new patient who has the earliest registration time. This is reasonable, because it is well known that giving priority to patients with short consultation times will mitigate congestion. If the physician is an expert in category 2, the scheduling system may follow a similar procedure to determine the next patient to be seen using the same coefficient δ. This heuristic may also be extended to more than two patient categories.

Under the modified benchmark policy, the scheduling system will assign a patient to a slow physician only if the case is relatively urgent, thus making better use of each physician's expertise. Such a simple heuristic may lead to considerable performance improvement when physicians are heterogeneous. In Table 2, the global FCFS, new patients first, and benchmark policies do not differentiate physicians in making scheduling decisions. Under these policies, a physician's average work rate is identical to that in the previous example, and therefore, the system's performance does not show much difference. We evaluate the modified benchmark policy for $\epsilon = 3$, 6, and 10 minutes. Following the same rule in deciding whether to serve a new patient, the modified policy does not perform better than the benchmark policy in terms of door-to-provider time. Instead, it has the advantage of reducing lengths of stay by allowing physicians to serve patients in a more efficient way. We

Table 2. Performance Comparison of Door-to-Provider Times ($\bar{V}$) and Lengths of Stay ($\bar{L}$) Under Different Scheduling Policies, with Arrival Rate 15.2 Patients Per Hour and Four Heterogeneous Physicians

Scheduling policy	$\bar{V}$	$\bar{L}$	% $\bar{V} > 30$	% $\bar{L} > 200$	% Violations
Global FCFS	38.69	53.19	46.60	0.36	46.60
New patients first	3.88	72.19	0.12	8.14	8.14
Benchmark, $\epsilon = 1$	20.74	64.62	36.66	6.06	38.78
Benchmark, $\epsilon = 3$	20.06	63.32	25.84	6.20	28.92
Benchmark, $\epsilon = 6$	18.65	63.00	14.38	6.24	18.51
Benchmark, $\epsilon = 10$	16.39	63.12	11.54	7.04	13.75
Modified benchmark, $\epsilon = 3$, $\delta = 0.5$	21.36	50.24	28.62	1.32	28.93
Modified benchmark, $\epsilon = 3$, $\delta = 1$	21.16	49.14	29.46	1.18	29.46
Modified benchmark, $\epsilon = 3$, $\delta = 2$	21.21	48.98	31.92	1.34	31.92
Modified benchmark, $\epsilon = 6$, $\delta = 1$	20.04	49.11	24.18	1.38	24.42
Modified benchmark, $\epsilon = 10$, $\delta = 1$	17.67	49.06	18.92	2.12	18.92
HRS, $T_L = 90$, $T_U = 120$	11.63	56.68	10.14	1.82	10.76
HRS, $T_L = 100$, $T_U = 130$	9.99	56.70	6.76	1.30	6.76
HRS, $T_L = 105$, $T_U = 130$	9.42	54.23	6.68	1.16	6.72

test $\delta = 0.5$, 1, and 2 for $\epsilon = 3$ minutes, and we test $\delta = 1$ for $\epsilon = 6$ and 10 minutes. With a larger threshold value, the modified policy will send more patients to fast physicians, shortening the mean length of stay. However, with an increased δ, the modified policy may also postpone more consultations of new patients, although their door-to-provider times are about to exceed the safety limit. Therefore, increasing δ may neither shorten the mean door-to-provider time nor reduce door-to-provider violations. One may improve the performance on door-to-provider times by increasing ϵ, which however, may worsen the performance on lengths of stay.

As in the previous example, the hybrid approach can achieve a more balanced performance in complying with the mandatory targets. It outperforms the modified benchmark policy in terms of door-to-provider time while maintaining a comparable level of length of stay violations. We believe that this advantage also stems from the capability of evaluating the influence of entire distributions. Using the hybrid approach, the scheduling system may send patients to fast physicians with more appropriate timing, thus striking a desirable balance among all performance measures. The e-companion has more simulation results, where at a reasonable computational expense, the hybrid robust-stochastic approach outperforms other scheduling policies in a wide range of parameter regimes.

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 of the patient scheduling problem, by which we obtain a near-optimal solution to the corresponding P model problem at a significantly lower computational expense. Using this approach and real-time patient flow data, we developed a dynamic scheduling algorithm for making recommendations about the next patient to be seen by each available physician. Our hybrid robust-stochastic approach allows for practical features and outperforms existing scheduling policies in numerical experiments. The capability of scheduling in the presence of heterogeneous physicians, in particular, is a major advantage of this approach. In the future, we may include additional cost structures in the hybrid robust-stochastic formulation to answer more questions arising from patient flow management in EDs.

The proposed hybrid formulation may provide a computationally tractable approach to solving optimization problems in stochastic networks with delay or throughput time constraints. Such problems often arise from healthcare systems where service requirements are time sensitive, including patient transfer from EDs to inpatient wards (Mandelbaum et al. 2012, Shi et al. 2016), ambulance deployment (McLay and Mayorga 2013, Maxwell et al. 2014, Chong et al. 2016), and health examinations (Baron et al. 2017). Similar problems may arise from transportation systems, including 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, Jepsen et al. 2008).

In a recent paper, Jaillet et al. (2016) extended the proposed hybrid formulation to a class of satisficing problems that are typically computationally intractable. In that paper, a set of sufficient conditions is identified for a hybrid formulation to be equivalent to the corresponding P model. Unfortunately, those conditions do not apply to the patient scheduling problem in this paper. How to quantify the loss of optimality from the hyperrectangular approximation of uncertainty sets is an open problem for our future research.

Recent advances in distributionally robust optimization may also enable us to convert P model problems into tractable forms. As pointed out by Hanasusanto et al. (2015) and Hanasusanto et al. (2017), one may use distributionally robust formulations to mitigate the intractability of evaluating high-dimensional integrals (i.e., when the distribution of a random vector belongs to certain ambiguity sets, one may obtain the worst case probability of the random vector being in a given polyhedron by solving a linear or conic program). By those techniques, we may also convert the patient scheduling problem into a mixed integer program. Compared with the hybrid approach, a distributionally robust formulation would be particularly useful when either distributional information or patient flow data are limited.

Acknowledgments

The authors would like to thank the associate editor and the referees for their thoughtful comments and constructive suggestions, which led to a significantly improved paper. In particular, they are grateful to the anonymous referee who proposed the modified benchmark policy. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of the Singapore Ministry of Education or the Singapore Government.

Endnotes

[1] Level 4 is assigned to nonemergency 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.

[2] The modified benchmark policy was proposed by an anonymous referee.

References

Armony M, Israelit S, Mandelbaum A, Marmor YN, Tseytlin Y, Yom-Tov GB (2015) On patient flow in hospitals: A data-based queueing-science perspective. *Stochastic Systems* 5(1):146–194.

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

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

Baron O, Berman O, Krass D, Wang J (2017) Strategic idleness and dynamic scheduling in an open-shop service network: Case study and analysis. *Manufacturing Service Oper. Management* 19(1):52–71.

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

Bertsimas D, Sim M (2004) The price of robustness. *Oper. Res.* 52(1): 35–53.

Bertsimas D, Gamarnik D, Rikun AA (2011) Performance analysis of queueing networks via robust optimization. *Oper. Res.* 59(2): 455–466.

Bertsimas DJ, van Ryzin G (1993) Stochastic and dynamic vehicle routing in the Euclidean plane with multiple capacitated vehicles. *Oper. Res.* 41(1):60–76.

Charnes A, Cooper WW (1963) Deterministic equivalents for optimizing and satisficing under chance constraints. *Oper. Res.* 11(1): 18–39.

Chong KC, Henderson SG, Lewis ME (2016) The vehicle mix decision in emergency medical service systems. *Manufacturing Service Oper. Management* 18(3):347–360.

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

de Bruin AM, Bekker R, van Zanten L, Koole GM (2010) Dimensioning hospital wards using the Erlang loss model. *Ann. Oper. Res.* 178(1):23–43.

de Ruiter FJCT, Brekelmans RCM, den Hertog D (2016) The impact of the existence of multiple adjustable robust solutions. *Math. Programming A* 160(1–2):531–545.

Delage E, Iancu DA (2015) Robust multi-stage decision making. Aleman DM, Thiele AC, eds. *The Operations Research Revolution*, INFORMS TutORials in Operations Research (INFORMS, Catonsville, MD), 20–46.

Doytchinov B, Lehoczky J, Shreve S (2001) Real-time queues in heavy traffic with earliest-deadline-first queue discipline. *Ann. Appl. Probab.* 11(2):332–378.

Fisher ML, Jörnsten KO, Madsen OBG (1997) Vehicle routing with time windows: Two optimization algorithms. *Oper. Res.* 45(3): 488–492.

Gilboy N, Tanabe P, Travers D, Rosenan AM (2011) *Emergency Severity Index (ESI): A Triage Tool for Emergency Department Care*, version 4 (AHRQ Publications, Rockville, MD).

Green LV, Soares J, Giglio JF, Green RA (2006) Using queueing theory to increase the effectiveness of emergency department provider staffing. *Academic Emergency Medicine* 13(1):61–68.

Hanasusanto GA, Roitch V, Kuhn D, Wiesemann W (2015) A distributionally robust perspective on uncertainty quantification and chance constrained programming. *Math. Programming B* 151(1):35–62.

Hanasusanto GA, Roitch V, Kuhn D, Wiesemann W (2017) Ambiguous joint chance constraints under mean and dispersion information. *Oper. Res.* 65(3):751–767.

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

Huang J, Carmeli B, Mandelbaum A (2015) Control of patient flow in emergency departments, or multiclass queues with deadlines and feedback. *Oper. Res.* 63(4):892–909.

Jaillet P, Jena SD, Ng TS, Sim M (2016) Satisficing awakens: Models to mitigate uncertainty. Working paper, National University of Singapore, Singapore.

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

Laporte G, Louveaux FV, van Hamme L (2002) An integer *L*-shaped algorithm for the capacitated vehicle routing problem with stochastic demands. *Oper. Res.* 50(3):415–423.

Maglaras C, Van Mieghem JA (2005) Queueing systems with lead-time constraints: A fluid-model approach for admission and sequencing control. *Eur. J. Oper. Res.* 167(1):179–207.

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

Maxwell MS, Ni EC, Tong C, Henderson SG, Topaloglu H, Hunter SR (2014) A bound on the performance of an optimal ambulance redeployment policy. *Oper. Res.* 62(5):1014–1027.

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

McLay LA, Mayorga ME (2013) A dispatching model for server-to-customer systems that balances efficiency and equity. *Manufacturing Service Oper. Management* 15(2):205–220.

Nemirovski A, Shapiro A (2006) Convex approximations of chance constrained programs. *SIAM J. Optim.* 17(4):969–996.

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

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

Saghafian S, Hopp WJ, Van Oyen MP, Desmond JS, Kronick SL (2012) Patient streaming as a mechanism for improving responsiveness in emergency departments. *Oper. Res.* 60(5):1080–1097.

Saghafian S, Hopp WJ, Van Oyen MP, Desmond JS, Kronick SL (2014) Complexity-augmented triage: A tool for improving patient safety and operational efficiency. *Manufacturing Service Oper. Management* 16(3):329–345.

Sanchez M, Smally AJ, Grant RJ, Jacobs LM (2006) Effects of a fast-track area on emergency department performance. *J. Emergency Medicine* 31(1):117–120.

Seow KT, Dang NH, Lee DH (2010) A collaborative multiagent taxi-dispatch system. *IEEE Trans. Automation Sci. Engrg.* 7(3):607–616.

Shi P, Chou MC, Dai JG, Ding D, Sim J (2016) Models and insights for hospital inpatient operations: Time-dependent ED boarding time. *Management Sci.* 62(1):1–28.

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

Spacey SA, Wiesemann W, Kuhn D, Luk W (2012) Robust software partitioning with multiple instantiation. *INFORMS J. Comput.* 24(3):500–515.

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

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

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

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

Zhang X, Kamgarpour M, Georghiou A, Goulart P, Lygeros J (2017) Robust optimal control with adjustable uncertainty sets. *Automatica* 75:249–259.