Submitted to
manuscript

Joint Patient Selection and Scheduling under No-Shows: Theory and Application in Proton Therapy

Soroush Saghafian
Harvard Kennedy School, Harvard University, Cambridge, MA 02138, soroush saghafian@hks.harvard.edu

Nikolaos Trichakis
Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA 02139, ntrichakis@mit.edu

Ruihao Zhu
Institute for Data, Systems, and Society, Massachusetts Institute of Technology, Cambridge, MA 02139, rzhu@mit.edu

Helen A. Shih
Francis H. Burr Proton Therapy Center, Massachusetts General Hospital, Boston, MA 02114, hshih@mgh.harvard.edu

We study how to admit and schedule heterogeneous patients by using simple, interpretable, yet effective policies when capacity is scarce, no-show behavior is patient- and time-dependent, service duration and reward are deterministic but patient-dependent, and overtime is costly. Our work is motivated by the afore-mentioned operational challenges that typically face adopters of new technologies in the healthcare sector. We anchor our study on a partnership with the proton therapy center of Massachusetts General Hospital (MGH), which offers a new radiation technology for cancer patients. We formulate the problem as a non-linear integer optimization problem. However, since the solution to this formulation lacks both tractability and interpretability, to be relevant to practice, we limit our study to simple and interpretable policies. In particular, we propose a simple index-based rule and derive analytical performance guarantees for it. We also calibrate our model using empirical data from our partner hospital, and conduct a series of experiments to evaluate the performance of our proposed policy under practical circumstances. The analytical perfor-mance guarantees and our numerical experiments demonstrate (a) the strong performance of the proposed policies, and (b) their robustness to various practical considerations (e.g., to potential misspecification of no-show probabilities). Our results show that our proposed policy, despite being a simple and interpretable index-based rule, is capable of improving performance by about 20% at an organization such as MGH, and of delivering results that are not far from being optimal across a wide range of parameters that might vary between organizations. This suggests that the proposed policy can be viewed as an effective "one-fits-all" capacity allocation rule that can be used in a variety of environments in which operational challenges such as no-shows and overtime costs need to be navigated using simple and interpretable rules.

Key words: appointment scheduling, non-monotone submodular maximization, no-shows, index rules

1. Introduction

Motivation. Adoption of new treatment technologies often enables organizations in the healthcare sector to drastically improve their service quality and generate additional value. At the same time, adoption of such technologies typically requires substantial investments (*e.g.*, in new equipment or skilled labor). It is, therefore, crucial for adopters to allocate their new-technology-enabled service to users in a way that makes the best use of their installed capacity. On the surface, this crucial task is facilitated by the ample demand that flocks to adopters. Specifically, due to the advantages that such new technologies offer to users, organizations that adopt them typically face a high demand level compared to the installed capacity. This enables them to be more selective about which users to admit and when to schedule them to receive the service. However, these admission and scheduling decisions are impeded by common practical considerations that often arise in such contexts, most notably *no-shows*, *overtime costs*, and the need for *interpretable allocation rules.*

No-Shows. Users who are scheduled to receive service but do not show up can cause serious problems for adopters of new technologies particularly in light of their elevated opportunity and unused capacity costs. Whereas user's no-show propensity might be reduced by the new technology's higher service quality, it could also be offset by potentially increased scheduling lead times owing to increased demand. The consequences of no-shows is particularly compounded in services in which capacity allocation requires some advance coordination and planning with users, which make it impossible for the adopters to allocate the capacity assigned to a no-show user to a new one. For example, patients accepted to be treated with a new treatment technology need to be contacted in advance, informed, checked for their insurance and other documents, and scheduled for treatment starting a specific date. While waiting for their scheduled appointments, patients might decide to use the old treatment technology instead of waiting longer for the new one, especially if their health condition starts to degrade. Thus, even though capacity is allocated to a patient, s/he might not show up to use it. Furthermore, the lengthy advance planning and coordination that is required to shift the allocated capacity to another patient means that the allocated capacity will be wasted, since last minute replacements are often impossible in practice.

In view of the aforementioned considerations, selecting which users to serve from the set of interested users and when to schedule them to receive the service needs to be decided upon in a way that accounts for the induced no-show behavior of scheduled users. Of note, accurate predictions of who is likely to "show up" and the related time sensitivities, therefore, can exceedingly help adopters of new technologies to make better allocation and scheduling decisions. However, such accurate predictions are rarely available to adopters, since they deal with a newly introduced technology for which there is not enough data to reliably estimate show-up probabilities. Lack of data availability also impedes the ability to identify important consumer features that can serve as strong show-up predictors.

Overtime Costs. Arguably, the most common approach to compensate for the risk of wasted capacity due to no-shows is overbooking.[1] The flip side of overbooking is that it could lead to overtime operations, which could be costly for the service provider. In particular, for adopters in certain services, access to flexible or slack capacity on demand to cover overtime needs might be limited or very expensive. A hospital offering a new treatment technology, for example, might need to hire additional skilled labor that is knowledgeable about the new treatment, or compensate existing staff for overtime. Therefore, scheduling decisions need to carefully account for potential overtime costs.

Interpretable Allocation Rules. Allocation of scarce, highly valuable resources tends to be contentious in nature as it could have important implications for the allocatees' welfare. Consequently, it is often highly desirable for the allocation rules to be transparent and interpretable, so that they can be easily understood and trusted by the allocatees. This is especially important in sectors such as healthcare, because denying or providing treatment to patients, and sometimes even delaying the treatment for them due a prioritization rule in place, could be a life-or-death related decision (Bertsimas et al. 2013, Saghafian et al. 2014). Therefore, most adopters of disruptive new technologies tend to be reluctant to rely on complex "black-box" type algorithms as their allocation rules, and rather prefer to make use of interpretable and easy-to-implement rules instead.

Our Study. In this paper, we develop a procedure to assist organizations that face the above-mentioned issues with making two interwoven decisions: given a set of heterogeneous

[1] Penalizing users who do not show up is often not an effective mechanism, since such penalties cannot be very large due to a variety of regulatory constraints (among others). Thus, even when imposed, such penalties are not large enough to offset the opportunity cost accrued due to unused expensive technology. Moreover, in many non-profit organizations (*e.g.*, some hospitals), there is no tangible financial gain for the service provider that can be gained by imposing a penalty: the wasted capacity simply implies loss of some social good (*e.g.*, treatment that could be offered to a different patient).

users, and some limited service capacity over a time window, (a) who to allocate the scarce capacity to (*i.e.*, an admission decision), and (b) for those admitted, when to serve them (*i.e.*, scheduling decisions). Our study of these decisions under the foregoing issues is particularly motivated by the situation at the Proton Therapy center of our partner hospital, Massachusetts General Hospital (MGH).

By using protons rather than x-rays, Proton Therapy offers a superior technology for treating cancer compared to the traditional radiation therapy. Specifically, Proton Therapy offers two important advantages compared to the traditional x-ray-based radiation therapy: (a) more radiation delivered to the malignant tumor, and (b) less radiation delivered to the healthy tissues surrounding the tumor. In addition, Proton Therapy typically causes fewer and less severe side effects such as low blood counts, fatigue, and nausea. Due to these advantages, demand for Proton Therapy among cancer patients is currently extremely high. However, since the technology is relatively new, only a few facilities in the United States currently offer it (including a center at our partner hospital, MGH), and capacity at each of these facilities is fairly limited. Due to this typical high demand-to-capacity ratio, some patients face long service lead times further prolonged by the lengthy insurance approval and proton therapy planning process, in total taking a few (if not many) weeks. Because patients waiting to receive treatment might seek outside options (*e.g.*, traditional radiation therapy), MGH faces costly last minute cancelations (*i.e.*, no-shows).[2] Furthermore, compensating for the wasted capacity due to such cancelations through overbooking often translates to significant overtime costs at MGH. Finally, since decisions on who should be accepted to receive treatment via this new technology and when to serve accepted patients are sensitive in nature (e.g., could be a matter of life or death), MGH administrators are reluctant to implement "black-box" rules. Instead, they are in need of allocation rules that can be easily interpreted and described to patients, providers, and stakeholder.

Our Approach. We follow a step-by-step approach. Motivated by the situation at our partner hospital, we start by developing a framework to study joint admission and scheduling decisions for a given set of heterogeneous users under the risk of (a) patient- and

[2] For simplicity, we refer to all last minute cancelations as "no-shows." However, the reader should note that in Proton Therapy there might be several reasons for such cancelations, and some of them can be beyond the patient's discretion not to come. Example for last minute cancelations include medical (*e.g.*, sudden changes in medical conditions indicating a need for more chemotherapy or more time to recover from surgery or chemotherapy) and financial (*e.g.*, a much cheaper or logistically more convenient treatment option becoming available).

time-dependent no-shows, and (b) overtime operations. We then analyze the optimal policy, and find it to be complex and not interpretable for implementation in practice. Thus, as part of our step-by-step approach, we try to first gain an understanding of the complex problem faced by our partner hospital by considering a simplified model that can capture the main trade-offs discussed above. This allows us to develop a simple heuristic rule that is expected to work-well in practice. We then relax some of the simplification assumptions of our model, provide extensions, and make use of simulations calibrated with hospital data to test the performance of the proposed heuristic rule under more realistic scenarios that better represent the complex environment at our partner hospital.

A particular example of our step-by-step approach is the fact that we begin our analysis by gaining insights into the main trade-offs by assuming that each patient only requires one visit. We then relax this assumption and provide an extension by considering more realistic situations in which some patients require multiple (and a heterogenous number of) visits. Finally, we note that hospital administrators need to consider other practical issues that may exist in their practice but do not systematically affect the main trade-offs we study in this paper. For example, some Proton Therapy Centers may prefer to reserve a small portion of their capacity for specific use (e.g., a small block of time of one of their gantry rooms might be reserved for pediatric patients under anesthesia). We highlight that such preferences (a) can be easily incorporated into our heuristic rule[3], and (b) are relatively case-by-case without any systematic effect on the main driving forces we strive to study.

Our Contributions. Our main contributions can be summarized as follows.

- We develop a simple, interpretable, and easy-to-implement index-base rule that could be used to tackle joint admission and scheduling decisions for heterogeneous users under the risk of patient- and time-dependent no-shows, and overtime operations, faced by various hospitals, including MGH.

- We provide theoretical performance guarantee for our proposed index-based rule. From a technical perspective, our work extends the available results for the Generalized Assignment Problem (GAP). This is because, in contrast to the literature on GAP, we deal with overtime costs, and hence, an integer optimization problem with a nonlinear objective

[3] For example, reserved capacity can be first deducted from the available capacity, and the related patients can be removed from the set of patients awaiting capacity allocation. Our proposed allocation rule can then be used for the remaining patients and capacity.

function. Another key difference with the available studies on GAP is that we seek to find a policy that is simple and interpretable, can allow for no-shows, and is also robust to potential misspecifications of no-shows.

- We provide simulation-based evidence using hospital data for the effectiveness of our proposed policy. In addition, we find this policy to be robust to input data misspecifications (*e.g.*, patient- and time-dependent no-show probabilities, which are inherently hard to calibrate in practice). Our simulation experiments also reveal that this policy performs well under a wide range of factors that could vary among hospitals. Thus, besides being an easy-to-implement and effective policy, our proposed policy can be viewed as a "one-fits-all" rule that can be used in a variety of hospitals.
- Employing machine learning and predictive analytics, we also shed light on patient characteristics and useful learning approaches that can be utilized to effectively predict no-show risks. This can be valuable in practice, since predicting no-show risks and incorporating them in admission and scheduling decisions can yield substantial benefits. Predicting no-show, however, is a challenging task, since no-show risks often depend on a variety of patient characteristics as well as the delay in offering the capacity (*i.e.*, are patient- and time-dependent). Despite this challenge, our machine learning approaches show promising results for predicting no-show risks, and thus, offer new predictive tools that can be implemented in practice.

2. Model and Related Studies

We start by developing a simplified model that can capture the main trade-offs discussed in Section 1. In doing so, we take advantage of some of the observations we have made based on collaboration with our partner hospital (one of the authors of this paper is the medical director of Proton Therapy Center of our partner hospital). Before we formally introduce the model, we make three notable observations.

First, as we discuss in more in detail in Section 5, while patients submit their requests/applications in a dynamic manner, patient applications are batched and are reviewed only periodically and with a fixed review cycle (e.g., four weeks). That is, a group of physicians review all the collected patient applications (e.g., medical documents, history, etc.) in the beginning of each cycle and evaluate their suitability for Proton Therapy. Since reviewing applications require extensive evaluations, this makes "online" (i.e., on-the-spot) decision-making impossible. Instead, the practice involves collecting all the applications and making decisions in an "offline" manner. Once it is decided which patients should be accepted (among those reviewed in that cycle), accepted patients are then scheduled to receive treatment. Those who are reviewed but not accepted (i.e., are rejected) are informed. These rejected patients typically seek other treatment options (e.g., traditional X-ray therapy) and do not apply again, since waiting further can be detrimental. In addition, reviewed applications will not be re-reviewed in another cycle. Thus, decisions made at each review cycle do not have a systematic impact on those of the next review cycle. These facts allow us to focus on studying the problem faced by the practitioners at the beginning of each review cycle in isolation (i.e., as a static problem as opposed to a dynamic one).

Second, as is detailed in Section 5, there are various features in Proton Therapy that prevent assigning capacity that is freed up due to a last-minute no-show to a different patient. For instance, delivering treatment in Proton Therapy requires a physician (e.g., a radiation oncologist) to have enough time to develop a radiation treatment plan along with a dosimetry and medical physics team. In addition, prior to delivering the treatment, multiple other preprocessing steps need to be taken (e.g., clinical determination of region to be treated, quality assurance testing, and peer review of treatment plan, among others). Due to these preprocessing steps, no-shows of scheduled patients almost always result in wasted scheduled capacity. Thus, unlike many other healthcare appointment scheduling studies, we consider a model in which no-shows do not free-up capacity.

Third, we note that no-show probabilities in Proton Therapy mainly depend on when a patient's first visit is scheduled, and are not affected by his/her follow-up visits—if any. That is, once the treatment delivery process for a patient starts, the patient adherence to using subsequent appointments is very high.

With these observations from our partner hospitals in mind, we proceed by developing a simple model that captures the main trade-offs in admitting and scheduling patients. To this end, we start by assuming that each patient requires only one visit. The extension of the results gained to multiple visits is relatively straightforward, and we defer the related discussion and analysis to Section 6.

Consider a facility that has limited capacity to provide some medical service over some future time window that spans T time periods. A time period could be, for example, one day. On each such future time period there are C available service time slots. If service runs in excess of this capacity, overtime cost is incurred at a rate of θ per time slot. Service on each time period cannot run for more than $\overline{C}$ time slots under any circumstances. These capacity and cost structures reflect that staffing is usually a primary part of expenses for medical services. That is, given that schedules for staffing have to be made ahead of time, the capacity C should then be understood as the nominal capacity that has been already "paid for" in advance (we hereafter often refer to it as the "soft capacity"), and $\overline{C}$ as a physical constraint on the available technology and/or a regulatory limit on staffing.

There are N patients who seek to be admitted and scheduled for the service. Each patient belongs to one of K different classes, indexed by $k \in [K]$.[4] Let λ_k be the number of patients who belong to the kth class. Providing service to a patient of class k requires $l_k \in \mathbb{Z}$ time slots and generates a reward of $r_k \geq 0$.[5] A patient scheduled to receive service at time period t may or may not show up, depending on the class s/he belongs to (*i.e.*, his/her patient characteristics) and the time t (*i.e.*, the number of periods s/he has to wait to receive the service). Specifically, let $p_k(t)$ be the probability that a patient of class k shows up for service if scheduled for time period t. The mapping $p_k : [T] \to [0,1]$ is assumed to be non-increasing to reflect preference for receiving service earlier. Notably, although for a fixed time t no-shows might be subject to individual patient preferences, herein we assume that they are predominantly driven by the patient's class, which reflects common patient characteristics (*e.g.*, medical urgency). Patients who do not show up irrevocably depart the system, and no reward is collected. The parameters $\{r_k, \lambda_k, l_k, p_k(\cdot)\}_{k=1}^K$ are deterministic, class-dependent, and can be estimated from data. In Section 5, we discuss the calibration of all the model's parameters (via a case study) using data that we have collected from our partner hospital, MGH.

The facility's Decision Maker (DM) needs to choose which patients to admit and when to schedule them. Let $x_{k,t}$ denote the number of patients of class k scheduled for service at time period t (for $k \in [K]$ and $t \in [T]$). The *expected reward* to be collected by providing service at time period t is

$$\sum_{k \in [K]} p_k(t) r_k x_{k,t}, \tag{1}$$

[4] For any positive integer Z, we denote the set $1, 2, \ldots, Z$ with $[Z]$.

[5] In practice, these reward parameters are estimated using a variety of medical factors that determine the suitability of treatment for the patient. See Section 5, where we discuss in detail estimation of rewards and other parameters of our model.

and the *expected overtime cost* to be incurred at time period t is

$$\theta\mathbb{E}\left[\sum_{k\in[K]}\text{Binomial}\left(x_{k,t},p_k(t)\right)l_k - C\right]^+, \tag{2}$$

where $[\cdot]^+ = \max\{\cdot,0\}$ and Binomial(χ,ζ) is a binomial random variable with parameters (number of trials, success probability) = (χ,ζ). The *expected profit* to be made at time period t is the difference between the expected reward and the expected overtime cost. To introduce some notation for ease of presentation, let $G_t:\mathbb{Z}^K\to\mathbb{R}$ map the numbers of patients of each class scheduled for service at time period t, $x_{1,t},\ldots,x_{K,t}$, into the expected profit for that time period, *i.e.*,

$$G_t\left(x_{1,t},\ldots,x_{K,t}\right) := \sum_{k\in[K]}p_k(t)r_kx_{k,t} - \theta\mathbb{E}\left[\sum_{k\in[K]}\text{Binomial}\left(x_{k,t},p_k(t)\right)l_k - C\right]^+. \tag{3}$$

Similarly, let $G:\mathbb{Z}^{KT}\to\mathbb{R}$ map all patient admission and scheduling decisions, $\boldsymbol{x} = \{x_{k,t}\}_{k\in[K],t\in[T]}$, into the DM's expected profit, *i.e.*,

$$G\left(\boldsymbol{x}\right) := \sum_{t\in[T]}G_t\left(x_{1,t},\ldots,x_{K,t}\right). \tag{4}$$

The DM's problem is to make joint patient admission and scheduling decisions so as to maximize the expected profit. It can be formulated as the following nonlinear integer optimization problem

$$\begin{aligned}
\text{maximize}\quad & G\left(\boldsymbol{x}\right) && (5)\\
\text{subject to}\quad & \sum_{t\in[T]}x_{k,t}\le\lambda_k \qquad \forall k\in[K] && (6)\\
& \sum_{k\in[K]}l_kx_{k,t}\le\overline{C} \qquad \forall t\in[T] && (7)\\
& x_{k,t}\in\{0,1,\ldots,N\} \qquad \forall k\in[K],\forall t\in[T], && (8)
\end{aligned}$$

with variable $\boldsymbol{x}\in\mathbb{Z}^{KT}$.

Because G is a nonlinear, complex function, solving the optimization problem (5) – (8) is intractable for even moderate problem sizes. However, even if exact solutions to this problem were attainable, they would be of limited practical relevance due to lack of *interpretability.* In particular, a patient admittance approach that uses optimization problems such as $(5)-(8)$ operates essentially as a "black-box" that is nearly impossible to explain to or communicate with patients and/or physicians. Indeed, as noted in the Introduction, interpretability of admission and scheduling rules in utilizing a new technology is highly desirable in practice, if not necessary.

In addition to interpretability, our goal is to devise a policy with the following desiderata. First, to be implementable in practice, the policy needs to be computationally efficient and *scalable* in order to accommodate large-scale problem instances. Second, the policy needs to provably perform well, *i.e.*, have analytical *performance guarantees* vis-a-vis the optimal value of problem $(5)-(8)$. Finally, because it is often hard to accurately calibrate no-show probabilities (*e.g.*, due to lack of large-scale data caused by the fact that the technology is new and has not been offered for a sufficiently long-period of time), the policy's performance needs to be *robust* against potential misspecification of no-show probabilities.

To achieve our goal, we focus our attention on devising *index policies* that allow for making joint admission and scheduling decisions. By design, and owing to their simplicity, index policies are both interpretable and scalable. In the analysis that follows, we derive an index policy that also enjoys analytical performance guarantees, is robust against potential misspecification of no-show probabilities, and performs very well in numerical experiments calibrated with data from our partner hospital. Prior to doing so, however, we first briefly review the studies that are related to our work.

2.1. Related Studies

Owing to their pervasiveness and ubiquitous use in service operations, scheduling techniques have been studied in a large body of the literature. Herein, we make no attempt to survey the literature, but rather focus on papers that are closest to ours.

Healthcare Operations. Patient scheduling under no-shows has received a lot of attention in healthcare operations. Cayirli and Veral (2003) and Gupta and Denton (2008) provide a broad literature review of some earlier works in this stream. Most of the existing studies on patient scheduling under no-shows consider only homogeneous no-show probabilities (Kaandorp and Koole 2007, Hassin and Mendel 2008, LaGanga and Lawrence 2012). In Luo et al. (2012), the authors consider a patient scheduling problem with no-shows and service interruptions. Feldman et al. (2014) studies the inter-day appointment scheduling problem with homogeneous patients under no-shows and patient preferences.

Because the general problem is computationally intractable, they provide an optimal policy for the static model, and propose a heuristic solution for the dynamic model. The key difference of our approach is that we seek to derive interpretable index policies with performance guarantees.

Another different but closely related problem is scheduling of jobs with time varying status, motivated by disaster response scenarios (Argon et al. 2008, Chan et al. 2013). Different from these studies, we consider overtime costs and derive performance guarantees, alongside a robustness analysis. Master et al. (2016) considers a discrete time, multi-server system with jobs whose values decay as time elapses. Because the problem is intractable, they propose and analyze the performance of several approximation algorithms. Our model is closely related to that of Master et al. (2016), but with several notable differences; we consider overtime operations as well as multiple capacity constraints, which introduce new challenges as one needs to simultaneously and carefully balance profit generation and capacity consumption.

Other related studies within the healthcare operations literature include Patrick et al. (2008), Wang and Gupta (2014), Helm and Van Oyen (2014), Diamant et al. (2018), Saghafian et al. (2019), and Saure et al. (2020). Patrick et al. (2008) designs approximate dynamic programming method to schedule patients with different priorities to a diagnostic facility in a public health-care setting. Wang and Gupta (2014) studies nurse staffing with heterogeneous absenteeism (no-shows). Helm and Van Oyen (2014) considers the problem of assigning elective patients with different rewards to different hospital units under capacity constraints. Diamant et al. (2018) develops a dynamic model of patient scheduling with rewards and no-shows in which a clinic assigns patients to an appointment day but delays the decision of which assessments patients undergo until it is observed who arrives. Saghafian et al. (2019) studies the problem of assigning patients admitted through the ED to hospital inpatient units, where patients have different rewards, waiting costs, and service needs (capacity use). Saure et al. (2020) studies the problem of scheduling multi-priority, multi-class patients with random service duration.

Finally, for other studies related to patient scheduling with time-dependent no-shows, we refer to Kong et al. (2019), Liu et al. (2009), Zeng etl. (2010), Begen and Queyranne (2011), Zacharias and Pinedo (2013), Zacharias and Yunes (2019) and the references therein. These studies are, however, mainly motivated by medical appointments in which

(a) there is a single patient type (*e.g.*, Liu et al. (2009), Begen and Queyranne (2011), Zacharias and Yunes (2019)), and (b) admission is rather exogenous and does not play an important role (*e.g.*, appointments of a primarily care physician (Zeng etl. 2010, Zacharias and Pinedo 2013)). As noted before, our focus in this study is on joint admission and scheduling decisions among a pool of heterogenous potential users, and we are motivated by settings in which the scarce capacity of a new technology (*e.g.*, proton therapy) needs to be allocated to appropriately selected users at appropriate times.

Generalized Assignment Problem. When the overtime cost is large enough and overbooking is always detrimental, our problem reduces to maximizing reward, and hence, becomes closely related to the Generalized Assignment Problem (GAP) (Chekuri and Khanna 2006, Fleischer et al. 2006, Feige and Vondrák 2006). Computing the optimal solution for GAP is in general NP-hard, but efficient approximation schemes exist. Fleischer et al. (2006) proposes an $\frac{e-1}{e}$-approximation algorithm to solve the problem. Feige and Vondrák (2006) further improves the approximation ratio to $\left(\frac{e-1}{e}+\epsilon\right)$ for some $\epsilon > 0$. However, both of them require exponential preprocessing time (for the value oracle) and complicated rounding techniques. Cohen et al. (2006) also proposes an efficient combinatorial local search algorithm to solve the problem to a worse $\left(\frac{1}{2}-\epsilon\right)$-approximation.

A key difference between our study and the literature on GAP is that we deal with overtime costs, and hence, an integer optimization problem with a nonlinear objective function. Another key difference with the aforementioned studies is that we seek to find a policy that is simple and interpretable, can allow for no-shows, and is also robust to potential misspecifications of no-shows. As noted earlier, these features are typically important for adopters of new technologies, and our focus on them is particularly motivated by the situation at our partner hospital.

Submodular Maximization under a Knapsack Constraint. When the overtime cost is moderate, overbooking may be profitable. In that case, as we will see, the objective function of our problem can be approximated by a (non-monotone) submodular function, and thus, relates to the problem of submodular maximization under a knapsack constraint. For monotone submodular maximization under a knapsack constraint, a simple marginal reward/weight with a fixed scheme is a $\frac{e-1}{2e}$-approximation (Khuller et al. 2019, Krause and Guestrin 2005, Thibaut Horel 2015). This technique combined with a preprocessing step, which conducts an exhaustive search of all feasible solutions with small cardinality, can further improve the approximation ratio to $\frac{e-1}{e}$. Nevertheless, this is impractical in our case as it can drastically increase the computational cost given the large number of patients. For non-monotone submodular maximization under a knapsack constraint, Lee et al. (2009) proposes a $\frac{1}{5}-\epsilon$-approximation algorithm, and Kulik et al. (2013) develops a randomized $\frac{1}{e}-\epsilon$-approximation. Similarly, Feige et al. (2011) presents approximation algorithms with approximation ratios ranging from $\frac{1}{4}$ to $\frac{1}{2}$. However, these algorithms are not suitable for our problem as they typically involve a sophisticated rounding technique, and thus, severely lack interpretability and transparency.

Treatment Planning in Radiation Therapy. Several papers in the literature address appropriate design of treatment plans in radiation therapy. For this stream of research, we refer interested readers to Bortfeld et al. (2008), Chan and Misic (2013), and the references therein. Treatment plans are, however, designed once the decision to allocate the capacity to the patient is made. Thus, our work differs from this stream of literature in that we focus on the more strategic level decisions of capacity use (admission and scheduling) as opposed to the operational level decisions related to the design of treatment plans for admitted patients.

3. The MAX-RATE Admission and Scheduling Policy

In this section, we describe our proposed joint admission and scheduling policy, which we term `MAX-RATE` policy. This policy is a dynamic index-based rule that prioritizes classes and sequentially assigns available slots to patients from the prioritized class.

At a high level, the policy works as follows. It sweeps through the time periods in ascending order, $t = 1, \dots, T$, and for each of them it sequentially schedules patients following an index rule. The index is class-specific and is based on an approximation of the incremental expected profit of scheduling a patient from each class, which is dynamically updated to reflect the remaining capacity and the expected overtime. Specifically, as formally described in Algorithm 1, for each time period t, the `MAX-RATE` policy observes $\lambda_{k,t}$, the period's number of remaining patients in class k ($\lambda_{k,1}$ is initialized to λ_k, *i.e.*, line 3 of Algorithm 1), and keeps track of the running expected and maximum total service duration, C' and C'' (both initialized to 0 at the beginning of each time period, *i.e.*, line 5 of Algorithm 1). Using C', it first computes the following index for time period t and each patient class k

that has remaining patients (*i.e.*, line 7 of Algorithm 1. Here, the notion $\mathbb{I}\{\cdot\}$ denotes the indicator function.)

$$k\text{th class index } := \frac{p_k(t)r_k}{l_k} - \theta\left(\frac{[C' + p_k(t)l_k - C]^+ - [C' - C]^+}{l_k}\right),$$

and then schedules a patient from the class with the highest index value. This process runs until either the hard capacity constraint is met or scheduling any additional patient yields a negative index (*i.e.*, lines 8 and 9 of Algorithm 1). Finally, the algorithm compares the expected total profit of the selected patients against the expected profit of the single most profitable patient and picks the one with higher expected reward (*i.e.*, lines 14 and 15 of Algorithm 1).

Algorithm 1 `MAX-RATE` policy

1: **input:** The patient classes $\{r_k.\lambda_k, l_k, p_k(\cdot)\}_{k\in[K]}$, the unit overtime cost θ, the scheduling horizon T, the soft and hard capacity C and $\overline{C}$;
2: **output:** The admission control and scheduling decision $\{x_{k,t}\}_{k\in[K]t\in[T]}$;
3: Initializing $\lambda_{k,1} \leftarrow \lambda_k, x_{k,t} \leftarrow 0, \qquad \forall k \in [K]\forall t \in [T]$;
4: **for** $t = 1, \dots, T$ **do**
5:     Setting $C' \leftarrow 0, C'' \leftarrow 0, P_t \leftarrow 0$;
6:     **while** True **do**
7:         Find the next patient from a non-empty class to be scheduled: $k' \leftarrow \arg\max_{k\in[K]} \frac{p_k(t)r_k - \theta\left([C'+p_k(t)l_k-C]^+ - [C'-C]^+\right)}{l_k}\mathbb{I}\{\lambda_{k,t} > 0\}$;
8:         **if** $C'' + l_{k'} \le \overline{C}$ and $p_{k'}(t)r_{k'} - \theta\left([C' + p_{k'}(t)l_{k'} - C]^+ - [C' - C]^+\right) > 0$ **then**
9:             Schedule one more class k' patient for time period t : $x_{k',t} \leftarrow x_{k',t} + 1$, and update the expected capacity consumption $C' \leftarrow C' + p_{k'}(t)l_{k'}$, the hard capacity consumption $C'' \leftarrow C' + l_{k'}$, and the expected approximate profit $P_t \leftarrow P_t + p_{k'}(t)r_{k'} - \theta\left([C' + p_{k'}(t)l_{k'} - C]^+ - [C' - C]^+\right)$;
10:         **else**
11:             Break;
12:         **end if**
13:     **end while**
14:     **if** $\max_{k\in[K]}\left(p_k(t)r_k - \theta[p_k(t)l_k - C]^+\right)\mathbb{I}\{\lambda_{k,t} > 0\} \ge P_t$ **then**
15:         Find the single most profitable patient $k' \leftarrow \arg\max_{k\in[K]}\left(p_k(t)r_k - \theta[p_k(t)l_k - C]^+\right)\mathbb{I}\{\lambda_{k,t} > 0\}$, and fix the schedule $(x_{1,t}, \dots, x_{k',t}, \dots, x_{K,t}) \leftarrow (0, \dots, 1, \dots, 0)$;
16:     **end if**
17:     Update $\lambda_{k,t+1} \leftarrow \lambda_{k,t} - x_{k,t}\ \forall k \in [K]$;
18: **end for**
19: **return** $\{x_{k,t}\}_{k\in[K]t\in[T]}$;

The proposed policy is evidently simple and interpretable. Scheduling priority of a patient for a time period, as dictated by the index, bears the following rather intuitive explanation. Scheduling of a patient of class k can be viewed as an "investment" of l_k slots.

The index score then comprises two terms. The first term is the expected reward per slot for such investment, and resembles a knapsack-style index score. The second term provides a simple and intuitive approximation of the expected overtime cost per slot. In this way, the second term captures the salient overtime cost in our model without negatively affecting either interpretability or scalability. Note that the approximation of the overtime cost is also suitably designed to ensure strong performance.

Given that the `MAX-RATE` policy is evidently both interpretable and scalable for use in practice, we now switch our focus to the remaining desired properties, namely performance and robustness.

4. Performance Analysis

In this section, we provide a theoretical guarantee for the performance of the `MAX-RATE` policy. For our analysis, we use the following definition.

Definition 1. For a maximization problem $M(\cdot)$, given an input instance $\mathcal{I}$, denote by $\boldsymbol{x}$ and $\boldsymbol{x}^*$ solutions returned by a policy $\mathcal{S}$ and by an optimal policy, respectively. We say that policy $\mathcal{S}$ returns an $(\alpha(\mathcal{I}), \beta(\mathcal{I}))-$approximate solution if $M(\boldsymbol{x}) \geq \alpha(\mathcal{I}) M(\boldsymbol{x}^*) + \beta(\mathcal{I})$, and a $\gamma-$approximate solution if $M(\boldsymbol{x}) \geq \gamma M(\boldsymbol{x}^*)$, for all $\boldsymbol{x}$.

Among the steps we follow so as to analyze the `MAX-RATE` policy, we approximate the expected profit (4) with an upper bound that can be obtained by exchanging the order of the expectation $\mathbb{E}[\cdot]$ and the max operators $[\cdot]^+$ in the calculation of the expected overtime cost. Let $P_t : \mathbb{Z}^K \to \mathbb{R}$ provide our expected profit approximation for each time period $t \in [T]$, *i.e.*, given $x_{1,t}, \ldots, x_{K,t}$, the approximate profit for period t is

$$P_t\left(x_{1,t}, \ldots, x_{K,t}\right) := \sum_{k \in [K]} p_k(t) r_k x_{k,t} - \theta \left[\sum_{k \in [K]} p_k(t) l_k x_{k,t} - C \right]^+ . \tag{9}$$

Also, let $P : \mathbb{Z}^K \to \mathbb{R}$ provide the resulting total approximate profit:

$$P(\mathbf{x}) := \sum_{t \in [T]} P_t\left(x_{1,t}, \ldots, x_{K,t}\right).$$

An approximation of the optimization problem (5) – (8) can then be obtained as

$$\begin{aligned} \text{maximize} \quad & P(\boldsymbol{x}) && (10) \\ \text{subject to} \quad & \sum_{t \in [T]} x_{k,t} \leq \lambda_k \qquad \forall k \in [K] && (11) \end{aligned}$$

$$\sum_{k\in[K]} l_k x_{k,t} \le \overline{C} \qquad \forall t\in[T] \tag{12}$$

$$x_{k,t}\in\{0,1,\ldots,N\} \qquad \forall k\in[K], \forall t\in[T]. \tag{13}$$

The `MAX-RATE` policy can be thought of as sequentially "filling up" the schedule for each time period, starting from the first one (motivated by the show-up probability functions being non-increasing in time). Thus, we can think of the policy as decomposing the original problem and approximately solving a series of optimization problems, one for each period t:

$$\begin{aligned}
\text{maximize} \quad & P_t(\boldsymbol{x}) && (14)\\
\text{subject to} \quad & x_{k,t}\le \lambda_{k,t} \qquad \forall k\in[K] && (15)\\
& \sum_{k\in[K]} l_k x_{k,t}\le \overline{C} && (16)\\
& x_{k,t}\in\{0,1,\ldots,N\} \qquad \forall k\in[K], && (17)
\end{aligned}$$

for time period $t\in[T]$, where $\lambda_{k,t}$ represents the number of patients left in class $k\in[K]$ after scheduling allocations have been performed for time periods $1,\ldots,t-1$.

At a high level, our performance analysis proceeds in three steps. In the first step, we characterize "how well" the index score of the `MAX-RATE` policy performs when solving an instance of the decomposed problem (14)-(17). To this end, we leverage the submodularity of P_t—a property that can be readily verified (Krause and Guestrin 2005). If P_t were further monotone with respect to each of its arguments, a standard greedy algorithm would provide us with an optimal solution to the decomposed problem. However, P_t is not necessarily monotone as scheduling one more patient may incur more overtime cost than reward. Nevertheless, we exploit the piecewise linear structure of the profit functions using a novel technique, and derive for the index score we use an approximation ratio in solving the decomposed problem. As a second step, we characterize the performance loss due to the time decomposition. In particular, we utilize an inductive argument to analyze the approximation ratio of the `MAX-RATE` policy in solving optimization problem (10)-(13). In the third step, we bound the loss due to approximating $G(\boldsymbol{x})$ via $P(\boldsymbol{x})$. Finally, a salient point of the analysis is whether the overtime cost θ is high enough, in particular higher than the maximal expected reward of a single patient, *i.e.*,

$$\theta\ge\overline{\theta}:=\max_{k\in[K]} p_k(1) r_k,$$

so that no overtime is warranted under any circumstances. Putting all these pieces together, we arrive at the following performance guarantee result for our proposed policy (the proof is included in the Appendix).

THEOREM 1. *The* **MAX-RATE** *policy returns a solution* $\boldsymbol{x}$ *with* $\left(\frac{e-1}{3e-1}, \frac{(e-1)T\theta C}{3e-1} + \frac{2(e-1)\sum_{t\in[T-1]}\theta\left[\max_{k\in[K]} p_k(t)\overline{C}-C\right]^+}{3e-1}\right)$*−approximate expected profit for optimization problem* (5) − (8)*, i.e.,*

$$G(\boldsymbol{x}) \geq \frac{e-1}{3e-1}G(\boldsymbol{x}^*) - \frac{(e-1)T\theta C}{3e-1} - \frac{2(e-1)\sum_{t\in[T]}\theta\left[\max_{k\in[K]} p_k(t)\overline{C}-C\right]^+}{3e-1},$$

where $\boldsymbol{x}^*$ *is an optimal solution to (5)-(8), if the overtime cost is low, i.e.,* $\theta < \overline{\theta}$*. If the overtime cost is high, i.e., if* $\theta \geq \overline{\theta}$*, we have*

$$G(\boldsymbol{x}) \geq \frac{1}{3}G(\boldsymbol{x}^*).$$

The approximation factors in the theorem above compare favorably with other factors obtained in the literature for similar types of problems. To this end, let us compare the approximation guarantees we provide for our problem with those provided for the classical GAP, which is a much simpler problem whereby no overtime considerations are present. As we remarked in our review of related papers, state-of-the-art algorithms for GAP achieve guarantees of $\frac{e-1}{e}$ or $\frac{1}{2}$, depending on their complexity. Nevertheless, all of them involve complicated computation procedures lacking interpretability. The difference between these and our coefficient, $\frac{e-1}{3e-1}$, can then be attributed to (a) the increased complexity of dealing with a nonlinear objective G that accounts for overtime, and (b) the fact that we limit the policy space to simple, interpretable index rules. These additional challenges also come at a cost of additive terms in our guarantee, for example, $\frac{e-1}{3e-1}\theta TC$. It is important to note, however, that under practical circumstances, these terms are likely to be insignificant and dwarfed by $\frac{e-1}{3e-1}G(\boldsymbol{x}^*)$. To see this, note that TC is the total number of time slots available and recall that θ is the per-time-slot overtime cost. If we write $G(\boldsymbol{x}^*) =$ (average reward per time slot at optimality) $\times\, TC$, the comparison of the two terms then boils down to the comparison between the average reward we can optimally extract versus the overtime cost, per time slot available. However, in practice the former is likely to be much larger than the latter. In Proton Therapy, for example, rewards are associated with saving lives. Nonetheless, even if the overtime cost becomes high, the additive terms in our guarantees vanish entirely. In that case, we simply obtain a factor of $\frac{1}{3}$, which in part reflects our limiting of the policy space to simple, interpretable index rules.

4.1. Effect of Misspecification of Show-Up Probabilities

Among the model parameters, show-up probabilities $p_t(k)$ are often the most challenging to accurately estimate from data. In particular, as noted earlier, it is very likely that the true show-up probabilities, $\tilde{p}_k(\cdot)$, would deviate in practice from the ones estimated using data. This can occur for a variety of reasons, including (a) lack of large-scale data caused by the fact that the technology is new and has not been in use for a sufficiently long period of time, and (b) the fact that these probabilities depend on both t and k. It is, therefore, desirable to explore how robust the `MAX-RATE` policy is to potential misspecifications of the show-up probabilities, $p_k(t)$.

To consider such misspecifications, we assume that the exact show-up probability functions, $\tilde{p}_k(\cdot)$, satisfy

$$\tilde{p}_k(t) = \min\{\xi_{k,t} p_k(t), 1\}, \quad \forall k \in [K], t \in [T], \tag{18}$$

where $\xi_{k,t}$ are unknown perturbation parameters, assumed to take values in the interval $[1-\epsilon, 1+\epsilon]$, for some $\epsilon \geq 0$. Since in practice longer scheduling delays almost never yield higher show-up likelihoods, we further assume that monotonicity in t holds for both the true show-up probability functions and the perturbed ones, and denote by E the set of all perturbation parameters that satisfy these assumptions.

Under this model of uncertainty, we consider the following robust counterpart to our original optimization problem:

$$\text{maximize} \qquad \min_{\{\xi_{k,t}\}_{k\in[K]t\in[T]}\in E} \sum_{t\in[T]} \sum_{k\in[K]} \tilde{p}_k(t) r_k x_{k,t} - \theta \mathbb{E}\left[\sum_{k\in[K]} \text{Binomial}\left(x_{k,t}, \tilde{p}_k(t)\right) l_k - C\right]^+ \tag{19}$$

$$\text{subject to} \qquad \sum_{t\in[T]} x_{k,t} \leq \lambda_k \qquad \forall k \in [K] \tag{20}$$

$$\sum_{k\in[K]} l_k x_{k,t} \leq \overline{C} \qquad \forall t \in [T] \tag{21}$$

$$x_{k,t} \in \{0,1,\ldots,N\} \qquad \forall k \in [K] t \in [T] \tag{22}$$

The above robust optimization problem has a *maximin* objective function. That is, it seeks a scheduling solution that would maximize the worst-case (with respect to all possible perturbations) expected profit. The next result characterizes the performance of the `MAX-RATE` policy for this setting.

THEOREM 2. *For the problem* (19)-(22)*:*

- *if* $\theta < \max_{k\in[K]}(1+\epsilon)p_{k,1}r_k$*,* `MAX-RATE` *policy provides a* $\left(\frac{(1-\epsilon)(e-1)}{(1+\epsilon)(3e-1)}, const\right)-$ *approximate solution, where* $const = \frac{(8e\epsilon-2\epsilon+2(e-1))\sum_{t\in[T]}\theta\left[\max_{k\in[K]}p_k(t)\overline{C}-C\right]^+}{(1+\epsilon)(3e-1)} + \frac{2\epsilon[(4-\epsilon)e+\epsilon-2]TC}{(1-\epsilon)(1+\epsilon)(3e-1)} + \frac{(e-1)T\theta C}{3e-1}$*;*
- *if* $\theta \geq \max_{k\in[K]}(1+\epsilon)p_{k,1}r_k$*,* `MAX-RATE` *policy provides a* $\frac{1-\epsilon}{3(1+\epsilon)}-$*approximate solution.*

The approximation guarantees in Theorem 2 illustrate how the performance of the proposed index policy depends on potential misspecifications of the show-up probabilities, as captured by the parameter ϵ. Note that for $\epsilon = 0$, the approximation guarantees recover precisely those presented in Theorem 1.

5. Case Study: Proton Therapy Treatment Admission and Scheduling

To gain further insights into the performance and advantages of the `MAX-RATE` policy for implementation in practice, we conduct in-depth numerical performance analyses, using both real and synthetic data. For the former, we utilize a data set that we have collected from our partner hospital, Massachusetts General Hospital (MGH).

In summary, we find that our proposed policy strikes a favorable balance between interpretability and performance. In particular, being an index-based scoring policy, it is almost as simple and has the same interpretable scoring-based format as the current practice at MGH (which we describe in detail below). Importantly, our results show that it yields about 20% performance improvement in expected clinical benefits—an estimate we arrived at by using real data and the same performance metrics as used by MGH. Furthermore, we find that the `MAX-RATE` policy has a suboptimality gap that ranges between 2% and 10%, demonstrating that interpretability and good performance are *not* mutually exclusive. Finally, our sensitivity analyses using synthetic data reveal that the performance of the `MAX-RATE` policy is also robust to main environmental factors that might vary across organizations (*e.g.*, demand-to-capacity ratio, no-show probabilities, and overtime costs).

In what follows, we first briefly describe the process of proton therapy treatment at MGH. We then introduce our data, parameter estimation procedures, and analyses.

Proton Therapy Treatment Process. To be treated via proton therapy, patients are required to "apply" in advance. That is, in consultation with their physicians, they decide to seek proton therapy treatment and submit all the required documents. Subsequently, there are a series of steps that need to be completed prior to treatment commencing. First, each application is reviewed in detail so as to evaluate the patient's suitability for proton therapy and expected clinical benefits. Second, once a patient is accepted for treatment, the staff try to ensure that s/he is either covered through insurance and/or is able to pay the associated costs out-of-pocket. Once these steps are done, the physician (*e.g.*, a radiation oncologist) begins to develop a radiation treatment plan with a dosimetry and medical physics team that involves multiple steps from clinical determination of region to be treated to quality assurance testing and peer review of treatment plan. These processes are lengthy in time, but are essential before the patient can start the treatment.

Because all the aforementioned required steps prior to treatment take at least a couple of weeks, there is usually a delay between when the patient applies and when s/he is scheduled to start the treatment. Lack of enough proton machines compared to the demand creates further scheduling delays. Of note, these required processes also prohibit compensating for no-shows by assigning the capacity to another patient: when a patient who is scheduled to receive treatment does not show up, last minute replacements are not possible. Patients who do not show up to receive their treatment often seek treatment from other resources (*e.g.*, traditional x-ray radiation). This is, for example, the case for patients who no longer can wait, those whose health condition suddenly degrades, and those who receive advice against proton therapy after applying for it. Similar to those patients who are declined to receive the service, no-show patients almost never apply again.

Proton Therapy Admission Process at MGH. The number of prospective patients far exceed available proton therapy treatment capacity at MGH. Applications are reviewed periodically by a panel of expert oncologists, medical physicists, and dosimetrists who evaluate appropriateness of each case with a collective determination made. That is, applications are collected over some period at MGH, which we refer to as "cycle," and then the panel meets at the end of each cycle to make a decision.

This elaborate and complex decision-making process is currently guided by a scoring system developed by MGH, which assigns each applicant with a prioritization score, termed the "capstone" score. The capstone score was designed to reflect the incremental medical benefit a patient would have by receiving proton radiation, as opposed to conventional therapies.[6] In other words, the capstone score approximates the utility that the center

[6] To develop the capstone scoring system, the MGH Proton Center conducted a series of questionnaires, in which physicians were asked to rank hypothetical patients based on the efficacy that they expected proton therapy to have on the patients. Using these as input data, a regression model was fit by the center so as to calculate this score for each applicant.

would derive by providing service to that patient based solely on clinical grounds. Notably, no operational considerations, such as capacity consumption, or no-show probability, are used in the calculation of this score. By and large, the panel tends to prioritize and schedule patients using the capstone scoring formula.

Data Set. Our data set spans a period of 481 days in 2016-2017, and includes information about all 1,153 patients who were reviewed to be treated at the MGH Proton Therapy Center during this period. Due to incomplete data entries, we omit data on five of these 1,153 patients. Moreover, 75 patients out of the remaining 1,148 patients had data entry errors (*e.g.*, appointment dates earlier than the application date), and hence, we omit them as well. Thus, our final data set includes information about 1,073 patients. For each patient, the data includes information about the patient's application date, demographic features such as age, gender, residency location, and medical features such as Karnofsky score,[7] comorbidities, and prior radion therapy records, among others. For admitted patients, the data also includes appointment date, service duration, and an indicator of whether they showed up or not. The capstone score is another piece of information available for each patient in our data set, which we utilize as the "reward" parameter in our model. As remarked above, this score reflects medical benefits, and tends to increase as the patient's life year gains by the therapy are expected to increase, and also as the patient's alternative treatment options beyond proton therapy become limited or less effective.

Scheduling Delays. For each admitted patient, we define the scheduling delay as the difference between her scheduled appointment date at the center and her application date (for simplicity, we also refer to application date as "arrival date"). Figure 1 illustrates the boxplot of these scheduling delays, and shows that the majority of the patients have delays in the range of 7 to 154 days (a week to five months). Furthermore, it can be seen from Figure 1 that the first and third quartiles of the scheduling delay are 27 days and 74 days, respectively.

Experimental Setup. As noted earlier, the current practice at our partner hospital involves periodic meetings (with a fixed number of days between meetings as the cycle) among a review panel to determine the patients that should be accepted and scheduled. To design and perform our experiments, we attempt to make assumptions that best represent

[7] Karnofsky score is a measure between 0 (death) and 100 (perfect health), and is typically used in medicine to inform on a patient's general well-being.

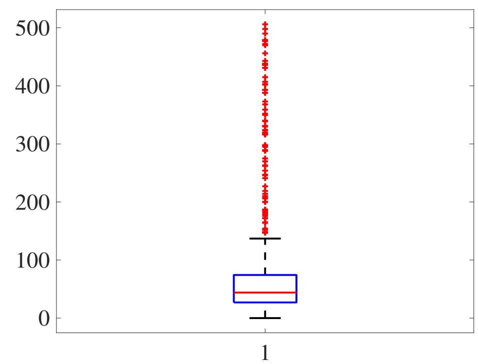


Figure 1 **Boxplot for scheduling delays**

this and other aspects of the practice at our parent hospital. In particular, following the current practice, we divide our experiment's horizon (481 days) into cycles of equal time periods. For the ith cycle, $(i \geq 3)$, admission and scheduling decision are made at the end of $(i-2)$th cycle, while treatment preparations, such as the ones outlined in the treatment process discussion above, take place during the $(i-1)$th cycle. Since decisions made at the end of each cycle involve patients with arrival dates during that cycle, this makes the minimum and maximum scheduling delay one and three cycles, respectively. Because in our data set the first and third quantile of scheduling delay are 27 and 74 days, respectively, we set the length of a cycle to be 30 days (*i.e.*, a month) so that $T = 30$ and each patient experiences scheduling delay in the range of $[30, 90]$, which is close to the range $[27, 74]$ that we observe from our data set (see the boxplot in Figure 1).

Estimating Show-Up Probabilities and the Number of Patient Classes. In our dataset, about 15% of the patients do not show up in the scheduled time. In order to accurately estimate our show-up probability functions, we take advantage of the information available in our data. For a class k patient, we consider the function $p_k(t)$ as a survival function. We then use clustering and regression methods to estimate these survival functions. Specifically, we first perform Cox regression and identify features that are statistically significant in estimating show-up probabilities. We then perform K-means clustering (with different K values representing the total number of patient classes) using these features. We next calculate the weighted Area Under the Curve (AUC) to compare and choose the best number of clusters. Finally, we consider patients belonging to each cluster in the final result to be within the same class. That is, the best clustering result is used to identify patient classes, where each cluster corresponds to a distinct patient class., and thereby, capture the effect of overtime costs.

Calculating Soft Capacity. Since our data set does not indicate the duration of overtime operations, we start our analyses by assuming that the system runs with no overtime. This allows us to directly calculate the total operating minutes on each day using our data set. We assume that available capacity equals to this number of total operating minutes on each day. It can be readily seen that this calculation underestimates the actual capacity. Making available to our policy this calculated capacity enables our study then to report conservative estimates about our policy's actual performance. As we will describe later, we also perform simulation analyses by allowing overtime.

Estimating Service Duration for Rejected Patients. Our data set only includes service duration for patients who are accepted/admitted by our partner hospital. That is, the potential service duration of the patients that are rejected is not observable to us, simply because they were not admitted (and no treatment plan was designed for them) at our partner hospital. Since an alternative policy such as the `MAX-RATE` policy might choose to accommodate some of these patients, it is essential to estimate the service duration of all patients (*i.e.*, both admitted and rejected patients). To this end, we treat the service duration of rejected patients as missing values, and make use of the MICE (Multivariate Imputation by Chained Equations) package of R to impute and estimate their values.

Fair Comparison with Current Practice. As noted earlier, with a cycle time of 30 days, the majority of the patients scheduled for the ith ($i \geq 3$) cycle in our data set arrive during the $(i-2)$th cycle. While this captures the majority of the patients, this is not a flawless assumption: in practice, occasionally some patients have a longer scheduling delay. To perform fair comparisons with the current practice (as a benchmark), and to ensure that our comparisons are not biased toward our proposed algorithm, for such occasional cases that are scheduled for the ith ($i \geq 3$) cycle but have an arrival date prior to the $(i-2)$th cycle, we assume a random arrival during the $(i-2)$th cycle. This allows us to perform a fair comparison with current practice, and report (slightly) conservative values for the improvements achieved due to our proposed algorithm.

Performance Metrics. We measure reward from providing service to a patient using the patient's capstone score, as it was designed to approximate the center's utility and medical benefits from treating that patient. Then, we measure the performance of the

K	1	2	3	4	5	6	7
Weighted AUC	0.9377	0.9457	0.9520	0.9586	0.9037	0.8348	0.7763
Improv.	11.0222%	11.0127%	11.0129%	19.6289%	21.3440%	13.2720%	11.1573%

Table 1 Weighted AUC and expected Improvement for different number of clusters

`MAX-RATE` policy in terms of percentage improvement in "profit" (*i.e.*, expected reward minus expected cost):

$$\text{Improvement} = \frac{\text{Profit of } \texttt{MAX-RATE} \text{ policy} - \text{Profit of current practice}}{\text{Profit of current practice}}. \tag{23}$$

Herein, in the absence of overtime, the profit of each policy can be readily calculated as the sum of the expected rewards corresponding to the respective patients admitted by that policy, whereby the expectation is taken with respect to the associated show-up probabilities. We report the expected value, standard deviation, and the 95% confidence interval for the percentage improvement (calculated over 5,000 iterations). Further, we report the expected percentage optimality gap (OPT gap) of the `MAX-RATE` policy:

$$\text{OPT gap} = \frac{\text{Optimal Profit} - \text{Profit of } \texttt{MAX-RATE} \text{ policy}}{\text{Optimal Profit}},$$

where the optimal profit is obtained via the optimal value of (5)-(8). Note that in the absence of overtime, *i.e.*, $\theta = 0$, calculating the optimal value is tractable for the scale of our instances.

Results. Our data analyses show that the following vector of patient characteristics is most significant in predicting show-up probabilities:

(research origin, comorbidities, prior RT, capstone score, service duration).

Thus, we perform K-means clustering on these features and vary K from one to seven. The results are shown in Table 1. Setting a weighted AUC threshold of 90%, we observe from this table that we should divide the patients into 4 or 5 classes to achieve the maximum profit improvement while having a good clustering result (weights represent the percentage of total patients that fall in each cluster). We also include the show up probability obtained from the data and the estimated show up probability for $K = 4$ and 5 in Section C of the appendix. To further validate this, we measure the standard deviation, 95% confidence interval, and the OPT Gap for the cases with $K = 4$ or 5 clusters (see Table 2).

In summary, as it can be seen from the results presented in Table 2, the proposed `MAX-RATE` policy procedure is significantly better than the current practice, delivering a

K	Weighted AUC	Improv.	Std.	95% C.I.	OPT gap
4	0.96	19.62%	0.0067	[19.60%,19.64%]	8.31%
5	0.90	21.34%	0.0095	[21.31%,21.37%]	10.35%

Table 2 **Detailed measurements for $K = 4$ and $K = 5$ clusters**

profit improvement of approximately 20%. In addition, it has a relatively low optimality gap, between 8–10%. Since the optimal policy is complex and hard to implement in practice, this suggest that `MAX-RATE` policy offers an effective and yet easy-to-implement alternative for jointly making admission control and scheduling decisions.

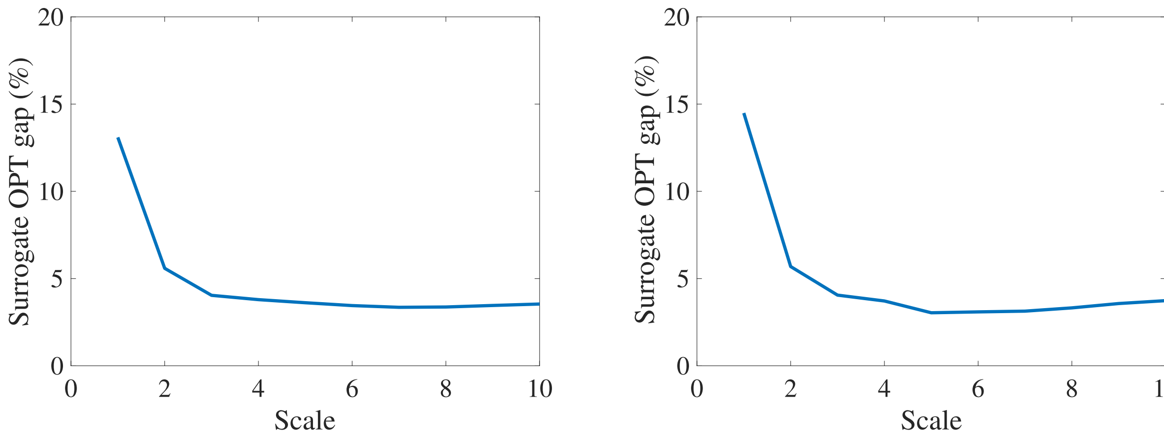
5.1. Overtime Considerations

As noted earlier, the data set from our partner hospital (MGH) does not include information about overtime operations. Since the `MAX-RATE` policy can also incorporate overtime operations, we scale up our data set (in terms of λ_k's) and use it to examine the effect of overtime. We do so by considering a scale factor that represents the ratio of the number of patients in the scaled data set to that of the original one. We vary the scale factor from 1 to 10, and set the hard capacity to 1.5 times of the soft capacity of each day, and consider an overtime cost of $\theta = 2$.

For the purposes of measuring the optimality gap of our policy, note that calculating the optimal profit via (5)-(8) now becomes intractable. Therefore, we instead calculate the optimal value of (10)-(13), which provides an upper bound for the optimal profit, given that $P(\boldsymbol{x}) \geq G(\boldsymbol{x})$, for all $\boldsymbol{x}$ (see Corollary 1 in the appendix). For the `MAX-RATE` policy, we calculate its expected profit by simulating multiple sample paths of show-ups. We then obtain the following

$$\begin{aligned}\text{Surrogate OPT Gap} &= \frac{\text{Upper bound on Optimal Profit} - \text{Expected profit of } \texttt{MAX-RATE} \text{ policy}}{\text{Upper bound on Optimal Profit}} \\ &= \frac{P(\tilde{\boldsymbol{x}}) - G(\boldsymbol{x})}{P(\tilde{\boldsymbol{x}})},\end{aligned} \tag{24}$$

where $\tilde{\boldsymbol{x}}$ is an optimal solution of (10)-(13) and $\boldsymbol{x}$ is the solution returned by the `MAX-RATE` policy. One immediate observation is that the surrogate OPT Gap serves as an upper bound for the OPT gap. Figures 2(a) and 2(b) depict the surrogate OPT gap of the `MAX-RATE` policy when the number of clusters (*i.e.*, patient classes) are 4 and 5, respectively. As it can be seen from these figures, the `MAX-RATE` policy continues to have strong performance in the presence of overtime. In particular, it has a fairly low surrogate optimality gap for reasonable levels of the scale factor. Since the surrogate OPT gap is an upper bound for



(a) Surrogate OPT gap with $K=4$ clusters (b) Surrogate OPT gap with $K=5$ clusters

Figure 2 Surrogate OPT gaps on the scaled data set

Scale	1	2	3	4	5	6	7	8	9	10
Reward ($\times 10^4$)	4.78	6.09	6.61	6.89	7.11	7.30	7.44	7.58	7.71	7.83
Overtime cost ($\times 10^3$)	1.00	1.14	1.47	1.46	1.53	1.57	1.59	1.63	1.67	1.77

Table 3 Profit breakdown for $K=4$

Scale	1	2	3	4	5	6	7	8	9	10
Reward ($\times 10^4$)	4.51	5.90	6.47	6.80	7.06	7.24	7.38	7.50	7.61	7.70
Overtime cost ($\times 10^3$)	1.00	1.17	1.43	1.46	1.55	1.53	1.60	1.66	1.73	1.71

Table 4 Profit breakdown for $K=5$

the actual optimality gap, this gives us confidence about the performance of the `MAX-RATE` policy. We also include the detailed breakdown of the profits, in terms of rewards and overtime cost, in Tables 3 and 4.

5.2. Sensitivity Analysis

To go beyond the case study at our partner hospital, and investigate the suitability of the `MAX-RATE` policy more broadly, we now perform various sensitivity analyses. To this end, instead of using our data set, which may only represent the environment at our partner hospital, we design a new test suite and include various levels for the main parameters that vary among hospitals. In particular, we start by creating a representative base case scenario, and then vary its number of patient classes, rewards, show-up probability functions, soft and hard capacities, and overtime cost. This allows us to measure the effect of these main parameters on the optimality gap of the `MAX-RATE` policy, which we measure as before using the surrogate optimality gap (24). We run each scenario for 100 iterations.

To gain deeper insights and have a benchmark, we also measure the performance of another index-style policy with the following ratio as its index:

$$\frac{\text{Expected reward of class } k}{\text{Service duration of class } k} = \frac{p_k(t) r_k}{l_k}.$$

This index is the classical "knapsack-style" index and is a popular heuristic for problems like ours.

Base Case. In the base case, we set the number of patient classes to $K = 5$, and assume each class has 300 patients. Each patient class has a reward drawn uniformly random from the interval $[50, 150]$, and an inverse Weibull show-up probability function with random parameters (a_k, b_k). The show-up probability function of class k is then

$$p_k(t) = \exp\left(-\left(\frac{t}{a_k}\right)^{b_k}\right) \tag{25}$$

and the hazard function of class k is

$$h_k(t) = \frac{b_k}{a_k^{b_k}} t^{b_k - 1}. \tag{26}$$

In this setting, the class k's show-up probability function is increasing in a_k, and the hazard function is increasing in delay if $b_k > 1$. However, if $b_k < 1$, the hazard function is decreasing in delay. If $b_k = 1$, the hazard function is constant (*i.e.*, corresponds to a exponentially distributed show-up probability function). To cover a wide range of scenarios, we assume that b_1, b_3 and b_5 are drawn uniformly from $(0, 1]$ while b_2 and b_4 are drawn uniformly from $[1, 20]$. We also choose the a_k values uniformly from $[0, 50]$.

We assume service duration for class $1, 2, \ldots, 5$ to be $30, 40, \ldots, 70$, respectively. All the patients arrive on the first day of the cycle. The soft capacity and hard capacity for each day are set to 1,260 minutes (7 hours×3) and 2,160 minutes (12 hours×3). We also start our analysis by considering an overtime cost $\theta = 2$. In order to make a fair comparison, we compare the expected profit of the optimal policy and the `MAX-RATE` policy via OPT Gap for one cycle (30 days). Some summary statistics for the base case setting are shown in Table 5, where *load factor* is defined as:

$$\text{Load Factor} = \frac{\text{Total service duration of all arriving patients}}{\text{Total soft capacity during the cycle}}. \tag{27}$$

Total #patients	1,500
Average service duration (mins)	50
Total soft capacity (mins)	37,800
Load Factor	198.41%

Table 5 Summary statistics for the base case setting

The Effect of Load Factor. We start our sensitivity analyses by considering the effect of load factor. To examine this effect, we vary the number of arriving patients from 1,500 to 6,000 with a step size of 500 while keeping the populations of each class the same. The results are shown in Figure 3(a), which indicates that the surrogate OPT Gap drops from about 18% to as low as 5% as we increase the load factor from about 200% to 800%. This is due to the fact that, as the number of patients in each class increases, the `MAX-RATE` policy can identify the most profitable patient class more effectively. Notably, our policy outperforms the knapsack-style policy consistently across the entire range of load factors. These results suggest that our proposed `MAX-RATE` policy would likely be particularly effective for hospitals in which demand compared to the available capacity is high. As noted earlier, since Proton Therapy is a relatively new radiation technology with many advantages compared to traditional X-ray therapy most hospitals that offer it already face high demand compared to their capacity. Thus, we expect that the proposed `MAX-RATE` policy can offer significant benefits to many proton therapy centers.

The Effect of the Number of Patient Classes. To see the effect of the number of patient classes, we increase the number of classes from 5 to 25 with a step size of 5. In each increment, we keep the service duration of the 5 new classes as $30, 40, 50, 60$, and 70, and the total number of patients unchanged to maintain the same load factor. The results presented in Figure 3(b) show that the surrogate OPT Gap increases only about 8% as the number of classes increases from 5 to 25. This indicates that the `MAX-RATE` policy is relatively robust to the number of patient classes. Finally, we observe a significantly better performance for the `MAX-RATE` policy compared to the knapsack-style index, an advantage that persists across the range of patient classes considered.

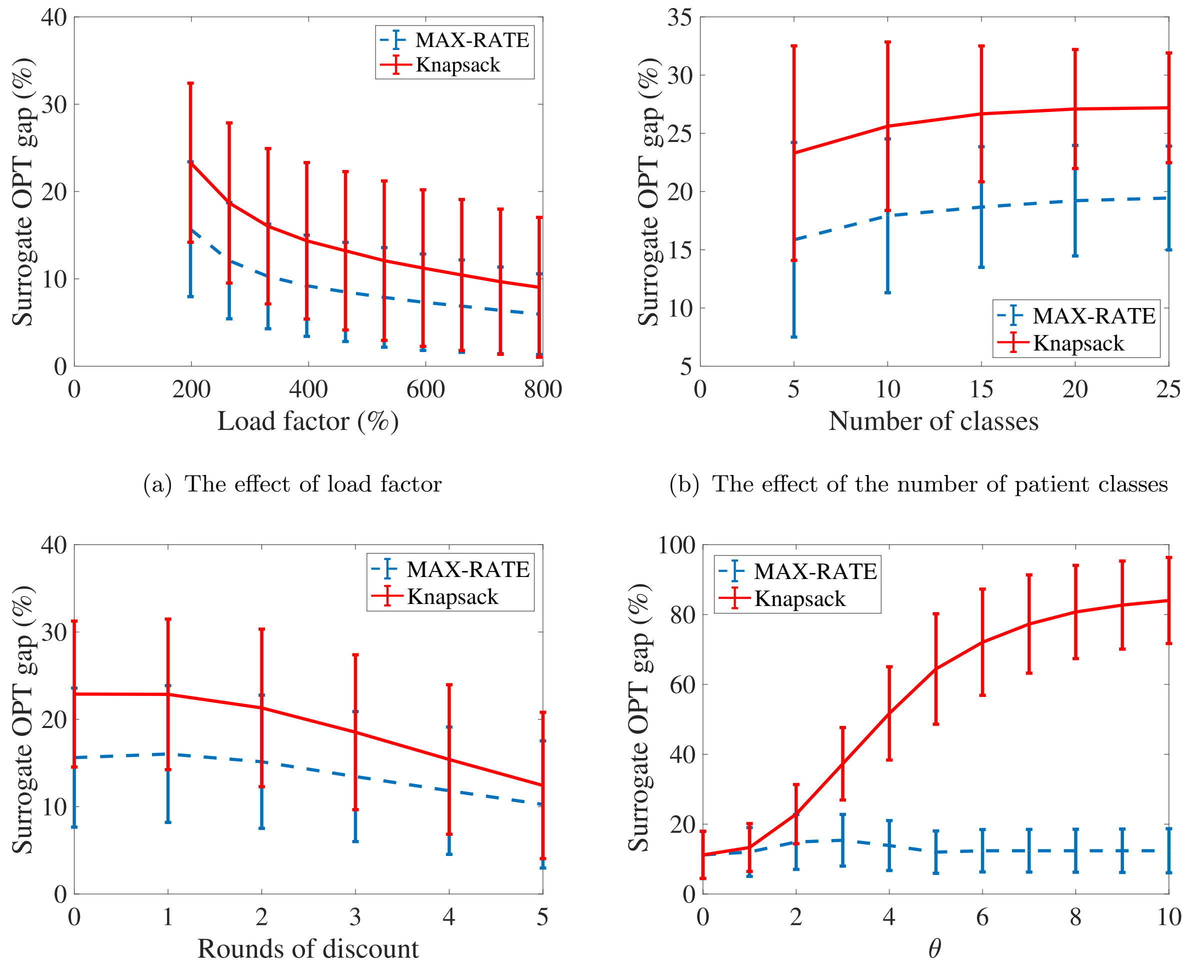
The Effect of No-Shows. Another factor that varies among hospitals is the no-show rate of patients. Specifically, some hospitals have a low number of no-shows while others face significant number of no-shows. To see the effect of no-shows, we decrease the parameter a_k in the show-up probability function for each class k by iteratively multiplying it by $\alpha = 0.8$ (for a maximum number of five times). The results presented in Figure 3(c) show that the surrogate OPT Gap decreases very slowly as the value of show-up probabilities decreases, which suggest that the `MAX-RATE` policy is relatively robust to high no-show probabilities. In addition, for small (large) no-show probabilities, we observe that the `MAX-RATE` policy yields a significant (modest) improvement over the knapsack-style index.

The Effect of Overtime Cost. The cost of overtime operations depends on various factors. As a result, there is a variation among hospitals in terms of the compensations and other expenses that they incur due to overtime operations. To examine the effect of the overtime cost (parameter θ), we increase it from 0 to 10 with a step size of 1. From the results presented in Figure 3(d), we observe that the surrogate OPT Gap first increases from 14% to 18% as θ increases from 0 to 2. It then drops from 18% to 11%, and remains at 10% for any $\theta \geq 5$. This is because, when the overtime cost is sufficiently small, the loss due to overtime is also small. However, as the overtime cost increases, both the optimal policy and the `MAX-RATE` policy reduce the number of patients scheduled during overtime, which in turn shrinks the surrogate OPT Gap. Once the overtime cost exceeds a threshold, use of overtime operations becomes significantly costly, and thus, the surrogate OPT Gap remains constant for both policies. These findings match our earlier analytical results which state that the surrogate OPT Gap is larger when the value of θ is moderate than when it is 0 or higher than some threshold $\bar{\theta}$. Importantly, our results also indicate that the `MAX-RATE` policy is robust to the overtime cost parameter. In sharp contrast, the knapsack-style index (which does not account for overtime) is highly sensitive and performers well only when the value of θ is low.

6. Extension: Multiple Visits

Put together, our analysis and numerical experiments so far reveal that the `MAX-RATE` policy is an effective and yet easy-to-implement policy. Importantly, its strong performance is fairly robust to various factors that vary among hospitals (*e.g.*, utilization, number of patient classes, show-up probabilities, and overtime cost).

In this section, we showcase how one can readily generalize our policy to accommodate multiple visits, and perform numerical analysis that demonstrate that the `MAX-RATE` policy retains strong performance under this extension as well. To this end, suppose that class k patients require m_k visits. The vast majority of proton therapy protocols that include multiple visits, require them to occur on consecutive days. Thus, we assume that if a



(a) The effect of load factor

(b) The effect of the number of patient classes

(c) The effect of no-show probabilities

(d) The effect of θ

Figure 3 The effect of various parameters on the surrogate OPT Gap

class k patient is scheduled to receive service at time period t, s/he will consume l_k time slots at time periods $t, t+1, \ldots, t+m_k-1$. Notably, this assumption is without loss of generality, and the algorithm can be extended in a straightforward manner if treatment protocols prescribe a different visit frequency. As we previously discussed, patients who show up for their first appointment, continue to do so for their subsequent appointments. The approximate profit maximization problem can thus be re-formulated as

$$\text{maximize} \quad \sum_{t\in[T]}\sum_{k\in[K]} p_k(t) r_k x_{k,t} - \theta \left[\sum_{k\in[K]} p_k(t) l_k \sum_{\tau=\max\{1,t-m_k+1\}}^{t} x_{k,\tau} - C \right]^+ \tag{28}$$

$$\text{subject to} \quad \sum_{t\in[T]} x_{k,t} \le \lambda_k \qquad \forall k \in [K] \tag{29}$$

$$\sum_{k\in[K]} l_k \sum_{\tau=\max\{1,t-m_k+1\}}^{t} x_\tau \le \overline{C} \qquad \forall t\in[T] \tag{30}$$

$$x_{k,t}\in\{0,1,\ldots,N\} \qquad \forall k\in[K], \forall t\in[T]. \tag{31}$$

Here, different than the single visit case, in computing the overtime of each time period, we have to take into account the patients scheduled in previous time periods. For example, $\sum_{\tau=\max\{1,t-m_k+1\}}^{t} x_{k,\tau}$ is the number of class k patients that would show up on time period t. These include class k patients scheduled on time periods $\max\{1,t-m_k+1\},\ldots,t$.

To extend our `MAX-RATE` policy for this setting, we consider the following generalized policy, which we term `MAX-RATE.M` policy: for each time step $t=1,\ldots,T$, the policy keeps track of the running expected total service durations $C'(t),\ldots,C'(T)$ (all initialized to 0) for remaining time periods $t,\ldots,T$. It then computes the following index for time period t and each patient class k that has remaining patients

$$k\text{th class index} := \frac{p_k(t)r_k}{m_k\cdot l_k} - \theta\left(\sum_{t'=t}^{t+m_k-1}\frac{[C'(t')+p_k(t')l_k - C]^+ - [C'(t')-C]^+}{m_k\cdot l_k}\right),$$

and then schedules a patient from the class with the highest index value. This process runs until either the hard capacity constraint is met or scheduling any additional patient yields a negative index. Notably, this is a straightforward generalization of our previous index policy, whereby we just now account for the multiple visits that take place.

We numerically evaluate the `MAX-RATE.M` policy with the same setting as that of Section 5.2. In particular, we compare the surrogate OPT gap against a generalized knapsack-type index policy with the following ratio as its index:

$$\frac{\text{Expected reward of class } k}{\text{Service duration of class } k} = \frac{p_k(t)r_k}{m_k\cdot l_k}.$$

We allow the number of visits of each class to be randomly sampled from the sets $\{1,2,\ldots,2i+1\}$ for $i=0,1,\ldots,9$, so that the averaged load factor varies from about 198.41% to 1984.1% with a step size of 198.41%. The results, depicted in Figure 4, illustrate that `MAX-RATE.M` policy retains strong performance, and considerably outperforms the knapsack-type index policy. Also, we note that both policies' surrogate OPT gap decrease as the average number of visits per patient increases. This is because, as the average number of visits per patient increases, the problem gets gradually closer to a single-day setting, which is shown to enjoy a better performance guarantee in Theorem 3 in the Appendix.

Remark 1 (Inter-Day Waiting Cost). We note that our method could be extended to further include inter-day waiting cost (Truong 2005, Zacharias et al. 2021). Suppose each patient of class k has a deterministic and class-dependent inter-day waiting cost ν_k. Then, we can modify the first term of (28) to $\sum_{t\in[T]}\sum_{k\in[K]} p_k(t)(r_k-\nu_k)x_{k,t}$ to accommodate this.

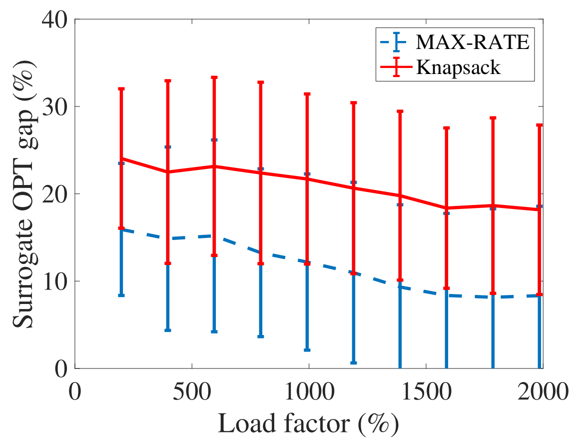


Figure 4 The effect of multiple visits on the surrogate OPT Gap

7. Conclusions

Adoption of new technologies enable firms in the healthcare sector to materially improve the quality of the service they offer. Because adoption often requires significant resources, healthcare organizations strive to utilize the available capacity as efficiently as possible. In this paper, motivated by the introduction of proton therapy—a new technology that provides superior treatment for many cancer patients—at our partner hospital (Massachusetts General Hospital, MGH), we studied the problem of admitting and scheduling patients for service in a way that addresses operational issues that arise in this context.

In particular, we presented a model of allocating service capacity in the presence of (a) patient- and time-dependent no-show behaviors, and (b) overtime operations. To make admission and scheduling decisions, we limited ourselves to simple and interpretable index rules that can be implemented in practice. We proposed a simple index policy, which balances the expected benefit from providing service to patients with the risk of overtime cost. For this policy, we derived analytical performance guarantees that compare favorably with existing results in the literature for the simpler class of generalized assignments problems.

Furthermore, we conducted in-depth numerical performance analyses using both empirical data from MGH and synthetic data. The analyses revealed that simple rules of the type we propose are capable of efficiently balancing performance and interpretability, and hence, are good candidates for use in practice. Specifically, we found that, while simple and interpretable, our proposed policy was able to (a) substantially improve upon current policies used at MGH, and (b) yield results that are not too far from being optimal. In addition, our results revealed that our proposed policy is robust to a variety of factors that are hard to estimate (*e.g.*, no-show probabilities) and/or might vary from hospital to hospital (*e.g.*, overtime cost, demand-to-capacity ratio, etc.). Thus, it offers a *"one-size-fits-all"* rule that can be robustly used in practice. Given the importance of devising simple, interpretable, and robust policies that can effectively allocate scarce capacity of new technologies to consumers, we hope that future research continues our efforts in this vein.

References

Argon, N.T., Ziya, S., Righter, R. (2008). Scheduling impatient jobs in a clearing system with insights on patient triage in mass casualty incidents. Probability in the Engineering and Informational Sciences 22(3): 301–332.

Begen, M., Queyranne, Q. (2011) Appointment Scheduling with Discrete Random Durations. Mathematics of Operations Research 36(2):240-257.

Bertsimas, D., Farias, V., Trichakis, N. (2013). Fairness, efficiency, and flexibility in organ allocation for kidney transplantation. Operations Research 61(1): 73–87.

Bortfeld, T., Chan, T.C.Y., Trofimov, A., Tsisiklis, J.N. (2008). Robust management of motion uncertainty in intensity-modulated radiation therapy. Operations Research 56(6): 1461–1473.

Buyukkoc, C., Varaiya, P., Walrand, J. (1985). The $c\mu$ Rule Revisited. Advances in Applied Probability 17(1) 237–238.

Cayirli T., Veral E. (2003). Outpatient scheduling in health care: A review of literature. Production Oper. Management 12(4): 519-549.

Chan, T.C.Y., Micic, V.V. (2013). Adaptive and robust radiation therapy optimization for lung cancer. European Journal of Operational Research 231(3): 745–756.

Chan, C.W., Farias, V.F., Escobar G. (2015). The impact of delays on service times in the intensive care unit. Columbia Business School, Working Paper.

Chan, C.W., Green, L.V., Lu, Y., Leahy, N., Yurt, R. (2013). Prioritizing burn-injured patients during a disaster. Manufacturing and Service Operations Management 15(2): 170–190.

Chekuri C., Khanna S. (2006). A Polynomial Time Approximation Scheme for the Multiple Knapsack Problem. SIAM Journal on Computing 35(3), 713–728.

Cohen R., Katzir L., Raz D. (2006). An Efficient Approximation for the Generalized Assignment Problem. Information Processing Letters 100, 162–166.

Diamant, A., Milner, J., Quereshy, F. (2018). Dynamic Patient Scheduling for Multi-Appointment Health Care Programs. Production and Operations Management, 27(1), 58–79.

Feige U, Vondrák J. (2006). Approximation algorithms for combinatorial allocation problems: Improving the factor of $1-1/e$. Proc. 47th Annual Symposium on Foundations of Computer Science. (ACM, New York), 667–676.

Feige U, Mirrokni V., Vondrák J. (2011). Maximizing Non-monotone Submodular Functions. SIAM J. on Computing.

Feldman J., Liu N., Topaloglu H., Ziya S. (2014). Appointment Scheduling Under Patient Preference and No-Show Behavior. Operations Research 62(4) 794–811.

Feldman M., Naor J., Schwartz R. (2011). A unified continuous greedy algorithm for submodular maximization. Proc. 2011 IEEE 52nd Annual Sympos. Foundations Comput. Sci. (IEEE, Piscataway, NJ), 570–579.

Fleischer L., Goemans M., Mirrokni V., Sviridenko M. (2006). Tight Approximation Algorithms for Maximum General Assignment Problems. Proc. 17th Annual ACM-SIAM Symposium on Discrete Algorithm. (SIAM, Philadelphia), 611–620.

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

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

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

Horel T. (2015). Notes on Greedy Algorithms for Submodular Maximization.

Kaandorp, G. C., Koole G. (2007). Optimal outpatient appointment scheduling. Health Care Management Science 10(3) 217–229.

Khuller, S., Moss, A., Naor, J. 1999. The Budgeted Maximum Coverage Problem. Information Processing Letters.

Kong, Q., Li, S., Liu, N., Teo, C.P., Yan, Z. 2019. Appointment Scheduling Under Time-Dependent Patient No-Show Behavior. Management Science (forthcoming).

Korula, N., Pal, G. (2009). Algorithms for Secretary Problems on Graphs and Hypergraphs. Proceedings of the 36th Internatilonal Collogquium on Automata, Languages and Programming: Part II Pages 508–520.

Krause, A., Guestrin, C. (2005). A Note on the Budgeted Maximization of Submodular Functions. http://reports-archive.adm.cs.cmu.edu/anon/cald/CMU-CALD-05-103.pdf.

Kulik A., Shachnai H., Tamir T. (2013). Approximations for Monotone and Nonmonotone Submodular Maximization with Knapsack Constraints. Mathematics of Operations Research 38(4): 729–739. https://doi.org/10.1287/moor.2013.0592.

LaGanga, L., Lawrence S. R. (2012). Appointment overbooking in health care clinics to improve patient service and clinic performance. Production and Operations Management 21(5): 874–888.

Lee, J., Mirrokni V., Nagarajan V., Sviridenko M. (2009). Non-Monotone Submodular Maximization under Matroid and Knapsack Constraints. Proc. 41st Annual ACM Symposium on Theory of Computing: 323–332.

Liu, N., Ziya, S., Kulkarni, V. (2009) Dynamic Scheduling of Outpatient Appointments Under Patient No-Shows and Cancellations. Manufacturing & Service Operations Management 12(2):347-364.

Luo, J., V. G. Kulkarni, S. Ziya. (2012). Appointment scheduling under patient no-shows and service interruptions. Manufacturing and Service Operations Management 14(4): 670–684.

Master, N., Chan, C.W., Bambos, N. (2016). Myopic policies for non-preemptive scheduling of jobs with decaying value. Probability in the Engineering and Informational Sciences 1–36. doi: 10.1017/S0269964816000474.

Patrick, J., Puterman, M., Queyranne, M. (2008) Dynamic Multipriority Patient Scheduling for a Diagnostic Resource. Operations Research 56(6):1507-1525.

Saghafian, S., Hopp, W.J., Van Oyen, M.P., Desmond, J.S., Kronick, S.L. (2014). ComplexityAugmented Triage: A Tool for Improving Patient Safety and Operational Efficiency. Manufacturing and Service Operations Management 16 (3): 329–345

Saghafian S., Kilinc, D., S.J. Traub (2019). Dynamic Assignment of Patients to Primary and Secondary Inpatient Units: Is Patience a Virtue? Working Paper, Harvard University (HKS Working Paper No. RWP17-010).

Saure, A., Begen, M., Patrick J. (2020). Dynamic multi-priority, multi-class patient scheduling with stochastic service time., European Journal of Operational Research, Volume 280, Issue 1, 2020, 254–265,

Schaefer, M (1976). Note on the k-Dimensional Jensen Inequality. The Annals of Probability Vol. 4, No. 3, pp. 502-504.

Simic, S. (2009). On An Upper Bound for Jensen's Inequality. Journal of Inequalities. Pure and Applied Mathematics.

Truong, V.-A. (2005). Optimal advance scheduling. Management Science 61(7) 1584–1597.

Wang, W. Y., Gupta, D. (2014). Nurse absenteeism and staffing strategies for hospital inpatient units. Manufacturing and Service Operations Management, 16(3), 439–454.

Zacharias, C., Liu, N., Begen, M.A. (2021). Dynamic inter-day and intra-day scheduling. Working Paper.

Zacharias, C., Pinedo, M. (2019). Appointment Scheduling with No-Shows and Overbooking. Production and Operations Management 23(5):788-801.

Zacharias, C., Yunes, T. (2019). Multimodularity in the Stochastic Appointment Scheduling Problem with Discrete Arrival Epochs. Management Science 66(2):744-763.

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