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

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

Melvyn Sim  
Department of Analytics and Operations, NUS Business School, National University of Singapore, Singapore 119245  
melvynsim@nus.edu.sg

Meilin Zhang  
School of Business, Singapore University of Social Sciences, Singapore 599494  
zhangmeilin@suss.edu.sg

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

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

1. Introduction

Emergency department (ED) crowding and the consequential delays have been a worldwide issue and received considerable attention from governments, public media, and academic communities. ED crowding compromises the quality of and access to emergency care, putting patients at great risk of treatment errors. Numerous studies have revealed an association between crowding and increased morbidity and mortality in EDs (McHugh 2013). For hospitals, ED crowding damages their public reputation and incurs revenue loss due to ambulance diversion and patients’ leaving without being
seen. As pointed out by Rabin et al. (2012), widespread crowding also impedes hospitals’ ability to achieve national safety and quality goals, compromises the healthcare system, and limits the regional capacity for disaster response. In 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 CMS. To address the crowding issue, governments and regulatory organizations may set mandatory targets for emergency care. In 2005, England’s National Health Service mandated that 98% of ED patients must be treated and either discharged home or admitted to an inpatient ward within four hours of arrival. The implementation of this “four-hour rule” greatly improved the percentage of patients spending less than four hours in EDs, from 77.3% in 2002–2003 to 97.2% in 2008–2009 (Weber et al. 2011).

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

The focus of this paper is patient scheduling in EDs. From a modeling perspective, an ED can be viewed as a queueing network with medical units being the nodes, patients being the customers, and beds, medical staff, and equipments being the servers (Armony et al. 2015). Aside from prioritized customers and time-sensitive service requirements, this network is characterized by frequent returning routes of customers, i.e., after their initial consultations by a physician, most patients would undergo medical tests and return to the same physician before eventually being discharged or hospitalized. Although emergency physicians are required to provide treatment for a broad spectrum of illnesses and injuries, their expertise and work rates differ from one another. In
other words, the servers of this network are heterogeneous. When there are multiple patients waiting
to be seen, their respective physicians and the sequence of their consultations must be carefully
scheduled in order to meet the stringent door-to-provider and length-of-stay targets. However,
the aforementioned features, including the complex network structure, server heterogeneity, highly
uncertain patient arrival processes, and time-sensitive service requirements, all pose challenges in
solving the patient scheduling problem.

To address this problem in a practical setting, we propose a hybrid robust–stochastic approach
to exploiting patient flow data for real-time patient scheduling. Our intention is to maximize the
percentage of patients whose door-to-provider times and lengths of stay are within the manda-
tory targets. Since the patient arrival pattern is highly variable, we would refrain from making
assumptions, such as the arrival rate and the interarrival 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; see (6)–(7) and the discussion therein.
With the joint probability of target attainment being the objective, such an optimization problem
was first studied by Charnes and Cooper (1963), who termed this formulation the $P$-model. 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 non-convex 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 its consultation times being feasible. The schedule associated with the obtained
uncertainty set is the optimal solution to the hybrid formulation. For the computational reason,
we restrict the family of uncertainty sets to a collection of hyperrectangles. Then, under the inde-
pendence 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; see Section 7 for more details.

The hybrid robust–stochastic approach is of both practical and methodological importance. First, although the hybrid formulation is essentially a mixed integer program, solving this problem is practically efficient and allows real-time scheduling in EDs. As a dynamic approach driven by data, it allows for practical features such as a time-varying patient arrival process, general consultation time distributions, and heterogeneous physicians. In the literature on scheduling of queueing networks, these features are generally absent from 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. Besides patient scheduling, similar problems arise from other stochastic systems with time-sensitive service requirements; see Section 8 for 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 non-anticipative 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-face letters, respectively. Calligraphic letters are used for sets, such as $\mathcal{I}$, and we use $|\mathcal{I}|$ for the cardinality of the set. Random variables and vectors are denoted with a tilde mark, such as $\tilde{s}$ and $\tilde{s}$. 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; see Section 7.3 for comparison between these two approaches.

Although the aforementioned queueing models are able to represent basic operational characteristics of an ED, they may be overly simplistic and incapable of capturing some salient features. For a queueing model to be analytically tractable, one may require probabilistic assumptions such as exponential interarrival and service time distributions, stationary arrival processes, and homogeneous servers. As pointed out by Bertsimas et al. (2011), performance analysis of queueing networks is largely unsolvable without these assumptions. However, as the ED environment is complex and changes frequently, such assumptions may not be appropriate. 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; see, 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. In order 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 are 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 in order 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; see 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, so 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 in order 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 at 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, while others may undergo diagnostic tests, such as X-rays and blood tests, or to receive treatment by a nurse. When the test result is ready or the treatment is completed, the patient will return to the waiting area and become a returning patient, waiting to be examined by the same physician. A patient may see the same physician several times before eventually being discharged or hospitalized.
Figure 1 Queueing network model of patient flow in an ED.

The scheduling system will determine the assignment of patients to each physician and the order of their consultations. A new patient can be assigned to any available physician, while a returning patient must be seen by the physician he consulted initially. We assume that the scheduling system does not manage patients who are waiting for tests or treatment by a nurse, since those patients are typically served on the first-come, first-served (FCFS) basis. A patient scheduling system is critical for the mitigation of ED crowding, because simple prioritization rules are incapable of balancing door-to-provider times and lengths of stay. If physicians give priority to new patients 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. On the other hand, giving priority to returning patients can effectively shorten the queue length at 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 makes recommendations when a physician completes a consultation or a new patient 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 $J$ be the set of physicians, $\mathcal{N}$ the set of new patients, $\mathcal{C}$ the set of patients being seen, and $\mathcal{R}$ 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 patient being seen by physician $j$ and $\mathcal{R}_j$ 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 $\mathcal{I} = \mathcal{W} \cup \mathcal{C}$ the set of patients in the ED excluding those sent to tests or treatments.

For $i \in \mathcal{I}$ and $j \in \mathcal{J}$, let $\tilde{s}_{ij}$ be the consultation time of patient $i$ if he would be seen by physician $j$. For $i \in \mathcal{C}_j$, $\tilde{s}_{ij}$ 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 $\tilde{s}_{ij}$. Since the physicians may be heterogeneous, even though $i \in \mathcal{I}$ 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 whose values are taken from a finite set of positive numbers $S_{ij} = \{s_{ij}(1), \ldots, s_{ij}(N_{ij})\}$. We use $\underline{s}_{ij}$ and $\overline{s}_{ij}$ to denote the smallest and greatest numbers in $S_{ij}$. Let $\tilde{s} = (\tilde{s}_{ij})_{i \in \mathcal{I}, j \in \mathcal{J}}$ be the random vector of all these consultation times. Then, $\tilde{s}$ takes values from the product space $S = \prod_{i \in \mathcal{I}, j \in \mathcal{J}} S_{ij}$.

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 $i$. Since returning patients must be seen by their initial physicians, the assignment should satisfy

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

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,

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

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

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

The pair of assignment and sequencing decisions $(\varphi, \Phi)$ 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 \( s = (s_{ij})_{i \in I, j \in J} \in \mathcal{S} \), the waiting time of patient \( k \in \mathcal{W} \) can be obtained by

\[
w_k(s, (\varphi, \Phi)) = \sum_{\ell \in C_{\varphi(k)}} s_{\ell, \varphi(k)} + \sum_{\ell \in \Theta(k)} s_{\ell, \varphi(k)}, \tag{4}
\]

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 \( C_{\varphi(k)} = \emptyset \)) 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 \( \tau_i \). For a new patient \( i \in \mathcal{N} \), \( \tau_i \) 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 \( i \in \mathcal{R} \), \( \tau_i \) is specified by the scheduling system in order 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, since 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 \( \pi: \mathcal{S} \rightarrow \mathcal{A} \), 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

\[
\max \quad \mathbb{P}(w_i(\tilde{s}, \pi(\tilde{s})) \leq \tau_i : i \in \mathcal{W}) \\
\text{s.t.} \quad \pi \in \mathcal{V}. \tag{5}
\]

This formulation, however, cannot be implemented because to determine the admissible schedule, one is required to know the realization of \( \tilde{s} \) in advance. To fix this issue, we should confine feasible solutions to (5) within the set of non-anticipative arrangements, which do not rely on future information to determine the patients to be seen when physicians become available. In order to specify a non-anticipative arrangement, we need to determine the assignment and sequencing decisions in a sequential manner. Let \( w(1) \leq w(2) \leq \cdots \) 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 the physicians have been with their current patients, etc. Using the cumulative information, we may define non-anticipative arrangements through a recursive procedure. Since the construction of non-anticipative arrangements is generally
complicated, we leave the details to the e-companion. Let $\mathcal{V}_1$ be the set of all non-anticipative arrangements. We may obtain an optimal non-anticipative arrangement by solving the problem
\[
\max \ P(w_i(\tilde{s}, \pi(\tilde{s})) \leq \tau_i : i \in \mathcal{W})
\]
\[\text{s.t.} \quad \pi \in \mathcal{V}_1. \]
\[(6)\]

The $P$-model problem (6) turns out to be intractable, owing in part to the curse of dimensionality induced by the recursive structure of non-anticipative arrangements. To simplify the computation, we may further confine feasible solutions to (5) within the set of static arrangements, i.e.,
\[
\mathcal{V}_0 = \{\pi \in \mathcal{V} : \pi(s_1) = \pi(s_2) \text{ for } s_1, s_2 \in \mathcal{S}\}.
\]

The next proposition states that static arrangements are non-anticipative.

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

Since 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
\[
\max \ P(w_i(\tilde{s}, \mu) \leq \tau_i : i \in \mathcal{W})
\]
\[\text{s.t.} \quad \mu \in \mathcal{A}. \]
\[(7)\]

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 $C_{\varphi^\dagger(i)} = \emptyset$ 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 at 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, even though solving (7) is much simpler than solving (6). Under an admissible schedule, evaluating the joint probability in (7) involves multi-dimensional 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 $\mathbb{R}_+^{|I||J|} \times \mathcal{A}$. Under $\mu \in \mathcal{A}$, the set

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

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

$$\max \mathbb{P}(\tilde{s} \in \mathcal{X}(\mu)) \quad \text{s.t.} \quad \mu \in \mathcal{A},$$

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

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

because the entries of $\tilde{s}$ 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 whose associated convex polyhedron has the largest probability measure induced by $\tilde{s}$. If the probability measure of a convex polyhedron is large, we may expect the polyhedron to contain a hyperrectangular subset whose probability measure is also large. Conversely, if we can find an admissible schedule whose associated convex polyhedron contains a “large” hyperrectangle, we may also expect the polyhedron itself to be “large”. Hence, instead of searching for the admissible schedule that has the “largest” convex polyhedron, we would find the admissible schedule whose associated convex polyhedron has the “largest” hyperrectangular subset. Since it is far easier to evaluate the probability measure of a hyperrectangle, the patient scheduling problem would be more computationally amiable under this formulation.

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

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

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

$$\max \mathbb{P}(\tilde{s} \in Q) \quad \text{s.t.} \quad w_k(s, \pi(s)) \leq \tau_k, \quad k \in \mathcal{W}, \; s \in \mathcal{Q} \quad Q \in \mathcal{H}, \; \pi \in \mathcal{V}_1.$$ (9)
From a robust optimization perspective, $\mathcal{H}$ can be regarded as a family of uncertainty sets for $\tilde{s}$, and the specific uncertainty set $Q$ can be adjusted within $\mathcal{H}$ using different arrangements. The objective function in (9) involves random variables, while the constraints are based on a robust optimization formulation whose uncertainty set is adjustable. Hence, we refer to this formulation as a hybrid robust–stochastic approach.

Since the representation of non-anticipative 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 $\tilde{\sigma} = (\tilde{\sigma}_1, \ldots, \tilde{\sigma}_M)$ 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 $F$ to denote the collection of closed hyperrectangular subsets of $\mathbb{R}^M$ that are bounded from above, i.e., $F = \{ \prod_{k=1}^M (-\infty, b_k] : (b_1, \ldots, b_M) \in \mathbb{R}^M \}$. Let $\mathcal{D}$ be a nonempty set and $\mathcal{G}$ the collection of functions from $\mathbb{R}^M$ to $\mathcal{D}$. Let $\theta$ be a function from the product space $\mathbb{R}^M \times \mathcal{D}$ to $\mathbb{R}^L$ such that for $\sigma = (\sigma_1, \ldots, \sigma_M) \in \mathbb{R}^M$ and $\nu \in \mathcal{D}$, $\theta(\sigma, \nu)$ is nondecreasing in each $\sigma_k$. Then for a given $\eta \in \mathbb{R}^L$, let us consider the following problem

$$\max \ P(\tilde{\sigma} \in \mathcal{B})$$
$$\quad \text{s.t.} \quad \theta(\sigma, \psi(\sigma)) \leq \eta, \quad \sigma \in \mathcal{B}$$
$$\quad \mathcal{B} \in \mathcal{F}, \quad \psi \in \mathcal{G}.$$  \hspace{1cm} (10)

Theorem 1 provides a simplified form of (10).

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

$$\max \ \sum_{k=1}^M \ln G_k(b_k)$$
$$\quad \text{s.t.} \quad \theta(b, \nu) \leq \eta, \quad b = (b_1, \ldots, b_M)$$
$$\quad b \in \mathbb{R}^M, \quad \nu \in \mathcal{D}.$$  \hspace{1cm} (11)

Let $\mathcal{B}^*$ be the hyperrectangular subset in $\mathcal{F}$ with boundary value $b^*$ and $\psi^*$ the constant function with $\psi^*(\sigma) = \nu^*$ for $\sigma \in \mathbb{R}^M$. Then, $(\mathcal{B}^*, \psi^*)$ 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 $(d^*, \mu^*) \in \mathcal{S} \times \mathcal{A}$ be an optimal solution to the following problem:

$$\max \ \sum_{i \in I} \sum_{j \in J} \ln F_{ij}(d_{ij})$$
$$\quad \text{s.t.} \quad w_k(d, \mu) \leq \tau_k, \quad k \in W, \quad d = (d_{ij})_{i \in I, j \in J}$$
$$\quad d \in \mathcal{S}, \quad \mu \in \mathcal{A}.$$  \hspace{1cm} (12)
Let $Q^*$ be the hyperrectangular uncertainty set in $H$ with boundary value $d^*$ and $\pi^*$ the static arrangement with $\pi^*(s) = \mu^*$ for $s \in S$. Then, $(Q^*, \pi^*)$ 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, since the worst case occurs only when all consultation times take their largest values, we would be required to examine admissible schedules under this single scenario only. It is also worth mentioning that existing robust optimization formulations with adjustable uncertainty sets, such as the budget of uncertainty 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 $(Q, \pi)$ 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 non-adjustable counterpart, yielding better objective values. Corollary 1, however, implies that we may solve a non-adjustable 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 *non-static*, non-anticipative arrangements. As pointed out by de Ruiter et al. (2016), an adjustable solution may outperform the non-adjustable 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. Since the representation of non-anticipative arrangements is complicated, finding such an adjustable solution is generally difficult (see 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. Since 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., $\underline{s} = (\underline{s}_{ij})_{i \in I, j \in J}$. More specifically, we may solve the following optimization problem

$$
\begin{align*}
\min & \quad \alpha \\
\text{s.t.} & \quad w_k(\underline{s}, \mu) \leq \tau_k + \alpha, \quad k \in W, \quad \underline{s} = (\underline{s}_{ij})_{i \in I, j \in J} \\
& \quad \mu \in \mathcal{A}, \\
\end{align*}
$$

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

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

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

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

$$x_{ij1} = 1 \quad \text{for } i \in C_j \text{ and } j \in J.$$\hspace{1cm} (14)

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

$$\sum_{j \in J} |I| \sum_{\ell = 1}^{\ell + 1} x_{ij\ell} = 1 \quad \text{for } i \in R_j \text{ and } j \in J.$$\hspace{1cm} (15)

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

$$\sum_{j \in J} \sum_{\ell = 1}^{\ell + 1} x_{ij\ell} = 1 \quad \text{for } i \in W,$$\hspace{1cm} (16)

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

$$\sum_{i \in I_j} x_{ij\ell} \leq 1 \quad \text{for } j \in J \text{ and } \ell = 1, \ldots, |I|,$$\hspace{1cm} (17)

where $I_j = N \cup C_j \cup R_j$ 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, so empty positions can appear only at the end of the queue. Since the $\ell$th position by physician $j$ has a patient if and only if equality holds in (17), the above constraint is equivalent to

$$\sum_{i \in I_j} x_{ij(\ell + 1)} \leq \sum_{i \in I_j} x_{ij\ell} \quad \text{for } j \in J \text{ and } \ell = 1, \ldots, |I| - 1.$$\hspace{1cm} (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 $\{x_{ij\ell} : i \in I, \ j \in J, \ \ell = 1, \ldots, |I|\}$ that satisfies (14)–(18).
Consider the delay constraints in (12). Put
\[
\bar{\tau} = \max_{j \in J} \left( \sum_{i \in I_j} \bar{s}_{ij} - \min_{i \in I_j} \bar{s}_{ij} \right)
\] (19)
where \( \bar{s}_{ij} \) is the greatest value \( \tilde{s}_{ij} \) 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
\[
\sum_{\ell=1}^{m} \sum_{i \in I_j} x_{ij\ell} \cdot d_{ij} \leq \sum_{i \in W_j} x_{ij(m+1)} \cdot (1 - \sum_{i \in W_j} x_{ij(m+1)}) \cdot \bar{\tau} \quad \text{for } j \in J \text{ and } m = 1, \ldots, |I| - 1,
\] (20)
where \( W_j = N \cup R_j \) 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 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 \( \sum_{i \in W_j} x_{ij(m+1)} = 1 \), and inequality (20) becomes
\[
\sum_{\ell=1}^{m} \sum_{i \in I_j} x_{ij\ell} \cdot d_{ij} \leq \sum_{i \in W_j} x_{ij(m+1)} \cdot \tau_i.
\]
Since the sum on the right side is equal to the delay target, this inequality is the delay constraint for the \((m + 1)\)st patient. If there is no patient in the \((m + 1)\)st position, \( \sum_{i \in W_j} x_{ij(m+1)} = 0 \) and the first sum on the right side of (20) becomes zero. Then, inequality (20) turns out to be
\[
\sum_{\ell=1}^{m} \sum_{i \in I_j} x_{ij\ell} \cdot d_{ij} \leq \bar{\tau},
\]
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 \( \{ u_{ij\ell} \geq 0 : i \in I_j, \ j \in J, \ \ell = 1, \ldots, |I| - 1 \} \), we may write (20) into two separate inequalities
\[
\sum_{\ell=1}^{m} \sum_{i \in I_j} u_{ij\ell} \leq \sum_{i \in W_j} x_{ij(m+1)} \cdot \tau_i + \left( 1 - \sum_{i \in W_j} x_{ij(m+1)} \right) \cdot \bar{\tau} \quad \text{for } j \in J \text{ and } m = 1, \ldots, |I| - 1
\] (21) and \( u_{ij\ell} \geq x_{ij\ell} \cdot d_{ij} \) for \( i \in I_j, \ j \in J, \) and \( \ell = 1, \ldots, |I| - 1 \). Since \( x_{ij\ell} \in \{0, 1\} \) and \( d_{ij} \leq \bar{s}_{ij} \), the latter inequality is equivalent to
\[
u_{ij\ell} \geq d_{ij} - (1 - x_{ij\ell}) \cdot \bar{s}_{ij} \quad \text{for } i \in I_j, \ j \in J, \ and \ \ell = 1, \ldots, |I| - 1.
\] (22)
Then, inequalities (21)–(22) specify the delay constraints in canonical form.

We have obtained a set of constraints for the patient scheduling problem, given by (14)–(18) and (21)–(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 \( i \in I, j \in J, \) and \( n = 1, \ldots, N_{ij} \). The hybrid optimization problem (9) can be written as the following mixed integer program

\[
\begin{align*}
\max & \quad \sum_{i \in I} \sum_{j \in J} \sum_{n=1}^{N_{ij}} y_{ij}(n) \cdot g_{ij}(n) \\
\text{s.t.} & \quad \text{constraints (14)–(18) and (21)} \\
& \quad u_{ij\ell} \geq \sum_{n=1}^{N_{ij}} y_{ij}(n) \cdot s_{ij}(n) - (1 - x_{ij\ell}) \cdot \bar{s}_{ij}, \quad i \in I, j \in J, \ell = 1, \ldots, |I| - 1 \\
& \quad \sum_{n=1}^{N_{ij}} y_{ij}(n) = 1, \quad i \in I, j \in J \\
& \quad x_{ij\ell}, y_{ij}(n) \in \{0, 1\}, \quad i \in I, j \in J, \ell = 1, \ldots, |I|, n = 1, \ldots, N_{ij} \\
& \quad u_{ij\ell} \geq 0, \quad i \in I, j \in J, \ell = 1, \ldots, |I| - 1.
\end{align*}
\]

6. Dynamic Scheduling Using the Hybrid Approach

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

To solve the dynamic scheduling problem by the hybrid approach, we 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, while that for returning patients is their lengths of stay. Assume that a decision iteration is triggered at time \( t \). For \( i \in W \), 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 \( i \in N \), 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 exceeding 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 satisfies \( D_i < T_{i1} < \cdots < T_{iB} < K_i \) for \( i \in R \), where \( T_{im} \) is interpreted as the mandatory limit for the duration from patient \( i \)'s arrival at the waiting area until he is seen by the physician for the \((m+1)st\) time. If a returning patient \( i \in R \) 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. Since 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; see Section 7.3 for 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. Since a dynamic programming formulation would be intractable, time inconsistency is practically inevitable in solving the patient scheduling problem. See Delage and Iancu (2015) for time consistency issues arising from robust multi-stage 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. While their condition appears stable, these patients require timely treatment to resolve their acute symptoms. When the ED is getting crowded, this group of patients will be the most likely to suffer from prolonged waiting. 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

---

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

In the numerical study, we categorize all level-3 cases in the data set 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. Since 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 four to six physicians working for level-3 patients in each eight-hour shift. Although emergency physicians are required to provide treatment for a wide range of illnesses and injuries, their expertise and work rates differ from one another. Among 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 appear 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, while the variability in consultation times by physician 3 is much less than that by any other physician. In order for a patient scheduling approach to be relevant to practice, the heterogeneity of physicians must be taken into account.

7.2. Comparison with Sample Average Approximations

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

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

Then, $\mathbb{E}(\tilde{\chi}(\mu)) = \mathbb{P}(w_k(\tilde{s}, \mu) \leq \tau_k : k \in \mathcal{W})$. Let \( \{s_1, \ldots, s_N\} \), where \( s_n = (s^*_{ij})_{i \in I, j \in J} \) for \( n = 1, \ldots, N \), be a sample of consultation time vectors independently taken from the distribution of \( \tilde{s} \) 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 \( \mu \) can be approximated by \( \sum_{n=1}^{N} \chi_n(\mu)/N \). Using this fact, we formulate an approximate problem for (7) by

\[
\begin{align*}
\max & \quad \sum_{n=1}^{N} z_n \\
\text{s.t.} & \quad w_k(s_n, \mu) \leq z_n \cdot \tau_k + (1 - z_n) \cdot \bar{\tau}_n, \quad k \in W, \quad n = 1, \ldots, N \\
\mu \in \mathcal{A}, & \quad z_n \in \{0, 1\}, \quad n = 1, \ldots, N
\end{align*}
\]

where

\[
\bar{\tau}_n = \max_{j \in J} \left( \sum_{i \in I_j} s_{ij}^n - \min_{i \in I_j} s_{ij}^n \right)
\]

is an upper bound of patient waiting times. When \( z_n = 0 \), the inequality in (24) always holds for all \( k \in W \) and \( \mu \in \mathcal{A} \); when \( z_n = 1 \), the inequality holds for all \( k \in W \) 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 six physicians and twenty patients in the ED. Eight and twelve patients belong to the first and second categories, respectively. There are three new patients, twelve returning patients, and five patients being seen by physicians. Each physician has two returning patients waiting to be seen. We assume that all new patients have the same delay target \( \tau_N \) and all returning patients have the same delay target \( \tau_R \). All patients will leave the ED when they complete their current consultations. The six physicians have identical work capabilities, so 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 re-sampled 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 \( N \). Since the computation time of the SAA
method appears to increase exponentially with the sample size, we use a logarithmic scale for computation time in the figure. 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 size \(N = 20, 40, 100,\) and \(200\). If the sample size is small, the solutions exhibit great variability in performance across different realizations, while most of them are not satisfactory. Increasing the sample size can stabilize and improve the performance of solutions, yet 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 6 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, so 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 to that by the HRS solution only when the sample size is large. While 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. In order to achieve comparable performance, it may take hours to obtain a solution from an SAA realization, while 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 employ 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. Since the major concern is the crowding of level-3 patients, we set the safety limit for all patients’ door-to-provider times to be $D = 30$ minutes and set the mandatory target for all patients’ lengths of stay to be $K = 200$ minutes. In the numerical experiments below, we generate a stream of 5,000 patients arriving at the ED according to a Poisson process with rate 15.2 patients per hour. For the convenience of simulation, we assume that all medical tests and treatments by a nurse take no time. 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, while 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 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%.

In order to solve the mixed integer program (23) sequentially, we need to specify mandatory limits for returning patients when a decision iteration is triggered. Since all patients have identical requirements for door-to-provider times and lengths of stay, we would 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. Since 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

\[
T_1 = T_L + (T_U - T_L) \frac{|\mathcal{N}|}{|\mathcal{N}| + |\mathcal{R}^{(1)}|},
\]

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 \(\Delta_1\) and \(\Delta_2\) and set \(T_2 = T_1 + \Delta_1\) and \(T_3 = T_2 + \Delta_2\). For the convenience of implementation, we assume that \(\Delta_1\) and \(\Delta_2\) 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, so 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. Since medical tests and treatments are assumed to be instantaneous, a physician will be kept working on each patient until all consultations of this patient are completed. In this case, the global FCFS policy is equivalent to the *returning-patients-first* policy.
New-patients-first: When a physician finishes a consultation, the scheduling system will assign the new patient who has the earliest registration time to 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 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. employed the shortest-deadline-first policy for new patients, which is reduced to the FCFS policy when all patients have the same door-to-provider 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. 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 at the waiting area finding one or more physicians available, the scheduling 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 at 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 in order for the ED to meet the stringent time constraints. Although the percentage of patients who
Table 1  Performance comparison of door-to-provider times ($V$) and lengths of stay ($L$) under various scheduling policies, with arrival rate 15.2 patients per hour and four homogeneous physicians.

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

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 $\epsilon = 1, 3, 6,$ and $10$ 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 $\epsilon$ 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, in order for the overall computation time to be comparable to that of the hybrid approach. While the SAA method requires longer computation times than the benchmark policy, its performance is merely comparable to that of the benchmark policy with $\epsilon = 3$ minutes. Since 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 distributions, 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 the moment, should we send the category-2 patient to this “slow” physician, or keep the category-2 patient waiting and send a category-1 patient so that the physician can be working at a “fast” rate? In this case, a trade-off must be made between preventing an immediate time violation and reducing future crowding. 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 not good at category 2; the other two physicians are more experienced in category 2 but not familiar with category 1. In the simulation, we generate original consultation times using the empirical distributions according to patients’ categories, while the actual consultation time of a patient depends on the specific physician. If the physician is an expert in the patient’s category, the actual consultation time will be 80% of the original time; otherwise, the actual consultation time will be 120% of the original time. All other simulation settings are the same as in the previous example. In addition to the scheduling policies mentioned earlier, we also consider the following scheduling policy that is modified from the benchmark policy.\(^2\)

\(^2\) The modified benchmark policy was proposed by an anonymous referee.
Table 2  Performance comparison of door-to-provider times (V) and lengths of stay (L) under different scheduling policies, with arrival rate 15.2 patients per hour and four heterogeneous physicians.

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

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. Since 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 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 systems may follow a similar procedure to determine the next patient to be seen using the same coefficient $\delta$. 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, so 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 test $\delta = 0.5, 1,$ and 2 for $\epsilon = 3$ minutes, and 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. On the other hand, with an increased $\delta$, the modified policy may also postpone more consultations of new patients, even though their door-to-provider times are about to exceed the safety limit. Therefore, increasing $\delta$ 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 $\epsilon$, 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. Please refer to the e-companion for 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, e.g., patient transfer from EDs to inpatient wards (Mandelbaum et al. 2012 and Shi et al. 2016), ambulance deployment (McLay and Mayorga 2013, Maxwell et al. 2014, and Chong et al. 2016), and health examinations (Baron et al. 2017). Similar problems may arise from transportation systems, e.g., taxi dispatching (Seow et al. 2010), electric vehicle charging management (Yilmaz and Krein 2013), and vehicle routing with stochastic demands and time windows (Bertsimas and van Ryzin 1993, Fisher et al. 1997, Laporte et al. 2002, and Jepsen et al. 2008).

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 are 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, 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. As 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 lead to a significantly improved paper. In particular, we are grateful to the anonymous referee who proposed the modified benchmark policy.

The work of S. He was supported in part by Singapore Ministry of Education Academic Research Fund MOE2017-T2-1-012. The work of M. Sim was supported in part by Singapore Ministry of Education Social Science Research Thematic Grant MOE2016-SSRTG-059. 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.

References


Data-Driven Patient Scheduling in Emergency Departments: A Hybrid Robust–Stochastic Approach (E-Companion)

In Section EC.1, we define non-anticipative arrangements and present the proof of Proposition 1. All other proofs are provided in Section EC.2. We include additional numerical experiments in Section EC.3.

EC.1. Non-anticipative Arrangements and Proof of Proposition 1

Consider the patient scheduling problem at a specific time. Without loss of generality, let us assume that the scheduling procedure starts at time zero, when there is at least one physician available. We use $w^{(1)} \leq w^{(2)} \leq \cdots$ to denote the moments when physicians are available to start a consultation, with $w^{(1)} = 0$. For $i \in W$, if patient $i$ starts his consultation at time $w^{(k)}$, the patient’s waiting time is also equal to $w^{(k)}$, which should be kept below the delay target $\tau_i$. We assume that if a physician becomes idle after time zero, she must keep idle and is not allowed to take any patient in the future. Such a physician is said to be inactive. This assumption ensures that there are no breaks before consultations, so patient waiting times are given by (4). Clearly, breaks before consultations should be avoided in optimization.

For notational convenience, let us assume $\mathcal{I} = \{1, \ldots, I\}$ and $\mathcal{J} = \{1, \ldots, J\}$, where $I$ and $J$ are the numbers of patients and physicians, respectively. We use $W$ to denote the number of waiting patients, i.e., $W = |W|$. Let us introduce some variables to describe the state of the queueing network prior to time $w(k)$. Put $\bar{\mathcal{I}} = \mathcal{I} \cup \{-1, 0\}$. We use a $J$-dimensional vector $\gamma(k) = (\gamma_j(k))_{j \in \mathcal{J}}$, with $\gamma_j(k) \in \bar{\mathcal{I}}$, to denote the consultation status of the physicians. We take $\gamma_j(k) = i$ for $i \in \mathcal{I}$ if patient $i$ is being seen by physician $j$ before $w(k)$, take $\gamma_j(k) = -1$ if physician $j$ is idle but may take patients in the future, and take $\gamma_j(k) = 0$ if physician $j$ is inactive and will not take any patient. Note that only when $w(k) = 0$, is it possible that $\gamma_j(k) = -1$, because we require that a physician who becomes idle after time zero must be inactive from then on. For $j \in \mathcal{J}$, we use $a_j(k)$ to denote the age of consultation by physician $j$ before time $w(k)$, i.e., $a_j(k)$ is the amount of time physician $j$ has been with the current patient, with the convention that $a_j(k) = 0$ if physician $j$ is idle or inactive. The vector of consultation ages is $a(k) = (a_j(k))_{j \in \mathcal{J}}$. We use a $W$-dimensional vector of binary entries, $\beta(k) = (\beta_i(k))_{i \in W}$ with $\beta_i(k) \in \{0, 1\}$, to specify the status of the patients who were initially waiting, i.e., $\beta_i(k) = 0$ if patient $i$ is still waiting prior to $w(k)$ and $\beta_i(k) = 1$ if patient $i$ started consultation before $w(k)$.

Let $q(k)$ be the physician who becomes available at $w(k)$. Using cumulative information about patients and physicians at the time, we would determine the patient to be seen by physician $q(k)$.
We use $\xi(k)$ to denote this patient, who must be selected from the set of patients eligible to be seen by physician $q(k)$ at time $w(k)$. To represent this set, let us put
\begin{equation}
\mathcal{U}_j(\hat{\beta}) = \{i \in \mathcal{W}_j : \hat{\beta}_i = 0\} \text{ for } j \in \mathcal{J} \text{ and } \hat{\beta} = (\hat{\beta}_i)_{i \in \mathcal{W}} \in \{0,1\}^{\mathcal{W}},
\end{equation}
where $\mathcal{W}_j = \mathcal{N} \cup \mathcal{R}_j$ is the set of waiting patients eligible to be seen by physician $j$, so $\mathcal{U}_j(\hat{\beta})$ is the set of patients eligible to be seen by physician $j$ when the status of the patients is given by $\hat{\beta}$. Then, we must have
\begin{equation}
\xi(k) \in \mathcal{U}_{q(k)}(\beta(k)) \cup \{0\}.
\end{equation}
We take $\xi(k) = 0$ if physician $q(k)$ will not see any patient at $w(k)$, in which case the physician will become inactive. Since all waiting patients must be seen, a physician is not allowed to be inactive if there are waiting patients who cannot be seen by other physicians (e.g., the physician’s returning patients). To describe this constraint, let us put
\begin{equation}
\mathcal{M}(j, \hat{\gamma}) = \{\ell \in \mathcal{J} : \ell \neq j, \hat{\gamma}_\ell \neq 0\} \text{ for } j \in \mathcal{J} \text{ and } \hat{\gamma} = (\hat{\gamma}_\ell)_{\ell \in \mathcal{J}} \in \tilde{\mathcal{I}}^J,
\end{equation}
which, given the consultation status $\hat{\gamma}$, is the set of physicians still eligible to take patients in the future, with physician $j$ being excluded. Then, $\xi(k)$ must satisfy
\begin{equation}
\xi(k) \neq 0 \text{ if } \mathcal{U}_{q(k)}(\beta(k)) \not\subseteq \bigcup_{j \in \mathcal{M}(q(k),\gamma(k))} \mathcal{W}_j.
\end{equation}
Put
\begin{align*}
\vec{w}(k) &= (w(1), \ldots, w(k)), \quad \vec{\gamma}(k) = (\gamma(1), \ldots, \gamma(k)), \\
\vec{a}(k) &= (a(1), \ldots, a(k)), \quad \vec{\beta}(k) = (\beta(1), \ldots, \beta(k)),
\end{align*}
all of which are known at time $w(k)$. We may specify the state of the queueing network at any time before $w(k)$ using these vectors. Therefore, $(\vec{w}(k), \vec{\gamma}(k), \vec{a}(k), \vec{\beta}(k))$ may represent the history of the queueing network before $w(k)$. To determine the patient to be seen at $w(k)$, we should specify a Borel function
\begin{equation}
f_k : \mathcal{J} \times \mathbb{R}_+^k \times \tilde{\mathcal{I}}^{kJ} \times \mathbb{R}_+^{kJ} \times \{0,1\}^{kW} \to \mathcal{W},
\end{equation}
where $\mathcal{W} = \mathcal{W} \cup \{0\}$. When physician $q(k)$ is available, the next patient to be seen is given by
\begin{equation}
\xi(k) = f_k(q(k), \vec{w}(k), \vec{\gamma}(k), \vec{a}(k), \vec{\beta}(k)),
\end{equation}
which is subject to (EC.2)–(EC.3).

We next specify constraints on the state variables. We use $\vec{r}_j(k)$ to denote the remaining consultation time by physician $j$ before $w(k)$, where $\vec{r}_j(k) = 0$ if physician $j$ is either idle, inactive, or
about to complete a consultation. The corresponding vector is \( \tilde{r}(k) = (\tilde{r}_j(k))_{j \in \mathcal{J}} \). In general, we do
not know the remaining consultation time of a patient who is being seen.

Let \( \mathbb{F}_k \) be the set of Borel functions from the product space \( \mathcal{J} \times \mathbb{R}_+^k \times \bar{T}^kJ \times \mathbb{R}_+^kJ \times \{0,1\}^kW \) to the space \( \bar{W} \) such that each \( f_k \in \mathbb{F}_k \) satisfies

\[
f_k(j, \bar{w}_k, \bar{\gamma}_k, \bar{a}_k, \bar{\beta}_k) \in \mathcal{U}_j(\bar{\beta}_k) \cup \{0\} \tag{EC.5}
\]

and

\[
f_k(j, \bar{w}_k, \bar{\gamma}_k, \bar{a}_k, \bar{\beta}_k) \neq 0 \quad \text{if} \quad \mathcal{U}_j(\bar{\beta}_k) \not\subseteq \bigcup_{\ell \in \mathcal{M}(j, \gamma_k)} W_\ell, \tag{EC.6}
\]

where \( \bar{\gamma}_k = (\gamma_1, \ldots, \gamma_k) \) and \( \bar{\beta}_k = (\beta_1, \ldots, \beta_k) \). Using a sequence of functions \( \{f_k \in \mathbb{F}_k : k \in \mathbb{N}\} \), we may determine both the next patient to be seen and the state variables at \( w(1), w(2), \ldots \) as follows. Initially, we take

\[
w(1) = 0, \quad \beta_i(1) = 0, \quad a_j(1) = 0, \quad \text{for } i \in \mathcal{W} \text{ and } j \in \mathcal{J}. \tag{EC.7}
\]

Both \( q(1) \), the first available physician, and \( \gamma(1) \), the physicians’ consultation status before \( w(1) \), are also given as initial conditions. Without loss of generality, we assume that

\[
\gamma_j(1) \neq 0 \quad \text{for all } j \in \mathcal{J}, \tag{EC.8}
\]

so no physicians are inactive at the beginning. The remaining consultation time by physician \( j \) before \( w(1) \) is given by

\[
\tilde{r}_j(1) = \begin{cases} 
0 & \text{if } j = q(1), \\
\tilde{s}_{\gamma_j(1),j} & \text{if } j \neq q(1),
\end{cases} \tag{EC.9}
\]

where \( \tilde{s}_{\gamma_j(1),j} = 0 \) if \( \gamma_j(1) = -1 \).

For \( k \in \mathbb{N} \), physician \( q(k) \) becomes available at \( w(k) \) and will take patient \( \xi(k) \) determined by (EC.4). We update each physician’s consultation status by

\[
\gamma_j(k+1) = \begin{cases} 
\xi(k) & \text{if } j = q(k), \\
\gamma_j(k) & \text{if } j \neq q(k),
\end{cases} \tag{EC.10}
\]

and the status of the patients by

\[
\beta_i(k+1) = \begin{cases} 
1 & \text{if } i = \xi(k), \\
\beta_i(k) & \text{if } i \neq \xi(k).
\end{cases} \tag{EC.11}
\]

After sending patient \( \xi(k) \) to physician \( q(k) \), the remaining consultation time by each physician is

\[
\tilde{r}_j'(k+1) = \begin{cases} 
\tilde{s}_{\xi(k),j} & \text{if } j = q(k), \\
\tilde{r}_j(k) & \text{if } j \neq q(k),
\end{cases} \tag{EC.12}
\]
where \( \tilde{s}_{\xi(k),j} = 0 \) if \( \xi(k) = 0 \). Using these remaining consultation times, we may determine the amount of time until the next physician is available, which is given by

\[
v(k + 1) = \min \{ \tilde{r}_j(k + 1) : j \in J, \; \gamma_j(k + 1) \neq 0 \}.
\] (EC.13)

Then, the next available physician is

\[
q(k + 1) = \min \{ j \in J : \tilde{r}_j(k + 1) = v(k + 1), \; \gamma_j(k + 1) \neq 0 \}
\] (EC.14)

and we will determine the next patient at time

\[
w(k + 1) = w(k) + v(k + 1).
\] (EC.15)

If the set in (EC.13) is empty, we take \( v(k + 1) = 0 \) and choose an arbitrary value for \( q(k + 1) \).

Before time \( w(k + 1) \), the remaining consultation time by physician \( j \) is

\[
\tilde{r}_j(k + 1) = (\tilde{r}_j(k + 1) - v(k + 1))^+,
\] (EC.16)

and the age of consultation is

\[
a_j(k + 1) = \begin{cases} 
    v(k + 1) & \text{if } j = q(k) \text{ and } \gamma_j(k + 1) \neq 0, \\
    a_j(k) + v(k + 1) & \text{if } j \neq q(k) \text{ and } \gamma_j(k + 1) \neq 0, \\
    0 & \text{if } \gamma_j(k + 1) = 0.
\end{cases}
\] (EC.17)

With the updated state variables, we can go back to (EC.4) and determine the next patient to be seen. This procedure is repeated until all entries of \( \gamma(k) \) become zero, i.e., all physicians become inactive. Since each time we either send a patient to the available physician or let the physician be inactive, all physicians must become inactive by repeating the procedure at most \( I + J \) times. In other words, it suffices to consider functions \( f_1, \ldots, f_{I+J} \) for patient scheduling. Since we take \( v(k+1) = 0 \) and \( q(k+1) = j \) for any \( j \in J \) if the set in (EC.13) is empty (i.e., all physicians are inactive at time \( w(k) \)), we will have \( w(k) = w(k+1) = \cdots = w(I+J) \). In this case, all patients have completed their consultations by \( w(k) \), and we must have \( \xi(k) = \xi(k+1) = \cdots = \xi(I+J) = 0 \) by (EC.5). Similarly, all state variables are also well defined by (EC.10)–(EC.17).

By this recursive procedure, we obtain a sequence of vectors \( \{(q(k),\xi(k)) : k = 1,\ldots,I+J\} \). For each \( i \in W \), there is a unique integer \( k \) such that \( \xi(k) = i \). We may thus define a function \( \zeta : W \to \{1,\ldots,I+J\} \) such that \( w(\zeta(i)) \) is the moment patient \( i \) starts his consultation. Then, we can uniquely determine an admissible schedule \( (\varphi, \Phi) \in A \) by

\[
\varphi(i) = q(\zeta(i)) \quad \text{and} \quad \Phi(i) = \{ \xi(k) : q(k) = \varphi(i), \; k < \zeta(i) \} \quad \text{for } i \in W.
\] (EC.18)

Using a sequence of functions \( \{f_k \in F_k : k = 1,\ldots,I+J\} \) together with (EC.4) and (EC.7)–(EC.17), we may thus uniquely determine an admissible schedule for any realization of consultation times.
In other words, such a sequence of functions uniquely defines an arrangement \( \pi \in \mathcal{V} \). We refer to an arrangement \( \pi \in \mathcal{V} \) as a non-anticipative arrangement if there exists a sequence of functions \( \{ f_k \in \mathbb{F}_k : k = 1, \ldots, I + J \} \) that defines \( \pi \) through (EC.4) and (EC.7)–(EC.18). We use \( \mathcal{V}_1 \) to denote the set of all non-anticipative arrangements. Clearly, \( \mathcal{V}_1 \subset \mathcal{V} \). In addition, all static arrangements are non-anticipative.

**Proof of Proposition 1.** We only need to prove \( \mathcal{V}_0 \subset \mathcal{V}_1 \). To this end, let us construct a sequence of functions \( \{ f_k \in \mathbb{F}_k : k = 1, \ldots, I + J \} \) that leads to an identical admissible schedule for all realizations of consultation times.

Let \( (\varphi, \Phi) \in \mathcal{A} \) be an arbitrary admissible schedule. For \( j \in \mathcal{J} \) and \( \beta \in \{0,1\}^W \), put

\[ K_j(\hat{\beta}) = \{ i \in \mathcal{W} : \varphi(i) = j, \hat{\beta}_i = 0, \hat{\beta}_{\ell} = 1 \text{ for } \ell \in \Phi(i) \}, \]

which, by (3), has at most one element. When the status of the patients is specified by \( \hat{\beta} \), \( K_j(\hat{\beta}) \) contains the next patient to be seen by physician \( j \) under the admissible schedule. We use \( \kappa_j(\hat{\beta}) \) to denote the element of \( K_j(\hat{\beta}) \) if it is nonempty.

For \( (j, \bar{w}_k, \bar{\gamma}_k, \bar{\alpha}_k, \bar{\beta}_k) \in \mathcal{J} \times \mathbb{R}_+^k \times \mathbb{I}^{k,j} \times \mathbb{R}_+^{k,j} \times \{0,1\}^{kW} \), define

\[ f_k(j, \bar{w}_k, \bar{\gamma}_k, \bar{\alpha}_k, \bar{\beta}_k) = \begin{cases} \kappa_j(\bar{\beta}_k) & \text{if } K_j(\bar{\beta}_k) \neq \emptyset, \\ 0 & \text{if } K_j(\bar{\beta}_k) = \emptyset, \end{cases} \]

the value of which depends only on \( j \) and \( \beta_k \). Let us examine (EC.5)–(EC.6) to verify \( f_k \in \mathbb{F}_k \).

If \( K_j(\beta(k)) = \emptyset \), \( \beta_i(k) = 1 \) for all \( i \in \mathcal{W} \) such that \( \varphi(i) = j \), i.e., a physician may become inactive only if all patients assigned to her under the admissible schedule have started their consultations according to (EC.20). By (EC.1) and (EC.19), \( K_j(\beta_k) \subset \mathcal{U}_j(\beta_k) \), so \( f_k \) satisfies (EC.5). If there exists \( j \in \mathcal{J} \) and \( k \in \mathbb{N} \) such that

\[ \mathcal{U}_j(\beta_k) \not\subset \bigcup_{\ell \in M(j,\gamma_k)} \mathcal{W}_\ell, \]

there must be some \( h \in \mathcal{U}_j(\beta_k) \) who cannot be seen by physicians other than \( j \) after \( w(k) \). Suppose that \( \varphi(h) \neq j \). Then, physician \( \varphi(h) \) must be inactive at \( w(k) \) because she is not eligible to take patient \( h \). Hence, all patients assigned to physician \( \varphi(h) \) under the given admissible schedule have started their consultations by \( w(k) \), which, however, contradicts the fact that \( \beta_h(k) = 0 \). Then, we have \( \varphi(h) = j \). Since \( \beta_h(k) = 0 \), we deduce that \( K_j(\beta(k)) \neq \emptyset \). By (EC.20), \( f_k \) must satisfy (EC.6).

Using the sequence of functions \( \{ f_k \in \mathbb{F}_k : k = 1, \ldots, I + J \} \), we will send all patients to the physicians by repeating the recursive procedure at most \( I + J - 1 \) times. If \( \varphi(i) = j \) for \( i \in \mathcal{W} \), it follows from (EC.19)–(EC.20) that patient \( i \) must be sent to physician \( j \). For \( i_1, i_2 \in \mathcal{W} \) such that \( \varphi(i_1) = \varphi(i_2) = j \) and \( \Phi(i_1) \subset \Phi(i_2) \), it also follows from (EC.19) that if \( i_2 \in K_j(\beta(k)) \) for some \( k \in \mathbb{N} \), we must have \( \beta_{i_1}(k) = 1 \), so patient \( i_1 \) must be seen before patient \( i_2 \). Therefore, the admissible schedule \((\varphi, \Phi)\) is preserved regardless of the realization of consultation times. \( \square \)
Exploiting the information on patients and physicians that accumulates over time, we would formulate a multi-stage optimization problem to obtain a set of functions \( \{f_k \in \mathbb{F}_k : k = 1, \ldots, I+J\} \) that maximizes the joint probability of meeting all delay targets. Note that if \( \xi(k) \neq 0 \), \( w(k) \) is equal to the waiting time of patient \( \xi(k) \). To allow for \( \xi(k) = 0 \) in our formulation, we put \( \tau_0 = \bar{\tau} \), where \( \bar{\tau} \) is given by (19), so that \( w(k) \leq \tau_0 \) for all \( k = 1, \ldots, I+J \). The optimal functions can be obtained by solving the following problem

\[
\begin{align*}
\max \quad & \mathbb{P}(w(k) \leq \tau_{\xi(k)} : k = 1, \ldots, I+J) \\
\text{s.t.} \quad & \text{initial conditions (EC.7)-(EC.9),} \\
& \text{constraints (EC.4) and (EC.10)-(EC.17), } j \in J, \quad k = 1, \ldots, I+J \\
& f_k \in \mathbb{F}_k, \quad k = 1, \ldots, I+J.
\end{align*}
\]

(EC.21)

Since a set of functions \( \{f_k \in \mathbb{F}_k : k = 1, \ldots, I+J\} \) defines a non-anticipative arrangement through (EC.4) and (EC.7)-(EC.17), the above problem is equivalent to the \( P \)-model problem (6). Hence, in order to find an optimal non-anticipative arrangement, we may first solve (EC.21) and obtain a set of optimal functions \( \{f_k \in \mathbb{F}_k : k = 1, \ldots, I+J\} \). Then using these functions, we may obtain an optimal non-anticipative arrangement by (EC.4) and (EC.7)-(EC.18).

EC.2. Other Proofs

Proof of Theorem 1. Let \( B = \prod_{k=1}^M (-\infty, b_k] \) be an element in \( \mathcal{F} \), where \( b = (b_1, \ldots, b_M) \in \mathbb{R}^M \) is the boundary value. Since \( \{\tilde{\sigma}_1, \ldots, \tilde{\sigma}_M\} \) is a set of independent random variables, maximizing \( \mathbb{P}(\tilde{\sigma} \in B) \) is equivalent to maximizing

\[
\ln \mathbb{P}(\tilde{\sigma} \in B) = \ln \prod_{k=1}^M \mathbb{P}(\tilde{\sigma}_k \leq b_k) = \ln \prod_{k=1}^M G_k(b_k) = \sum_{k=1}^M \ln G_k(b_k).
\]

Let \( (\hat{B}, \hat{\psi}) \) be an optimal solution to (10) and \( (b^*, \nu^*) \) an optimal solution to (11), where \( \hat{B} \in \mathcal{F} \), \( \hat{\psi} \in \mathcal{G} \), \( b^* = (b_1^*, \ldots, b_M^*) \in \mathbb{R}^M \), and \( \nu^* \in \mathcal{D} \). If the feasible set of (10) is nonempty, we must have \( \theta(\sigma, \hat{\psi}(\sigma)) \leq \eta \) for all \( \sigma \in \hat{B} \). Let \( \hat{b} = (\hat{b}_1, \ldots, \hat{b}_M) \in \mathbb{R}^M \) be the boundary value of \( \hat{B} \). By taking \( \nu = \hat{\psi}(\hat{b}) \), we have \( \theta(\hat{b}, \nu) \leq \eta \), so \( (\hat{b}, \nu) \) is a feasible solution to (11). This implies that

\[
\ln \mathbb{P}(\tilde{\sigma} \in \hat{B}) = \sum_{k=1}^M \ln G_k(\hat{b}_k) \leq \sum_{k=1}^M \ln G_k(b_k^*). \tag{EC.22}
\]

Conversely, if the feasible set of (11) is nonempty, we have \( \theta(b^*, \nu^*) \leq \eta \). Let \( \psi^* \) be the constant function that satisfies \( \psi^*(\sigma) = \nu^* \) for all \( \sigma \in \mathbb{R}^M \). Let \( B^* \) be the hyperrectangular subset in \( \mathcal{F} \) with boundary value \( b^* \). Since \( \sigma_k \leq b_k^* \) for \( \sigma = (\sigma_1, \ldots, \sigma_M) \in B^* \), the monotonicity of \( \theta \) implies that \( \theta(\sigma, \psi^*(\sigma)) \leq \theta(b^*, \psi^*(b^*)) \leq \eta \) for all \( \sigma \in B^* \), so \( (B^*, \psi^*) \) is a feasible solution to (10). This implies that

\[
\sum_{k=1}^M \ln G_k(b_k^*) = \ln \mathbb{P}(\tilde{\sigma} \in B^*) \leq \ln \mathbb{P}(\tilde{\sigma} \in \hat{B}). \tag{EC.23}
\]

By (EC.22)-(EC.23), we conclude that (11) is an equivalent form of (10). \( \square \)
Proof of Corollary 1. Consider the optimization problem
\[
\max_{\tilde{s} \in Q} P(\tilde{s} \in Q) \\
\text{s.t. } w_k(s, \pi(s)) \leq \tau_k, \quad k \in W, \ s \in Q
\]
where $\mathcal{V}$ is the set of all arrangements. For any $\mu \in A$ and all $k \in W$, it follows from (4) that $w_k(s, \mu)$ is nondecreasing in each $s_{ij}$, where $s = (s_{ij})_{i \in I, j \in J}$. By Theorem 1, the above problem is equivalent to (12), which allows us to complete the proof using Proposition 1.

Proof of Proposition 2. Consider the equivalent form (12). If it has a feasible solution $(d, \mu)$, then by (4), $w_k(\tilde{s}, \mu) \leq w_k(d, \mu) \leq \tau_k$ for $k \in W$, because $s_{ij} \leq d_{ij}$ for all $i \in I$ and $j \in J$. Hence, $\alpha^* \leq 0$. Let $\mu^*$ be the optimal solution to (13). Conversely, if $\alpha^* \leq 0$, we have $w_k(\tilde{s}, \mu^*) \leq \tau_k + \alpha^* \leq \tau_k$ for $k \in W$, which implies that $(\tilde{s}, \mu^*)$ is a feasible solution to (12).

Proof of Theorem 2. Since $y_{ij}(n) \in \{0, 1\}$ and $\sum_{n=1}^{N_{ij}} y_{ij}(n) = 1$, there is a unique $n \in \{1, \ldots, N_{ij}\}$ such that $y_{ij}(n) = 1$ for $i \in I$ and $j \in J$. If $d_{ij} = s_{ij}(m)$ for some $m = 1, \ldots, N_{ij}$, by taking $y_{ij}(m) = 1$ and $y_{ij}(n) = 0$ for $n \neq m$, we have
\[
d_{ij} = y_{ij}(m) \cdot s_{ij}(m) = \sum_{n=1}^{N_{ij}} y_{ij}(n) \cdot s_{ij}(n),
\]
which allows us to rewrite (22) into the first inequality in (23). Since
\[
\ln F(d_{ij}) = \ln F\left(\sum_{n=1}^{N_{ij}} y_{ij}(n) \cdot s_{ij}(n)\right) = \sum_{n=1}^{N_{ij}} y_{ij}(n) \cdot \ln F(s_{ij}(n)) = \sum_{n=1}^{N_{ij}} y_{ij}(n) \cdot g_{ij}(n),
\]
we can also rewrite the objective function in (12) into that in (23). The rest of the proof follows from the construction of the constraints (14)–(18) and (21)–(22).

EC.3. More Experiments on Dynamic Patient Scheduling

In this section, we provide more simulation results when physicians are heterogeneous. We modify simulation parameters in the example reported in Table 2 to evaluate the performance of the hybrid approach.

We first examine a case with arrival rate 14.7 patients per hour and all other simulation settings being identical to the example in Table 2. The numerical results are provided in Table EC.1. In this example, the ED is not as heavily loaded as in the previous example. If we use a patient scheduling policy that does not differentiate physicians, such as the global FCFS, new-patient-first, or benchmark policy, the resulting traffic intensity will be 90.49%, as compared with that of 93.57% in Table 2. As in the previous example, the hybrid approach achieves a good balance for all performance measures.
Table EC.1 Performance comparison of door-to-provider times ($V$) and lengths of stay ($L$) under different scheduling policies, with arrival rate 14.7 patients per hour and four heterogeneous physicians.

<table>
<thead>
<tr>
<th>Scheduling policy</th>
<th>$\bar{V}$</th>
<th>$\bar{L}$</th>
<th>$% \ V &gt; 30$</th>
<th>$% \ L &gt; 200$</th>
<th>$%$ Violations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global FCFS</td>
<td>28.26</td>
<td>42.92</td>
<td>33.34%</td>
<td>0.12%</td>
<td>33.34%</td>
</tr>
<tr>
<td>New-patients-first</td>
<td>3.61</td>
<td>61.60</td>
<td>0.04%</td>
<td>6.30%</td>
<td>6.30%</td>
</tr>
<tr>
<td>Benchmark, $\epsilon = 3$</td>
<td>15.87</td>
<td>51.34</td>
<td>17.56%</td>
<td>4.14%</td>
<td>18.24%</td>
</tr>
<tr>
<td>Modified benchmark, $\epsilon = 3$, $\delta = 1$</td>
<td>16.37</td>
<td>42.38</td>
<td>13.10%</td>
<td>1.16%</td>
<td>13.10%</td>
</tr>
<tr>
<td>Modified benchmark, $\epsilon = 6$, $\delta = 1$</td>
<td>16.03</td>
<td>42.81</td>
<td>11.36%</td>
<td>2.20%</td>
<td>11.82%</td>
</tr>
<tr>
<td>Modified benchmark, $\epsilon = 10$, $\delta = 1$</td>
<td>14.69</td>
<td>43.19</td>
<td>9.90%</td>
<td>2.54%</td>
<td>10.15%</td>
</tr>
<tr>
<td>HRS, $T_L = 90$, $T_U = 120$</td>
<td>10.35</td>
<td>52.26</td>
<td>9.18%</td>
<td>1.70%</td>
<td>9.18%</td>
</tr>
<tr>
<td>HRS, $T_L = 100$, $T_U = 130$</td>
<td>9.82</td>
<td>52.65</td>
<td>8.00%</td>
<td>1.60%</td>
<td>8.20%</td>
</tr>
<tr>
<td>HRS, $T_L = 105$, $T_U = 130$</td>
<td>9.46</td>
<td>50.70</td>
<td>6.52%</td>
<td>1.78%</td>
<td>6.52%</td>
</tr>
</tbody>
</table>

Table EC.2 Performance comparison of door-to-provider times ($V$) and lengths of stay ($L$) under different scheduling policies, with arrival rate 15.8 patients per hour and four heterogeneous physicians.

<table>
<thead>
<tr>
<th>Scheduling policy</th>
<th>$\bar{V}$</th>
<th>$\bar{L}$</th>
<th>$% \ V &gt; 30$</th>
<th>$% \ L &gt; 200$</th>
<th>$%$ Violations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global FCFS</td>
<td>63.59</td>
<td>78.11</td>
<td>67.94%</td>
<td>1.18%</td>
<td>67.94%</td>
</tr>
<tr>
<td>New-patients-first</td>
<td>4.59</td>
<td>123.25</td>
<td>0.22%</td>
<td>25.06%</td>
<td>25.06%</td>
</tr>
<tr>
<td>Benchmark, $\epsilon = 3$</td>
<td>24.68</td>
<td>96.86</td>
<td>36.52%</td>
<td>15.36%</td>
<td>42.70%</td>
</tr>
<tr>
<td>Modified benchmark, $\epsilon = 3$, $\delta = 1$</td>
<td>26.89</td>
<td>57.26</td>
<td>39.38%</td>
<td>1.66%</td>
<td>39.38%</td>
</tr>
<tr>
<td>Modified benchmark, $\epsilon = 6$, $\delta = 1$</td>
<td>24.33</td>
<td>58.39</td>
<td>29.96%</td>
<td>2.92%</td>
<td>30.82%</td>
</tr>
<tr>
<td>Modified benchmark, $\epsilon = 10$, $\delta = 1$</td>
<td>21.45</td>
<td>62.41</td>
<td>23.70%</td>
<td>4.62%</td>
<td>25.16%</td>
</tr>
<tr>
<td>HRS, $T_L = 90$, $T_U = 120$</td>
<td>18.05</td>
<td>70.38</td>
<td>19.30%</td>
<td>3.20%</td>
<td>19.98%</td>
</tr>
<tr>
<td>HRS, $T_L = 100$, $T_U = 130$</td>
<td>14.65</td>
<td>70.08</td>
<td>14.90%</td>
<td>3.06%</td>
<td>15.62%</td>
</tr>
<tr>
<td>HRS, $T_L = 105$, $T_U = 130$</td>
<td>16.31</td>
<td>72.11</td>
<td>18.94%</td>
<td>3.34%</td>
<td>19.26%</td>
</tr>
</tbody>
</table>

Then, we increase the arrival rate to 15.8 patients per hour, which results in a traffic intensity of 97.25%. In such a heavily loaded system, all three policies that do not differentiate physicians perform poorly. In contrast, the modified benchmark policy and the hybrid approach are able to make use of physicians’ expertise efficiently, thus delivering much better performance. The hybrid approach achieves the most balanced performance again in complying with the performance targets.

Table EC.3 summarizes the performance under different scheduling policies with randomly generated targets for each patient’s door-to-provider time and length of stay. More specifically, the
Table EC.3  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, four heterogeneous physicians, and randomly generated targets.

<table>
<thead>
<tr>
<th>Scheduling policy</th>
<th>$\bar{V}$</th>
<th>$\bar{L}$</th>
<th>% $V &gt; \tilde{\tau}_d$</th>
<th>% $L &gt; \tilde{\tau}_\ell$</th>
<th>% Violations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global FCFS</td>
<td>38.69</td>
<td>53.19</td>
<td>34.34%</td>
<td>0.46%</td>
<td>34.34%</td>
</tr>
<tr>
<td>New-patients-first</td>
<td>3.88</td>
<td>72.19</td>
<td>0.32%</td>
<td>8.68%</td>
<td>8.68%</td>
</tr>
<tr>
<td>Benchmark, $\epsilon = 3$</td>
<td>18.59</td>
<td>65.46</td>
<td>27.24%</td>
<td>6.52%</td>
<td>28.24%</td>
</tr>
<tr>
<td>Modified benchmark, $\epsilon = 3$, $\delta = 1$</td>
<td>20.82</td>
<td>52.30</td>
<td>25.24%</td>
<td>1.52%</td>
<td>25.24%</td>
</tr>
<tr>
<td>Modified benchmark, $\epsilon = 6$, $\delta = 1$</td>
<td>19.92</td>
<td>50.78</td>
<td>22.04%</td>
<td>1.98%</td>
<td>23.12%</td>
</tr>
<tr>
<td>Modified benchmark, $\epsilon = 10$, $\delta = 1$</td>
<td>17.23</td>
<td>49.67</td>
<td>18.32%</td>
<td>2.38%</td>
<td>19.66%</td>
</tr>
<tr>
<td>HRS, $T_L = 90$, $T_U = 120$</td>
<td>12.32</td>
<td>59.35</td>
<td>9.33%</td>
<td>1.92%</td>
<td>10.68%</td>
</tr>
<tr>
<td>HRS, $T_L = 100$, $T_U = 130$</td>
<td>11.58</td>
<td>57.62</td>
<td>7.85%</td>
<td>1.62%</td>
<td>8.32%</td>
</tr>
<tr>
<td>HRS, $T_L = 105$, $T_U = 130$</td>
<td>10.24</td>
<td>56.02</td>
<td>7.56%</td>
<td>1.38%</td>
<td>8.65%</td>
</tr>
</tbody>
</table>

door-to-provider time target is generated following a uniform distribution from 20 minutes to 40 minutes and the length-of-stay target is generated following a uniform distribution from 180 minutes to 220 minutes. In this table, we use $\tilde{\tau}_d$ and $\tilde{\tau}_\ell$ to denote the randomly generated door-to-provider and length-of-stay targets, respectively. All other simulation settings are identical to the example in Table 2. In Table EC.3, the performance of each policy is generally comparable to that in Table 2, where the door-to-provider time target and the length-of-stay target are set to be 30 and 200 minutes, respectively, for all patients. The mean door-to-provider time and the mean length of stay of the hybrid robust–stochastic approach are slightly longer when the targets are randomly generated. This phenomenon implies that the hybrid approach tends to “penalize” patients with longer delay targets to maximize the joint probability of all patients meeting their targets. The hybrid approach still outperforms the modified benchmark policy.

Table EC.4 summarizes the performance under different policies with a periodic, time-varying Poisson arrival process. We assume that the patient arrival rate changes every hour, but within each hour, the arrival rate is constant. We also assume that the arrival rate is periodic with the period being one day. Figure EC.1 illustrates the relative frequencies of patient arrivals over different hours of a day in the anonymous hospital. In this numerical example, we assume that the time-varying patient arrival rate follows the same pattern as in the figure, where the average arrival rate over each day is 13.2 patients per hour. All other simulation settings are identical to the example in Table 2. In Table EC.4, the performance of each policy is similar to that in previous cases. The percentage of patients whose lengths of stay exceed 200 minutes under the hybrid approach is slightly greater than that under the modified benchmark policy. This is because the hybrid approach would strike
a balance between door-to-provider times and lengths of stay. It achieves much better performance in terms of door-to-provider times than the modified benchmark policy.

A case with six heterogeneous physicians is also considered. Three of them are experts in treating cases in category 1 but not good at category 2, while the other three are more experienced in category 2 but not familiar with category 1. We assume that the patient arrival rate is constant, equal to 18.0 patients per hour. All other simulation settings are identical to the example in Table 2. In this table, the performance of each policy is also similar to that in the previous examples.

Table EC.4 Performance comparison of door-to-provider times ($V$) and lengths of stay ($L$) under different scheduling policies, with a time-varying arrival process (the average arrival rate over a day is 13.2 patients per hour) and four heterogeneous physicians.

<table>
<thead>
<tr>
<th>Scheduling policy</th>
<th>$\bar{V}$</th>
<th>$\bar{L}$</th>
<th>% $V &gt; 30$</th>
<th>% $L &gt; 200$</th>
<th>% Violations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global FCFS</td>
<td>34.59</td>
<td>49.85</td>
<td>48.24%</td>
<td>0.12%</td>
<td>48.24%</td>
</tr>
<tr>
<td>New-patients-first</td>
<td>2.28</td>
<td>69.48</td>
<td>0.00%</td>
<td>9.64%</td>
<td>9.64%</td>
</tr>
<tr>
<td>Benchmark, $\epsilon = 3$</td>
<td>18.66</td>
<td>58.15</td>
<td>17.40%</td>
<td>4.22%</td>
<td>20.14%</td>
</tr>
<tr>
<td>Modified benchmark, $\epsilon = 3$, $\delta = 1$</td>
<td>21.79</td>
<td>41.65</td>
<td>30.20%</td>
<td>1.06%</td>
<td>30.20%</td>
</tr>
<tr>
<td>Modified benchmark, $\epsilon = 6$, $\delta = 1$</td>
<td>21.87</td>
<td>40.65</td>
<td>25.63%</td>
<td>1.65%</td>
<td>25.63%</td>
</tr>
<tr>
<td>Modified benchmark, $\epsilon = 10$, $\delta = 1$</td>
<td>21.46</td>
<td>39.48</td>
<td>24.36%</td>
<td>1.86%</td>
<td>24.36%</td>
</tr>
<tr>
<td>HRS, $T_L = 90$, $T_U = 120$</td>
<td>11.63</td>
<td>52.34</td>
<td>11.12%</td>
<td>2.23%</td>
<td>12.82%</td>
</tr>
<tr>
<td>HRS, $T_L = 100$, $T_U = 130$</td>
<td>10.83</td>
<td>52.66</td>
<td>9.92%</td>
<td>2.56%</td>
<td>10.23%</td>
</tr>
<tr>
<td>HRS, $T_L = 105$, $T_U = 130$</td>
<td>9.94</td>
<td>53.89</td>
<td>9.63%</td>
<td>2.35%</td>
<td>11.30%</td>
</tr>
</tbody>
</table>
Table EC.5  Performance comparison of door-to-provider times ($V$) and lengths of stay ($L$) under different scheduling policies, with arrival rate 18.0 patients per hour and six heterogeneous physicians.

<table>
<thead>
<tr>
<th>Scheduling policy</th>
<th>$\bar{V}$</th>
<th>$\bar{L}$</th>
<th>$% V &gt; 30$</th>
<th>$% L &gt; 200$</th>
<th>$%$ Violations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global FCFS</td>
<td>24.23</td>
<td>39.46</td>
<td>33.50%</td>
<td>0.00%</td>
<td>33.50%</td>
</tr>
<tr>
<td>New-patients-first</td>
<td>1.56</td>
<td>54.71</td>
<td>0.00%</td>
<td>5.10%</td>
<td>5.10%</td>
</tr>
<tr>
<td>Benchmark, $\epsilon = 3$</td>
<td>15.34</td>
<td>43.26</td>
<td>9.92%</td>
<td>1.76%</td>
<td>10.92%</td>
</tr>
<tr>
<td>Modified benchmark, $\epsilon = 3$, $\delta = 1$</td>
<td>18.11</td>
<td>36.00</td>
<td>24.50%</td>
<td>0.36%</td>
<td>24.50%</td>
</tr>
<tr>
<td>Modified benchmark, $\epsilon = 6$, $\delta = 1$</td>
<td>19.40</td>
<td>34.62</td>
<td>23.89%</td>
<td>0.63%</td>
<td>23.89%</td>
</tr>
<tr>
<td>Modified benchmark, $\epsilon = 10$, $\delta = 1$</td>
<td>19.87</td>
<td>35.22</td>
<td>24.92%</td>
<td>0.67%</td>
<td>24.92%</td>
</tr>
<tr>
<td>HRS, $T_L = 90$, $T_U = 120$</td>
<td>9.55</td>
<td>38.36</td>
<td>5.32%</td>
<td>1.36%</td>
<td>6.25%</td>
</tr>
<tr>
<td>HRS, $T_L = 100$, $T_U = 130$</td>
<td>9.12</td>
<td>37.45</td>
<td>4.68%</td>
<td>1.76%</td>
<td>6.08%</td>
</tr>
<tr>
<td>HRS, $T_L = 105$, $T_U = 130$</td>
<td>9.23</td>
<td>38.23</td>
<td>4.83%</td>
<td>1.82%</td>
<td>6.13%</td>
</tr>
</tbody>
</table>