manuscript

Advance Admission Scheduling via Resource Satisficing

Minglong Zhou, Melvyn Sim
Department of Analytics & Operations, NUS Business School,
minglong_zhou@u.nus.edu, melvynsim@nus.edu.sg

Shao-Wei Lam
Singapore General Hospital, Health Services Research, lam.shao.wei@singhealth.com.sg

We study the problem of advance scheduling of ward admission requests in a public hospital, which affects the usage of critical resources such as operating theaters and hospital beds. Given the stochastic arrivals of patients and their uncertain usage of resources, it is often infeasible for the planner to devise a risk-free schedule to meet these requests without violating resource capacity constraints and creating negative effects that include healthcare overtime, longer patient waiting times, and even bed shortages. The difficulty of quantifying these costs and the need to safeguard against their overutilization lead us to propose a *resource satisficing* framework that renders the violation of resource constraints less likely and also diminishes their impact whenever they occur. The risk of resource overutilization is captured by our resource satisficing index (RSI), which is calibrated to reflect a risk-adjusted utilization rate for a better interpretation to the healthcare planner. Unlike the expected utilization rate, RSI is risk sensitive and serves to better mitigate the risks of overutilization. Our satisficing approach aims to balance out the overutilization risks by minimizing the largest RSIs among all resources and time periods, which, under our proposed *partial adaptive scheduling policy*, can be formulated and solved via a converging sequence of mixed-integer linear optimization problems. A computational study establishes that our approach reduces resource overutilization risks to a greater extent than does the benchmark method using the first fit (FF) heuristic.

Key words: Robustness optimization, robust optimization, satisficing, advance admission scheduling
History: December 31, 2019

1. Introduction

The scheduling of patient appointment and admission is an integral part of any hospital's daily operations. Broadly speaking, the appointment scheduling process includes intraday scheduling and multiday scheduling. The former involves designing the detailed sequence and start times of services given a set of patients that need to be served. The latter focuses on the matter of admitting patients and scheduling them to clinical procedures, possibly in the near future. In this paper we focus on the advance scheduling of patient admission, which integrates bed management and multiday scheduling. The advance scheduling process adaptively determines the patients' admission

dates subject to such resource constraints as the availability of beds, operating theaters, diagnostic imaging equipment and manpower. Advance scheduling is an important mechanism for ensuring a hospital's smooth operation (Chakraborty et al. 2010), and good scheduling balances resource utilization and reduces patient dissatisfaction (Lowery and Martin 1989). We describe a general advance admission scheduling problem where patients are dynamically admitted to different days and may subsequently go through some clinical procedures such as consultation, surgery, magnetic resonance imaging, and so on. Admitting patients requires hospital beds, and subsequent clinical procedures also use a variety of resources. Therefore, advance admission decisions can significantly influence the usage of different resources on each day. The planner's problem is to schedule, on a daily basis, patients to future dates for admission subject to the resource constraints such as bed availability. In our context, this admission scheduling system has two general objectives: (i) to schedule all patients on the day of their arrival to future dates subject to a specific waiting time limit; and (ii) to safeguard the system against overutilization of various relevant resources. These resources include—but are not limited to—operating theaters, hospital beds and diagnostic imaging equipment. Overutilization of resources corresponds to exceeding resource capacity, which is often associated with loss of service for patients and incurring overtime for healthcare workers. In many situations, patients are routed to non-primary wards or temporary beds when bed shortages happen (Dai and Shi 2018), which compromizes the quality of the care. In addition, healthcare overtime can threaten patient safety, nurse turnover, and physician burnout (see *e.g.,* Rogers et al. 2004, Stimpfel et al. 2012, Shanafelt et al. 2017).

It is difficult, in general, to derive a good scheduling policy for such realistic advance admission scheduling problems. Challenges persist for two major reasons: the state space is large and the objectives are hard to quantify. First, scheduling decisions are made online. At the moment when patients are scheduled, future information (*e.g.*, future arrivals, no-shows) is unobserved. Scheduling decisions amount to firm commitments, and they cannot be modified when future information becomes available. Problems of this sort typically have a huge state space. A dynamic programming model that minimizes expected costs can be computationally intractable even for very small instances (Liu et al. 2010). Hence for a general, real-world advance admission scheduling problem, it is impossible to account for—much less anticipate—all future uncertainties and dynamics. Second, a clear objective function is hard to define. The healthcare system's main concerns are the value of care and operational constraints. The former refers to the trade-offs between the quality and costs of care, but the quality of overall care cannot be easily translated into monetary values. From this it follows that the trade-off between quality and cost is elusive. Therefore, we may not have accurate cost information for using traditional approaches such as those in multi-criteria optimization. As regards the latter concern, we remark that resource constraints associated with the availability

of shared resources impose restrictions on daily operations. Because of the stochastic nature of resource usage, it is impossible to deliver a risk-free schedule that precludes any resource overutilization in daily operations. That is, there is always some chance that the capacity of resources will be exceeded. A good scheduling plan should avoid, to the greatest extent possible, the overutilization of any resource. To safeguard the system from resource overutilization, it is desirable to consider both the likelihood *and* the magnitude of overutilization—factors that jointly determine the risk of overutilization.

In practice, the hospital's objective is to meet operational targets in its daily operations. More specifically, they wish to serve all requesting patients while mitigating the risk of overutilization. On the ground, many hospitals' current practice is based mainly on a first-come, first-served policy. This policy resembles the first fit (FF) algorithm, which schedules a patient to the first available day that fits him. In the context of the online bin packing problems, the FF algorithm also admits a good competitive ratio (see *e.g.,* Johnson 1973, Johnson et al. 1974). Yet the planner must expect that its resource capacities will not be exceeded. We aim to improve the current practice and develop a computationally tractable model that carefully accounts for and mitigates the risk of resource overutilization. Another cause of resource overutilization risk— besides the stochastic usage of resources—is the uncertain nature of future arrivals and random no-shows. To some extent, the model should be able to anticipate these factors.

To address these issues collectively, we propose a *resource satisficing* framework that reduces the risk of resource overutilization, which is defined via our *resource satisficing index* (RSI). The term "satisficing" is a portmanteau of "satisfy" and "suffice". This notion, introduced by Simon (1959), aims to achieve feasibility under conditions of uncertainty. In many real-world problems, the decision maker may not seek to optimize the profit or the cost. Instead, the goal is to operate—as much as possible—within particular resource constraints, or to find a solution that "satisfices" those constraints. Simon (1959) provides an example of selling a house and the agent, after an exploration phase in which she learns about the climate of her housing market, has identified a benchmark that would be used as a comparator in her decision making. In the spirit of satisficing, we use a benchmarked admission heuristic as a comparator, which reflects how the hospital administrators articulate the trade-offs concerning the risks of overutilization for various resources, and our goal is to reduce overutilization risks. The satisficing approach fits well in the healthcare context and circumvents the need for the healthcare administrators to accurately extract the various cost parameters and articulating the trade-offs between monetary and non-monetary outcomes.

To tackle the uncertainty of future events, we propose to anticipate a small number of future requests; we call these "potential patients". Along with assigning the actual patients on hand, we also propose a *partial adaptive scheduling policy* to anticipate several patients that may arrive in

the future. These patients are unobserved and their types are (for the moment) unknown. Our tractable scheduling policy can adapt its solution to the actual realization of patient types and, more importantly, scale well computationally with respect to the number of the potential patients. An intuitive benefit of considering some amount of potential patients is that it helps the planner intentionally leave spaces in the more flexible days (that is compatible with more types of patients) and also leave spaces for patients with high priority or severe conditions.

Literature review

We refer interested readers to Gupta (2007), Gupta and Denton (2008), Guerriero and Guido (2011), and May et al. (2011) for comprehensive reviews of appointment scheduling that cover such topics as intraday scheduling and multi-day scheduling. Our paper focuses on multi-day scheduling, which include allocation scheduling and advance scheduling. *Allocation scheduling* focuses on designing each day's allocation capacity. After observing all the patients on a waiting list, the decision variable is the number of patients on that list to serve during the day in question. Unattended patients remain on the waiting list. Gerchak et al. (1996) characterize an optimal policy for allocating capacity to regular and emergency patients, and Min and Yih (2010) extend those results for multiple patient classes. There are many works in this stream, including Huh et al. (2013), Min and Yih (2014), and Wang and Truong (2018). Wang and Truong study a multi-priority online scheduling policy for allocation scheduling problems, where a waiting list can extend over several days. The appointment system's capacity can be increased by incurring some overload cost, and the algorithm balances the waiting cost and the overload cost. Samiedaluie et al. (2017) study patient admission policies (allocation scheduling) in a neurology ward where there are multiple types of patients with different medical characteristics. They consider only hospital bed resource, and aim to reduce disutility from patient waiting.

In our context, appointment requests must be processed on the same day; hence a multi-day waiting list is unacceptable. Our problem constitutes an advance scheduling problem. *Advance scheduling* schedules subjects to future days in response to their requests, and these assignments are usually irrevocable. Few scholars have considered algorithms or policy related to advance scheduling problems. Some papers in this area include Liu et al. (2010), Liu et al. (2011), Souki and Rebai (2012), Feldman et al. (2014), and Truong (2015). Patrick et al. (2008) develop heuristics based on approximate dynamic programming for scheduling patients of different priorities. Liu et al. (2010) derive heuristics based on Markov Decision Process for scheduling homogeneous consultation appointments while assuming that the service time is deterministic. The authors are mainly concerned with time-varying no-shows and cancellation behavior. In our case, most of the risks stem from the random usage of resources such as service times. Truong (2015) studies an optimal online

policy for a two-class advance scheduling problem. She uses a dynamic programming approach and shows that the problem reduces to an allocation scheduling problem for which a structural policy exists. Yet that paper's results need not hold when there are more than two patient classes. Liu et al. (2019) extend the model to a more general setting and incorporate patients' length-of-stays. Recently, several works study patient appointment in a multistage and multidisciplinary healthcare network (see, *e.g.,* Diamant et al. 2018, Wang et al. 2019). Wang et al. study patient appointment scheduling in a multistage healthcare network, where patients go through a list of stations stochastically. They consider a dynamic programming formulation and schedule patient appointment slots under a myopic policy.

Our paper is also related to the bed management and elective admission literature (see *e.g.,* Kao and Tung 1981) as the advance admission decisions must account for the risk of bed shortages. Helm and Van Oyen (2014) study the hospital admission scheduling and control problem. They propose a Poisson-arrival-location model and formulate the strategic planning problem as mixed-integer linear optimization problem. They develop linearizing approximations to their metrics capturing patient blocking, where they assume a waiting list and model a repeated admission cycle. Shi et al. (2016) use the queueing approach and study the inpatient flow problem, and their model fits the data well. These works typically focus on the planning phase and consider steady state decisions.

This problem is also related to the online bin packing problem. The hospital's current practice is based on the first fit algorithm, which has a modest competitive ratio in the context of online bin packing (see *e.g.,* Johnson et al. 1974, Yao 1980, Seiden 2002). More recently, Gupta and Bandi (2018) develop a robust online algorithm for scheduling surgery requests that incorporates historical information of the arrival sequence. Wang et al. (2018) develop an algorithm for online revenue management problems that resemble, to a large extent, advance scheduling problems. They assume that a random number of customers of different types will arrive sequentially during the planning horizon. These customers use a deterministic amount of resources and must be assigned "on the fly" as they arrive.

Besides the appointment scheduling literature, our research is closely related also to convex measures of risk used in the area of stochastic finance (see *e.g.,* Foellmer and Schied 2002, Foellmer and Knispel 2011). Our risk metric is motivated by the satisficing measure (see *e.g.,* Aumann and Serrano 2008, Brown and Sim 2009, Brown et al. 2012) that reflects the risk of not achieving a target. The decision maker must specify this target beforehand. In our context, the target can be easily defined based on resource capacities. The satisficing framework has previously been used to deal not only with scheduling but also with portfolio selection and vehicle routing (see *e.g.,* Hall et al. 2015, Jaillet et al. 2016, Qi 2017). Recently, Long et al. (2019) discuss a unifying robustness optimization framework using the concept of satisficing. Related to our work, Xie et al. (2018)

develop a new *bed shortage index* (BSI), a variant of Aumann and Serrano (2008) riskiness index, which is closely related to the bed occupancy rate but captures more facets of the risk of bed shortage. Likewise, we are motivated to propose a new variant of the metric that generalizes BSI that can be applied to different types of resource usage including operating theaters, hospital beds, diagnostic imaging equipment, among others.

Contributions

Our paper's main contributions are as follows.

- We propose a satisficing framework that can address the a multidisciplinary advance admission scheduling problem. Our framework serves to mitigate the overutilization risk of multiple resources in a real-world advance admission scheduling problem with multiple patient types. Our proposed satsificing model circumvents the need for clearly defining a trade-off between monetary and nonmonetary objectives. This crucial practical advantage aligns with the hospital's objective. Our satisficing model also delivers new insights into resource satsificing problems. As far as we know, this is the only paper that considers multiple patient types and stochastic usage of multiple resources in addressing the advance admission scheduling problem.
- The satisficing objective of the framework is based on our proposed resource satisficing index (RSI), which generalizes the bed shortage index (BSI) of Xie et al. (2018). While the BSI is calibrated to coincide with expected utilization when the random usage is Poisson distributed, RSI has the flexibility of choosing different reference probability distributions that are commonly associated with different types of random resource usage. By minimizing the largest RSIs among all resources and time periods, we effectively mitigate the risk of exceeding the capacity for each resource. RSI also extends to regions where resources cannot be satisfied in expectation, which is relevant in the advance admission scheduling problem because of the need to temporarily overload the system.
- We propose the partial adaptive scheduling policy, which anticipates future uncertainties to some extent, while keeping the problem computationally tractable in practice. The optimization problem can then be formulated and solved via a converging sequence of mixed-integer optimization problems, which in practice can be solved effortlessly using state-of-the-art commercial solvers. Computation results show that our heuristic mitigates the risk of exceeding the resource capacities.

The rest of this paper proceeds as follows. In Section 2, we introduce the advance admission scheduling problem and describe the resource usage associated with different sources of patients. We also introduce the partial adaptive scheduling policy. In section 3, we propose and motivate

the resource satisficing index (RSI), which is a satisficing measure. Section 4 begins by formulating the resource satisficing optimization Problem (4). We then identify the key subproblem (5), and develop an efficient algorithm for solving it. We also discuss an extension to incorporate distributional ambiguity. In Section 5 we establish that our heuristic outperforms current practice, and we conclude by summarizing in Section 6.

Notation. We use the calligraphic font (*e.g.*, $\mathcal{A}$) to denote sets, and vectors are denoted with boldface lowercase letters (*e.g.*, $\boldsymbol{x}$). We use $[N]$ to denote the running index $\{1,2,3,\ldots,N\}$ for a known integer N. Random variables are marked with a tilde (*e.g.*, $\tilde{z}$). We adopt the convention that $\inf\emptyset=+\infty$, where $\emptyset$ is the empty set. We use $\mathbb{1}$ to represent the indicator function; thus $\mathbb{1}(\mathcal{C})=1$ if the set $\mathcal{C}$ is nonempty or $\mathbb{1}(\mathcal{C})=0$ if $\mathcal{C}$ is empty. We use $\mathbb{E}_{\mathbb{P}}$ to denote the expectation with respect to a probability measure $\mathbb{P}$. We use $\mathcal{G}$ to denote the ambiguity set of probability distributions. We use $\mathcal{P}_0(\Omega)$ to denote the set of all probability distributions on a set Ω of scenarios. Given a random variable $\tilde{\nu}\sim\mathbb{P}$ and a set $\mathcal{W}$, we denote $\tilde{\nu}\in\mathcal{W}$ to represent $\mathbb{P}[\tilde{\nu}\in\mathcal{W}]=1$ for all $\mathbb{P}\in\mathcal{G}$. For two random variables $\tilde{v}.\tilde{u}$, we use $\tilde{v}\leq\tilde{u}$ to represent state-wise dominance. Similarly, for some given constant $u\in\mathbb{R}$, we use $\tilde{\nu}\leq u$ to represent $\mathbb{P}[\tilde{\nu}\leq u]=1$.

2. Advance Admission Scheduling

On any particular day, admission requests are received from a set $\mathcal{A}$ of *requesting patients*, *e.g.,* the set of patients that requested for admission through the call center on this particular day. We wish to schedule them into diffrerent days on a planning horizon with T days in total. We have K different types of resources, which include operating theater usage and hospital beds, among others.

We use $\tilde{s}^d_{\ell k}$ to denote the random usage of the kth resource, d days after admitting the ℓth patient. To illustrate the modeling flexibility, suppose the ℓth requesting patient would be scheduled for surgery on the second day of admission and the kth resource corresponds to the operating theatre usage, then we will set $\tilde{s}^d_{\ell k}=0$ for $d\neq 1$ and $\tilde{s}^1_{\ell k}$ denotes the random surgery time performed on the day after admission. If the resource k corresponds to the bed usage, then $\tilde{s}^d_{\ell k}$ takes the value of one if he is still warded after d days or zero otherwise. The patient must be discharged after $\bar{d}$ days, hence for all resource k, $\tilde{s}^d_{\ell k}=0$ when $d>\bar{d}$. Note that it is easy to embed patient ℓ's no-show within $\tilde{s}^d_{\ell k}$ for which the realized resource usage would be zero, *i.e.*, $\{\tilde{s}^d_{\ell k}=0\}$ for all d and k.

Besides the requesting patients in $\mathcal{A}$, there are three other sources of patients, namely, *scheduled*, *emergency*, and *potential* that would influence the usage of recourses. Scheduled patients are those who are already admitted or scheduled for admission in the future, while emergency patients are unscheduled ones who are admitted on arrivals. We use a single random variable $\tilde{w}_{tk}$ to denote the

random usage of the kth resource at day t by scheduled and emergency patients. As a concrete example, let $\mathcal{D}_t$ denote the set of patients who have been admitted or scheduled for admission on day t, with $t \leq 0$ referring to patients admitted $-t$ days prior to day zero (*i.e.,* the current day) who are not discharged at $t = 0$. Then, for all $t \in [T], k \in [k]$, we have

$$\tilde{w}_{tk} = \sum_{\tau=t-\bar{d}}^{t} \sum_{\ell \in \mathcal{D}_\tau} \tilde{s}_{\ell k}^{t-\tau} + \sum_{\tau=1}^{t} \sum_{j=1}^{\tilde{n}_\tau} \tilde{v}_{jk}^{t-\tau} \tag{1}$$

where $\tilde{n}_\tau$ represents the random number of emergency arrivals on day τ and $\tilde{v}_{jk}^{t-\tau}$ denotes the usage of the kth resource by the jth emergency patient after admitted for $t - \tau$ days.

We have discussed how to model the resource usage by requesting patients, scheduled patients, and emergency patients. Now, it remains to address future elective arrivals and the corresponding scheduling decisions. The advance scheduling problem is a dynamic one; ideally, one should account for all future materializations of elective arrivals and the corresponding scheduling decisions. However, for tractability, people often use approximations. Wang et al. (2019) show in simulation that myopic policy already performs well in their dynamic scheduling setting. In this paper, we consider only a fixed number of *potential patients* to approximate future dynamics. Potential patients are those who may arrive in the future and will need to be scheduled for admission in some days in $[T]$. Since these patients are currently unobserved, we must anticipate possible scenarios, particularly the types of future arriving patients for which the distributions of respective resource usage could be determined empirically. A planner is uncertain about the types of potential patients. There is no restriction on the definition of type, which can be determined based on, *inter alia*, the discipline of surgery requested, patient characteristics and patient preferences. The incorporation of potential patients can serve to account for additional priority among patient types, flexibility among physicians, *etc...*

We let R be the total number of possible patients' types and the discrete random variable $\tilde{r}_i$ on support $[R]$ represents the random type of the ith potential patient for $i \in [I]$. The distribution of this random type can be obtained from historical data. A potential patient is of type $r \in [R]$ with probability p_r; that is, $\mathbb{P}[\tilde{r}_i = r] = p_r$ for all $i \in [I]$ and all $r \in [R]$. We use $\tilde{u}_{ikr}^d$ to represent the random usage of the kth resource by the ith potential patient, d days after his scheduled admission if his type is r. The distribution of $\tilde{u}_{ikr}^d$ can be characterized in a similar fashion as for the requesting patients. Notice that we do not necessarily consider all patient types, *i.e.,* $[R]$ can be just a subset of patient types. This is because we only need to anticipate potential patients with higher priority, who need to be admitted within a short amount of time. We adopt a heuristic approach and anticipate only a fixed number I of potential patients. The number I should be chosen based on validation and simulation. Here, we remark that we can let the number of potential patients be a random variable, $\tilde{I}$ – we can still attain an exact and tractable reformulation.

For the purpose of generality, we do not assume a single known probability distribution for random resource usage. Instead, we incorporate distributional ambiguity and consider a set $\mathcal{G}$ of possible (joint) probability distributions that characterize the system's random resource usage. In particular, $\mathcal{G}$ would be a singleton if the distribution is unique, which is the case if we know the underlying probability distribution.

ASSUMPTION 1. *For any distribution in $\mathcal{G}$ associated with the system's random parameters, we assume that the random variables $\tilde{r}_1, \ldots, \tilde{r}_I$ are independent and identically distributed. In addition, the random variables associated with the usage of resource k at time period t are independently distributed for different patients and that their moment generation functions exist.*

Partially adaptive scheduling policy

There are two sets of decision variables. First, $\boldsymbol{x} \in \{0,1\}^{|\mathcal{A}|\times T}$ is the *here-and-now* scheduling decision for the requesting patients $\mathcal{A}$; here $x_{\ell t} = 1$ if we schedule the ith requesting patient to day $t \in [T]$. Second, the *wait-and-see* scheduling policy $\bar{y}_{it} : [R]^i \to \{0,1\}$, for all $i \in [I], t \in [T]$, which is nonanticipative and corresponds to the scheduling decision for the ith potential patients as a function of the types of patients that have arrived prior to the ith patient. Observe that the fully adaptive scheduling policy, $\bar{y}_{it}$ would be associated with an exponential number of decision variables in I. Instead, to circumvent this *curse of dimensionality*, we propose the *partial adaptive scheduling policy*, which can be encoded as the function $\hat{y}_{it} : [R] \to \{0,1\}$, for all $i \in [I], t \in [T]$ so that $\hat{y}_{it}(r) = 1$ if we schedule the ith potential patient to day t when her realized type is r. To characterize this scheduling policy, we can define the decision variable, $\boldsymbol{y} \in \{0,1\}^{I\times T\times R}$ so that

$$\hat{y}_{it}(\tilde{r}_i) = \sum_{r\in[R]} y_{itr} \mathbb{1}(\tilde{r}_i = r),$$

where $y_{itr} = 1$ if we schedule the ith potential patient to day $t \in [T]$ if her type is r. This partial adaptive scheduling policy can also be adopted when there are scheduling decisions associated with emergency patients. Observe that the number of decision variables associated with the partial adaptive scheduling policy is polynomial in the sizes of I.

The total random usage $\tilde{\nu}_{tk}$ of resource k on day t is

$$\tilde{\nu}_{tk} = \tilde{w}_{tk} + \sum_{\tau\in[t]} \sum_{\ell\in\mathcal{A}} \tilde{s}_{\ell k}^{t-\tau} x_{\ell\tau} + \sum_{\tau\in[t]} \sum_{i\in[I]} \sum_{r\in[R]} \tilde{u}_{ikr}^{t-\tau} y_{i\tau r} \mathbb{1}(\tilde{r}_i = r) \quad \forall k \in [K], \forall t \in [T]. \tag{2}$$

Let Γ_{tk} be the capacity of resource $k \in [K]$ for day $t \in [T]$. In our model, if the usage of any type of resources $\tilde{\nu}_{tk}$ exceeds Γ_{tk} then we speak of encountering an overutilization in day t for resource k. For a concrete example, suppose $K = 2$ and the two resources are operating theater and hospital

Table 1 Summary of notations

Symbol	Description
$\mathcal{A}$	Set of requesting patients
T	Length of the horizon
I	Number of potential patients
R	Number of patient types
K	Number of resources
$\tilde{\nu}_{tk}$	Random usage of the kth resource on day t
Γ_{tk}	Capacity of the kth resource on day t
$\tilde{w}_{tk}$	Random usage of the kth resource on day t by admitted and emergency patients
$\tilde{s}^d_{\ell k}$	Random usage of the kth resource by the ℓth patient, d days after his admission
$\tilde{u}^d_{ikr}$	Random usage of the kth resource by the ith potential patient, d days after his admission, and if his type is r
$x_{\ell t}$	Admission decision of the ℓth requesting patient on day t
y_{itr}	Admission decision of the ith potential patient on day t if his type is r
ϕ_k	Benchmark utilization rate of the kth resource

bed. Then, let $\tilde{\nu}_{t1}$ represent the total surgery time required in day t, and let $\tilde{\nu}_{t2}$ represent the total number of beds required in day t. Hence, if $\tilde{\nu}_{t1}$ exceeds Γ_{t1}, then that day encounters overtime. Likewise, if $\tilde{\nu}_{t2}$ exceeds Γ_{t2}, then that day encounters a bed shortage.

We summarize the key notations in Table 1.

3. Resource Satisficing Index

Because $\tilde{\nu}_{tk}$ is random, we do not necessarily restrict $\tilde{\nu}_{tk} \leq \Gamma_{tk}$ for all realization of $\tilde{\nu}_{tk}$. Hence this inequality amounts to only a "soft" constraint whereby we must safeguard the system from overutilization as much as possible. We define a risk metric that makes good use of distributional information to describe and hedge against the risk of resource overutilization.

Let $\mathcal{Z}$ be the set of all random usages (*e.g.,* non-negative real-valued functions) defined on a set Ω of scenarios. For any risk metric $\rho\colon \mathcal{Z} \times \mathbb{R}_+ \to \mathbb{R}$, the term $\rho(\tilde{\nu}, \Gamma)$ describes the overutilization risk associated with the random usages, $\tilde{\nu} \in \mathcal{Z}$ in exceeding the capacity, Γ. Inspired by the the riskiness index of Aumann and Serrano (2008) and the adversarial impact measure in Long et al. (2019), we propose the resource satisficing index (RSI) to evaluate overutilization risk of a random resource usage, $\tilde{\nu}$, with respect to its capacity, $\Gamma > 0$ for which its value is zero in the absence of such risk. We introduce the original riskiness index before our resource satisficing index.

Definition 1 (Riskiness Index). The Aumann and Serrano riskiness index $\varphi[\tilde{\xi}]$ of a random variable $\tilde{\xi}$ representing an uncertain monetary loss of a gamble is the reciprocal of the absolute

risk aversion (ARA) of an individual with constant ARA who is indifferent to taking that gamble. By extending their definition to incorporate ambiguity aversion, we define the riskiness index as:

$$\varphi[\tilde{\xi}] = \inf\left\{\alpha \geq 0 : C_\alpha[\tilde{\xi}] \leq 0\right\} \tag{3}$$

where

$$C_\alpha[\tilde{\xi}] \triangleq \begin{cases} \inf\left\{u \;\middle|\; \tilde{\xi} \leq u\right\} & \text{if } \alpha = 0 \\ \alpha \log\left(\sup\limits_{\mathbb{P}\in\mathcal{G}} \mathbb{E}_{\mathbb{P}}\left[\exp\left(\tilde{\xi}/\alpha\right)\right]\right) & \text{if } \alpha \in (0,\infty) \\ \sup\limits_{\mathbb{P}\in\mathcal{G}} \mathbb{E}_{\mathbb{P}}\left[\tilde{\xi}\right] & \text{if } \alpha = \infty \end{cases}$$

denotes the worst-case certainty equivalent under an exponential utility function with risk tolerance parameter $\alpha \in \mathbb{R} \cup \{\infty\}$.

Notably, riskiness index is a class of satisficing measure characterized in Brown and Sim (2009) and Brown et al. (2012), which provides a performance guarantee on the probability and expectation of capacity violation. Specifically, for a capacity, Γ, and any magnitude of violation $\varepsilon > 0$:

$$\begin{aligned} &\mathbb{P}\left[\tilde{\xi} \geq \Gamma + \varepsilon\right] \leq \exp\left(-\tfrac{\varepsilon}{\varphi\left[\tilde{\xi}-\Gamma\right]}\right) \quad \forall \mathbb{P} \in \mathcal{G}; \\ &\mathbb{E}_{\mathbb{P}}\left[(\tilde{\xi}-\epsilon)^+\right] \leq \tfrac{\varphi\left[\tilde{\xi}-\Gamma\right]}{e} \exp\left(-\epsilon/\varphi\left[\tilde{\xi}-\Gamma\right]\right) \quad \forall \mathbb{P} \in \mathcal{G}, \end{aligned}$$

indicating that when the riskiness index is smaller, both theoretical bounds decrease much faster with ε. Although the probability bound is a trivial one when $\epsilon = 0$, we still have a guarantee on the expected capacity violation. In addition, the riskiness index belongs to the class of adversarial impact measures characterized in Long et al. (2019), which admits the following guarantee:

$$\mathbb{E}_{\mathbb{P}}\left[\tilde{\xi}\right] - \Gamma \leq \varphi\left[\tilde{\xi}-\Gamma\right] D_{KL}(\mathbb{P}\|\hat{\mathbb{P}}), \quad \forall \mathbb{P} \in \mathcal{P}_0(\Omega), \forall \hat{\mathbb{P}} \in \mathcal{G},$$

where $D_{KL}(\mathbb{P}\|\hat{\mathbb{P}})$ is the KL-divergence of $\mathbb{P}$ from $\hat{\mathbb{P}}$. Therefore, the expected resource overutilization under any arbitrary distribution is bounded by the riskiness index times the KL-divergence of this distribution from the ambiguity set.

Due to the potential difficulties of interpretation, we foresee challenges of introducing the riskiness index in the context of hospital operations. In contrast, the expected utilization rate is ubiquitous in the healthcare literature, which, despite its limitations, is often used as a proxy to infer the risk of overutilization. Our goal is to introduce a variant of the riskiness index that can be associated with expected utilization rate so that it is more interpretable to hospital administrators.

To address interpretability, Xie et al. (2018) develop a new *bed shortage index* (BSI) to evaluate the steady-state bed shortage risk via a bijective function mapping of the Aumann and Serrano (2008) riskiness index that is calibrated to coincide with expected utilization when the random usage is Poisson distributed, which is a natural probability distribution for characterizing the

emergency arrivals of patients to be warded. However, this distribution may not be appropriate for other types of random resource usage, such as the stochasticity of surgical times associated with operating theaters. Hence, we define the resource satisficing index (RSI), which has the flexibility of choosing different reference probability distributions that are commonly associated with different types of random resource usage.

DEFINITION 2 (RESOURCE SATISFICING INDEX). Given a random usage, $\tilde{\nu} \in \mathcal{Z}$ and capacity $\Gamma > 0$. For a given calibration function $\Phi_\Gamma : [0,\infty] \mapsto [0,1]$, parameterized by Γ and satisfies

1. $\Phi_\Gamma(0) = 0$,
2. $\Phi_\Gamma(\infty) = 1$, and
3. $\Phi_\Gamma(\alpha)$ is continuous and increasing in $\alpha \in [0,\infty]$,

the resource satisficing index (RSI) $\rho\colon \mathcal{Z} \times \mathbb{R}_+ \to \mathbb{R}$, is defined as

$$\rho(\tilde{\nu},\Gamma) \triangleq \begin{cases} \Phi_\Gamma(\varphi[\tilde{\nu}-\Gamma]) & \text{if } \sup\limits_{\mathbb{P}\in\mathcal{G}} \mathbb{E}_{\mathbb{P}}[\tilde{\nu}] \le \Gamma, \\ \sup\limits_{\mathbb{P}\in\mathcal{G}} \mathbb{E}_{\mathbb{P}}[\tilde{\nu}/\Gamma] & \text{otherwise.} \end{cases}$$

In contrast with the riskiness index and BSI, RSI also extends to regions where resources cannot be satisfied in expectation under ambiguity, *i.e.,* $\sup\limits_{\mathbb{P}\in\mathcal{G}} \mathbb{E}_{\mathbb{P}}[\tilde{\nu}] > \Gamma$. In advance scheduling in a public hospital, where it is mandatory to accommodate all the patients, it is sometime necessary to temporarily overload some of the resources to avoid infeasibility. Under such circumstances, RSI would coincide with expected utilization and avoid decision indifference whenever resources could not be met in expectation under ambiguity. When $\sup\limits_{\mathbb{P}\in\mathcal{G}} \mathbb{E}_{\mathbb{P}}[\tilde{\nu}] \le \Gamma$, by incorporating a *calibration function*, we transform the original riskiness index to a number in domain $[0,1]$. The calibration function should be chosen so that the RSI is closely related to the expected utilization rate. Since different types of resource usage are more naturally associated with different reference probability distributions, the calibration function should vary for different reference distributions. In a multiple-resource setting, different calibration functions can serve to normalize the overutilization risk associated with different resources. We next propose the calibration functions for some common families of reference probability distributions. This calibration renders the RSI closely related to the expected utilization rate, resulting in a more interpretable metric for healthcare administrators.

THEOREM 1. *For a given type of resource with capacity* Γ*, let* $\tilde{\nu}_\mu$ *represent a non-negative random resource usage, whose distribution belongs to a family of probability distributions parameterized by* μ *such that*

1. $C_\infty[\tilde{\nu}_\mu] = \mu$,

2. $C_0\left[\tilde{\nu}_\mu\right] > \Gamma$ *for all* $\mu > 0$, *and*
3. $C_\alpha\left[\tilde{\nu}_\mu\right]$ *is increasing in* μ *for given* $\alpha \in (0,\infty)$.

Define

$$g_\alpha(\mu) \triangleq C_\alpha\left[\tilde{\nu}_\mu\right].$$

Then, the calibration function $\Phi_\Gamma : [0,\infty] \mapsto [0,1]$,

$$\Phi_\Gamma(\alpha) \triangleq \begin{cases} 0 & \text{if } \alpha = 0 \\ g_\alpha^{-1}(\Gamma)/\Gamma & \text{if } \alpha \in (0,\infty) \\ 1 & \text{if } \alpha = \infty \end{cases}$$

is continuous and increasing in $\alpha \in [0,\infty]$. *Moreover,*

$$\rho\left(\tilde{\nu}_\mu, \Gamma\right) = \sup_{\mathbb{P}\in\mathcal{G}} \mathbb{E}_{\mathbb{P}}\left[\tilde{\nu}_\mu/\Gamma\right].$$

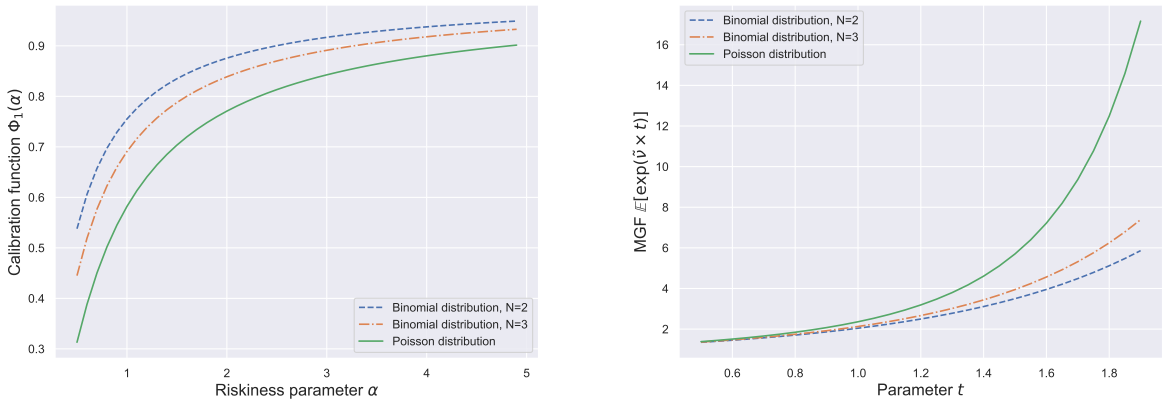
Proof. The proof can be found in Appendix A. ■

By Theorem 1, we can derive the calibration functions based on common distributions, such that the RSI of a random variable from these family of distributions would coincide with its expectation. For illustration, we provide some of these calibration functions in Table 2, and the derivation can be found in Appendix B.

Table 2 Calibrating functions for common distribution families.

Probability distribution of random usage, $\tilde{\nu}_\mu$	Calibration function, $\Phi_\Gamma(\alpha)$
Exponential distribution	$\frac{\alpha}{\Gamma}\left(1-\exp(-\Gamma/\alpha)\right)$
Gamma distribution with shape parameter κ	$\frac{\alpha\kappa}{\Gamma}\left(1-\exp(-\Gamma/(\alpha\kappa))\right)$
Poisson distribution	$\frac{1}{\alpha(\exp(1/\alpha)-1)}$
Binomial distribution with parameter N, $N > \Gamma$	$\frac{N(\exp(\Gamma/(\alpha N))-1)}{\Gamma(\exp(1/\alpha)-1)}$
Ambiguous distribution with mean μ and support parameter, $D > \Gamma$	$\frac{D(\exp(\Gamma/\alpha)-1)}{\Gamma(\exp(D/\alpha)-1)}$

For illustration, we plot calibration functions based on three reference distributions in Figure 1(a). As we can see, for the same value of riskiness parameter α, the calibration function based on Poisson distribution gives the lowest RSI value. This is because the Poisson distribution is "riskier" than Binomial distribution with parameter $N = 2, 3$, in terms of the moment generating function, when they all have the same mean. In fact, the moment generating function of a Poisson



(a) Calibration functions (b) Moment generating functions

Figure 1 Comparison of calibration functions and moment generating functions for three reference distributions.

distribution is always above that of a Binomial distribution with the same mean. For example, in Figure 1(b), we plot the moment generating functions for the three distributions, fixing the expected values to be the same ($\mu = 0.5$ in this instance). Therefore, when one chooses Poisson distribution as the reference distribution (*i.e.,* benchmarking on the Poisson distribution), any overutilization risk will appear to be more acceptable, compared to benchmarking on a Binomial distribution.

In practice, the distribution of a random usage does not necessarily belong to a common family of distributions as described in Theorem 1, *e.g.,* a summation of exponentially distributed random variables with different means does not belong to any such common family. In these cases, we will *not* be able to calibrate the RSI so that it coincides with the expected utilization rate. Nevertheless, we can still calibrate the RSI using an appropriate reference distribution and interpret the RSI as a "risk-adjusted" utilization rate, benchmarking on the reference distribution. Intuitively speaking, the RSI of a random variable would exceed or fall below its expected utilization rate depending on whether its distribution is "riskier" or "safer" than the reference distribution in terms of moment generating function. As a concrete example, let's consider a resource capacity $\Gamma = 1$ and a random resource usage, $\tilde{\nu}_\mu$, which follows a Poisson distribution with rate $\mu \leq 1$. By Theorem 1, we can calculate the riskiness index associated with any rate μ, and this allows us to calculate the RSI under different calibration function. Figure 2 illustrates the RSI of this Poisson random variable under three different calibration functions based on Binomial distribution and Poisson distribution. By Theorem 1, the RSI calibrated with Poisson distribution coincide with the expected utilization rate. For the same mean value μ, the RSI is the highest when calibrated using a Binomial distribution with $N = 2$. As we have discussed, this is because the Binomial distribution with parameter $N = 2$ is the "safest" among the three in terms of moment generating function;

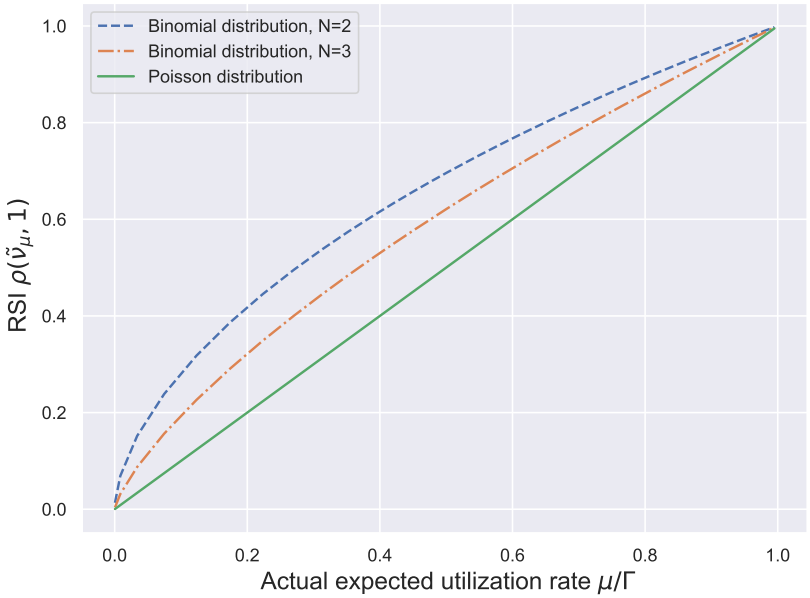


Figure 2 RSI calibrated with different reference distributions when the true distribution is a Poisson distribution.

therefore, the overutilization risk appears to be high. In the simulation study, we will discuss how to choose appropriate calibration functions in a data-driven way, benchmarking on the empirical distributions of resource usage.

THEOREM 2. *The RSI is a lower semi-continuous measure that has the following properties. For any* $\Gamma > 0$*, and random usages,* $\tilde{\nu}, \tilde{\nu}_1, \tilde{\nu}_2$*, we have*

1. *Monotonicity:* $\rho(\tilde{\nu}_1, \Gamma) \geq \rho(\tilde{\nu}_2, \Gamma)$ *if* $\tilde{\nu}_1 \geq \tilde{\nu}_2$.
2. *Quasi-convexity:* $\rho(\lambda\tilde{\nu}_1 + (1-\lambda)\tilde{\nu}_2, \Gamma) \leq \max\{\rho(\tilde{\nu}_1, \Gamma), \rho(\tilde{\nu}_2, \Gamma)\}$.
3. *Excess utilization: if* $\sup_{\mathbb{P}\in\mathcal{G}} \mathbb{E}_{\mathbb{P}}[\tilde{\nu}] > \Gamma$*, then* $\rho(\tilde{\nu}, \Gamma) = \sup_{\mathbb{P}\in\mathcal{G}} \mathbb{E}_{\mathbb{P}}[\tilde{\nu}/\Gamma] > 1$.
4. *Risk-free: if* $\tilde{\nu} \leq \Gamma$*, then* $\rho(\tilde{\nu}, \Gamma) = 0$.

Proof. The proof of Theorem 2 can be found in Appendix A. ■

Monotonicity states that if a random usage is less than another, then the former would be more preferred. Quasi-convexity is a common property that is consistent with preference under risk aversion, and also enables the underlying decision problem to be solved more efficiently. Note that quasi-convexity is preserved when we define the RSI to coincide with the expected utilization rate whenever the latter exceeds one, which is the third property. In contrast, the BSI of Xie et al. (2018) is indifferent to all random usages that are overutilized in expectation. In advance admission

scheduling, especially in a public hospital, the planner may temporarily overutilize resources to accommodate all requesting (arriving) patients instead of turning them away. For instance, more than 20% of the days in our simulation study have overutilized resources. Hence, as a decision criterion, RSI, which is sensitive the degree of overutilization, is more effective than BSI to address the advance admission scheduling problem. The last property states that the RSI is zero (*i.e.,* the most preferred) when the resource capacity will not be violated. The property rules out the expected resource utilization rate as a coherent metric for evaluating the risk of its overutilization.

An illustration of the RSI contours is shown in Figure 3. Each point in the figure represents a different surgery with a random surgery time $\tilde{\nu}$ that follows a two-point distribution. The actual surgery time is outcome 1 (the x-axis) with probability 0.7 and outcome 2 (the y-axis) with probability 0.3. Suppose the operating theater can only operate for $\Gamma = 1$ hour and the RSI is calibrated with exponential distribution. The solid lines are the RSI contours. The dashed line represents all random usages with an expected usage of 0.7. Consider all random usages along the dashed line, which have the same mean. The expected utilization rate cannot distinguish them, although it is clear that it is more risky towards the top left-hand corner (along the dashed line) because variance increases in that direction. The RSI, on the other hand, prefers those having lower variance, as the value is increasing toward the top left-hand corner (along the dashed line). Even though the distribution here is not an exponential distribution, the RSI is still closely related to the expected utilization rate – they are on a similar scale. The former deviates from the latter because it accounts for risks based on distributional information, as we can see the RSI increases with the variance in this example. This figure also illustrates the various properties listed in Theorem 2.

Before formally spelling out our satisficing model, we illustrate how the RSI can improve scheduling decisions. The only resource we consider in this example is the operating theater (*i.e.,* the maximum number of hours an operating theater can operate daily is $\Gamma = 10$ hours). Suppose there are two days into which we want to schedule four patients, and only one operating theater can be used on each day. Suppose further that the surgery times are

$$\tilde{\nu}_1 \sim \begin{cases} 6 & \text{w.p. } 0.5, \\ 4 & \text{w.p. } 0.5; \end{cases} \qquad \tilde{\nu}_2 \sim \begin{cases} 5 & \text{w.p. } 0.5, \\ 1 & \text{w.p. } 0.5; \end{cases} \qquad \nu_3 = 4; \qquad \nu_4 = 3.5.$$

We consider two models for scheduling four patients to these two days. The first model minimizes the largest expected utilization rate among the two days, whereas the second model minimizes the largest RSI among those days. The respective optimal scheduling decisions can be derived. In the first model, we assign patients 1 and 2 to one day and assign patients 3 and 4 to the other day. The probability that this arrangement will result in overtime is 0.25. In the second model, we assign patients 1 and 4 to one day and patients 2 and 3 to the other. In this scenario, we will never encounter overtime even in the worst case. So by considering our proposed risk metric, we can improve scheduling efficiency.

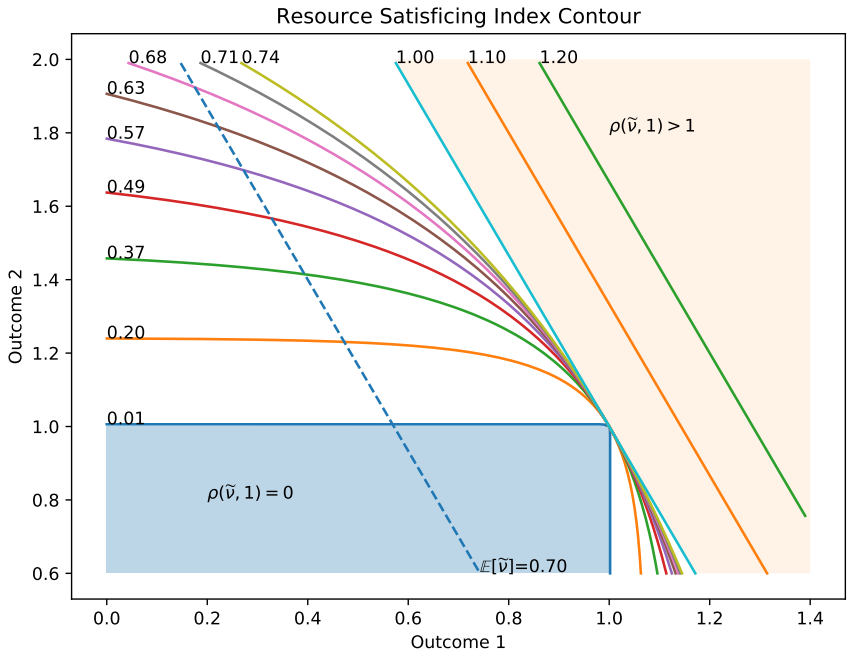


Figure 3 **The actual usage is either outcome 1 with probability 0.7 or outcome 2 with probability 0.3.**

4. Resource Satisficing Optimizing Problem

Admission schedule can significantly impact on the usage of resources over the planning horizon. Since all requesting patients must be served, a reasonable approach is to optimize the admission schedule that will spread out the risks of overutilization by minimizing the peak utilization rate over the planning horizon and for all resources (see *e.g.,* Teow et al. 2007, Meng et al. 2015).

To capture the overtilization risk, we minimize the maximum weight adjusted RSI over the planning horizon and for all resources as follows:

$$\max\left\{\frac{\rho_1(\tilde{\nu}_{11},\Gamma_{11})}{\phi_1},\ldots,\frac{\rho_K(\tilde{\nu}_{TK},\Gamma_{TK})}{\phi_K}\right\},$$

where ρ_k is the RSI associated with resource k, which has its own calibration function, and ϕ_k is the corresponding normalization parameter. We should not treat all resource usage uniformly, because some resource usages are inherently riskier than the others. Therefore, the flexibility of using different calibration functions is necessary when dealing with multiple resources. In our simulation study, we use a data driven approach to calibrate the RSI based on the benchmarked admission heuristic. This criterion ensures an equitable distribution of overutilization risks among all days and for all types of resources, hence reducing risk concentrations that may have severe negative impact on treatment outcomes. At the same time, the normalization parameters provide additional flexibility for the planner to vary the tradeoffs among different resources. Observe that if all random

resource usages have the same distributions associated with the corresponding RSIs' calibration functions, then $\rho_k(\tilde{\nu}_{tk}, \Gamma_{tk}) = \mathbb{E}_{\mathbb{P}}[\tilde{\nu}_{tk}/\Gamma_{tk}]$ and the objective would become one that minimizes the maximum normalized expected utilization rate. Hence, speaking intuitively, these normalization parameters may be selected to align with the hospital's desired expected utilization rates for the different resources.

MIP formulation, bisection search and Benders' decomposition

We model our resource satisficing advance admission scheduling problem as follows:

$$\begin{aligned} \beta^* = \min\ & \beta \\ \text{s.t. } & \rho_k(\tilde{\nu}_{tk}, \Gamma_{tk}) \le \beta\phi_k \quad \forall t \in [T],\ \forall k \in [K], \\ & (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X}, \end{aligned} \tag{4}$$

where

$$\mathcal{X} = \left\{ (\boldsymbol{x}, \boldsymbol{y}) \;\middle|\; \begin{array}{l} \displaystyle\sum_{t\in[T]} x_{\ell t} = 1\ \forall \ell \in \mathcal{A}, \\ x_{\ell t} \in \{0,1\}\ \forall \ell \in \mathcal{A},\ \forall t \in [T], \\ \displaystyle\sum_{t\in[T]} y_{itr} = 1\ \forall i \in [I],\ \forall r \in [R], \\ y_{itr} \in \{0,1\}\ \forall i \in [I],\ \forall t \in [T],\ \forall r \in [R], \\ \text{other mixed-integer constraints on } \boldsymbol{x}, \boldsymbol{y} \end{array} \right\}.$$

Apart from the assignment constraints in $\mathcal{X}$, we could include, among other things, patient availability on different dates, patient preferences to physicians, and the compatibility between patients and physicians. For potential patients with unknown type, we can also incorporate type-specific constraints, such as type-r patients can be admitted only on particular days. These constraints are practically important, because some types are high priority patients, *i.e.,* they can only wait for much fewer days than T.

We derive a solution method that solve Problem (4) via a sequence of mixed-integer optimization problems (MIPs), which can be applied using commercial solvers. This solution framework mainly uses bisection search and Benders decomposition. An important subroutine is checking whether there exists some feasible solution $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X}$ to Problem (4) for a fixed β. This feasibility problem is derived in the following theorem.

THEOREM 3. *There exists some feasible solution* $(\boldsymbol{x},\boldsymbol{y})\in\mathcal{X}$ *to Problem* (4) *for a fixed* β *if and only if the optimal solution* γ^* *to Problem* (5) *satisfies* $\gamma^*\le 0$.

$$\begin{aligned}
\min\ & \gamma \\
\text{s.t. }\ & w_{tk}(\infty)+\sum_{\ell\in\mathcal{A}}\sum_{\tau\in[t]} x_{\ell\tau}s_{\ell k}^{t-\tau}(\infty)+\sum_{i\in[I]}\sum_{\tau\in[t]}\sum_{r\in[R]} y_{i\tau r}u_{itrk}^{t-\tau}(\infty)-\Gamma_{tk}\beta\phi_k\le\gamma \quad \forall k\in\mathcal{K}_1, t\in[T], \\
& w_{tk}(\alpha_{tk})+\sum_{\ell\in\mathcal{A}}\sum_{\tau\in[t]} x_{\ell\tau}s_{\ell k}^{t-\tau}(\alpha_{tk})+\sum_{i\in[I]} f_{itk}(\alpha_{tk},\boldsymbol{y}_i)-\Gamma_{tk}\le\gamma \quad \forall k\in\mathcal{K}_2, t\in[T], \\
& (\boldsymbol{x},\boldsymbol{y})\in\mathcal{X},
\end{aligned} \tag{5}$$

where $\mathcal{K}_1=\{k:\beta\phi_k\ge 1\}$, $\mathcal{K}_2=\{k:\beta\phi_k<1\}$, $\boldsymbol{y}_i=(y_{i11},\ldots,y_{iTR})'$; *in addition,*

$$\alpha_{tk}\triangleq\Phi^{-1}_{\Gamma_{tk}}(\beta\phi_k),\ w_{tk}(\alpha)\triangleq C_\alpha\left[\tilde{w}_{tk}\right],\ s^d_{\ell k}(\alpha)\triangleq C_\alpha\left[\tilde{s}^d_{\ell tk}\right],\ u^d_{itrk}(\alpha)\triangleq C_\alpha\left[\tilde{u}^d_{ikr}\right],$$

and

$$f_{itk}(\alpha,\boldsymbol{y}_i)\triangleq\alpha\log\left(\sum_{r\in[R]} p_r\exp\left(\sum_{\tau\in[t]} y_{i\tau r}\frac{u^{t-\tau}_{itrk}(\alpha)}{\alpha}\right)\right).$$

Proof. The proof of Theorem 3 can be found in Appendix A. ■

Theorem 3 identifies a convex reformulation of the subproblem we need to solve within each iteration of a bisection search. The first set of $T|\mathcal{K}_1|$ constraints in Problem (5) are affine, and the second set of $T|\mathcal{K}_2|$ constraints are affine in $\boldsymbol{x}$ and convex in $\boldsymbol{y}$. This facilitates using a subgradient method to evaluate it. Observe that, unlike the conventional robust optimization models, the worst case distributions in $\mathcal{G}$ does not depend on the decisions $\boldsymbol{x},\boldsymbol{y}$ – it only appears in the functional C_α. In other words, we can calculate the worst-case certainty equivalent, *e.g.,* $w_{tk}(\alpha)$, before optimizing over the scheduling decisions, which greatly simplifies the problem. Interested readers can further refer to, *e.g.,* Ben-Tal et al. (2013), Wiesemann et al. (2014), Esfahani and Kuhn (2018).

The solution to Problem (5) guides the bisection search. Recall that β^* is the true optimal solution to Problem (4). If $\gamma^*\le 0$ in Problem (5) for fixed β, then $\beta\ge\beta^*$ and we can further reduce β; otherwise, $\beta<\beta^*$ and we must increase β. Hence, it follows that we can conduct a bisection search to solve Problem (4) as long as we can devise a subroutine that solves Problem (5) efficiently. Note that Problem (5) is a mixed-integer optimization problem with non-linear convex constraints. We invoke the following subgradient result.

PROPOSITION 1. *For any* $\boldsymbol{y}\in\mathbb{R}^{T\times R}$,

$$f_{itk}(\alpha,\boldsymbol{y})=\max_{\boldsymbol{v}\in\mathbb{R}^{T\times R}}\left\{f_{itk}(\alpha,\boldsymbol{v})+\sum_{r\in[R]}\sum_{\tau\in[t]} g^{\tau}_{itrk}(\alpha,\boldsymbol{v})\left(y_{\tau r}-v_{\tau r}\right)\right\} \tag{6}$$

where

$$g^{\tau}_{itrk}(\alpha,\boldsymbol{v}) = \frac{p_r u^{t-\tau}_{itrk}(\alpha)\exp\left(\sum\limits_{\tau_1\in[t]} v_{\tau_1 r} u^{t-\tau_1}_{itrk}(\alpha)/\alpha\right)}{\sum\limits_{m\in[R]} p_m \exp\left(\sum\limits_{\tau_2\in[t]} v_{\tau_2 m} u^{t-\tau_2}_{itmk}(\alpha)/\alpha\right)}$$

is the first order derivative of $f_{itk}(\alpha,\boldsymbol{v})$ with respect to $v_{\tau r}$. The worst-case of Problem (6) *occurs when $\boldsymbol{v}=\boldsymbol{y}$.*

Proof of Proposition 1. The proof follows trivially from the convexity of $f_{itk}(\alpha,\boldsymbol{y})$ with respect to $\boldsymbol{y}$. □

Therefore, Problem (5) can be equivalently represented with only affine constraints. We then solve this problem by way of the Benders decomposition (BD) algorithm described next.

BD Algorithm.

1. Initialize $\mathcal{Y}=\emptyset$ and input fixed β.
2. Solve the following subproblem:

$$\begin{aligned}
\min\ & \gamma \\
\text{s.t. }\ & w_{tk}(\infty)+\sum_{\ell\in\mathcal{A}}\sum_{\tau\in[t]} x_{\ell\tau}s^{t-\tau}_{\ell k}(\infty)+\sum_{i\in[I]}\sum_{\tau\in[t]}\sum_{r\in[R]} y_{i\tau r}u^{t-\tau}_{itrk}(\infty)-\Gamma_{tk}\beta\phi_k\le\gamma \quad \forall k\in\mathcal{K}_1, t\in[T], \\
& w_{tk}(\alpha_{tk})+\sum_{\ell\in\mathcal{A}}\sum_{\tau\in[t]} x_{\ell\tau}s^{t-\tau}_{\ell k}(\alpha_{tk})+\sum_{i\in[I]} f_{itk}(\alpha_{tk},\boldsymbol{v}_i) \\
& \qquad +\sum_{i\in[I]}\sum_{\tau\in[t]}\sum_{r\in[R]} g^{\tau}_{itrk}(\alpha_k,\boldsymbol{v}_i)\,(y_{i\tau r}-v_{i\tau r})-\Gamma_{tk}\le\gamma \quad \forall k\in\mathcal{K}_2, t\in[T], \boldsymbol{v}\in\mathcal{Y}, \\
& (\boldsymbol{x},\boldsymbol{y})\in\mathcal{X}.
\end{aligned} \tag{7}$$

   If $\gamma^*\le 0$, then record the optimal decisions $(\boldsymbol{x}^*,\boldsymbol{y}^*)$ and go to the next step. Otherwise, we can conclude that $\beta<\beta^*$, and we terminate this Benders decomposition procedure.

3. Check if

$$w_{tk}(\alpha_k)+\sum_{\ell\in\mathcal{A}}\sum_{\tau\in[t]} x^*_{\ell\tau}s^{t-\tau}_{\ell k}(\alpha_k)+\sum_{i\in[I]} f_{itk}(\alpha_k,\boldsymbol{y}^*_i)-\Gamma_{tk}\le 0 \quad \forall k\in\mathcal{K}_2, t\in[T].$$

   If it is true, go to the next step. Otherwise, include $\boldsymbol{y}^*$ in $\mathcal{Y}$ and return to step 2.

4. We conclude that $\beta\ge\beta^*$ and terminate this procedure.

There are two reasons why the BD algorithm will terminate within a finite number of iterations. First, the set $\mathcal{X}$ is finite. Second, the optimal decision $(\boldsymbol{x}^*,\boldsymbol{y}^*)$ from any BD iteration is always a new member to $\mathcal{Y}$. To see this, suppose the optimal decision $(\boldsymbol{x}^*,\boldsymbol{y}^*)$ from any BD iteration is

already in $\mathcal{Y}$. Then, by Proposition 1, it just means $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ must be feasible in Problem (5), which is a contradiction because the BD algorithm would have already terminated before this iteration.

We have devised a subroutine above that solves Problem (5) efficiently for any fixed β. Then it follows from the preceding discussion that we can conduct the bisection search and finally evaluate β^* at any level of accuracy. In Appendix C, we discuss the MIP formulation when the number of potential patients, $\tilde{I}$, is random.

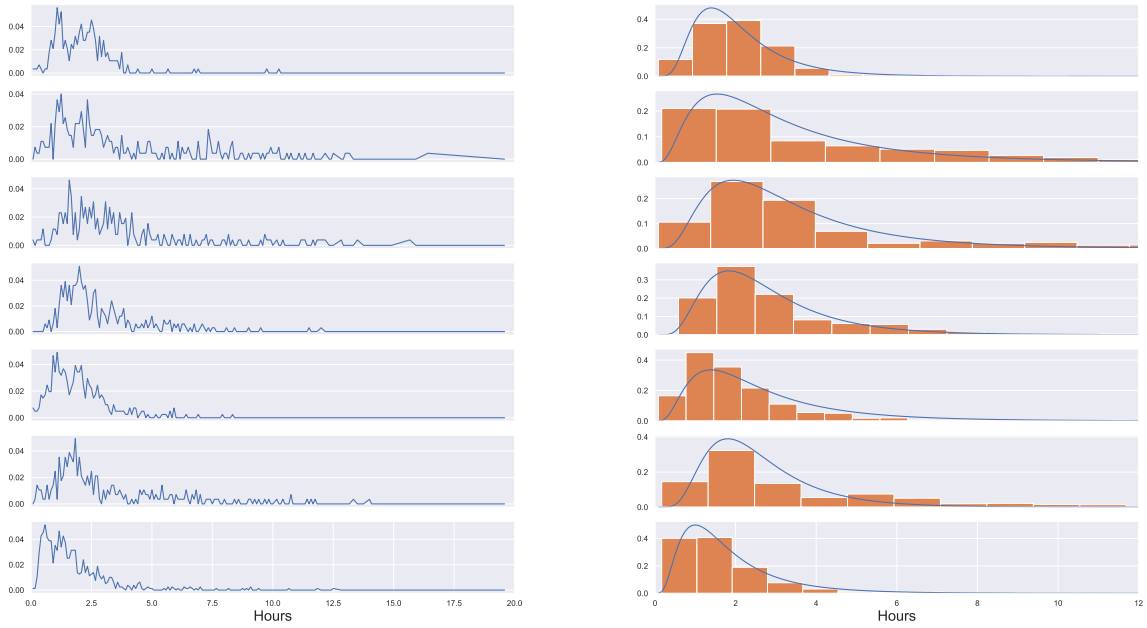
5. Simulation Study

In this section, we conduct a simulation study on our model (4), focusing on two resources, namely, operating theater (OT) and hospital beds usages. We first illustrate how to determine the calibration functions for different resources in a data-driven way. Then, we run simulations and illustrate that our model can reduce overutilization for multiple resources altogether without compromizing the number of patients served.

The setting and data

We collaborate with a public hospital in Singapore who seeks to optimize the OT usage, *e.g.,* balance the overall operating theater utiliztaion and reduce overtime. Improving OT usage requires improving the allocation of patient surgeries to different operating theaters and different days. For example, allocating several surgeries with high variance on surgery duration into one session can lead to high variance in utilization rate and severe overtime. To characterize OT resource usage, we obtain the empirical distribution on surgery duration of 16 different surgery disciplines. For illustration, we plot histograms for several disciplines in Figure 4.

To illustrate our model in an advance admission scheduling context, we focus on both the OT resource (*i.e.,* total surgery duration) and the bed availability, which is subject to patients' random length of stay. More specifically, we consider a setting where patients stay in the hospital for a random number of days after admission, and they go through a surgery on the day of admission. In our study, we consider four patient types (disciplines), each with Poisson distributed arrivals, truncated lognormal distributed (six hours maximum) surgery times and geometric distributed length-of-stays. The parameters are summarized in Table 3. The choice of lognormal distribution is guided by the empirical surgery duration distributions, and the associated parameters, μ and σ, are directly obtained from the data. The arrival probability is also obtained from the data, which is effectively the proportion of each patient type among the four chosen disciplines. We do not have data on the distributions of length-of-stay. For conveniency, we assume they follow geometric distributions. In addition, the arrivals of emergency patients on any day follow a Poisson distribution and the rates on different days are shown in Table 4, which is motivated by empirical evidence that there are more emergency patients on Mondays and Tuesdays, and fewer on Sundays.



(a) Raw data on surgery duration (b) Histogram of surgery duration distribution

Figure 4 A sample of empirical distributions of surgery duration.

Table 3 Summary of the distributional information

	Type			
Statistic	1	2	3	4
Surgery time parameter e^{μ}	1.83	2.13	0.84	1.31
Surgery time parameter σ	0.57	0.74	0.99	0.74
Hazard rate of length of stay	0.5	0.55	0.35	0.45
Probability of type	0.228	0.236	0.356	0.180

Table 4 Setup: Rate of emergency patients

	Weekday						
	Mon.	Tue.	Wed.	Thu.	Fri.	Sat.	Sun.
Rate of emergency patients	0.8	0.8	0.6	0.6	0.7	0.5	0.4

We set the bed capacity as $\Gamma_1 = 8$ beds and the daily operation theater duration as $\Gamma_2 = 8$ hours. Let λ_t denote the arrival rate of emergency patients on day t, the certainty equivalent of

the resource usage (1) by admitted and emergency patients can be written as:

$$\begin{aligned} C_\alpha[\tilde{w}_{tk}] &= \sum_{\tau=t-\bar{d}}^{t} \sum_{\ell \in \mathcal{D}_\tau} C_\alpha\left[\tilde{s}_{\ell k}^{t-\tau}\right] + \sum_{\tau=1}^{t} C_\alpha\left[\sum_{j=1}^{\tilde{n}_\tau} \tilde{v}_{jk}^{t-\tau}\right] \\ &= \sum_{\tau=t-\bar{d}}^{t} \sum_{\ell \in \mathcal{D}_\tau} C_\alpha\left[\tilde{s}_{\ell k}^{t-\tau}\right] + \sum_{\tau=1}^{t} \alpha\lambda_\tau \left(\exp(v_k^{t-\tau}) - 1\right), \end{aligned}$$

where

$$v_k^d = \log\left(\sum_{r\in[R]} p_r \exp\left(\frac{C_\alpha[\tilde{u}_{rk}^d]}{\alpha}\right)\right).$$

The inner certainty equivalent $C_\alpha[\tilde{u}_{rk}^d]$ can be calculated because the random usage term $\tilde{u}_{rk}^d$ has a known distribution. For instance, for bed usage, $\tilde{u}_{11}^d \sim \text{Bernoulli}(0.5^d)$.

To derive a steady-state result, in all following instances, we simulate the hospital's operations over one thousand days and then calculate average statistics for eight hundred days in the middle of that time span. In all comparisons, we use the same sample path to generate all arrivals and resource usages. In addition, every patient must be scheduled, *i.e.,* we do not reject any admission request.

Benchmarking against an admission heuristic

In the spirit of satisficing, we use a benchmarked admission heuristic as a comparator, and reduce the risk of resource overutilization of this comparator. Our benchmark in this study is the FF algorithm, which assigns patients in a first-come, first-served fashion while ensuring that the expected resource utilization rate does not exceed one. When overutilization is inevitable, we permit higher expected utilization rate to accommodate all requests in the FF algorithm. Though we focus on the FF algorithm, which is ubiquitous owing to its practicality, in general, we can use other admission heuristics adopted by the hospital.

We first implement FF algorithm and analyze the resulting usage of hospital beds and OT resource. In the first instance, we let the arrival rate of admission request be 2.7 and implement FF algorithm over one thousand days to obtain the empirical distributions of the daily usage of both resources. We provide the histograms of the empirical usage distributions in Figure 5 and Figure 6. The average utilization rate of OT resource is 74.5% and that of bed resource is 75.1%.

The empirical distributions describe resource overutilization risks under FF algorithm. The goal is to reduce overutilization risks of all resources simultaneously while referencing the empirical resource usage distributions. In other words, we do not treat bed usage and OT usage uniformly; instead, the RSIs would depend on their empirical usage distributions under the benchmark policy. However, we cannot directly use empirical distributions as reference distributions because they do not have corresponding calibration functions. For each resource, we choose a common reference

Figure 5 Histogram of daily surgery time (FF algorithm)

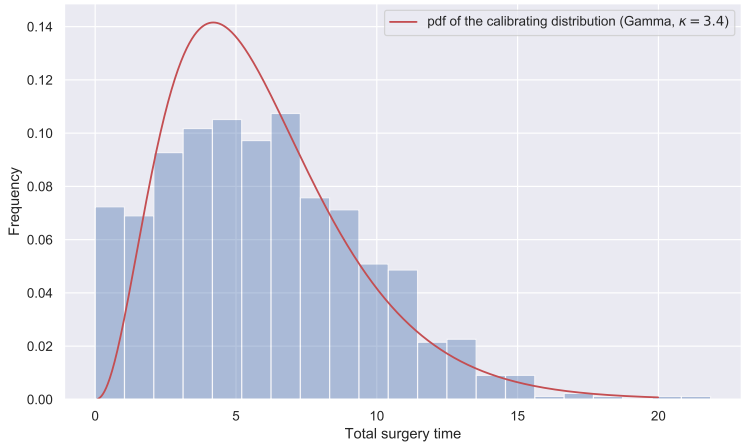
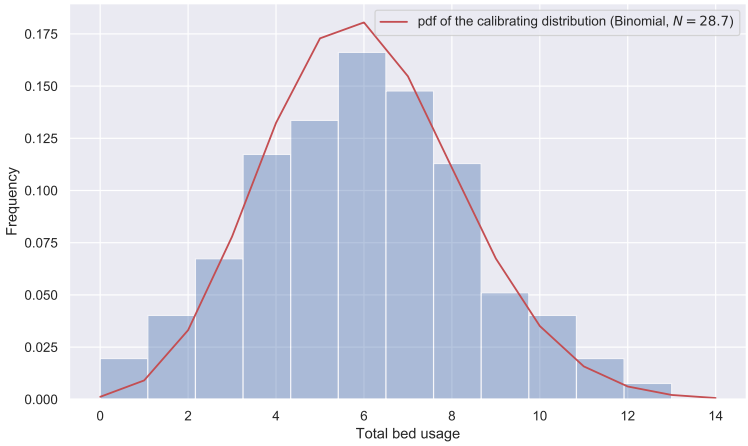


Figure 6 Histogram of daily bed usage (FF algorithm)



distribution from Table 2, which best matches its empirical usage distribution in terms of riskiness. The random bed usage is a counting process and hence, we use the Binomial distribution with parameter N as the reference distribution. The OT usage is associated with surgery time and we use Gamma distribution, with parameter κ as the reference distribution. We next describe how to choose appropriate parameters N and κ to match the empirical usage distributions.

By Equation (3), we calculate the riskiness index of the empirical bed usage under FF algorithm. We want to find a Binomial distribution that has the same mean and riskiness index as the empirical bed usage. In other words, we choose the parameter N such that the Binomial distribution with parameters N and $p = 0.751/N$ would have the same riskiness index as the empirical bed usage. By the result in Table 2, the calibrating parameter is set to $N = 28.7$. Similarly, we calibrate the daily surgery time with Gamma distribution. The calibrating parameter is set to $\kappa = 3.4$ so that the

Gamma distribution with shape parameter $\kappa = 3.4$ and scale parameter $\theta = 0.745/\kappa$ would have the same mean and riskiness index as the empirical OT usage. In Figure 5 and Figure 6, we also plot the corresponding reference distributions used in the calibrations. As we have expected, since the empirical and reference distributions have the same riskiness indices and expected utilization rates, their distributions also match well.

We solve Problem (4) by using the respective calibration functions and setting the reference utilization rates ϕ_1, ϕ_2 to be the average utilization rates under FF algorithm, so that the weight adjusted RSIs are at unit levels under the FF algorithm. Table 5 summarizes the performance comparison under this case.

Table 5 Performance comparison: Advance admission with bed management. Case 1.

		Satisficing		
Method	FF	$I=0$	$I=1$	$I=2$
Average overtime per day	43.5 min	35.3 min	33.4 min	32.9 min
Prob. of overtime	28.2%	24.3%	22.9%	21.8%
Prob. of overtime $>$ 1 hr	19.2%	16.7%	15.8%	15.8%
Prob. of overtime $>$ 2 hr	14.4%	10.3%	11.0%	10.1%
Prob. of overtime $>$ 3 hr	9.7%	7.3%	7.7%	7.1%
Proportion of days with bed shortage	13.7%	10.8%	9.6%	9.7%
Proportion of days with bed shortage $\geq$ 2 beds	7.7%	4.8%	3.9%	4.8%

As we can see in Table 5, we provide significant improvements in all metrics compared to the FF algorithm. In addition, considering some potential patients provides additional benefits. In practice, the number of potential patients, I should be chosen based on validation through simulations. In this instance, the overall performance on the variety of metrics when $I = 2$ is better compared to $I = 0$. However, we observed in our simulation study that larger values of I may not necessary lead to greater improvement. It helps to have smaller values of I because, as shown in Table 6, the computation time increases rapidly as I increases.

We vary the patient arrival rate slightly to verify the performance under different traffic intensity. For each case, the calibration functions are chosen according to our preceding discussion. The

Table 6 Average computation time for $I = 0, 1, 2, 3, 4$.

	$I=0$	$I=1$	$I=2$	$I=3$	$I=4$
Computational time (in seconds)	0.3	0.9	1.7	3.4	7.4

performance comparison is summarized in Table 7 when the average utilization rates of OT resource and bed are 68.0% and 66.2% respectively. The performance comparison is summarized in Table 8 when the average utilization rates of OT resource and bed are 83.4% and 83.3% respectively. From these cases, we find that anticipating more potential patients is not necessarily helpful when the average utilization is very high or very low, and when there is no difference in the priority of patients.

Table 7 Performance comparison: Advance admission with bed management. Case 1 (lighter).

		Satisficing		
Method	FF	$I=0$	$I=1$	$I=2$
Average overtime per day	38.3 min	27.2 min	24.8 min	25.7 min
Prob. of overtime	24.2%	18.5%	18.4%	17.8%
Prob. of overtime $>$ 1 hr	18.5%	13.3%	13.1%	11.6%
Prob. of overtime $>$ 2 hr	12.8%	8.4%	7.6%	8.4%
Prob. of overtime $>$ 3 hr	8.1%	4.8%	4.5%	5.5%
Proportion of days with bed shortage	6.8%	4.4%	4.7%	5.4%
Proportion of days with bed shortage $\geq$ 2 beds	2.2%	1.3%	1.1 %	1.1%

Table 8 Performance comparison: Advance admission with bed management. Case 1 (heavier).

		Satisficing		
Method	FF	$I=0$	$I=1$	$I=2$
Average overtime per day	52.8 min	44.2 min	43.6 min	45.4 min
Prob. of overtime	30.4%	30.0%	27.6%	30.2%
Prob. of overtime $>$ 1 hr	22.9%	22.4%	20.7%	21.9%
Prob. of overtime $>$ 2 hr	15.9%	14.5%	14.4%	14.4%
Prob. of overtime $>$ 3 hr	10.7%	9.3%	10.0%	9.6%
Proportion of days with bed shortage	17.8%	14.0%	14.2%	14.5%
Proportion of days with bed shortage $\geq$ 2 beds	8.0%	6.7%	6.9 %	5.8%

We also consider another setup where different days are compatible with different types of patients as shown in Table 9. Table 10 summarizes the performance comparison where the average utilization rates of OT resource and bed are 72.6% and 71.8% respectively. As we can see in Table 10, our performance is the best when $I=2$. Compared to the FF algorithm, we see significant improvements in all metrics.

Table 9 Setup: Acceptable patient types (case 2)

	Weekday						
	Mon.	Tue.	Wed.	Thu.	Fri.	Sat.	Sun.
Compatible types	1,2,3,4	1,3,4	1,2,3,4	1,3,4	1,2,3,4	1,3,4	1,3,4

Table 10 Performance comparison: Advance admission with bed management. Case 2.

		Satisficing		
Method	FF	$I=0$	$I=1$	$I=2$
Average overtime per day	44.8 min	37.3 min	35.7 min	34.2 min
Prob. of overtime	25.5%	24.6%	24.2%	23.2%
Prob. of overtime $>$ 1 hr	18.9%	17.7%	16.8%	15.5%
Prob. of overtime $>$ 2 hr	13.5%	10.3%	10.9%	9.9%
Prob. of overtime $>$ 3 hr	9.5%	6.8%	6.8%	6.6%
Proportion of days with bed shortage	12.9%	8.4%	9.4%	7.7%
Proportion of days with bed shortage $\geq$ 2 beds	5.1%	3.8%	3.8%	2.7%

6. Conclusion

Advance admission scheduling in general, and especially when the usage of relevant resources is stochastic, presents a difficult challenge. Past theoretical results do not directly apply to a realistic advance admission scheduling problem such as the one we describe. Given a public hospital's real-world problem, we propose a *resource satisficing* framework for the advance scheduling of patient admission and appointment. After developing a resource satisficing index, we use it to safeguard the system from the risk—in terms of both likelihood and magnitude—of resource overutilization. Because our approach incorporates multiple patient types, we can cluster patients beforehand in a data-driven way. Patient types can be clustered based on arbitrary characteristics: patients' demographic information, the discipline of surgery requested, patients' preferences for or compatibility with physicians or dates, and the condition (severity) of patients. Hence, we can better distinguish among differences (*e.g.*, arrival pattern, surgery time distribution, no-show behavior) among different patient types. By considering a certain number of potential patients of uncertain type, we aim to replace myopic scheduling decisions with more forward-looking decisions that account for the different characteristics among days, if there is any, and also reserve spaces for high priority patients. Our heuristic approach is computationally accessible and well suited to addressing realistic advance scheduling problems.

Acknowledgement. The authors would like to thank Shuangchi He for his valuable comments on an earlier version of this paper.

References

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

Ben-Tal, A., D. den Hertog, A. De Waegenaere, B. Melenberg, G. Rennen. 2013. *Robust solutions of optimization problems affected by uncertain probabilities*, vol. 59.

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

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

Chakraborty, S., K. Muthuraman, M. Lawley. 2010. Sequential clinical scheduling with patient no-shows and general service time distributions. *IIE Transactions* **42**(5) 354–366.

Dai, J.G., P. Shi. 2018. Inpatient bed overflow: An approximate dynamic programming approach. *Manufacturing and Service Operations Management, Forthcoming* .

Diamant, A., J. Milner, F. Quereshy. 2018. Dynamic patient scheduling for multi-appointment health care programs. *Production and Operations Management* **27**(1) 58–79.

Esfahani, P.M., D. Kuhn. 2018. Data-driven distributionally robust optimization using the wasserstein metric: performance guarantees and tractable reformulations. *Mathematical Programming* **171**(1) 115–166.

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

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

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

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

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

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

Gupta, D., C. Bandi. 2018. Operating-room staffing and scheduling URL `https://www.kellogg.northwestern.edu/faculty/bandi/OnlineStaffingAndScheduling.pdf`.

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

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

Helm, J.E., M.P. Van Oyen. 2014. Design and optimization methods for elective hospital admissions. *Operations Research* **62**(6) 1265–1282.

Huh, W.T., N. Liu, V-A. Truong. 2013. Multiresource allocation scheduling in dynamic environments. *Manufacturing & Service Operations Management* **15**(2) 280–291.

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

Johnson, D.S. 1973. Near-optimal bin packing algorithms. *PhD thesis, Massachusetts Institute of Technology* .

Johnson, D.S., A. Demers, J.D. Ullman, M.R. Garey, R.L. Graham. 1974. Worst-case performance bounds for simple one-dimensional packing algorithms. *SIAM Journal on Computing* **3**(4) 299–325.

Kao, E., G.G. Tung. 1981. Bed allocation in a public health care delivery system. *Management Science* **27**(5) 507–520.

Liu, N., V-A. Truong, X. Wang, B.R. Anderson. 2019. Integrated scheduling and capacity planning with considerations for patients' length-of-stays. *Production and Operations Management* .

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

Liu, Y., C. Chu, K. Wang. 2011. A new heuristic algorithm for the operating room scheduling problem. *Computers & Industrial Engineering* (3) 865–871.

Long, D.Z., M. Sim, M. Zhou. 2019. The dao of robustness URL `https://ssrn.com/abstract=3478930`.

Lowery, J.C., J.B. Martin. 1989. Evaluation of an advance surgical scheduling system. *Journal of Medical Systems* **12**(1) 11–23.

May, G.H., W.E. Spangler, D.P. Strum, L.G. Vargas. 2011. The surgical scheduling problem: Current research and future opportunities. *Production and Operations Management* **20**(3) 392–405.

Meng, F., J. Qi, M. Zhang, J. Ang, S. Chu, M. Sim. 2015. A robust optimization model for managing elective admission in a public hospital. *Operations Research* .

Min, D., Y. Yih. 2010. An elective surgery scheduling problem considering patient priority. *Computers and Operations Research* **37**(6) 1091–1099.

Min, D., Y. Yih. 2014. Managing a patient waiting list with time dependent priority and adverse events. *RAIRO - Operations Research* **48**(1) 53–74.

Patrick, J., M.L. Puterman, M. Queyranne. 2008. Dynamic multi-priority patient scheduling for a diagnostic resource. *Operations Research* **56**(6) 1507–1525.

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

Rogers, A.E., W.T. Hwang, L.D. Scott, L.H. Aiken, D.F. Dinges. 2004. The working hours of hospital staff nurses and patient safety. *Health Affairs* **23**(4) 202–212.

Samiedaluie, S., B. Kucukyazici, V. Verter, D. Zhang. 2017. Managing patient admissions in a neurology ward. *Operations Research* **65**(3) 635–656.

Seiden, S. 2002. On the online bin packing problem. *Journal of the ACM* **49**(5) 640–671.

Shanafelt, T., J. Goh, C. Sinsky. 2017. The business case for investing in physician well-being. *JAMA Internal Medicine* **177**(12) 1826–1832.

Shi, P., M.C. Chou, J.G. Dai, D. Ding, J. Sim. 2016. Models and insights for hospital inpatient operations: Time-dependent ed boarding time. *Management Science* **62**(1).

Simon, H.A. 1959. Theories of decision-making in economics and behavioral science. *The American Economic Review* **49**(3) 253–283.

Souki, M., A. Rebai. 2012. Heuristics for the operating theatre planning and scheduling. *Journal of Decision Systems* **19**(2) 225–252.

Stimpfel, A.W., D.M. Sloane, L.H. Aiken. 2012. The longer the shifts for hospital nurses, the higher the levels of burnout and patient dissatisfaction. *Health affairs* **31**(11) 2501–2509.

Teow, K.L., B.H. Heng, Y. Chong, J. Sim. 2007. Using mathematical programming to optimize bed occupancy by smoothing elective against emergency admissions. *NHG Annual Scientific Congress* **55**.

Truong, V-A. 2015. Optimal advanced scheduling. *Management Science* **61**(7) 1584–1597.

Wang, D., K. Muthuraman, D. Morrice. 2019. Coordinated patient appointment scheduling for a multistation healthcare network. *Operations Research* **67**(3) 599–618.

Wang, X., V-A. Truong. 2018. Multi-priority online scheduling with cancellations. *Operations Research* **66**(1) 104–122.

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

Wiesemann, W., D. Kuhn, M. Sim. 2014. Distributionally robust convex optimization. *Operations Research* **62**(6) 1358–1376.

Xie, J., G.G. Loke, M. Sim, S.W. Lam. 2018. The analytics of bed shortages: Coherent metric, prediction and optimization URL `http://dx.doi.org/10.2139/ssrn.3041878`.

Yao, A.C.C. 1980. New algorithms for bin packing. *Journal of the ACM* **27**(2).

Appendix

A. Proof of Results

Proof of Theorem 1. For a given Γ and a non-negative random resource usage $\tilde{\nu}_\mu$ whose distribution belongs to a family of distributions parameterized by the mean μ, the riskiness index $\varphi[\tilde{\nu}_\mu - \Gamma]$ can be written as:

$$\begin{aligned} \varphi[\tilde{\nu}_\mu - \Gamma] = \min\ & \alpha \\ \text{s.t. } & g_\alpha(\mu) \le \Gamma, \\ & \alpha \ge 0, \end{aligned} \tag{8}$$

where $g_\alpha(\mu) \triangleq C_\alpha[\tilde{\nu}_\mu]$ is as stated in the theorem. By the property of the certainty equivalent C_α, we have $\lim_{\alpha\to\infty} g_\alpha(\mu) = \mu$ and $\lim_{\alpha\to 0} g_\alpha(\mu) = C_0[\tilde{\nu}_\mu] > \Gamma$.

By the conditions in the theorem, it is easy to see that $g_\alpha(\mu)$ is continuous, decreasing in $\alpha \in (0, +\infty)$, and increasing in μ. Then, for any $\alpha \in (0, \infty)$, there exists some $\mu < \Gamma$ such that $g_\alpha(\mu) = \Gamma$, and we have $\alpha = \varphi[\tilde{\nu}_\mu - \Gamma]$.

Let $g_\alpha^{-1}(\Gamma)$ denote the inverse function of $g_\alpha(\mu)$ with respect to μ for $\alpha \in (0, \infty)$, which is also continuous for $\alpha \in (0, \infty)$. We define the calibration function $\Phi_\Gamma(\alpha) \triangleq \frac{g_\alpha^{-1}(\Gamma)}{\Gamma}$ for $\alpha \in (0, \infty)$. Then, for $\tilde{\nu}_\mu$ such that $\varphi[\tilde{\nu}_\mu - \Gamma] \in (0, \infty)$, which indicates $0 < \mu < \Gamma$, it follows that:

$$\mathbb{E}_{\mathbb{P}}[\tilde{\nu}_\mu / \Gamma] = \Phi_\Gamma(\varphi[\tilde{\nu}_\mu - \Gamma]). \tag{9}$$

Since $\lim_{\alpha\to+\infty} g_\alpha(\Gamma) = \Gamma$ and $g_\alpha^{-1}(\Gamma)$ is increasing in $\alpha \in (0, \infty)$, we have $g_\alpha^{-1}(\Gamma) < \Gamma$ for $\alpha < \infty$. In addition, $\lim_{\alpha\to\infty} g_\alpha^{-1}(\Gamma) = \Gamma$. We define $g_\infty^{-1}(\Gamma) = \Gamma$, which leads to $\Phi_\Gamma(\infty) = 1$ and $\Phi_\Gamma(\alpha) < 1$ for any $\alpha < \infty$. By the properties of the riskiness index, we know $\varphi[\tilde{\nu}_\Gamma - \Gamma] = \infty$. Hence, after defining this limiting case $\Phi(\infty) \triangleq 1$, equation (9) holds when $\mu = \Gamma$.

On the other hand, $g_0(\mu) = C_0[\tilde{\nu}_\mu] > \Gamma$ for all $\mu > 0$, which indicates that there does not exists $\mu > 0$ such that $g_0(\mu) = \Gamma$. Moreover, for any arbitrarily small $\alpha > 0$, there exists some $\mu > 0$ such that $g_\alpha(\mu) = \Gamma$ because the measure C_α is continuous and greater than or equals to the expectation. Therefore, $\lim_{\alpha\to 0} g_\alpha^{-1}(\Gamma) = 0$ and $g_\alpha^{-1}(\Gamma) > 0$ for $\alpha \in (0, \infty)$. To this end, we define $g_0^{-1}(\Gamma) \triangleq 0$, which leads to $\Phi(0) = 0$ and $\Phi(\alpha) > 0$ for $\alpha > 0$.

By this construction, the calibration function $\Phi : [0, \infty] \mapsto [0, 1]$,

$$\Phi(\alpha) \triangleq \begin{cases} 0 & \text{if } \alpha = 0 \\ g_\alpha^{-1}(\Gamma)/\Gamma & \text{if } \alpha \in (0, \infty) \\ 1 & \text{if } \alpha = \infty \end{cases}$$

is continuous and increasing in $\alpha \in [0, \infty]$. Moreover, when $\mu \le 1$, then $\rho(\tilde{\nu}_\mu, \Gamma) = \mathbb{E}_{\mathbb{P}}[\tilde{\nu}_\mu]$ follows from equation (9). When $\mu > 1$, $\rho(\tilde{\nu}_\mu, \Gamma) = \mathbb{E}_{\mathbb{P}}[\tilde{\nu}_\mu]$ follows by definition.

□

Proof of Theorem 2. The proof follows from the proof in Xie et al. (2018). Here we briefly talk about the proof idea. Readers can refer to Xie et al. (2018) for more details.

Define the set $\mathcal{H}[\tilde{\nu}] = \{\alpha \geq 0 : C_\alpha[\tilde{\nu}] \leq \Gamma\}$. We have $\inf \mathcal{H}[\tilde{\nu}] = \varphi[\tilde{\nu} - \Gamma]$. Observe that $\mathcal{H}(\tilde{\nu}) \neq \emptyset$ if $\sup_{\mathbb{P}\in\mathcal{G}} \mathbb{E}_{\mathbb{P}}[\tilde{\nu}] \leq \Gamma$. This claim follows because $C_\alpha[\tilde{\nu}]$ is continuous in $\alpha \in (0, \infty)$ and $C_\infty[\tilde{\nu}] = \sup_{\mathbb{P}\in\mathcal{G}} \mathbb{E}_{\mathbb{P}}[\tilde{\nu}]$. In addition, by definition, $\rho(\tilde{\nu}, \Gamma) > 1$ if $\sup_{\mathbb{P}\in\mathcal{G}} \mathbb{E}_{\mathbb{P}}[\tilde{\nu}] > \Gamma$; otherwise $\rho(\tilde{\nu}, \Gamma) \leq 1$.

1. *Monotonicity.* Suppose $\tilde{\nu}_1 \geq \tilde{\nu}_2$. If $\sup_{\mathbb{P}\in\mathcal{G}} \mathbb{E}_{\mathbb{P}}[\tilde{\nu}_2] > \Gamma$, then the monotonicity follows because the RSI becomes the expected utilization rate for both $\tilde{\nu}_1, \tilde{\nu}_2$. If $\sup_{\mathbb{P}\in\mathcal{G}} \mathbb{E}_{\mathbb{P}}[\tilde{\nu}_1] > \Gamma$ and $\sup_{\mathbb{P}\in\mathcal{G}} \mathbb{E}_{\mathbb{P}}[\tilde{\nu}_2] \leq \Gamma$, the result is trivially true by definition. Now we focus on the case where $\sup_{\mathbb{P}\in\mathcal{G}} \mathbb{E}_{\mathbb{P}}[\tilde{\nu}_1] \leq \Gamma$. For any $\alpha \in \mathcal{H}[\tilde{\nu}_1]$, we must have $\alpha \in \mathcal{H}[\tilde{\nu}_2]$. This indicates $\mathcal{H}[\tilde{\nu}_1] \subseteq \mathcal{H}[\tilde{\nu}_2]$. The result then follows by taking the infimum and observing that Φ_Γ is increasing in $\alpha \in [0, \infty]$.
2. *Quasi-convexity.* First we show that $\Phi_\Gamma(\varphi[\tilde{\nu} - \Gamma])$ is quasi-convex in $\tilde{\nu}$. For any $\alpha \in \mathcal{H}[\tilde{\nu}_1] \cap \mathcal{H}[\tilde{\nu}_2]$, by the convexity of $C_\alpha[\cdot]$, we must have $\alpha \in \mathcal{H}[\lambda\tilde{\nu}_1 + (1-\lambda)\tilde{\nu}_2]$. Taking the infimum then gives us the quasi-convexity of $\varphi[\tilde{\nu} - \Gamma]$ with respect to $\tilde{\nu}$. Then, by the monotonicity of Φ_Γ, we can conclude that $\Phi_\Gamma(\varphi[\tilde{\nu} - \Gamma])$ is quasi-convex in $\tilde{\nu}$. Then, it is easy to analyze that $\rho[\cdot]$ is quasi-convex, because the expectation measure is also quasi-convex.
3. *Excess utilization.* This follows from definition.
4. *Risk-free.* If $\tilde{\nu} \leq \Gamma$, then $C_0(\tilde{\nu}) \leq \Gamma$. Hence, $\rho(\tilde{\nu}, \Gamma) = 0$.
5. *Lower semi-continuity.* To show that lower semi-continuity holds, we consider any converging sequence such that $\lim_{n\to\infty} \tilde{\nu}_n = \tilde{\nu}$. We need to show that for any $a \in \mathbb{R}$, if $\rho(\tilde{\nu}_n, \Gamma) \leq a$ for all n, then $\rho(\tilde{\nu}, \Gamma) \leq a$. When $\sup_{\mathbb{P}\in\mathcal{G}} \mathbb{E}_{\mathbb{P}}[\tilde{\nu}] > \Gamma$, any converging sequence contains a converging sequence such that $\sup_{\mathbb{P}\in\mathcal{G}} \mathbb{E}_{\mathbb{P}}[\tilde{\nu}_n] > \Gamma$ for all n. Then, it's easy to see that lower semi-continuity holds in this piece. When $\sup_{\mathbb{P}\in\mathcal{G}} \mathbb{E}_{\mathbb{P}}[\tilde{\nu}] < \Gamma$, the lower semi-continuity follows from the lower semi-continuity of the riskiness index (see *e.g.,* Xie et al. 2018). In addition, the boundary case $\sup_{\mathbb{P}\in\mathcal{G}} \mathbb{E}_{\mathbb{P}}[\tilde{\nu}] = 1$ is contained in the lower piece where $\rho(\tilde{\nu}, \Gamma) \leq 1$. It is straightforward to see that lower semi-continuity holds at this boundary as well.

□

Proof of Theorem 3. First observe that we can solve 4 by using a bisection search on β. For any fixed β, we need to solve the following feasibility problem:

$$\begin{aligned} &\min\ 0 \\ &\text{s.t. } \rho_k(\tilde{\nu}_{tk}, \Gamma_{tk}) \leq \beta\phi_k \quad \forall t \in [T],\ \forall k \in [K], \\ &\qquad (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X}. \end{aligned} \tag{10}$$

For any $t \in [T], k \in [K]$, we first need to express the constraint $\rho_k(\tilde{\nu}_{tk}, \Gamma_{tk}) \leq \beta\phi_k$ explicitly. Observe that when $\beta\phi_k > 1$, the constraint $\rho_k(\tilde{\nu}_{tk}, \Gamma_{tk}) \leq \beta\phi_k$ becomes $\mathbb{E}_{\mathbb{P}}[\tilde{\nu}_{tk}/\Gamma_{tk}] \leq \beta\phi_k$. When

$\beta\phi_k = 1$, the constraint $\rho_k(\tilde{\nu}_{tk}, \Gamma_{tk}) \le 1$ is equivalent to $\mathbb{E}_{\mathbb{P}}[\tilde{\nu}_{tk}/\Gamma_{tk}] \le 1$ by definition. When $\beta\phi_k < 1$, the constraint $\rho_k(\tilde{\nu}_{tk}, \Gamma_{tk}) \le \beta\phi_k$ becomes $\varphi[\tilde{\nu}_{tk} - \Gamma_{tk}] \le \Phi^{-1}_{\Gamma_{tk}}(\beta\phi_k)$, which is equivalent to $C_{\Phi^{-1}_{\Gamma_{tk}}(\beta\phi_k)}[\tilde{\nu}_{tk} - \Gamma_{tk}] \le 0$. As stated in the theorem, we define

$$\alpha_{tk} \triangleq \Phi^{-1}_{\Gamma_{tk}}(\beta\phi_k), w_{tk}(\alpha) \triangleq C_\alpha[\tilde{w}_{tk}], s^d_{\ell k}(\alpha) \triangleq C_\alpha\left[\tilde{s}^d_{\ell k}\right], u^d_{itrk}(\alpha) \triangleq C_\alpha\left[\tilde{u}^d_{ikr}\right].$$

To distinguish different values of $\beta\phi_k$, we split the index set $[K]$ into two sets: $\mathcal{K}_1 = \{k : \beta\phi_k \ge 1\}$ and $\mathcal{K}_2 = \{k : \beta\phi_k < 1\}$. Then, for any $k \in \mathcal{K}_1$ and $t \in [T]$, the constraint $\rho_k(\tilde{\nu}_{tk}, \Gamma_{tk}) \le \beta\phi_k$ is equivalent to

$$w_{tk}(\infty) + \sum_{\ell\in\mathcal{A}}\sum_{\tau\in[t]} x_{\ell\tau} s^{t-\tau}_{\ell k}(\infty) + \sum_{i\in[I]}\sum_{\tau\in[t]}\sum_{r\in[R]} y_{i\tau r} u^{t-\tau}_{itrk}(\infty) \le \Gamma_{tk}\beta\phi_k. \tag{11}$$

By the independence assumption, for any $t \in [T], k \in \mathcal{K}_2$, the constraint $\rho_k(\tilde{\nu}_{tk}, \Gamma_{tk}) \le \beta\phi_k$ is equivalent to

$$w_{tk}(\alpha_{tk}) + \sum_{\ell\in\mathcal{A}} C_{\alpha_{tk}}\left[\sum_{\tau\in[t]} \tilde{s}^{t-\tau}_{\ell k} x_{\ell\tau}\right] + \sum_{i\in[I]} C_{\alpha_{tk}}\left[\sum_{r\in[R]}\sum_{\tau\in[t]} \tilde{u}^{t-\tau}_{ikr} y_{i\tau r}\mathbb{1}(\tilde{r}_i = r)\right] \le \Gamma_{tk}. \tag{12}$$

We need to reformulate constraint (17) further. For any $\ell \in \mathcal{A}$, $t \in [T]$, $k \in \mathcal{K}_2$, and $\tau \in [t]$, the term $C_\alpha\left[\sum_{\tau\in[t]} \tilde{s}^{t-\tau}_{\ell k} x_{\ell\tau}\right]$ can be equivalently represented as $\sum_{\tau\in[t]} x_{\ell\tau} s^{t-\tau}_{lk}(\alpha)$. This claim follows because:

$$\begin{aligned} C_\alpha\left[\sum_{\tau\in[t]} \tilde{s}^{t-\tau}_{\ell k} x_{\ell\tau}\right] &= \alpha\log\left(\sup_{\mathbb{P}\in\mathcal{G}} \mathbb{E}_{\mathbb{P}}\left[\exp\left(\sum_{\tau\in[t]} \tilde{s}^{t-\tau}_{\ell k} x_{\ell\tau}/\alpha\right)\right]\right) \\ &= \sum_{\tau\in[t]} x_{\ell\tau}\alpha\log\left(\sup_{\mathbb{P}\in\mathcal{G}} \mathbb{E}_{\mathbb{P}}\left[\exp\left(\frac{\tilde{s}^{t-\tau}_{\ell k}}{\alpha}\right)\right]\right) \\ &= \sum_{\tau\in[t]} x_{\ell\tau} s^{t-\tau}_{lk}(\alpha). \end{aligned}$$

The second equality follows from the binary nature of $x_{\ell\tau}$ for all $\ell \in \mathcal{A}, \tau \in [T]$ and the fact that $\sum_{\tau\in[t]} x_{\ell\tau} \le 1$ for all $t \in [T]$.

Then, for any $i \in [I], t \in [T], k \in \mathcal{K}_2$, the term

$$C_\alpha\left[\sum_{r\in[R]}\sum_{\tau\in[t]} \tilde{u}^{t-\tau}_{ikr} y_{i\tau r}\mathbb{1}(\tilde{r}_i = r)\right]$$

can be equivalently represented as $f_{itk}(\alpha, \boldsymbol{y}_i)$, where $\boldsymbol{y}_i = (y_{i11}, \ldots, y_{itR})'$ and $f_{itk}(\alpha, \boldsymbol{y}_i)$ is defined in Theorem 3 as:

$$f_{itk}(\alpha, \boldsymbol{y}_i) \triangleq \alpha\log\left(\sum_{r\in[R]} p_r \exp\left(\sum_{\tau\in[t]} y_{i\tau r}\frac{u^{t-\tau}_{itrk}(\alpha)}{\alpha}\right)\right).$$

This claim follows from the structure of the partial adaptive scheduling function together with the binary nature of y_{itr}. To see this:

$$\begin{aligned}
&C_\alpha\left[\sum_{r\in[R]}\sum_{\tau\in[t]}\tilde{u}_{ikr}^{t-\tau}y_{i\tau r}\mathbb{1}(\tilde{r}_i=r)\right]\\
&\quad=\alpha\log\left(\sum_{r\in[R]}p_r\sup_{\mathbb{P}\in\mathcal{G}}\mathbb{E}_{\mathbb{P}}\left[\exp\left(\frac{\sum_{\tau\in[t]}\tilde{u}_{ikr}^{t-\tau}y_{i\tau r}}{\alpha}\right)\right]\right)\\
&\quad=\alpha\log\left(\sum_{r\in[R]}p_r\exp\left(\sum_{\tau\in[t]}y_{i\tau r}\log\left(\sup_{\mathbb{P}\in\mathcal{G}}\mathbb{E}_{\mathbb{P}}\left[\exp\left(\frac{\tilde{u}_{ikr}^{t-\tau}}{\alpha}\right)\right]\right)\right)\right)\\
&\quad=\alpha\log\left(\sum_{r\in[R]}p_r\exp\left(\sum_{\tau\in[t]}y_{i\tau r}\frac{u_{itrk}^{t-\tau}(\alpha)}{\alpha}\right)\right)\\
&\quad=f_{itk}(\alpha,\boldsymbol{y}_i).
\end{aligned}$$

The second equality follows because $y_{i\tau r}\in\{0,1\}$ for all $i\in[I],\tau\in[T],r\in[R]$ and $\sum_{\tau\in[t]}y_{i\tau r}\le 1$ for all $i\in[I],r\in[R]$. Therefore, by constraint (11) and constraint (12), for any fixed β, Problem (10) if feasible if and only if $\gamma^*\le 0$, where

$$\begin{aligned}
\gamma^*=\min\ &\gamma\\
\text{s.t. }\ &w_{tk}(\infty)+\sum_{\ell\in\mathcal{A}}\sum_{\tau\in[t]}x_{\ell\tau}s_{\ell k}^{t-\tau}(\infty)+\sum_{i\in[I]}\sum_{\tau\in[t]}\sum_{r\in[R]}y_{i\tau r}u_{itrk}^{t-\tau}(\infty)-\Gamma_{tk}\beta\phi_k\le\gamma\quad\forall k\in\mathcal{K}_1,t\in[T],\\
&w_{tk}(\alpha_{tk})+\sum_{\ell\in\mathcal{A}}\sum_{\tau\in[t]}x_{\ell\tau}s_{\ell k}^{t-\tau}(\alpha_{tk})+\sum_{i\in[I]}f_{itk}(\alpha_{tk},\boldsymbol{y}_i)-\Gamma_{tk}\le\gamma\quad\forall k\in\mathcal{K}_2,t\in[T],\\
&(\boldsymbol{x},\boldsymbol{y})\in\mathcal{X}.
\end{aligned}$$

Therefore, the result follows. □

B. Derivation of Calibration Functions in Table 2

Recall that the riskiness index can be written as Problem (8). We need to find the inverse function $g_\alpha^{-1}(\Gamma)$.

1. Exponential distribution: the certainty equivalent can be written as:
$$g_\alpha(\mu) = \alpha \log\left(\frac{1}{1-\mu/\alpha}\right),$$
for which the inverse function is $g_\alpha^{-1}(\Gamma) = \alpha(1-\exp(-\Gamma/\alpha))$.

2. Gamma distribution with shape parameter N: Let $\mu = N\theta$ for some scale parameter θ, the certainty equivalent can be written as:
$$g_\alpha(\mu) = -\alpha N \log\left(1-\mu/(\alpha N)\right),$$
for which the inverse function is $g_\alpha^{-1}(\Gamma) = \alpha N(1-\exp(-\Gamma/(\alpha N)))$.

3. Poisson distribution: the certainty equivalent can be written as:
$$g_\alpha(\mu) = \alpha\mu\left(\exp(1/\alpha)-1\right),$$
for which the inverse function is $g_\alpha^{-1}(\Gamma) = \frac{\Gamma}{\alpha(\exp(1/\alpha)-1)}$.

4. Binomial distribution with N independent trials, $N > \Gamma$: the certainty equivalent can be written as:
$$g_\alpha(\mu) = \alpha N \log\left(1-\frac{\mu}{N}+\frac{\mu}{N}\exp(1/\alpha)\right),$$
for which the inverse function is $g_\alpha^{-1}(\Gamma) = \frac{N(\exp(\Gamma/(\alpha N))-1)}{(\exp(1/\alpha)-1)}$.

5. Ambiguous distribution with mean μ and support $[0, D]$, $D > \Gamma$: To express $g_\alpha(\mu)$, we first find the worst-case moment generating function
$$\sup_{\mathbb{P}\in\mathcal{G}} \mathbb{E}_{\mathbb{P}}\left[\exp\left(\frac{\tilde{\nu}}{\alpha}\right)\right].$$
By duality, it is equivalent to:
$$\begin{aligned} &\inf\ s_0 + s_1\mu \\ &\text{s.t. } s_0 + s_1\nu - \exp\left(\tfrac{\nu}{\alpha}\right) \geq 0 \quad \forall \nu \in [0, D], \end{aligned}$$
which is a robust optimization problem. Using standard robust optimization techniques, we get:
$$\sup_{\mathbb{P}\in\mathcal{G}} \mathbb{E}_{\mathbb{P}}\left[\exp\left(\frac{\tilde{\nu}}{\alpha}\right)\right] = 1-\frac{\mu}{D}+\frac{\mu}{D}\exp\left(1/\alpha\right).$$
It indicates
$$g_\alpha(\mu) = \alpha\log\left(1-\frac{\mu}{D}+\frac{\mu}{D}\exp\left(D/\alpha\right)\right),$$
for which the inverse function is $g_\alpha^{-1}(\Gamma) = \frac{D(\exp(\Gamma/\alpha)-1)}{\exp(D/\alpha)-1}$.

Then, the respective calibration functions are $g_\alpha^{-1}(\Gamma)/\Gamma$.

C. Considering a Random Number of Potential Patients

When we only consider a deterministic number of potential patients, I, the total random usage $\tilde{\nu}_{tk}$ of resource k on day t is given by (2). Now, if we let the number of potential patients be a random variable, $\tilde{I}$, then the total random usage $\tilde{\nu}_{tk}$ of resource k on day t becomes

$$\tilde{\nu}_{tk} = \tilde{w}_{tk} + \sum_{\tau\in[t]}\sum_{\ell\in\mathcal{A}} \tilde{s}_{\ell k}^{t-\tau} x_{\ell\tau} + \sum_{\tau\in[t]}\sum_{i\in[\tilde{I}]}\sum_{r\in[R]} \tilde{u}_{ikr}^{t-\tau} y_{i\tau r} \mathbb{1}(\tilde{r}_i = r) \quad \forall k\in[K],\, \forall t\in[T].$$

The only difference here is that we have a summation over $i\in[\tilde{I}]$. Here, we assume the support of $\tilde{I}$ is finite. In particular, we assume the support is $\mathcal{I} = \{1,\ldots,\bar{I}\}$, and we let $\delta_j := \mathbb{P}\left[\tilde{I} = j\right]$ for $j\in\mathcal{I}$. We model our resource satisficing advance admission scheduling problem as follows:

$$\begin{aligned} \beta^* = \min\ & \beta \\ \text{s.t. }\ & \rho_k(\tilde{\nu}_{tk}, \Gamma_{tk}) \le \beta\phi_k \quad \forall t\in[T],\ \forall k\in[K], \\ & (\boldsymbol{x},\boldsymbol{y})\in\mathcal{X}, \end{aligned} \tag{13}$$

where

$$\mathcal{X} = \left\{ (\boldsymbol{x},\boldsymbol{y}) \;\middle|\; \begin{array}{l} \displaystyle\sum_{t\in[T]} x_{\ell t} = 1\ \forall \ell\in\mathcal{A}, \\ x_{\ell t}\in\{0,1\}\ \forall \ell\in\mathcal{A},\, \forall t\in[T], \\ \displaystyle\sum_{t\in[T]} y_{itr} = 1\ \forall i\in[\bar{I}],\, \forall r\in[R], \\ y_{itr}\in\{0,1\}\ \forall i\in[\bar{I}],\, \forall t\in[T],\, \forall r\in[R], \\ \text{other mixed-integer constraints on } \boldsymbol{x},\boldsymbol{y} \end{array} \right\}.$$

THEOREM 4. *There exists some feasible solution $(\boldsymbol{x},\boldsymbol{y})\in\mathcal{X}$ to Problem* (13) *for a fixed β if and only if the optimal solution γ^* to Problem* (14) *satisfies $\gamma^*\le 0$.*

$$\begin{aligned} \min\ & \gamma \\ \text{s.t. }\ & w_{tk}(\infty) + \sum_{\ell\in\mathcal{A}}\sum_{\tau\in[t]} x_{\ell\tau} s_{\ell k}^{t-\tau}(\infty) + \sum_{i\in[I]}\sum_{\tau\in[t]}\sum_{r\in[R]} y_{i\tau r} u_{itrk}^{t-\tau}(\infty) - \Gamma_{tk}\beta\phi_k \le \gamma \quad \forall k\in\mathcal{K}_1, t\in[T], \\ & w_{tk}(\alpha_{tk}) + \sum_{\ell\in\mathcal{A}}\sum_{\tau\in[t]} x_{\ell\tau} s_{\ell k}^{t-\tau}(\alpha_{tk}) + h_{tk}(\alpha_{tk},\boldsymbol{y}) - \Gamma_{tk} \le \gamma \quad \forall k\in\mathcal{K}_2, t\in[T], \\ & (\boldsymbol{x},\boldsymbol{y})\in\mathcal{X}, \end{aligned} \tag{14}$$

where the notations are as defined in Theorem 3. In addition,

$$h_{tk}(\alpha,\boldsymbol{y}) := \alpha\log\left(\sum_{j\in[\bar{I}]}\delta_j \exp\left(\sum_{i\in[j]}\frac{f_{itk}(\alpha,\boldsymbol{y}_i)}{\alpha}\right)\right), \tag{15}$$

where $\boldsymbol{y} = (\boldsymbol{y}_1', \boldsymbol{y}_2', \ldots, \boldsymbol{y}_{\bar{I}}')'$.

Proof of Theorem 4. The proof is similar to the proof of Theorem 3. We define

$$\alpha_{tk} \triangleq \Phi^{-1}_{\Gamma_{tk}}(\beta\phi_k), w_{tk}(\alpha) \triangleq C_\alpha\left[\tilde{w}_{tk}\right], s^d_{\ell k}(\alpha) \triangleq C_\alpha\left[\tilde{s}^d_{\ell k}\right], u^d_{itrk}(\alpha) \triangleq C_\alpha\left[\tilde{u}^d_{ikr}\right],$$

and $\mathcal{K}_1 = \{k : \beta\phi_k \geq 1\}$, $\mathcal{K}_2 = \{k : \beta\phi_k < 1\}$. Then, for any $k \in \mathcal{K}_1$ and $t \in [T]$, the constraint $\rho_k\left(\tilde{\nu}_{tk}, \Gamma_{tk}\right) \leq \beta\phi_k$ is equivalent to

$$w_{tk}(\infty) + \sum_{\ell\in\mathcal{A}}\sum_{\tau\in[t]} x_{\ell\tau} s^{t-\tau}_{\ell k}(\infty) + \sum_{j\in\mathcal{I}}\sum_{i\in[j]}\sum_{\tau\in[t]}\sum_{r\in[R]} \delta_j y_{i\tau r} u^{t-\tau}_{itrk}(\infty) \leq \Gamma_{tk}\beta\phi_k. \tag{16}$$

By the independence assumption, for any $t \in [T], k \in \mathcal{K}_2$, the constraint $\rho_k\left(\tilde{\nu}_{tk}, \Gamma_{tk}\right) \leq \beta\phi_k$ is

$$w_{tk}(\alpha_{tk}) + \sum_{\ell\in\mathcal{A}} C_{\alpha_{tk}}\left[\sum_{\tau\in[t]} \tilde{s}^{t-\tau}_{\ell k} x_{\ell\tau}\right] + C_{\alpha_{tk}}\left[\sum_{i\in[\tilde{I}]}\sum_{r\in[R]}\sum_{\tau\in[t]} \tilde{u}^{t-\tau}_{ikr} y_{i\tau r}\mathbb{1}(\tilde{r}_i = r)\right] \leq \Gamma_{tk}. \tag{17}$$

By the proof of Theorem 3, for any $\ell \in \mathcal{A}$, $t \in [T]$, $k \in \mathcal{K}_2$, and $\tau \in [t]$, the term $C_\alpha\left[\sum_{\tau\in[t]} \tilde{s}^{t-\tau}_{\ell k} x_{\ell\tau}\right]$ can be equivalently represented as $\sum_{\tau\in[t]} x_{\ell\tau} s^{t-\tau}_{lk}(\alpha)$. Then, for any $t \in [T], k \in \mathcal{K}_2$, the term

$$C_\alpha\left[\sum_{i\in[\tilde{I}]}\sum_{r\in[R]}\sum_{\tau\in[t]} \tilde{u}^{t-\tau}_{ikr} y_{i\tau r}\mathbb{1}(\tilde{r}_i = r)\right]$$

can be equivalently written as

$$h_{tk}(\alpha, \boldsymbol{y}) := \alpha\log\left(\sum_{j\in[\bar{I}]} \delta_j \exp\left(\sum_{i\in[j]} \frac{f_{itk}(\alpha, \boldsymbol{y}_i)}{\alpha}\right)\right),$$

where $\boldsymbol{y} = (\boldsymbol{y}_1', \boldsymbol{y}_2', \ldots, \boldsymbol{y}_{\bar{I}}')'$, $\boldsymbol{y}_i = (y_{i11}, \ldots, y_{itR})'$ and $f_{itk}(\alpha, \boldsymbol{y}_i)$ is defined in Theorem 3. This claim follows from the structure of the partial adaptive scheduling function together with the binary nature of y_{itr}. To see this:

$$\begin{aligned}
&C_\alpha\left[\sum_{i\in[\tilde{I}]}\sum_{r\in[R]}\sum_{\tau\in[t]} \tilde{u}^{t-\tau}_{ikr} y_{i\tau r}\mathbb{1}(\tilde{r}_i = r)\right] \\
&\quad= \alpha\log\left(\sum_{j\in[\bar{I}]} \delta_j \sup_{\mathbb{P}\in\mathcal{G}} \mathbb{E}_{\mathbb{P}}\left[\exp\left(\frac{\sum_{i\in[j]}\sum_{r\in[R]}\sum_{\tau\in[t]} \tilde{u}^{t-\tau}_{ikr} y_{i\tau r}\mathbb{1}(\tilde{r}_i = r)}{\alpha}\right)\right]\right) \\
&\quad= \alpha\log\left(\sum_{j\in[\bar{I}]} \delta_j \exp\left(\sum_{i\in[j]} \log\left(\sup_{\mathbb{P}\in\mathcal{G}} \mathbb{E}_{\mathbb{P}}\left[\exp\left(\frac{\sum_{r\in[R]}\sum_{\tau\in[t]} \tilde{u}^{t-\tau}_{ikr} y_{i\tau r}\mathbb{1}(\tilde{r}_i = r)}{\alpha}\right)\right]\right)\right)\right) \\
&\quad= \alpha\log\left(\sum_{j\in[\bar{I}]} \delta_j \exp\left(\sum_{i\in[j]} \log\left(\sum_{r\in[R]} p_r \sup_{\mathbb{P}\in\mathcal{G}} \mathbb{E}_{\mathbb{P}}\left[\exp\left(\frac{\sum_{\tau\in[t]} \tilde{u}^{t-\tau}_{ikr} y_{i\tau r}}{\alpha}\right)\right]\right)\right)\right) \\
&\quad= \alpha\log\left(\sum_{j\in[\bar{I}]} \delta_j \exp\left(\sum_{i\in[j]} \log\left(\sum_{r\in[R]} p_r \exp\left(\sum_{\tau\in[t]} y_{i\tau r}\log\left(\sup_{\mathbb{P}\in\mathcal{G}} \mathbb{E}_{\mathbb{P}}\left[\exp\left(\frac{\tilde{u}^{t-\tau}_{ikr}}{\alpha}\right)\right]\right)\right)\right)\right)\right)
\end{aligned}$$

$$= \alpha \log \left(\sum_{j\in[\bar{I}]} \delta_j \exp \left(\sum_{i\in[j]} \log \left(\sum_{r\in[R]} p_r \exp \left(\sum_{\tau\in[t]} y_{i\tau r} \frac{u_{itrk}^{t-\tau}(\alpha)}{\alpha} \right) \right) \right) \right)$$
$$= \alpha \log \left(\sum_{j\in[\bar{I}]} \delta_j \exp \left(\sum_{i\in[j]} \frac{f_{itk}(\alpha, \boldsymbol{y}_i)}{\alpha} \right) \right).$$

The second equality follows from the independence assumption. The fourth equality follows because $y_{i\tau r} \in \{0,1\}$ for all $i \in [\bar{I}], \tau \in [T], r \in [R]$ and $\sum_{\tau\in[t]} y_{i\tau r} \le 1$ for all $i \in [\bar{I}], r \in [R]$. Therefore, by constraint (16) and constraint (17), a fixed β is feasible if and only if $\gamma^* \le 0$, where

$$\begin{aligned}
\gamma^* = \min\ & \gamma \\
\text{s.t. } & w_{tk}(\infty) + \sum_{\ell\in\mathcal{A}} \sum_{\tau\in[t]} x_{\ell\tau} s_{\ell k}^{t-\tau}(\infty) + \sum_{i\in[I]} \sum_{\tau\in[t]} \sum_{r\in[R]} y_{i\tau r} u_{itrk}^{t-\tau}(\infty) - \Gamma_{tk}\beta\phi_k \le \gamma \quad \forall k \in \mathcal{K}_1, t \in [T], \\
& w_{tk}(\alpha_{tk}) + \sum_{\ell\in\mathcal{A}} \sum_{\tau\in[t]} x_{\ell\tau} s_{\ell k}^{t-\tau}(\alpha_{tk}) + h_{tk}(\alpha_{tk}, \boldsymbol{y}) - \Gamma_{tk} \le \gamma \quad \forall k \in \mathcal{K}_2, t \in [T], \\
& (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X}.
\end{aligned}$$

□

Theorem 4 identifies a convex reformulation of the subproblem we need to solve within each iteration of a bisection search. The first set of $T|\mathcal{K}_1|$ constraints in Problem (14) are affine, and the second set of $T|\mathcal{K}_2|$ constraints are affine in $\boldsymbol{x}$ and convex in $\boldsymbol{y}$. Similar to Proposition 2, we can implement a subgradient method to evaluate Problem (14).

PROPOSITION 2. *For any* $\boldsymbol{y} \in \mathbb{R}^{\bar{I}\times T\times R}$,

$$h_{tk}(\alpha, \boldsymbol{y}) = \max_{\boldsymbol{v}\in\mathbb{R}^{\bar{I}\times T\times R}} \left\{ h_{tk}(\alpha, \boldsymbol{v}) + \sum_{i\in[\bar{I}]} \sum_{r\in[R]} \sum_{\tau\in[t]} H_{itrk}^{\tau}(\alpha, \boldsymbol{v}) \left(y_{i\tau r} - v_{i\tau r} \right) \right\}, \tag{18}$$

where

$$H_{itrk}^{\tau}(\alpha, \boldsymbol{v}) := \frac{\mathrm{d}h_{tk}(\alpha, \boldsymbol{v})}{\mathrm{d}v_{i\tau r}} = \frac{\sum_{j=i}^{\bar{I}} \delta_j \exp \left(\sum_{i\in[j]} f_{itk}(\alpha, \boldsymbol{v}_i)/\alpha \right) g_{itrk}^{\tau}(\alpha, \boldsymbol{v}_i)}{\sum_{j\in[\bar{I}]} \delta_j \exp \left(\sum_{i\in[j]} f_{itk}(\alpha, \boldsymbol{v}_i)/\alpha \right)}$$

is the first order derivative of $h_{tk}(\alpha, \boldsymbol{v})$ *with respect to* $v_{i\tau r}$. *The worst-case of Problem* (18) *occurs when* $\boldsymbol{v} = \boldsymbol{y}$.

Proof of Proposition 2. The proof follows trivially from the convexity of $h_{tk}(\alpha, \boldsymbol{y})$ with respect to $\boldsymbol{y} \in \mathbb{R}^{\bar{I}\times T\times R}$. □

Therefore, Problem (14) can be equivalently represented with only affine constraints. We then solve this problem by way of the Benders decomposition (BD) algorithm as described in the main text.