Health Care Management Science (2022) 25:682–709
https://doi.org/10.1007/s10729-022-09609-0

Quantile regression forests for individualized surgery scheduling

Arlen Dean[1] · Amirhossein Meisami[2] · Henry Lam[3] · Mark P. Van Oyen[1] · Christopher Stromblad[4] · Nick Kastango[4]

Received: 9 October 2019 / Accepted: 15 July 2022 / Published online: 18 August 2022

Abstract

Determining the optimal surgical case start times is a challenging stochastic optimization problem that shares a key feature with many other healthcare operations problems. Namely, successful problem solutions require using a vast array of available historical data to create distributions that accurately capture a case duration's uncertainty for integration into an optimization model. Distribution fitting is the conventional approach to generate these distributions, but it can only employ a limited, aggregate portion of the detailed patient features available in Electronic Medical Records systems today. If all the available information can be taken advantage of, then distributions individualized to every case can be constructed whose precision would support higher quality solutions in the presence of uncertainty. Our individualized stochastic optimization framework shows how the quantile regression forest (QRF) method predicts individualized distributions that are integrable into sample-average approximation, robust optimization, and distributionally robust optimization models for problems like surgery scheduling. In this paper, we present some related theoretical performance guarantees for each formulation. Numerically, we also study our approach's benefits relative to three other traditional models using data from Memorial Sloan Kettering Cancer Center in New York, NY, USA.

Keywords Operations research · Individualized learning · Surgery scheduling · Stochastic optimization · Robust optimization · Distributionally robust optimization

1 Introduction

The growth of personalized data from Electronic Health Records (EHR) and biomarkers has motivated us to investigate how to harness this individualization for problems in healthcare where the modeling of uncertainty is also critically required. We present a framework using quantile regression forests (QRF) to generate individualized distributions integrable into three optimizations paradigms. We demonstrate the effectiveness of our individualized optimization approach in terms of basic theory and practice. Specifically, we focus on operating room scheduling because it is exactly the type of problem that we expect can benefit greatly from our framework.

The operating room (OR) is well known to be both the most significant source of revenue and cost for hospitals [46]. As a significant driver of hospital operations and their efficiency, ORs are often affected by significant schedule volatility as they vary from low utilization to high levels of overtime [15, 37, 40]. It remains a challenge to schedule surgeries because surgical durations suffer from a high degree of variability due to the patient, procedure, and surgeon-related factors [27]. Studies on

✉ Arlen Dean
arlend@umich.edu

Amirhossein Meisami
meisami@adobe.com

Henry Lam
henry.lam@columbia.edu

Mark P. Van Oyen
vanoyen@umich.edu

Christopher Stromblad
stromblc@mskcc.org

Nick Kastango
KastangN@mskcc.org

[1] University of Michigan, Ann Arbor, MI, USA

[2] Adobe Inc, San Jose, CA, USA

[3] Columbia University, New York, NY, USA

[4] Memorial Sloan Kettering Cancer Center, New York, NY, USA

surgery scheduling, accordingly, often entail stochastic durations characterized by specified distributions as probabilistic estimators. These studies commonly defer to distributions on a population level where each patient is treated equally as coming from a single "generating" pool or a limited number of categories/types [14, 15, 35, 41]. The reason is it can be difficult to model the distribution of case times for every procedure type because there are a large number of distinct procedures, and many are known to have few observed instances [17, 36, 47]. Other studies have proposed Bayesian methods to obtain estimators even when data is limited [16, 18, 34]. These approaches still do not include, among others, patient-specific factors (e.g., comorbidities, body mass index). As a result, to our best knowledge, highly personalized case time distribution estimates have not been used for optimization in this domain.

Recently, decision-making problems like portfolio optimization [1, 21], assortment planning [26], and medical dosing [4] have introduced the use of individualized information by leveraging machine learning (ML) methods. Following this idea, we introduce individualized estimates via ML for use in optimization under uncertainty. Unlike the works above, we require the ML model to give distributional predictions to quantify the uncertainty individual to each case. Our proposed use of QRF accomplishes this as this method naturally outputs the prediction of an entire distribution. The benefit of ML methods, like QRF, is that they allow us to include patient, procedure, surgeon, and hospital factors for prediction even for unobserved feature values, and in a nonlinear manner beyond classical statistical tools [3]. Along this vein, we mention that, in the robust optimization paradigm, the closest work to ours is [50] who suggested several ML approaches to create uncertainty sets. However, in contrast to [50], our chosen ML approach of QRF accommodates not only robust optimization but also other optimization modeling methodologies.

We show that the common paradigms of optimization under uncertainty, specifically sample average approximation (SAA), robust optimization (RO), and distributionally robust optimization (DRO), are all amenable to incorporating QRF. These optimization approaches depend on a decision-maker's risk tolerance, and QRF can be integrated into the respective formulations to enhance efficacy. Among these paradigms, SAA represents a risk-neutral decision process as a stochastic optimization method that approximates an expectation objective by sampling [28, 45]. RO, which requires only a support set for the random variables, is risk-averse as it focuses on the worst-case outcomes [5, 7]. DRO lies somewhere between SAA and RO in that it considers the worst-case performances among partially informed probability distributions [6, 13, 53].

As a demonstration of our approach, we consider the problem of setting surgery start times. We formulate an SAA, RO, and DRO scheduling model using a QRF to navigate the difficult trade-off that arises from the stochasticity of patient case durations. Namely, surgeries occur in sequence, and so if a job's start time is set too late, then the job before it may finish with too much time left over. Such an event creates idling and can lead the last surgical case in the block to operate in overtime. On the other hand, a job scheduled too early likely cannot start on time because it needs to wait for the previous job to finish first. This delay can cause patient dissatisfaction and adverse health outcomes.

Theoretically, we illustrate how the statistical consistencies of our formulations (with respect to their target guarantees) follow the known properties of QRF. In the RO and DRO settings, we also study uncertainty sets naturally deduced from QRF and their tractabilities for the surgery problem. To complement these results, we conduct empirical experiments in collaboration with Memorial Sloan Kettering Cancer Center (MSKCC) in New York, NY, USA. We test our formulated models using real-patient data and numerically highlight the strengths of our QRF-optimization integrated framework.

The remainder of this paper is organized as follows: Section 2 introduces our setting and notation for surgery scheduling. Section 3 reviews QRF and Section 4 demonstrates how we integrate QRF into the three optimization approaches as well as their corresponding consistency results. Section 5 provides a numerical example to test our proposed methodology. We conclude with Section 6.

2 The surgery scheduling model

We consider a single operating room scheduling problem where the number of planned operations and their sequence for a given day is fixed. A common practice in surgery departments like those in cancer centers is that appointments for elective surgeries are made a few days before the surgery. Accordingly, we assume knowledge of the biometric characteristics of the incoming patients (namely the "features" in the predictive model). Our optimization model uses predicted distributions of the surgery durations, built from available patient features, to determine the best starting time for each surgery to account for the uncertainty of surgical durations. The model's objective captures the previously mentioned trade-off between the minimization of patient wait times and the occurrence of overtime. We chose to employ a more basic model because its analysis will show that QRF is suitable for use with the three optimization methodologies considered in this paper. This is aligned with other works in the scheduling literature (e.g., [20, 30, 35]) for clearer presentation of a proposed methodology. Moreover, we share a similar justification with the previous literature. Our objective can include idle times without significant deviation of notation if their cost rates are identical to those of the waiting times. Overall, the model setting is general enough to apply to problems in other areas where the adjustment of time allowances for a sequence of jobs is used for improved operational behavior or cost containment [42, 52].

We follow similar notations as in [14]. The model's decision is to select the start time of each surgery for a set of n elective surgeries in a given sequence on a given day for a given operating room. Let T be the scheduled closing time of the day and z_i define the random duration (respectively, the outcome) of the ith surgery. The decision maker has to set the surgery time allowance of the ith case, x_i, such that the first starts at time zero, the second at x_1, the third at $x_1 + x_2$, and so on. Assuming that the start time for the first surgery is zero, the start time of any subsequent surgery is scheduled at the sum of the time allowances of all previous surgeries. We also denote by $Z = (z_i)_{i=1,\ldots,n}$ and $X = (x_i)_{i=1,\ldots,n}$ the vectors of surgery durations and surgery time allowances. Based on these assumptions, the waiting time, w_i, which is defined as the difference between the scheduled surgery starting time and the actual starting time whenever the previous procedure ends, and the overtime l, can be represented as:

$$w_i = \max\{0, w_{i-1} + z_{i-1} - x_{i-1}\}, i = 2, \ldots, n \tag{1}$$

$$l = \max\{0, w_n + z_n - x_n\} \tag{2}$$

We assume that the first surgery always starts on-time and, hence, $w_1 = 0$. Through rescaling, without loss of generality, we assume each unit of waiting time costs one unit for any of the subsequent surgeries. Therefore, how long each patient personally has to wait from their appointed time until the start of their surgery is precisely the corresponding waiting time cost. Lastly, an overtime cost of ϕ per unit time is incurred if we run later than time T. Given the cost function $f(X, Z) = \sum_{i=2}^{n} w_i + \phi l$ and the definitions above, a surgery scheduling model, assuming hypothetically that Z are known, is constructed as:

$$\begin{aligned} \min_{x \in \mathcal{X}} \quad & f(X, Z) = \sum_{i=2}^{n} w_i + \phi l \\ s.t. \quad & w_{i+1} = \max\{0, w_i + z_i - x_i\} \qquad i = 1, \ldots, n-1 \\ & l = \max\{0, w_n + z_n - x_n\} \end{aligned} \tag{3}$$

where $\mathcal{X} = \{x_i \geq 0 \;\forall i, \sum_{i=1}^{n} x_i \leq T\}$ and $(x_1, \ldots, x_n)$ are the surgery time allowance decision variables. By a monotonicity argument on the objective function in terms of w_i's and l, it is standard to see that Eq. 3 can also be reformulated as

$$\begin{aligned} \min_{\substack{x \in \mathcal{X} \\ w,l}} \quad & f(X, Z) \\ s.t. \quad & w_2 \geq z_1 - x_1 \\ & w_{i+1} \geq w_i + z_i - x_i \qquad i = 2, \ldots, n-1 \\ & l \geq w_n + z_n - x_n \\ & w_i, l \geq 0 \qquad i = 2, \ldots, n \end{aligned} \tag{4}$$

where $(w_2, \ldots, w_n, l)$ are introduced as auxiliary decision variables. Note that the closing time T is used in the feasible region $\mathcal{X}$. For completeness, we show in Appendix 1 that idle times can be considered in this model by adding $\sum_{i=1}^{n-1}(x_i - z_i) + w_n$ to the objective function when their cost rates are identical to those of the waiting times. When the Z are stochastic, we can replace the objective of Eq. 3 as either $E[f(X, Z)]$ or $\min\{q : P(f(X, Z) \leq q) \geq 1 - \delta\}$, where $E[\cdot]$ and $P(\cdot)$ are the expectation and probability taken with respect to Z. The former is an expected value formulation, and the latter is a percentile formulation. Common approaches like SAA and RO provide approximate solutions to these formulations, as we describe in Section 4.

3 Conditional distributions and quantile regression forests

Under stochasticity, we need the distribution of each z_i to be tailored to the characteristics of each patient to ensure model Eq. 4 is solving a problem that considers the subtle differences across patient cases. Letting Ξ denote the patient's feature (e.g., gender, age etc.) vector in the feature space $\mathcal{B}$, we approximate $F(z|\xi) = F(z|\Xi = \xi)$, the distribution function of the surgery duration given $\Xi = \xi$.

Studies on machine learning (ML) methods for patient-specific surgical duration estimates have mainly focused on point estimators (i.e., the mean response) (e.g., [3, 44, 48]) and are typically not directly capable of predicting the probability distributions needed for optimization models involving uncertainty. Although bootstrap-type sampling can generate such predictions in basic linear models (see Section 3.5 of [23]), we consider elaborate quantile ML methods to be a more direct way of accomplishing this in complicated modeling environments. The simplest quantile method to estimate $F(z|\xi)$ is linear quantile regression (LQR) (see [29] for details), but it relies on linear assumptions. Thus, we are motivated to study QRF based on its ability to handle more complex high-dimensional data settings [38].

QRF extends from the widely known ensemble method of random forests (RF) (see [9] for details) and shares a similar procedural design. RF models are built to output mean predictions by bootstrap aggregating (or bagging) decision trees, which is known to achieve stable prediction [19] and possess variance reduction properties [10]. Leveraging this, rather than taking the mean of responses within the same leaf as the output of a tree, QRF takes the empirical distribution of the responses instead [33]. We refer to [38] for a more complete overview on QRF.

We close this section by stating that QRF recovers the true conditional distribution as the number of observations in the training data increases under appropriate regularity conditions. That is, the conditional distribution of z constructed from QRF is $\hat{F}(z|\xi)$, and it is known to converge in probability to $F(z|\xi)$ (see Theorem 1 of [38] for specific details). The next section will show how this consistency property can be leveraged to provide asymptotic guarantees for various optimization under uncertainty formulations that take into account individualized information.

4 Individualized optimization under uncertainty

This section presents our integration of QRF into three optimization formulations that capture the stochasticity of the surgery duration in an individualized fashion. We first consider an expected value formulation and SAA in Section 4.1. Then we move to a percentile formulation and RO in Section 4.2, and finally revisit the expected value formulation and DRO in Section 4.3.

We aim to provide theoretical justification for the validity of our approach by showing asymptotic guarantees for each method, as is the standard for algorithms of this nature. Through our analysis, we also hope to convey our reasoning on how ML methods can be incorporated into existing optimization modeling paradigms to enhance the quality of decisions, in particular, with respect to individualization. Moreover, our theoretical results suggest that the proposed approach can be widely applied to problems beyond just the one we consider in this paper. On the other hand, an admitted limitation of our analysis is that it characterizes only asymptotic behaviors. Therefore, our numerical results in the following section aim to offer empirical support for the practical use of our approach in the OR scheduling setting, where limited sample sizes commonly exist.

4.1 Expected value optimization and sample average approximation

Our first considered formulation is Eq. 3 but with an expected value objective function

$$E[f(X,Z)|\xi_1,\dots,\xi_n] = \int f(X,Z)\prod_{i=1}^{n} F(dz_i|\xi_i), \tag{5}$$

which minimizes the average overall scheduling cost conditional on the feature ξ_i of each patient, who uses a surgery duration z_i. In this objective, we assume the distributions of all patients are conditionally independent, and we approximate the distribution $F(z|\xi_i)$ by QRF, namely $\hat{F}(z|\xi_i)$.

For convenience, we denote $\hat{E}[\cdot|\xi_1,\dots,\xi_n]$ as the conditional expectation under independent $\hat{F}(z|\xi_i)$'s. Exact computation of the $\hat{E}[f(X,Z)|\xi_1,\dots,\xi_n]$ in this setting can be demanding. Thus, we use sample average approximation (SAA) to approximate the problem. In particular, we generate a set of scenarios under $\prod_{i=1}^{n}\hat{F}(z_i|\xi_i)$, denoted $\mathcal{S}=\{s_1,\dots,s_k\}$. Every scenario $s\in\mathcal{S}$ is associated with a realization of $Z(s)=(z_i(s))$ and with $w_1(s)=0$,

$$\begin{aligned} w_i(s) &= \max\{0, w_{i-1}(s)+z_{i-1}(s)-x_{i-1}\}, i=2,\dots,n\\ l(s) &= \max\{0, w_n(s)+z_n(s)-x_n\}. \end{aligned}$$

From Eq. 4 we can approximate the expected value problem via SAA as

$$\min_{\substack{x\in\mathcal{X}\\ w,l}} \sum_{k=1}^{K}\frac{1}{K}\Big(\sum_{i=2}^{n} w_i(s_k)+\phi l(s_k)\Big) \tag{6}$$

$$s.t.\quad w_{i+1}(s_k)\ge w_i(s_k)+z_i(s_k)-x_i \qquad \begin{aligned} &i=1,\dots,n-1,\\ &k=1,\dots,K\end{aligned} \tag{7}$$

$$l(s_k)\ge w_n(s_k)+z_n(s_k)-x_n \qquad k=1,\dots,K \tag{8}$$

$$w_i(s_k), l(s_k)\ge 0 \qquad \begin{aligned} &i=2,\dots,n,\\ &k=1,\dots,K\end{aligned} \tag{9}$$

where K is the number of scenarios independently generated in the SAA. Formulation Eq. 6 naturally combines the distributional prediction of QRF into the expected value minimization. As noted previously, although modeling case durations by distribution fitting is standard, utilizing the variety of data to capture the particularities across different case durations is more challenging. In this regard, unlike other approaches that may also consider nonidentical distributions in surgery scheduling, the QRF generates a distribution for every case unique to each patient as it is conditional on all their available health information. Our numerical study in Section 5.2 shows this is highly beneficial to solution quality. Under standard conditions, the SAA problem Eq. 6 has a solution and optimal value that converge to those of Eq. 5, as we formally state next.

Theorem 1 *Let $H(X)$ be the objective function Eq. 5 (suppressing $\xi_1,\dots,\xi_n$ for convenience), and H^* be the optimal value when solving Eq. 3 with Eq. 5 as the objective function. Let $\tilde{X}^*$ be an optimal solution to Eq. 6 where the scenarios are drawn from the distribution $\prod_{i=1}^{n}\hat{F}(z_i|\xi_i)$. Assume that z and ξ satisfy the conditions in Appendix 1, and moreover that z is bounded a.s. within $\mathcal{A}\subset\mathbb{R}_+$. We have $H(\tilde{X}^*)\xrightarrow{p}H^*$ as $K,N\to\infty$, where N is the observation size in building the QRF and K is the scenario size in SAA.*

For proof of Theorem 1 see Appendix 1.

4.2 Robust optimization

Robust optimization (RO) provides an alternative paradigm to account for the uncertainty in the surgery duration. In the RO paradigm, we replace the stochasticity with a so-called uncertainty set or ambiguity set, which is a deterministic set that, intuitively, captures the likely realization of the surgery duration (other interpretations are possible, e.g., [5], page 33 discussion point B).

Formulation Eq. 4 has resemblance with single-server queues, for which [2] has introduced an RO approach to estimate relevant quantities. Following their framework, we introduce $\underline{\Gamma}_k$ and $\overline{\Gamma}_k$ and assert that the surgery times belong to the uncertainty set

$$\tilde{\mathcal{U}} = \left\{ (z_1, z_2, ..., z_n) \;\middle|\; \underline{\Gamma}_{k,i} \leq \sum_{j=k}^{i-1} z_j \leq \overline{\Gamma}_{k,i}\,, \quad \forall k < i \leq n \right\} \tag{10}$$

The reason why we consider a set on $\sum_{j=k}^{i-1} z_j$, instead of other possible candidates (e.g., merely z_k themselves), is that the objective function, upon rewriting, depends explicitly on these quantities which gives rise to easy bounds. However, this convenient choice can be plausibly replaced by others, which result in convex optimization problems that can still be handled by standard solvers.

RO considers

$$\min_{x \in \mathcal{X}} \max_{z \in \tilde{\mathcal{U}}} \left(\sum_{i=2}^{n} w_i + \phi l \right) \tag{11}$$

where w_i and l satisfy Eqs. 1 and 2. Note that the waiting time of the ith patient can be expressed as

$$w_i = \max\{w_{i-1} + z_{i-1} - x_{i-1}, 0\} = \max_{1 \leq k < i} \left(\sum_{j=k}^{i-1} (z_j - x_j), 0 \right) \tag{12}$$

and it is also trivial to see that $l = w_{n+1}$. Putting these into Eq. 11 gives

$$\begin{aligned} \min_{x \in \mathcal{X}} \max_{z \in \mathcal{U}} \Bigg(& \sum_{i=2}^{n} \max_{1 \leq k < i} \left(\sum_{j=k}^{i-1} (z_j - x_j), 0 \right) \\ & + \phi \max_{1 \leq k < n+1} \left(\sum_{j=k}^{n} (z_j - x_j), 0 \right) \Bigg). \end{aligned} \tag{13}$$

Switching the order of maximizations, Eq. 13 is upper bounded by

$$\begin{aligned} \min_{x \in \mathcal{X}} \Bigg(& \sum_{i=2}^{n} \max_{1 \leq k < i} \max_{z \in \mathcal{U}} \left(\sum_{j=k}^{i-1} (z_j - x_j), 0 \right) \\ & + \phi \max_{1 \leq k < n+1} \max_{z \in \mathcal{U}} \left(\sum_{j=k}^{n} (z_j - x_j), 0 \right) \Bigg). \end{aligned} \tag{14}$$

The innermost maximization can be easily seen to be attained at the upper bounds imposed in $\mathcal{U}$, so that Eq. 14 is equivalent to

$$\begin{aligned} \min_{x \in \mathcal{X}} \Bigg(& \sum_{i=2}^{n} \max_{1 \leq k < i} \left(\overline{\Gamma}_{k,i-1} - \sum_{j=k}^{i-1} x_j, 0 \right) \\ & + \phi \max_{1 \leq k < n+1} \left(\overline{\Gamma}_{k,n} - \sum_{j=k}^{n} x_j, 0 \right) \Bigg). \end{aligned} \tag{15}$$

Now, denoting $Q_i = \max_{1 \leq k < i} \left(\overline{\Gamma}_{k,i-1} - \sum_{j=k}^{i-1} x_j, 0 \right)$, then Eq. 15 can be reformulated into the following linear program

$$\begin{aligned} \min_{x \in \mathcal{X}} \quad & \sum_{i=2}^{n} Q_i + \phi Q_{n+1} \\ s.t. \quad & Q_i \geq \overline{\Gamma}_{k,i-1} - \sum_{j=k}^{i-1} x_j \qquad i = 2, ..., n+1 \\ & \qquad\qquad\qquad\qquad\quad 1 \leq k < i \\ & Q_i \geq 0 \qquad\qquad\qquad\qquad\; i = 2, ..., n+1. \end{aligned} \tag{16}$$

The question remains how to calibrate $\overline{\Gamma}_{k,i}$'s. Using the idea of data-driven RO (e.g., [8]), one can set $\overline{\Gamma}_{k,i}$'s so that

$$\hat{P}\left(\underline{\Gamma}_{k,i} \leq \sum_{j=k}^{i-1} z_j \leq \overline{\Gamma}_{k,i}\,, \forall k < i \leq n \,\middle|\, \xi_1, \ldots, \xi_n \right) \geq 1 - \delta, \tag{17}$$

where $\hat{P}(\cdot|\xi_1, \ldots, \xi_n)$ refers to the probability under $\prod_{i=1}^{n} \hat{F}(z_i|\xi_i)$. Notice that in fact only $\overline{\Gamma}_{k,i}$'s are used; the $\underline{\Gamma}_{k,i}$'s can be dropped by adopting our subsequent analysis to a one-sided bound instead of two-sided in a straightforward manner. The guarantee of Eq. 17 can be translated to the optimal value of Eq. 11 and hence Eq. 16 provides an upper bound to

$$\min_{x \in \mathcal{X}} \min\{q : \hat{P}\big(f(X,Z) \leq q|\xi_1, \ldots, \xi_n\big) \geq 1 - \delta\}$$

Namely, the optimal $1 - \delta$ quantile of $f(X, Z)$ under $\prod_{i=1}^{n} \hat{F}(z_i|\xi_i)$. Theorem 2 below details a more elaborate version of this claim, taking into account the Monte Carlo noises that we discuss next.

To find $\underline{\Gamma}_{k,i}$ and $\overline{\Gamma}_{k,i}$, we find $\hat{\Gamma}$ such that

$$\hat{P}\left(\left|\frac{\sum_{j=k}^{i-1} z_j - \sum_{j=k}^{i-1} \hat{\mu}_j}{\sqrt{\sum_{j=k}^{i-1} \hat{\sigma}_j^2}}\right| \le \hat{\Gamma}, \forall k < i \le n \,\middle|\, \xi_1, .., \xi_n\right) \ge 1 - \delta \tag{18}$$

where $\hat{\mu}_j$ and $\hat{\sigma}_j^2$ are the mean and variance from $\hat{F}(\cdot|\xi_j)$. The left hand side of the expression inside the probability in Eq. 18 is a centered and normalized version of $\sum_{j=1}^{i-1} z_j$ that appears often in the central limit theorem. The choice of $\hat{\Gamma}$ in Eq. 18 then implies that one can choose

$$\underline{\Gamma}_{k,i} = \sum_{j=k}^{i-1} \hat{\mu}_j - \hat{\Gamma}\sqrt{\sum_{j=k}^{i-1} \hat{\sigma}_j^2}$$

and

$$\overline{\Gamma}_{k,i} = \sum_{j=k}^{i-1} \hat{\mu}_j + \hat{\Gamma}\sqrt{\sum_{j=k}^{i-1} \hat{\sigma}_j^2} \tag{19}$$

Now, to find the $\hat{\Gamma}$ that satisfies Eq. 18, we can use quantile estimation (e.g., [22]). Simulate, say K, i.i.d. copies of Z, and for each Z one can calculate

$$\max_{k<i\le n} \left|\frac{\sum_{j=k}^{i-1} z_j - \sum_{j=k}^{i-1} \hat{\mu}_j}{\sqrt{\sum_{j=k}^{i-1} \hat{\sigma}_j^2}}\right| \tag{20}$$

Find the $\lfloor K(1-\delta)\rfloor$-th order statistic of the K copies of Eq. 20, call it $\tilde{\Gamma}$. When K is large, $\tilde{\Gamma}$ will roughly satisfy Eq. 18. Note that, desirably, this calibration approach based on quantile estimation does not require a large n in Eq. 18, which could actually be relatively small in applications.

Using the above formulations and procedure gives the following:

Theorem 2 *Assume that z and ξ satisfy the conditions in Appendix 1. Let $\tilde{H}^*$ be the optimal value of formulation Eq. 16, with $\overline{\Gamma}_{k,i}$ calibrated from Eq. 19 and $\hat{\Gamma}$ approximated by $\tilde{\Gamma}$, the $\lfloor K(1-\delta)\rfloor$-th order statistic of Eq. 20 among the K samples of Z generated under $\prod_{i=1}^{n} \hat{F}(z_i|\xi_i)$. Assume that the true conditional distribution $F(z|\xi)$ is continuous and strictly increasing in z. Let H^* be the optimal value for $\min_{x\in\mathcal{X}} \min\{q : P\big(f(X,Z) \le q|\xi_1, \dots, \xi_n\big) \ge 1-\delta\}$. Then*

$$\liminf_{N\to\infty, K\to\infty} \tilde{H}^* \ge H^*$$

The proof of Theorem 2 is in Appendix 1.

Note $\min_{x\in\mathcal{X}} \min\{q : P\big(f(X,Z) \le q|\xi_1, \dots, \xi_n\big) \ge 1-\delta\}$ means minimizing the $(1-\delta)$-quantile of $f(X, Z)$. Theorem 2 stipulates that the RO formulation can be viewed as a conservative approximation of this optimization. Also, note that when the continuity assumption of $F(z|\xi)$ is removed, we can modify the above procedure slightly by inflating our obtained $\tilde{\Gamma}$ by an arbitrarily small constant, i.e., we use $\tilde{\Gamma} + \epsilon$ for some small ϵ, and our argument (detailed in the proof) will carry through to get the same guarantee in Theorem 2.

4.3 Distributionally robust optimization

We next consider distributionally robust optimization (DRO). This approach targets expected value objective function, like in the case of SAA, but under only partial information of the distributions. More specifically, it optimizes the worst-case expected value among all distributions that are in an uncertainty set or an ambiguity set which represents the partial information. In our scheme, we assume the quantiles of each patient's surgery duration distribution at a list of given probability levels are known. These information pieces can be drawn from the QRF, achieving individualization.

The motivation for using DRO is that its solution can be more robust to some hidden uncertainty. In our circumstance, for instance, imposing enough quantile information means we know the distribution of each patient's surgery duration distribution, but we do not assume any dependency structure among the patients. The DRO solution is thus robust against this hidden stochasticity that is not revealed by the individualization, which focuses only on the prediction for each patient.

More concretely, for each patient i, we choose a sequence $q_{i1} < q_{i2} < \cdots < q_{im}$, and set $r_{ij} = \hat{F}(q_{ij}|\xi_i)$ which can be inspected from the QRF. Roughly speaking, q_{ij} is the r_{ij}-th quantile of the duration distribution given ξ_i. For convenience, we assume that $r_{im} = 1$, so that the q_{im} is the upper limit of the support of the data. We consider the uncertainty set

$$\hat{\mathcal{U}} = \{P \in \mathcal{P} : P(z_i \le q_{ij}) = r_{ij},\ i = 1, .., n, j = 1, .., m\} \tag{21}$$

where $\mathcal{P}$ is the set of all probability distributions supported on $\mathbb{R}_+^n$. Here, the constraints indicate that the marginal r_{ij}-th quantiles for patient i are known to be q_{ij} under the QRF. They also include the information that the largest possible value of $\hat{F}(z|\xi_i)$ is estimated to be q_{im}.

We seek to solve

$$\min_{x\in\mathcal{X}} \max_{P\in\hat{\mathcal{U}}} E_P\big[f(X,Z)\big]. \tag{22}$$

We approach Eq. 22 using the technique in [35], which first replaces $f(X, Z)$ as the optimal value of a linear program (LP) given by

$$\begin{aligned} \max_{y} \quad & \sum_{i=1}^{n} (z_i - x_i) y_i \\ s.t. \quad & y_i - y_{i-1} \ge -1 \qquad 2 \le i \le n \\ & y_n \le \phi \\ & y_i \ge 0 \qquad \forall i = 1, ..., n. \end{aligned} \tag{23}$$

This can be derived from the definition of w_i and l in Eqs. 1 and 2, and considering the dual of the resulting LP, where $y_1, \dots, y_n$ are the dual variables. A slight change of notation to (23) also leads to the dual formulation (see Appendix 1) when considering idle times. Then problem Eq. 22 can be rewritten as

$$\min_{x \in \mathcal{X}} \max_{P \in \hat{\mathcal{U}}} E_P \Big[\max_{y \in \Omega} \sum_{i=1}^{n} (z_i - x_i) y_i \Big], \tag{24}$$

where Ω refers to the constraint set in Eq. 23, and $y = (y_1, \dots, y_n)$. We now focus on the inner maximization problem, $\max_{P \in \hat{\mathcal{U}}} E_P \big[\max_{y \in \Omega} \sum_{i=1}^{n} (z_i - x_i) y_i \big]$. The following lemma transforms it into a more manageable form:

Lemma 1 *The dual representation of the optimization* $\max_{P \in \hat{\mathcal{U}}} E_P \big[\max_{y \in \Omega} \sum_{i=1}^{n} (z_i - x_i) y_i \big]$ *is*

$$\begin{aligned} & \min_{\rho} \max_{y \in \Omega} \sum_{i=1}^{n} \max_{j=1,\dots,m} \left\{ (q_{ij} - x_i) y_i - \sum_{k=j}^{m} \rho_{ik} \right\} \\ & + \sum_{i=1}^{n} \sum_{j=1}^{m-1} r_{ij} \rho_{ij} + \sum_{i=1}^{n} \rho_{im}, \end{aligned} \tag{25}$$

where $\rho_{11}, \dots, \rho_{nm}$ are the dual variables corresponding to the quantile constraints.

For proof of Lemma 1 see Appendix 1.

By Lemma 1, problem Eq. 22 can be represented by

$$\begin{aligned} & \min_{x \in \mathcal{X}, \rho} \max_{y \in \Omega} \sum_{i=1}^{n} \max_{j=1,\dots,m} \left\{ (q_{ij} - x_i) y_i - \sum_{k=j}^{m} \rho_{ik} \right\} \\ & + \sum_{i=1}^{n} \sum_{j=1}^{m-1} r_{ij} \rho_{ij} + \sum_{i=1}^{n} \rho_{im}. \end{aligned} \tag{26}$$

We want to transform Eq. 26 to a linear program. Note that $\max_{y \in \Omega} \sum_{i=1}^{n} \max_{j=1,\dots,m} \left\{ (q_{ij} - x_i) y_i - \sum_{k=j}^{m} \rho_{ik} \right\}$ is a convex maximization problem over the polyhedron Ω, and it suffices to consider its extreme points for an optimal solution. This allows us to follow the method discussed in Proposition 2 of [35] to reduce it to solving an LP relaxation of an integer program using the following proposition.

Proposition 1 *Problem Eq. 22 can be reformulated as the following linear program*

$$\begin{aligned} \min_{x,\rho,\lambda} \quad & \sum_{i=1}^{n} \lambda_i + \sum_{i=1}^{n} \sum_{j=1}^{m-1} r_{ij} \rho_{ij} + \sum_{i=1}^{n} \rho_{im} \\ s.t. \quad & \mu_{io} \ge \left\{ q_{ij} \pi_{io} - \sum_{k=j}^{m} \rho_{ik} \right\} \qquad 1 \le i \le n \\ & \qquad\qquad\qquad\qquad\qquad\quad i \le o \le n+1 \\ & \qquad\qquad\qquad\qquad\qquad\quad 1 \le j \le m \\ & \sum_{i=g}^{\min\{o,n\}} \mu_{io} \le \sum_{i=g}^{\min\{o,n\}} x_i \pi_{io} + \lambda_i \qquad 1 \le g \le n \\ & \qquad\qquad\qquad\qquad\qquad\quad g \le o \le n+1 \\ & \sum_{i=1}^{n} x_i \le T \\ & x_i \ge 0 \qquad\qquad 1 \le i \le n, \end{aligned} \tag{27}$$

where $(\lambda_1, ..., \lambda_n)$ are the dual variables, $\pi_{io} = y_i$ for all $i \in [g, o]$ and any $g \le o$. Also, $\mu_{io} = \max_{j=1,\dots,m} \{ q_{ij} \pi_{io} - \sum_{k=j}^{m} \rho_{ik} \}$ for $o = 1, ..., n+1$.

For proof of Proposition 1 see Appendix 1.

The following shows that the DRO under the quantile information drawn from the QRF gives an asymptotic bound for the true expected value:

Theorem 3 *Assume that z and ξ satisfy the conditions in Appendix 1. Suppose that the true conditional distribution function $F(z|\xi)$ is continuous and strictly increasing in z. Let H^* be the optimal value of Eq. 3 with objective being Eq. 5, and $\hat{H}^*$ be the optimal value of Eq. 27. Then for any $\epsilon > 0$, we have*

$$P\big(H^* \le \hat{H}^* + \epsilon \big) \to 1 \tag{28}$$

as $N \to \infty$, where N is the data size in constructing the QRF.

For proof of Theorem 3 see Appendix 1.

Note that Theorem 3 does not require strong duality of the dual formulation in Eq. 25, as the upper bound in Theorem 3 can be obtained as long as Eq. 25 upper bounds its primal moment problem. However, we do need strong duality of an associated moment problem when the quantiles are set to be the true quantiles, and this property is guaranteed by the continuity and strict monotonicity of $F(z|\xi)$.

Finally, also note that, unlike SAA and RO discussed before, our DRO approach does not require running simulation from the QRF.

Table 1 Features (covariates) in the MSKCC dataset grouped into four main categories that were used for training the predictive models

Category	Features
Hospital	The primary surgeon, experience of the primary surgeon in years, number of publications of the primary surgeon, surgery room*, number of panels in the surgical case*, the sum of all RVU measures for the case*, the mean of all RVU measures for the case*, the maximum of all RVU measures for the case*, the minimum of all RVU measures for the case, number of procedures for the surgical case, indicator of robotic procedure*, the mean duration of the last 5 similar cases based on the primary procedure, the number of times the surgeon had a similar surgery in the 30 days, the number of times the surgeon had a similar surgery in the past, whether or not a particular equipment was needed for the surgery, and the type/amount of required equipment
Patient	Comorbidity measures such as obesity, depression, etc., Body Mass Index (BMI), weight, race, gender, age, indicator as inpatient or outpatient*, the number of days spent in the hospital prior to the surgery, the number of times patient underwent chemotherapy, the number of times patient underwent radiation therapy, the number of times patient had a CT scan, the number of times patient had previously underwent surgery*, and the number of days since the patient's last surgery*
Operational	The number of days between the day of surgery and the day the surgery was scheduled, the number of cases assigned to the surgeon on the day of surgery, the number of cases assigned to the surgery room on the day of surgery, the number of cases scheduled on the day of surgery*, the sequence number of the surgery on the surgeon's schedule, and the sequence number of the surgery in the surgery room
Temporal	The weekday of the surgery, the month of the surgery, and the year of the surgery

* denotes ten most important variables identified from the QRF model

Table 2 Characterization of operational data for URO service in MSKCC

Service	Surgeries	Surgeons	CPT Codes	Inpatient (%)	Mean Duration (SD)
URO	23,176	49	437	9,428 (40.7)	153.5 (129.2)

5 Numerical study

To examine the effectiveness of our approach, a numerical study was conducted using real-patient data from Memorial Sloan Kettering Canter (MSKCC), New York, NY, USA, one of the leading cancer treatment and research institutions in the world. The center has surgical operations across thirteen different services and in a total of forty different operating rooms. Our data contains the recorded surgeries for all services between 2010 to 2016 with 129,742 distinct patients, but for our analysis, we only used those in Urology (URO). Within the data, we considered hospital, patient, operational, and temporal factors as done by [39] to achieve highly accurate case duration predictions. Table 1 details the features corresponding to these factors and note that this includes patient-specific information like age, gender, race, BMI, treatment history, and comorbidities to achieve individualization.

The surgical durations used in our analysis are defined as the recorded times measured from wheels-in to wheels-out. While no outliers were thrown out, each case in our data has similar time stamps as those described by [34] and, likewise, observations with discrepancies between recorded patient and operation times were removed. Our final dataset contained 23,176 observations after also excluding patients with any missing time records. Table 2 summarizes the number of surgeries, unique primary surgeons, and unique current procedural terminology (CPT) codes, along with the number and proportion of inpatient surgeries and the historical durations of surgeries in the final dataset. Table 3 summarizes key demographics of patients in the data set.

5.1 Test setup

The individualized distributions from QRF were evaluated for each optimization model in Section 4 by comparing their performances against equivalent models using alternatively constructed distributions. There were three distribution constructing benchmarks considered for our experiments and analysis. The first was a single distribution fitted to all case durations in the training set. The second benchmark stratified the data and fitted a distribution by the primary surgeons since durations for the same procedure can vary significantly between surgeons. Linear quantile regression (LQR) was used as the third benchmark to ensure the QRF can be compared to a similar method. Fitting distributions

Table 3 Demographics of the patient data from the URO service

			Race, Number ofObservations (%)			
Service	Age (SD)	Female (%)	White	Black	Asian	Other
URO	63.1 (13.6)	5196 (22.4)	19,544 (84.3)	1,466 (6.3)	799 (3.4)	1,367 (5.9)

by CPT codes was not a considered benchmark because this approach follows closely with MSKCC's planned times, which are a considered benchmark in Section 5.4. MSKCC's scheduling system estimates these hospital planned times (H-PT) by using the median durations of the 20 most recent historical cases filtered by the CPT, then surgeon, and then operating room. The system will relax the filtering requirements for the operating room and surgeon if they do not yield any matches. Specific samples may also be excluded based on the system's defined outlier threshold and on occasion, these planned times can be overridden by a lead surgeon. We constructed predictions using the median from log-normal distributions fitted for each CPT (all had p-values > 0.1 using the Anderson-Darling test) and compared them with the H-PT. For CPTs with single instances, the lone value was considered the prediction. Our analysis presented in Table 13 of Appendix B found no statistical difference between the H-PT and CPT approach (see Appendix Table 14). Hence, we proceeded without considering CPT fitted distributions, given that we can infer their performance by examining the H-PT. We set 80% of the total available data for training/fitting each distribution approach and save the remaining 20% for use in evaluating the performance of the optimization methods. For the split, we used the most recently recorded observations in our data for the test set.

For the possible distribution to use, we considered the normal and log-normal distributions as they have been historically found to be good models for surgery durations [36, 56]. Note, these studies constructed distributions by conditioning on procedure type, while our distributions are not. For this reason, we fit a normal, log-normal, Weibull, and gamma distribution using all cases in the training set to verify if our setting is consistent with previous findings. Table 4 summarizes the estimated parameters for each model and their p-value from the Anderson-Darling (AD) test. The results indicate that each distribution was a poor fit which is not unexpected when the sample size is large, and the true distribution is not in the considered family. Since the log-normal model held the best p-value, we further checked how well this log-normal model would fit for the test set with the AD test and found it is an acceptable fit (p-value = 0.056). Lastly, we followed the procedure of [49] and obtained comparable results in that log-transformed data was considered a good fit in 71.1% of the Shapiro-Wilk tests while it was only 62.3% for non-transformed data. As a result, we proceeded with using the log-normal distribution to fit over all cases and cases grouped by the primary surgeon. The QRF model was built using 10-fold cross-validation for hyper-parameter tuning, while LQR requires no tuning. Both models used the same set of features described in Table 1 and no additional preprocessing was applied to the training set beyond encoding categorical variables for LQR.

Table 4 Parameter estimates and goodness of fit summary

Distribution	Parameter 1 (Std. Error)	Parameter 2 (Std. Error)	P-value
Log-Normal	4.63(0.006)	0.951 (0.004)	0.013
Normal	153.5 (0.826)	129.2 (0.584)	0.000
Weibull	1.19 (0.006)	162.5 (0.934)	0.009
Gamma	1.35 (0.011)	$0.0088\ (8.49 \cdot 10^{-5})$	0.004

5.2 Numerical results

We break down the results for the URO surgery schedules into regimes that correspond to the number of cases performed on a surgical day. Specifically, we organized the cases each day in the data set based on their operating room to yield schedules with the number of surgeries ranging from 2 to 11. A breakdown of the total number of surgeries corresponding to each case size is presented in Table 12 of Appendix B. The tables we show in this section further bin the size of schedules into pairs of two to generate five different regimes to summarize our results. For a breakdown of waiting and overtime costs for schedules of each case size separately, we refer to Appendix Figures 5, 6, 7, 8, 9 and 10. One should keep in mind that because OR time is extremely costly, most surgical days have as many cases as can be fit into the day. Therefore, regimes on average that consist of more surgeries in a day reflect lower average case durations.

The results to be presented in this section is based on the out-of-sample performance cost computed using (1) and (2) and the realized surgery durations from the test set. Consistent with [35], all results assume equal penalties for waiting and overtime, which is sufficient to illustrate the type of behavior each approach exhibits. For the intended OR surgical day length, T, we did not use MSKCC's official hours. Their hours include planned late rooms, which every day extends the official hours for a varied set of rooms, and this data is not collected consistently enough to be available. Instead, we chose T to be the sum of actual surgical duration data for a given day because this is the minimal possible time a room would run. MSKCC considers waste as any minute past what is necessary for the OR, and so it remains consistent with their view that everything past this defined T is overtime. While total planned time in each model's solution will be exactly equal to T, for our experiments, setting it this way still allows the realized objective costs to capture the accumulated inaccuracy of the distributions provided by a prediction approach for the individual cases. In Section 5.3, we use the median of historical cases for T to establish that our methodology is practical.

We start with the SAA framework discussed in Section 4.1. As a preliminary, we solved the problem repeatedly with sample sizes ranging from 10 to 6000, each with 5 replications, to evaluate the convergence of the SAA model's solutions. We use 90% of the training data to train our QRF model and select four instances of sizes 3,5,7 and 9 from the remaining 10% of the training data for validation. The validation set was split in the same manner as our test set, based on the chronological ordering of the cases. Appendix Figure 4 displays the trade-off between the size of samples drawn from QRF distributions and consistency of the solution cost for the SAA model. At 5000 samples, the model's objective cost converges with minimal variance, and its solution is good compared to the actual case durations from the validation data. Moreover, it was observed that 5000 samples was a generous quantity for use with the surgery distributions derived from the benchmark approaches. Therefore, to define the schedule of start times for each problem instance, we implemented SAA by drawing, for every surgery in that surgical day, 5000 samples from the surgery duration distributions derived from each approach considered (i.e., the L-Norm, L-Norm-PS, LQR, and QRF).

In Figure 1, the distribution of overall costs for each of the four models, stratified by the number of surgeries in the surgical day, is shown. Our individualized approach using QRF suggests lower scheduling costs (in minutes) than the other three models. More concretely, we can see from Table 5 that the improvement in the solution cost tends to increase for instances with a larger number of surgeries. For a global comparison, we computed a weighted average (by the number of instances) of the performance of each model across all the categories of cases per day. We found QRF

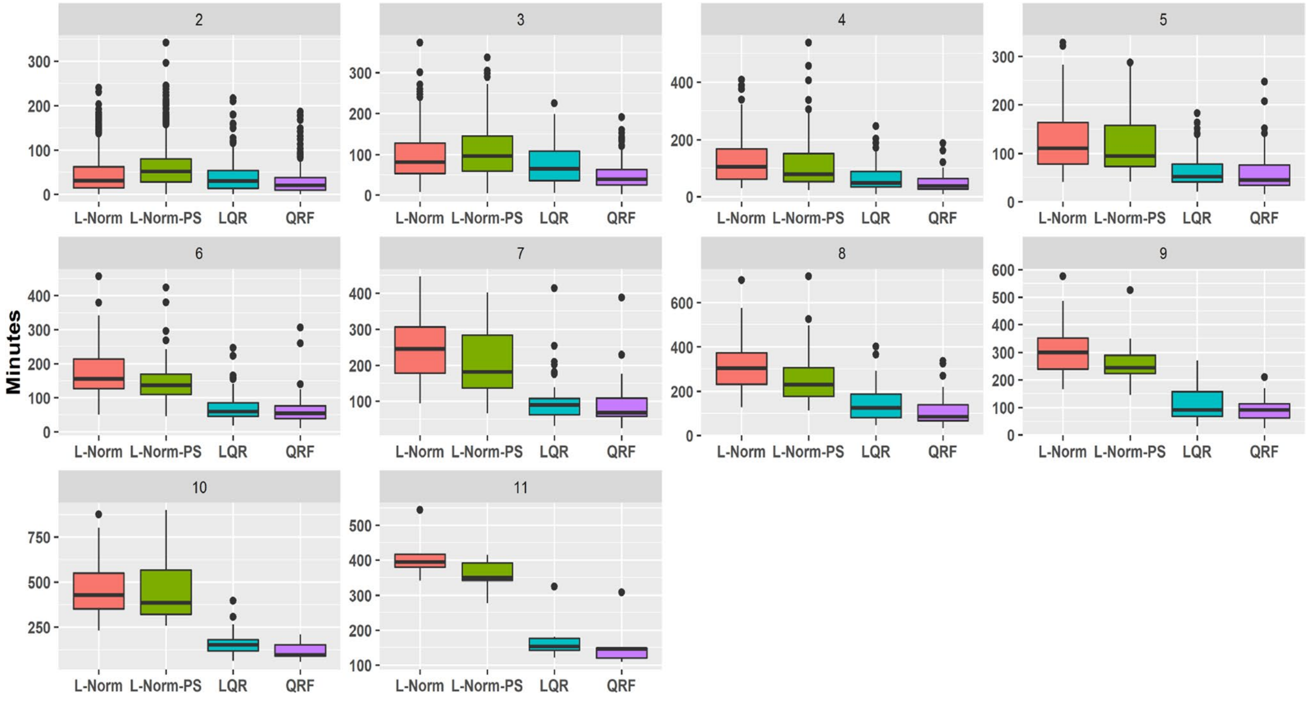


Fig. 1 Box-plots for the SAA method comparing the L-Norm, L-Norm-PS, LQR and QRF model's out-of-sample error for the different surgery regimes organized by the number of cases scheduled each day

Table 5 Weighted percentage improvement of the out-of-sample performance error when the QRF is used rather than the L-Norm, L-Norm-PS, LQR models in mean and percentiles for the Individualized Sample Average Approximation Optimization framework

		Number of Surgeries per Day				
Metric	Model	2-3	4-5	6-7	8-9	10-11
Mean	L-Norm	43.1	57.8	63.0	66.0	71.5
	L-Norm-PS	54.7	54.2	55.8	59.7	69.6
	LQR	28.0	16.8	11.4	20.2	23.7
75th percentile	L-Norm-PS	42.9	58.0	64.4	64.5	70.4
	L-Norm-PS	53.6	55.0	58.6	56.8	70.7
	LQR	33.7	16.8	4.7	26.9	15.3
95th percentile	L-Norm	45.1	60.0	57.4	52.7	70.9
	L-Norm-PS	51.9	57.8	53.0	48.4	70.4
	LQR	21.5	26.3	21.4	15.1	32.7

The percentiles metric corresponds to the schedule whose cost by the QRF model falls at that percentile

yields a 55.1%, 56.6%, and 23.1% lower average cost than L-Norm, L-Norm-PS, and LQR across all case sizes.

Analogous to the approach above for SAA, we next consider the RO framework discussed in Section 4.2. For each schedule, we implemented RO using uncertainty sets based on Eq. 10 derived from L-Norm, L-Norm-PS, LQR, and QRF. We set $K = 5000$ and $\delta = 0.05$ for each model. Figure 2 depicts the distributions of overall costs for each of the four models grouped by the number of surgeries each schedule has in the test set. Here, we see that the RO approach's conservative nature incurs a higher cost than the SAA approach. Table 6, however, still shows QRF again noticeably outperforms the L-Norm and L-Norm-PS models. The percentage improvement of QRF over these two models is also even more significant than under the SAA framework as surgery size grows. On the other hand, LQR is a closer competitor to QRF under RO where for schedules with two to three surgeries, it outperforms QRF by 1.2% on average. This is not necessarily unexpected, given that both LQR and QRF are ML approaches capable of predicting individualized distributions. The fact that QRF is only outperformed slightly by LQR for one regime but outperforms LQR for all other, larger-size regimes indicates our approach remains favorable, especially when handling larger-size schedules. Altogether, QRF leads to 64.2%, 68.8%, and 2.0% less total cost than L-Norm, L-Norm-PS, and LQR based on weighted averages over all regimes.

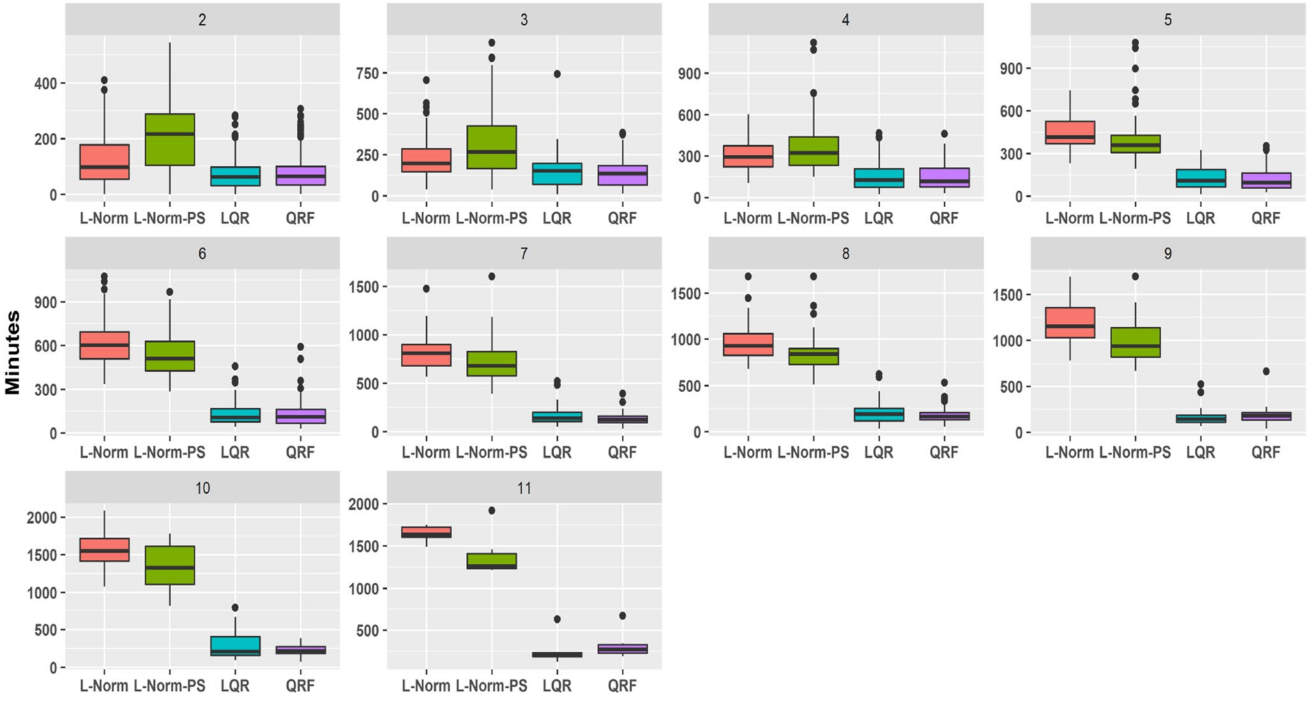


Fig. 2 Box-plots for the RO method comparing the L-Norm, L-Norm-PS, LQR and QRF model's out-of-sample error for the different surgery regimes organized by the number of cases scheduled each day

Table 6 Weighted percentage improvement of the out-of-sample performance error when the QRF is used rather than the L-Norm, L-Norm-PS, LQR models in mean and percentiles for the Individualized Robust Optimization framework

		Number of Surgeries per Day				
Metric	Model	2-3	4-5	6-7	8-9	10-11
Mean	L-Norm	38.9	63.2	80.8	82.2	83.8
	L-Norm-PS	62.0	64.3	78.1	79.4	81.1
	LQR	-1.2	5.8	7.0	4.3	12.8
75th percentile	L-Norm	41.2	57.7	79.2	82.0	83.2
	L-Norm-PS	62.7	56.4	77.2	78.7	81.5
	LQR	0.7	4.5	9.9	10.5	19.4
95th percentile	L-Norm	38.1	40.5	72.6	76.4	76.9
	L-Norm-PS	55.1	52.2	71.3	74.6	75.4
	LQR	-6.5	-0.9	5.4	22.9	33.8

The percentiles metric corresponds to the schedule whose cost by the QRF model falls at that percentile

Finally, we consider the DRO framework in Section 4.3. We implemented DRO with constraints dictated by percentiles $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95\}$ of the surgery duration distributions derived from L-Norm, L-Norm-PS, LQR and QRF. Figure 3 once again shows improvement of QRF over the other three models regarding the distributions of overall costs. In particular, Table 7 shows over all cases, QRF obtains a 67.1%, 62.8%, and 22.8% less out-of-sample cost than compared to L-Norm, L-Norm-PS, and LQR, respectively.

5.3 Sensitivity analysis of T

To understand the influence of the T parameter in models, we conduct sensitivity analysis to determine its effect on the out-of-sample performance. First, we rerun our models with scaling of the original T values used in our numerical experiments by $\{0.90, 0.95, 1.05, 1.10\}$. In Appendix B, Tables 20, 21 and 22 detail the results corresponding to this scaling for the SAA, RO, and DRO approaches. The results indicate a generally positive correlation between the QRF model's performance over the other benchmark models and the scaling value. The association was nonlinear for only the RO framework with LQR. Even these results, however,

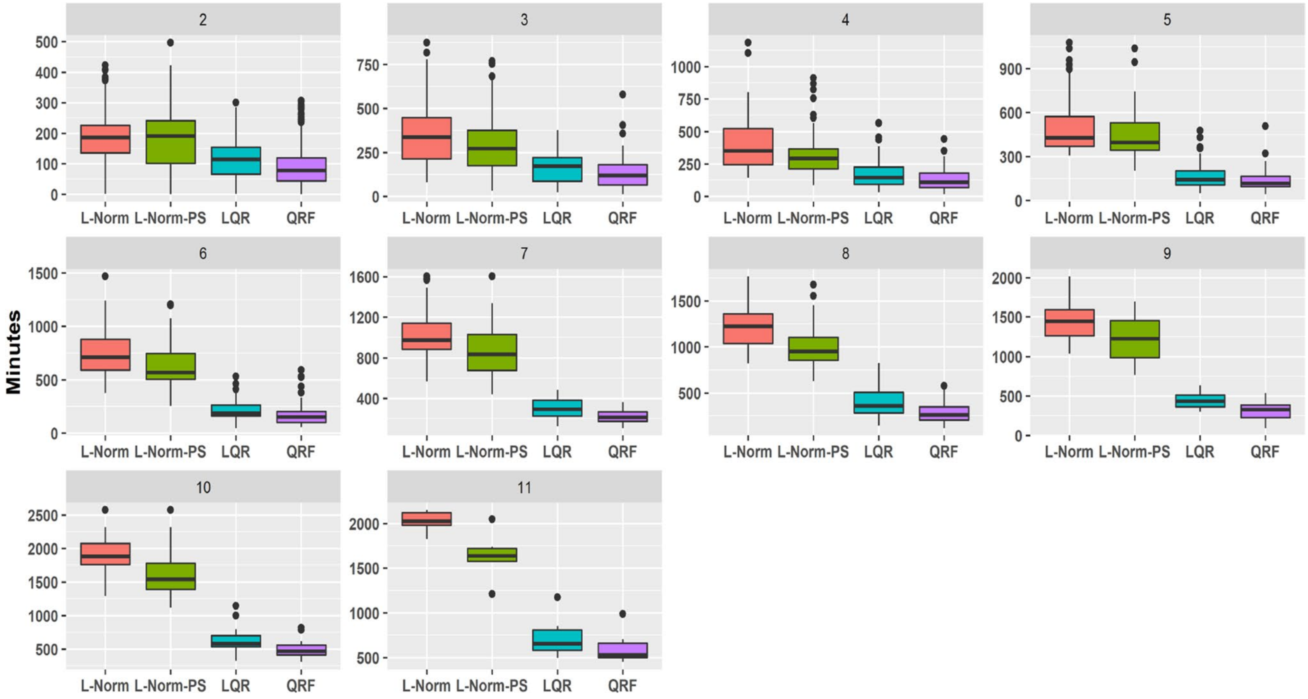


Fig. 3 Box-plots for the DRO method comparing the L-Norm, L-Norm-PS, LQR and QRF model's out-of-sample error for the different surgery regimes organized by the number of cases scheduled each day

Table 7 Weighted percentage improvement of the out-of-sample performance error when the QRF is used rather than the L-Norm, L-Norm-PS, LQR models in mean and percentiles for the Individualized Distributionally Robust Optimization framework

		Number of Surgeries Per Day				
Metric	Model	2-3	4-5	6-7	8-9	10-11
Mean	L-Norm	55.2	70.4	77.1	77.5	72.5
	L-Norm-PS	52.9	65.7	72.7	72.6	67.7
	LQR	22.5	20.6	21.9	27.7	20.6
75th percentile	L-Norm	51.8	68.4	76.5	75.0	71.9
	L-Norm-PS	51.1	61.2	73.3	70.4	66.7
	LQR	21.9	19.3	26.3	29.4	19.7
95th percentile	L-Norm	43.1	68.6	70.8	71.0	64.3
	L-Norm-PS	46.0	63.6	66.2	65.4	63.5
	LQR	7.7	28.0	7.5	18.5	21.2

The percentiles metric corresponds to the schedule whose cost by the QRF model falls at that percentile

show that the QRF approach performs better than LQR for larger values of *T*. Averaging across all case sizes, we find by varying *T* that *QRF* improves L-Norm, L-Norm-PS, and LQR for all considered optimization methodologies by no fewer than 35.0%, 33.8%, and 0.8%, respectively.

The findings indicate that QRF will still be beneficial so long as MSKCC can set T to reasonably reflect the cases each day. Next, we provide results to show that MSKCC can realistically see the benefits of utilizing QRF in practice. In particular, each model was rerun with T set to the historical median sum of actual surgical durations in the training data based on their case size. Table 23 in Appendix B presents the median values found from the training and test set.

Table 8 presents the percent improvement in using the QRF approach over other benchmarks in each optimization framework with our median-based *T*. Like previous results, QRF strongly improves upon the standard distribution fitting benchmarks across all optimization frameworks. The QRF demonstrated a better out-of-sample performance than L-Norm overall by 33.9% for SAA, 54.4% for RO, and 61.9% for the DRO framework. For L-Norm-PS, the QRF approach averaged across regimes was 30.2%, 59.0%, and 54.6% better under the SAA, RO, and DRO framework, respectively. LQR was the closest approach to QRF, having performed 1.1% better in the RO framework and 9.2% and 16.7% worse in the SAA and DRO framework. Even if MSKCC used the simple idea of medians to set *T*, the QRF approach can still typically provide superior results to other benchmarks. The results suggest that using a more sophisticated idea to estimate T more accurately will only lead to better performance with QRF.

5.4 Insights

For insight into the improvement numbers, we analyzed each of the four approaches in terms of prediction accuracy, separate of the optimization task. This is because for each of the percentage improvements shown in Tables 5, 6, and 7, all models were equal aside from their parameters. Note, the mean absolute error (MAE) and root-mean-square error (RMSE) presented in this section are in the unit of minutes and generally referenced as error(s).

Table 9 displays the MAE and RMSE from the true durations in the testing data for the median, min, and max values of the 5000 samples drawn as described in Section 5.2. The results displayed in Table 9 show that even the extremes of the drawn samples (i.e., the largest sampled value for each case) of the QRF is, on average, significantly closer to the true durations than L-Norm and L-Norm-PS. We see then QRF gives a distribution of the surgery time that has less variability, and this smaller variability, in turn, means smaller uncertainty in the optimization. As a result, the solution of the model needs to “hedge” a smaller variability of scenarios and is less conservative. This is the opposite of fitting a single or limited number of distributions from the population. Although every patient’s durations, when aggregated, do form a “pooled” distribution, patients and procedures correlate with the cases’ specifics. Thus, it is intuitive to see why having some patients share the same duration distribution

Table 8 Setting parameter T with historic median based on schedule case size. The results show the weighted percentage improvement of the out-of-sample performance error when the QRF is used rather than the L-Norm, L-Norm-PS, LQR models for each optimization framework

		Number of Surgeries per Day				
	Model	2-3	4-5	6-7	8-9	10-11
SAA	L-Norm	24.7	29.3	44.4	53.2	54.7
	L-Norm-PS	24.9	24.6	35.5	45.8	51.4
	LQR	10.5	5.4	5.0	14.2	9.2
RO	L-Norm	22.7	53.5	75.4	78.4	76.7
	L-Norm-PS	50.9	44.8	72.0	74.8	75.0
	LQR	-5.3	4.6	5.4	2.2	1.4
DRO	L-Norm	55.1	53.0	71.5	73.7	68.4
	L-Norm-PS	44.7	48.0	67.2	68.0	62.6
	LQR	16.8	8.9	19.6	20.8	17.7

Table 9 MAE (RMSE) in minutes of the median, minimum, and maximum values of the samples drawn from QRF, LQR, L-Norm, and L-Norm-PS

	MAE (RMSE) from Median of Samples	MAE (RMSE) from Min of Samples	MAE (RMSE) from Max of Samples
QRF	24.5 (40.2)	50.5 (76.7)	66.3 (85.8)
LQR	34.4 (55.3)	65.0 (95.4)	86.7 (108)
L-Norm	110 (142)	135 (188)	329 (352)
L-Norm-PS	81.4 (111)	98.6 (144)	272 (294)

with identical parameters can lead to more inaccurate modeling and hence poorer optimization solutions. For additional analysis of each distribution model on the test data see Tables 18 and 19 in Appendix B.

We further extended our analysis to include the MSKCC's planned times (H-PT) so that five approaches (QRF, LQR, L-Norm, L-Norm-PS, and H-PT) can be compared to the true durations in the test set. The median was taken as the prediction to be consistent with how MSKCC's planned times are derived when evaluating for accuracy. In addition, the median was also found to produce the lowest mean absolute error (MAE) for QRF and LQR. Table 10 provides the correlation, MAE, and RMSE associated with each approach. We also provide in Table 10 the MAE from the subset of predicted times that were greater than their observed duration (over predicted) and the subset of predicted times that were less than their observed duration (under-predicted). By ordering the MAE corresponding to each approach (excluding H-PT) from smallest to largest, a similar structure to the results in Tables 5, 6, and 7 can be seen in that QRF performs best, followed by LQR, L-Norm-PS, and L-Norm. Table 10 details the number of non-identical distributions constructed over our entire dataset to highlight the association between individualization and accuracy. Note, these are based on the quantiles for QRF and LQR, the distribution parameters for the log-normal distributions, and assumed for H-PT. The errors of the over-predicted and under-predicted groups further supports the notion that QRF produces fewer variable distributions.

As discussed previously, a known challenge in using CPT codes is that there exists a large number of them, a majority of which have limited or no historical data [16, 34]. The results of table 10 show QRF outperforms the H-PT, which is from a model stratified by CPT code. To more precisely understand the reason, we examined if QRF is better suited than H-PT for predicting the duration of procedures with little to no historical data. Before proceeding, we note that if a procedure has zero past observations, MSKCC's planning system will use similar surgeries performed by surgeons to impute a value for the H-PT. Likewise, ML methods can extrapolate for procedures with no historical observations by examining other relevant features. For analysis, we focused on CPT codes associated with ten or fewer historical observations and compared the predictive power of QRF versus H-PT in two breakdowns presented in Table 11. The first breakdown is predicted durations for procedures with at least one instance in the training data, while the second is for procedures with zero instances present in the training data.

The QRF slightly underperforms the H-PT when predicting the durations of previously unseen procedures but noticeably outperforms the H-PT if just a few records of a procedure are available. Interestingly, we find that all procedures with zero instances in the training data were part of secondary services scheduled through the URO department. Such events can occur when patients have urologic cancer that metastasized to regions like the thorax. As mentioned in Section 5.1, MSKCC's scheduling system primarily filters by CPT to predict surgery durations. This means secondary URO services could have jointly used CPTs with another department. Therefore, we reason that H-PT is likely better than QRF at predicting previously unseen procedures because H-PT uses historical records outside the URO department. It is still promising, however, that QRF could achieve statistically the same MAE as H-PT with fewer observations (t-test: p-value = 0.242).

A set of important/significant variables (see Table 17 in Appendix B) that showed up in both QRF and LQR were the

Table 10 Using the median corresponding to distributions from QRF, L-Norm, L-Norm-PS, LQR, and H-PT to Predict Case Durations

	Correlation	MAE (RMSE)	MAE (RMSE) of Over Predicted	MAE (RMSE) of Under Predicted	Num. of Non-identical Distributions Constructed
QRF	0.955	23.7 (39.2)	18.6 (28.9)	27.8 (46.4)	23,176
LQR	0.909	34.4 (55.6)	25.0 (37.2)	43.2 (68.7)	23,176
L-Norm	≈ 0	110 (142)	54.9 (57.9)	168 (193)	1
L-Norm-PS	0.539	81.4 (111)	65.8 (78.2)	112 (155)	49
H-PT	0.921	31.8 (52.5)	30.1 (51.8)	33.5 (52.7)	≥ 437

Table 11 Breakdown of QRF vs. H-PT predictive power for procedures with 10 or less total instances in the entire data

	Approach	Corr.	MAE (RMSE)	MAE (RMSE) for Over-Predicted	MAE (RMSE) for Under-Predicted
Min. One Instance in Training	H-PT	0.696	91.3 (136)	93.1 (143)	88.9 (128)
	QRF	0.851	56.9 (82.7)	43.1 (62.9)	71.1 (97.4)
Zero Instances in Training	H-PT	0.890	70.2 (94.5)	67.8 (88.1)	73.8 (104)
	QRF	0.878	72.1 (108)	43.9 (48.1)	84.3 (125)

relative value unit measurements (RVU). RVUs are standard scale measures currently used in the U.S. by the Centers for Medicare and Medicaid Services to determine physician fees (Hsiao et al., 1992). They are based mainly on the relative time, including pre-procedure, surgeons have previously taken to complete a specified service. As a result, RVUs are a proxy for the expected procedure duration that our QRF model can use without explicit knowledge of CPT codes. While the LQR model considered the primary surgeon as a significant variable, it is interesting that the QRF model did not. QRF's ability to accommodate complex, nonlinear associations makes it possible that our model also identifies surgeons inherently through trends in the number and value of RVUs along with other possible feature interactions. Thus, we believe our model incorporates the essential characteristics of the procedure and surgeon in large part through RVUs, and this allows it to make estimates even for cases without CPTs in the historical data.

6 Conclusion

The integration of decision-making and individualization is a critically important area of research for advancing healthcare delivery. In the context of surgery scheduling, the duration of each case in a given day varies due to their distinct characteristics, which leads to significant uncertainty. Previous studies have commonly captured this uncertainty by fitting case duration distributions at the aggregate level. However, this approach cannot generate individualized distributions to characterize the subtle differences in uncertainty across cases. Individualized distributions require conditioning on a patient's unique features.

This paper shows how QRF is an alternative method to distribution fitting that yields duration distributions individualized to every case. We present a framework to incorporate the tailored distributions generated from QRF into SAA, RO, and DRO models. As theoretical justification, reformulations and consistent statistical guarantees are derived for each optimization-under-uncertainty approach. We further conduct a case study using MSKCC data for empirical support of our framework. The numerical results reflect QRF and its individualized duration distributions can lead to optimization model solutions that significantly outperform distributions fitted at an aggregate and stratified level.

Our primary objective is to show the value of the QRF prediction method and the potential benefit individualized uncertainty modeling can bring to decision-making problems. Based on our numerical experiments, an alternative ML method to QRF can perform slightly better on specific case sizes depending on the optimization framework. To find the best ML model that works for a hospital, we would like to state that bootstrapping is a potentially viable way to obtain individualized prediction distributions from non-quantile-based ML models. This approach, however, may not always be appropriate. For instance, the formalized approach to using bootstrap samples with linear regression models is limited in requiring random errors to be homoscedastic [12]. One benefit of a quantile method like LQR is that it is known to outperform least-squares linear models when errors are non-Gaussian [29]. Even more appealing than LQR is QRF, given that it does not require any assumptions on the error distribution or the functional relationship between the outcome variable and its predictors. Ultimately, whether it is QRF, LQR, or bootstrap sampling, choosing which approach to use is a multifaceted modeling decision, and we are not advocating that QRF is always the best ML method to use. This paper strongly suggests that QRF, as a theoretically compatible approach for various optimization modeling frameworks, strikes a nice balance between model flexibility, computation, and performance. As a result, we believe any hospital not currently using individualized surgery schedules can easily obtain value in practice from individualized schedules through a QRF.

We foresee several potential directions to follow based on our study. It would be interesting to further understand how QRF predictions can impact solutions when the objective has waiting, idle, and overtime costs that are non-identical and how choosing the quantiles can trade off between idle and waiting times. Other penalty and cost types, and their sensitivities, can also be considered. We also plan on a follow-up study to understand the effectiveness of each optimization approach, their comparisons, and trade-offs in practice as we continue our collaboration with MSKCC. In this regard, we are excited by the future direction in researching how our individualized framework can be further enhanced to provide practical value in building decision-support tools. For surgery scheduling, this could include additional ML methods for setting parameters in the optimization models that depend on the complexity of the surrounding cases. As another direction, we could extend our approach to other healthcare problems such as chemotherapy infusion treatment sessions scheduling [11] or operating room planning [24]. Precision medicine and personalized, prescriptive analytics are evolving fields that can use sophisticated data-driven decision-making methods. We hope our framework and results encourage further exploration of individualized optimization under uncertainty in healthcare.

A: Proofs

A. 1: Proof of Theorem 1

By Conditions 1-5 in Appendix 1, we invoke Theorem 4 so that

$$\sup_{z\in\mathcal{A}} |\hat{F}(z|\xi_i) - F(z|\xi_i)| \xrightarrow{p} 0 \text{ as } N \to \infty, \tag{29}$$

for each i. Now, consider

$$
\begin{aligned}
&\hat{E}[f(X,Z)|\xi_1,\dots,\xi_n] - E[f(X,Z)|\xi_1,\dots,\xi_n]\\
&= \int f(X,Z)d\left(\prod_{i=1}^{n}\hat{F}(z_i|\xi_i) - \prod_{i=1}^{n}F(z_i|\xi_i)\right)\\
&= \int f(X,Z)d\left(\prod_{i=1}^{n}\hat{F}(z_i|\xi_i) - F(z_1|\xi_1)\prod_{i=2}^{n}\hat{F}(z_i|\xi_i)\right)\\
&+ \int f(X,Z)d\left(F(z_1|\xi_1)\prod_{i=2}^{n}\hat{F}(z_i|\xi_i)\right.\\
&\left.-F(z_1|\xi_1)F(z_2|\xi_2)\prod_{i=3}^{n}\hat{F}(z_i|\xi_i)\right)\\
&+ .. + \int f(X,Z)d\left(\prod_{i=1}^{n-1}F(z_i|\xi_i)\hat{F}(z_n|\xi_n) - \prod_{i=1}^{n}F(z_i|\xi_i)\right)
\end{aligned}
$$

by telescoping, which can be further written as

$$
\begin{aligned}
&\int H_1(X,z_1|\xi_2,\dots,\xi_n)d(\hat{F}(z_1|\xi_1) - F(z_1|\xi_1))\\
&+\int H_2(X,z_2|\xi_1,\xi_3,\dots,\xi_n)d(\hat{F}(z_2|\xi_2) - F(z_2|\xi_2))\\
&+ .. + \int H_n(X,z_n|\xi_1,\dots,\xi_{n-1})d(\hat{F}(z_n|\xi_n) - F(z_n|\xi_n))
\end{aligned}
\tag{30}
$$

where

$$
\begin{aligned}
&H_1(X,z_1|\xi_2,..,\xi_n) = \int f(X,Z)d\prod_{i=2}^{n}\hat{F}(z_i|\xi_i)\\
&H_2(X,z_2|\xi_1,\xi_3,..,\xi_n) =\\
&\qquad\qquad \int f(X,Z)d\left(F(z_1|\xi_1)\prod_{i=3}^{n}\hat{F}(z_i|\xi_i\right)\\
&\qquad\qquad \vdots\\
&H_n(X,z_n|\xi_1,..,\xi_{n-1}) = \int f(X,Z)d\left(\prod_{i=1}^{n-1}F(z_i|\xi_i)\right)
\end{aligned}
$$

In other words, $H_i(\cdot)$'s are the conditional expectation of $f(X,Z)$ given z_i, where the underlying distributions that generate the other z_j's are $\hat{F}$ for $j<i$ and F for $j>i$.

Next, with the assumption that z_i is bounded and $\mathcal{X}$ is compact, and thanks to the max-plus representation of $f(X,\cdot)$, we can verify that $H_i(\cdot,\cdot|\cdot)$ has uniformly bounded total variation, i.e.,

$$
\sup_{\substack{X\in\mathcal{X},\ \xi_j\in\Xi,\\ j=1,\dots,n,\ i=1,\dots,n}} \|H_i(X,\cdot|\xi_1,..,\xi_{i-1},\xi_{i+1},..,\xi_n)\|_{TV} \le C < \infty
$$

where $\|\cdot\|_{TV}$ is the total variation norm. Consequently, we have

$$
\begin{aligned}
&\int H_i(X,z_i|\xi_1,..,\xi_{i-1},\xi_{i+1},..,\xi_n)d(\hat{F}(z_i|\xi_i) - F(z_i|\xi_i))\\
&\le \|H_i(X,z_i|\xi_1,..,\xi_{i-1},\xi_{i+1},..,\xi_n)\|_{TV}\sup_{z_i\in\mathcal{A}}|\hat{F}(z_i|\xi_i) - F(z_i|\xi_i)|\\
&\le C\sup_{z_i\in\mathcal{A}}|\hat{F}(z_i|\xi_i) - F(z_i|\xi_i)|
\end{aligned}
$$

by using Section 7.2.2 Lemma B(ii) in [43] in the first inequality. Thus, from Eq. 30, we have

$$
\begin{aligned}
&|\hat{E}[f(X,Z)|\xi_1,\dots,\xi_n] - E[f(X,Z)|\xi_1,\dots,\xi_n]|\\
&\le C\sum_{i=1}^{n}\sup_{z_i\in\mathcal{A}}|\hat{F}(z_i|\xi_i) - F(z_i|\xi_i)| \xrightarrow{p} 0
\end{aligned}
\tag{31}
$$

by using Eq. 29 and Sluksky's Theorem.

Next, we argue that $\{f(X,\cdot): X\in\mathcal{X}\}$ is a Vapnik-Cervonenkis (VC) class of functions. This can be seen by iteratively applying Lemma 2.6.18 in [51] on the max-plus construction of $f(\cdot,\cdot)$ and noting that linear functions are VC. Then, using our assumption that z is bounded, by Theorem 2.8.1 in [51] and the remark at the end of Section 2.8.1 therein, we see that $\{f(X,\cdot): X\in\mathcal{X}\}$ is a uniform Glivenko-Cantelli (GC) class, meaning that

$$
\sup_{\hat{P}\in\mathcal{P}}\mathbb{P}_{\hat{P}}\left(\sup_{K\ge M}\sup_{X\in\mathcal{X}}|E_{\tilde{P}}[f(X,Z)] - E_{\hat{P}}[f(X,Z)]| > \epsilon\right) \to 0
\tag{32}
$$

as $M\to\infty$, where $\mathcal{P}$ is the set of all probability measures on Z where Z is bounded over $\mathcal{A}^n$, $E_{\hat{P}}[\cdot]$ denotes the expectation under $\hat{P}$, and $\mathbb{P}_{\hat{P}}(\cdot)$ generates K i.i.d. scenarios under $\hat{P}$ that lead to the empirical distribution $\tilde{P}$ used in the expectation $E_{\tilde{P}}[\cdot]$.

For convenience, we denote $\tilde{H}(X) = \tilde{E}[f(X,Z)|\xi_1,..,\xi_n]$ as the SAA objective function in Eq. 6 where the K scenarios are generated from $\prod_{i=1}^{n}\hat{F}(z_i|\xi_i)$. Denote $\hat{H}(X) = \hat{E}[f(X,Z)|\xi_1,\dots,\xi_n]$ as the objective function evaluated directly under $\prod_{i=1}^{n}\hat{F}(z_i|\xi_i)$, and $\hat{X}^*$ be an optimal solution to the formulation Eq. 3 but with this $\hat{H}(X)$ as the objective. Also denote X^* as the optimal solution to Eq. 3 with objective $H(X)$. We write

$$
\begin{aligned}
H(\tilde{X}^*)-H^* &= [H(\tilde{X}^*) - \tilde{H}(\tilde{X}^*)] + [\tilde{H}(\tilde{X}^*) - \tilde{H}(\hat{X}^*)]\\
&+[\tilde{H}(\hat{X}^*) - \hat{H}(\hat{X}^*)] + [\hat{H}(\hat{X}^*) - \hat{H}(X^*)]\\
&+[\hat{H}(X^*) - H(X^*)]
\end{aligned}
\tag{33}
$$

We analyze each term in Eq. 33. Note that the second and fourth terms are nonpositive by the optimality definition of $\tilde{X}^*$ and $\hat{X}^*$ with respect to $\tilde{H}(X)$ and $\hat{H}(X)$ respectively. For the last term, we have

$$
|\hat{H}(X^*) - H(X^*)| \le \sup_{X\in\mathcal{X}}|\hat{H}(X) - H(X)| \xrightarrow{p} 0 \ \text{ as } \ K\to\infty
\tag{34}
$$

thanks to Eq. 31. For the third term, we have

$$|\tilde{H}(\hat{X}^*) - \hat{H}(\hat{X}^*)| \le \sup_{X\in\mathcal{X}} |\tilde{H}(X) - \hat{H}(X)| \xrightarrow{p} 0 \ \text{ as } \ N \to \infty \tag{35}$$

uniformly over $\prod_{i=1}^{n} \hat{F}(z_i|\xi_i) \in \mathcal{P}$, in the sense of Eq. 32. For the first term, we have

$$H(\tilde{X}^*) - \tilde{H}(\tilde{X}^*) = [H(\tilde{X}^*) - \hat{H}(\tilde{X}^*)] + [\hat{H}(\tilde{X}^*) - \tilde{H}(\tilde{X}^*)]$$

and that

$$|H(\tilde{X}^*) - \hat{H}(\tilde{X}^*)| \xrightarrow{p} 0$$

and

$$|\hat{H}(\tilde{X}^*) - \tilde{H}(\tilde{X}^*)| \xrightarrow{p} 0$$

uniformly over $\prod_{i=1}^{n} \hat{F}(z_i|\xi_i) \in \mathcal{P}$, argued similarly as in Eqs. 34 and 35. Then, putting all the above together, using Eq. 33 we get

$$H(\tilde{X}^*) - H^* \xrightarrow{p} 0$$

as $K, N \to \infty$. Noting that $H(\tilde{X}^*) - H^*$ is nonnegative by the definition of H^*, this concludes our theorem.

A.2: Proof of Theorem 2

For convenience, denote

$$\mathcal{U}(\gamma) = \left\{ Z : \max_{k<i\le n} \left| \frac{\sum_{j=k}^{i-1} z_j - \sum_{j=k}^{i-1} \mu_j}{\sqrt{\sum_{j=k}^{i-1} \sigma_j^2}} \right| \le \gamma \right\} \tag{36}$$

where μ_j and σ_j^2 are the mean and variance under $F(z_j|\xi_j)$. Correspondingly, denote

$$\hat{\mathcal{U}}(\gamma) = \left\{ Z : \max_{k<i\le n} \left| \frac{\sum_{j=k}^{i-1} z_j - \sum_{j=k}^{i-1} \hat{\mu}_j}{\sqrt{\sum_{j=k}^{i-1} \hat{\sigma}_j^2}} \right| \le \gamma \right\}$$

where we recall that $\hat{\mu}_j$ and $\hat{\sigma}_j^2$ are the mean and variance under the QRF estimate $\hat{F}(z_j|\xi_j)$.

Denote Γ as the $(1-\delta)$-quantile of the random variable

$$\max_{k<i\le n} \left| \frac{\sum_{j=k}^{i-1} z_j - \sum_{j=k}^{i-1} \mu_j}{\sqrt{\sum_{j=k}^{i-1} \sigma_j^2}} \right|$$

Now, given a small positive constant ϵ, using the assumption that $F(z|\xi)$ is continuous and strictly increasing in z, we can find a small $\nu > 0$ such that

$$P\left(\max_{k<i\le n} \left| \frac{\sum_{j=k}^{i-1} z_j - \sum_{j=k}^{i-1} \mu_j}{\sqrt{\sum_{j=k}^{i-1} \sigma_j^2}} \right| \le \Gamma - \nu \right) \ge 1 - \delta - \epsilon \tag{37}$$

Now let $\hat{G}(\cdot|\xi_1, \dots, \xi_n)$ be the distribution function of the random variable

$$\max_{k<i\le n} \left| \frac{\sum_{j=k}^{i-1} z_j - \sum_{j=k}^{i-1} \hat{\mu}_j}{\sqrt{\sum_{j=k}^{i-1} \hat{\sigma}_j^2}} \right| \tag{38}$$

where the z_j are generated from $\hat{F}(z_j|\xi_j)$ independently. Correspondingly, let $\tilde{G}(\cdot|\xi_1, \dots, \xi_n)$ be the empirical distribution using K scenarios drawn from $\prod_{j=1}^{n} \hat{F}(z_j|\xi_j)$. Since the class of functions $\{I(\cdot \le t) : t \in \mathbb{R}\}$ is VC (e.g., Chapter 2.6, Problem 20 in [51]), by using the arguments similar to the proof of Eq. 32 for Theorem 1, we have

$$\sup_{u\in\mathbb{R}} |\tilde{G}(u|\xi_1, \dots, \xi_n) - \hat{G}(u|\xi_1, \dots, \xi_n)| \xrightarrow{p} 0$$

uniformly over $\hat{G}(\cdot|\xi_1, \dots, \xi_n) \in \mathcal{G}$, as $K \to \infty$, where $\mathcal{G}$ denotes the class of all possible probability measures. This implies that, given any small $\nu > 0$, we have, when K is large enough,

$$\hat{\mathcal{U}}(\tilde{\Gamma}) \supset \hat{\mathcal{U}}\left(\hat{\Gamma} - \frac{\nu}{2}\right) \tag{39}$$

uniformly over $\hat{G}(\cdot|\xi_1, \dots, \xi_n) \in \mathcal{G}$, where $\hat{\Gamma}$ is the smallest number (i.e., infimum) that satisfies Eq. 18, or alternately the quantile of Eq. 38 under $\prod_{j=1}^{n} \hat{F}(z_j|\xi_j)$. Now, by Conditions 1-5 in Appendix 1, we invoke Theorem 4 so that

$$\sup_{z\in\mathbb{R}_+} |\hat{F}(z|\xi_i) - F(z|\xi_i)| \xrightarrow{p} 0 \ \text{ as } \ N \to \infty, \tag{40}$$

for each i, which further implies that

$$\sup_{u\in\mathbb{R}_+} |\hat{G}(u|\xi_1, \dots, \xi_n) - G(u|\xi_1, \dots, \xi_n)| \xrightarrow{p} 0 \ \text{ as } \ N \to \infty$$

where $G(u|\xi_1, \dots, \xi_n)$ is the distribution function of the random variable Eq. 36 under $\prod_{j=1}^{n} F(z_j|\xi_j)$. Hence we have, when K is large enough,

$$\hat{\mathcal{U}}\left(\hat{\Gamma} - \frac{\nu}{2}\right) \supset \mathcal{U}(\Gamma - \nu) \tag{41}$$

Now, combining Eqs. 40 and 41, we have, as $K, N \to \infty$, that

$$\hat{\mathcal{U}}(\tilde{\Gamma}) \supset \mathcal{U}(\Gamma - \nu)$$

and hence

$$\tilde{H}^* = \min_{X\in\mathcal{X}} \max_{Z\in\hat{\mathcal{U}}(\tilde{\Gamma})} f(X, Z) \ge \min_{X\in\mathcal{X}} \max_{Z\in\mathcal{U}(\Gamma-\nu)} f(X, Z) \tag{42}$$

However, note that

$$P\left(f(X,Z) \le \max_{Z\in\mathcal{U}(\Gamma-\nu)} f(X,Z)\middle|\xi_1,\dots,\xi_n\right)$$
$$\ge P(Z\in\mathcal{U}(\Gamma-\nu)|\xi_1,\dots,\xi_n) \ge 1-\delta-\epsilon$$

where the first inequality is a direct implication from the property of RO, and the second inequality comes from our choice of ν in Eq. 37. Hence,

$$\max_{Z\in\mathcal{U}(\Gamma-\nu)} f(X,Z)$$
$$\ge \min\{q : P(f(X,Z)\le q|\xi_1,..,\xi_n)\ge 1-\delta-\epsilon\}$$

the right hand side being the $(1-\delta-\epsilon)$-quantile of $f(X, Z)$. From Eq. 42, we have

$$\tilde{H}^* \ge \min_{X\in\mathcal{X}} \min\{q : P(f(X,Z)\le q|\xi_1,\dots,\xi_n)\ge 1-\delta-\epsilon\}$$

Since ϵ is arbitrary, and the distribution function of $f(X, Z)$ is uniformly continuous because $\mathcal{X}$ is compact, we deduce further that

$$\tilde{H}^* \ge \min_{X\in\mathcal{X}} \min\{q : P(f(X,Z)\le q|\xi_1,..,\xi_n)\ge 1-\delta\} = H^*$$

which concludes our theorem.

A.3: Proof of Lemma 1

The dual formulation for θ is

$$\begin{aligned} \min_{\theta,\rho}\quad & \theta + \sum_{i=1}^{n}\sum_{j=1}^{m-1} r_{ij}\rho_{ij} + \sum_{i=1}^{n}\rho_{im} \\ s.t.\quad & \theta + \sum_{i=1}^{n}\sum_{j=1}^{m} I(z_i \le q_{ij})\rho_{ij} \ge f(X,Z) \quad \forall Z, \end{aligned} \tag{43}$$

where θ is the dual variable corresponding to $\int \mathrm{d}F(Z) = 1$. By analyzing the constraint, the model above can be simplified as

$$\theta \ge f(X,Z) - \sum_{i=1}^{n}\sum_{j=1}^{m} I(z_i \le q_{ij})\rho_{ij} \qquad \forall Z, \tag{44}$$

which is equivalent to

$$\theta \ge \max_{z}\max_{y\in\Omega}\left\{\sum_{i=1}^{n}(z_i-x_i)y_i - \sum_{i=1}^{n}\sum_{j=1}^{m} I(z_i \le q_{ij})\rho_{ij}\right\} \tag{45}$$

$$= \max_{y\in\Omega}\max_{z}\left\{\sum_{i=1}^{n}(z_i-x_i)y_i - \sum_{i=1}^{n}\sum_{j=1}^{m} I(z_i \le q_{ij})\rho_{ij}\right\} \tag{46}$$

$$= \max_{y\in\Omega}\max_{z}\left\{\sum_{i=1}^{n}\left((z_i-x_i)y_i - \sum_{j=1}^{m} I(z_i \le q_{ij})\rho_{ij}\right)\right\} \tag{47}$$

$$= \max_{y\in\Omega}\sum_{i=1}^{n}\max_{z_i}\left\{(z_i-x_i)y_i - \sum_{j=1}^{m} I(z_i \le q_{ij})\rho_{ij}\right\} \tag{48}$$

$$= \max_{y\in\Omega}\sum_{i=1}^{n}\max_{j=1,\dots,m}\max_{q_{i(j-1)}\le z_i\le q_{ij}}\left\{(z_i-x_i)y_i - \sum_{k=j}^{m}\rho_{ik}\right\} \tag{49}$$

$$= \max_{y\in\Omega}\sum_{i=1}^{n}\max_{j=1,\dots,m}\left\{(q_{ij}-x_i)y_i - \sum_{k=j}^{m}\rho_{ik}\right\}, \tag{50}$$

where Eq. 46 holds by the change of the order in maximizations. Equation 49 follows by defining upper and lower quantile values for any given duration, that is true $\forall z \in Z$. Finally, for any fixed j_i, Eq. 50 is immediate. Therefore, we have a lower bound on θ and since there is no other limits on θ, the minimization problem Eq. 43 can be stated as

$$\min_{\rho}\max_{y\in\Omega}\sum_{i=1}^{n}\max_{j=1,\dots,m}\left\{(q_{ij}-x_i)y_i - \sum_{k=j}^{m}\rho_{ik}\right\}$$
$$+\sum_{i=1}^{n}\sum_{j=1}^{m-1} r_{ij}\rho_{ij} + \sum_{i=1}^{n}\rho_{im}.$$

A.4: Proof of Proposition 1

It can be shown that for extreme points in Ω, y_n equals to 0 or $\phi > 0$ and also y_i equals to either 0 or $y_{i+1}+1$ for $i \le n-1$ [54, 55]. Hence, by noting the structure of the constraints in Ω, recursive application of the given relations for some $o = 1, \dots, n+1$ results in

$$y_i = \begin{cases} o-i, & 1 \le i \le o \le n; \\ n+\phi-i, & 1 \le i \le n, o = n+1 \end{cases}. \tag{51}$$

Given the recursive structure above, we can partition the integers $1, \dots, n+1$ into intervals such that $i \in [g, o]$ if and only if $i = o$ ($y_i = \pi_{io}$ for $i \in [g, o]$), which generates a one-to-one mapping of Ω's extreme points and a partition of the integers. Now, by introducing a binary variable t_{go} indicating whether $[g, o]$ is one of the partitions in $[1, n+1]$,

$$\max_{y\in\Omega}\sum_{i=1}^{n}\max_{j=1,\dots,m}\left\{(q_{ij}-x_i)y_i - \sum_{k=j}^{m}\rho_{ik}\right\}$$

can be reformulated as

$$\begin{aligned} \max_{t} \quad & \sum_{g=1}^{n+1}\sum_{o=g}^{n+1}\left(\sum_{i=g}^{o}\max_{j}\left\{(q_{ij}-x_i)\pi_{io}-\sum_{k=j}^{m}\rho_{ik}\right\}\right)t_{go} \\ s.t. \quad & \sum_{g=1}^{i}\sum_{o=i}^{n+1} t_{go} = 1\forall i \in \{1,...,n+1\} \\ & t_{go} \in \{0,1\} 1 \le g \le o \le n+1, \end{aligned} \tag{52}$$

where $(q_{j_{n+1}} - x_{n+1})\pi_{(n+1)o} - \sum_{k=j_{n+1}}^{m} \rho_{(n+1)k}$ equals $\pi_{(n+1)(n+1)} = 0$. Equation 52, due to the unimodularity of its constraint set, has a linear programming relaxation with an equivalent optimal objective value and binary optimal solution. Consequently, the dual of the linear relaxation of problem Eq. 52 can be written as

$$\begin{aligned} \min_{\lambda} \quad & \sum_{i=1}^{n}\lambda_i \\ s.t. \quad & \sum_{i=g}^{min\{o,n\}}\lambda_i \ge \sum_{i=g}^{min\{o,n\}}\max_{j}\left\{(q_{ij}-x_i)\pi_{io}-\sum_{k\ge j}^{m}\rho_{ik}\right\} \\ & \quad 1 \le g \le n, g \le o \le n+1, \end{aligned} \tag{53}$$

where $(\lambda_1, ..., \lambda_n)$ are the dual variables. Now, by incorporating Eq. 53 in problem Eq. 25, we have

$$\begin{aligned} \min_{x,\rho,\lambda} \quad & \sum_{i=1}^{n}\lambda_i + \sum_{i=1}^{n}\sum_{j=1}^{m-1} r_{ij}\rho_{ij} + \sum_{i=1}^{n}\rho_{im} \\ s.t. \quad & \sum_{i=g}^{min\{o,n\}}\max_{j=1,\dots,m}\left\{(q_{ij}-x_i)\pi_{io}-\sum_{k\ge j}^{m}\rho_{ik}-\lambda_i\right\} \le 0 \\ & 1 \le g \le n, g \le o \le n+1 \\ & \sum_{i=1}^{n} x_i \le T \\ & x_i \ge 0 1 \le i \le n, \end{aligned} \tag{54}$$

which, by introducing $\mu_{io} = \max_{j=1,\dots,m}\{q_{ij}\pi_{io} - \sum_{k=j}^{m}\rho_{ik}\}$, completes the transformation of problem Eq. 22 to its linear program equivalent given in Eq. 27. Refer to Proposition 2 in [35] for a detailed discussion.

A.5: Proof of Theorem 3

Consider $\hat{\mathcal{U}}$ in Eq. 21 and

$$U = \{P : P(z_i \le q_{ij}) = s_{ij}, i = 1,\dots,n, j = 1,\dots,m\}$$

where s_{ij} are such that $F(q_{ij}|\xi_i) = s_{ij}$, i.e., s_{ij} is the true conditional distribution at q_{ij} for patient i.

Since the true joint distribution $\prod_{i=1}^{n} F(z_i|\xi_i)$ lies in $\mathcal{U}$, we have

$$E[f(X,Z)] \le \max_{P\in\mathcal{U}} E_P[f(X,Z)]$$

and hence

$$H^* = \min_{X\in\mathcal{X}} H(X) \le \min_{X\in\mathcal{X}}\max_{P\in\mathcal{U}} E_P[f(X,Z)] \tag{55}$$

where $H(\cdot)$ is the objective function Eq. 5. Now, under Conditions 1-5 in Appendix 1, we invoke Theorem 4 to obtain $r_{ij} \xrightarrow{p} s_{ij}$ as $N \to \infty$. Next, we look at the duals of $\max_{P\in\mathcal{U}} E_P[f(X,Z)]$ and $\max_{P\in\hat{\mathcal{U}}} E_P[f(X,Z)]$, namely

$$\min_{\rho_{ij}j=1,\dots,m} \sum_{i,j} s_{ij}\rho_{ij} \text{ s.t. } \sum_{i,j} I(z_i \le q_{ij})\rho_{ij} \ge f(X,Z) \ \forall Z \tag{56}$$

and

$$\min_{\rho_{ij}j=1,\dots,m} \sum_{i,j} r_{ij}\rho_{ij} \text{ s.t. } \sum_{i,j} I(z_i \le q_{ij})\rho_{ij} \ge f(X,Z) \ \forall Z \tag{57}$$

respectively, which are upper bounds to $\max_{P\in\mathcal{U}} E_P[f(X,Z)]$ and $\max_{P\in\hat{\mathcal{U}}} E_P[f(X,Z)]$ by weak duality. Hence

$$\begin{aligned} & \min_{X\in\mathcal{X}}\max_{P\in\mathcal{U}} E_P[f(X,Z)] \\ & \le \min_{X\in\mathcal{X},\rho_{ij}}\left\{\sum_{i,j} s_{ij}\rho_{ij} : \sum_{i,j} I(z_i \le q_{ij})\rho_{ij} \ge f(X,Z) \ \forall Z\right\} \\ & \le \min_{X\in\mathcal{X},\rho_{ij}}\left\{\sum_{i,j} r_{ij}\rho_{ij} : \sum_{i,j} I(z_i \le q_{ij})\rho_{ij} \ge f(X,Z) \ \forall Z\right\} \\ & + \sum_{i,j}(s_{ij} - r_{ij})\rho^*_{ij} \end{aligned} \tag{58}$$

where $(\rho^*_{ij})_{ij}$ is a dual optimal solution to Eq. 56. Note that, since we assume $F(z|\xi)$ is continuous and strictly increasing, s_{ij}'s are all distinct, and one can verify that the vector $(s_{ij})_{ij}$ lies in the interior of the moment set $\{P \in \mathcal{P} : P(z_i \le q_{ij}) = s_{ij} : i = 1,\dots,n, j = 1,\dots,m\}$, which implies strong duality for optimization $\max_{P\in\mathcal{U}} E_P[f(X,Z)]$ (e.g., [31]). This then implies the existence of a dual optimal solution $(\rho^*_{ij})_{ij}$ that is finite.

Note that Eq. 58 is equal to

$$\hat{H}^* + \sum_{i,j}(s_{ij} - r_{ij})\rho^*_{ij} \tag{59}$$

since Eq. 57 is equivalent to Eq. 27 as they both dualize $\max_{P\in\mathcal{U}} E_P[f(X,Z)]$ (but with different representations). Thus, using our deduction that $r_{ij} \xrightarrow{p} s_{ij}$ as $N \to \infty$ and Eq. 55, we have

$$P(H^* \le \hat{H} + \epsilon) \to 1$$

as $N \to \infty$, for any given $\epsilon > 0$.

A.6: QRF consistency conditions

Let Ξ be the predictor variable with dimensionality p and $\mathcal{B}$ be the space in which Ξ lives. The following lists the assumptions from [38]:

Condition 1 $\mathcal{B} = [0,1]^p$ and Ξ is uniform on $[0,1]^p$.

Alternatively, one may assume that the density of covariates is bounded from above and below by positive constants.

Condition 2 The proportion of observations in a node, relative to all observations, is vanishing as the size of data increases, $(n \to \infty)$. The minimum number of observations in a node is non-decreasing in the limit as the size of data goes to infinity.

Condition 3 The probability that a variable is chosen for the splitpoint is bounded from below for every node by a positive constant. If a node is split, the split is chosen so that each of the resulting sub-nodes contains at least a proportion γ of the observations in the original node, for some $0 < \gamma \leq 0.5$.

Condition 4 There exists a constant L so that $F(z|\Xi = \xi)$ is Lipschitz continuous with parameter L, that is for all ξ, $\xi' \in \mathcal{B}$,

$$\sup_{y} |F(z|\Xi = \xi) - F(z|\Xi = \xi')| \leq L||\xi - \xi'||_1.$$

Condition 5 The distribution function $F(z|\Xi = \xi)$ is, for every $\xi \in \mathcal{B}$, strictly monotonically increasing in z.

Theorem 4 *Under the conditions listed in Appendix 1, it holds that, for every feature value $\xi \in \mathcal{B}$, the output of QRF, $\hat{F}(z|\xi)$, satisfies*

$$\sup_{z \in \mathbb{R}} |\hat{F}(z|\xi) - F(z|\xi)| \xrightarrow{p} 0 \quad N \to \infty.$$

where N denotes the number of i.i.d. observations, $F(z|\xi)$ is the conditional cumulative distribution function of the response variable and $\xrightarrow{p}$ denotes convergence in probability [38].

A.7: Idle time

Let s_i denote the idle (stand-by) time before the ith patient arrives. Consider the following modification to formulation (4) that includes the idle time variables

$$\begin{aligned} \min_{x \in \mathcal{X}} \quad & f(X,Z) = \sum_{i=2}^{n} w_i + s_i + \phi l \\ s.t. \quad & w_{i+1} \geq w_i + z_i - x_i \quad \forall i = 1, \dots, n-1 \\ & s_{i+1} \geq x_i - (w_i + z_i) \quad \forall i = 1, \dots, n-1 \\ & l \geq w_n + z_n - x_n \\ & w_i, l, s_i \geq 0 \quad \forall i = 2, \dots, n. \end{aligned} \tag{60}$$

We claim this is equivalent to the formulation below.

$$\begin{aligned} \min_{x \in \mathcal{X}} \quad & f(X,Z) = \sum_{i=1}^{n-1} (w_{i+1} - z_i + x_i) + w_n + \phi l \\ s.t. \quad & w_{i+1} \geq w_i + z_i - x_i \quad \forall i = 1, \dots, n-1 \\ & l \geq w_n + z_n - x_n \\ & w_i, l \geq 0 \quad \forall i = 2, \dots, n. \end{aligned} \tag{61}$$

Consider a feasible solution set of w, x, l for (61) and suppose that $s_{i+1} = w_{i+1} - (w_i + z_i - x_i)$. Since $w_{i+1} \geq w_i + z_i - x_i$ and $w \geq 0$, then for $i = 1$ to $n-1$, we have $s_{i+1} \geq 0$ and

$$\begin{aligned} w_{i+1} &\geq 0 \\ w_{i+1} - (w_i + z_i - x_i) &\geq x_i - (w_i + z_i) \\ s_{i+1} &\geq x_i - (w_i + z_i), \end{aligned}$$

which shows $w_{i+1} - (w_i + z_i - x_i)$ is a feasible solution for (60). It follows that the corresponding objective function is

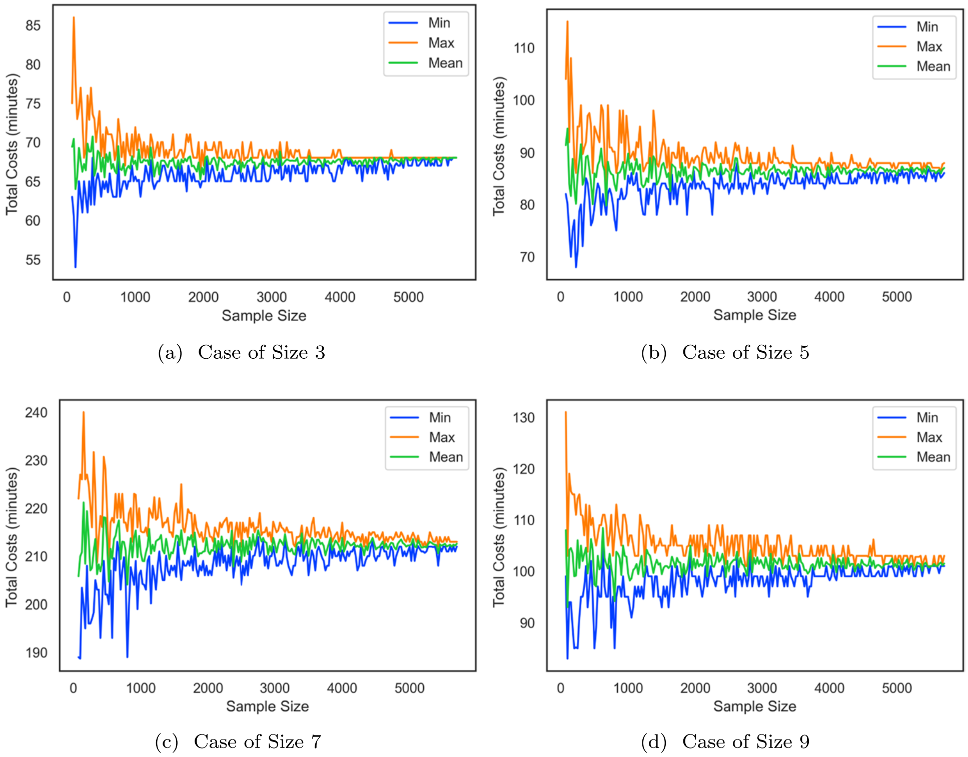
$$\begin{aligned} &= \sum_{i=2}^{n} w_i + s_i + \phi l \\ &= \sum_{i=1}^{n} w_i + \sum_{i=1}^{n-1} \Big(w_{i+1} - (w_i + z_i - x_i) \Big) + \phi l \\ &= \sum_{i=1}^{n} w_i - \sum_{i=1}^{n-1} (z_i - x_i) + w_n - w_1 + \phi l \\ &= \sum_{i=2}^{n} w_i - \sum_{i=1}^{n-1} (z_i - x_i) + w_n + \phi l \\ &= \sum_{i=1}^{n-1} (w_{i+1} - z_i + x_i) + w_n + \phi l, \end{aligned}$$

where the second to last inequality uses the fact that we assume $w_1 = 0$. Thus, we obtain our desired result.

Table 12 Summary of schedules at different case sizes used in numerical results

Case Size	2	3	4	5	6	7	8	9	10	11
Instance Count	716	190	67	62	66	49	45	21	16	6

Fig. 4 Convergence of model solution as the number of samples drawn increases



(a) Case of Size 3

(b) Case of Size 5

(c) Case of Size 7

(d) Case of Size 9

A.8: Dual formulation for idle time

Note that formulation (60) can be equivalently rewritten as

$$\begin{aligned} \min_{x\in\mathcal{X}} \quad & f(X,Z)=\sum_{i=1}^{n} w_i+s_i+\phi l \\ s.t. \quad & w_{i+1}-s_{i+1}=w_i+z_i-x_i \quad \forall i=1,\ldots,n-1 \\ & l\geq w_n+z_n-x_n \\ & w_i,l,s_i\geq 0 \quad \forall i=2,\ldots,n. \end{aligned} \tag{62}$$

The dual of this formulation is

$$\begin{aligned} \max_{y} \quad & \sum_{i=1}^{n}(z_i-x_i)y_i \\ s.t. \quad & y_i-y_{i-1}\geq -1 \qquad 2\leq i\leq n \\ & y_i\geq -1 \qquad 1\leq i\leq n-1 \\ & y_n\leq \phi. \end{aligned} \tag{63}$$

It can be seen from the structure of the LP that the extreme points correspond to $y_n\in\{\phi,-1\}$ and $y_i\in\{y_{i+1}+1,-1\}$ for $i=1,\ldots,n-1$. As shown by [25], we can construct a similar partition as the one used in Appendix 1 to complete the analysis of our DRO approach with uniform/identical idle time costs.

Table 13 MAE at median for CPT fitted distribution, H-PT, and QRF

	MAE
CPT	25.63
H-PT	25.12
QRF	20.78

Only observations with single procedures were included to avoid possible confounding

Table 14 Statistical tests comparing the median prediction from CPT fitted distributions and H-PT

	P-value
T-test	0.120
F-test	0.598

Only observations with single procedures were included to avoid possible confounding

B: Supplementary results

Figure 4, Tables 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, Figures 5, 6, 7, 8, 9, 10.

Table 15 Evaluation of QRF over its quantiles

	Quantile				
	0.05	0.25	0.5	0.751	0.95
% of Over-Predicted Cases	0.053	0.25	0.482	0.742	0.935
% of Under-Predicted Cases	0.947	0.75	0.498	0.249	0.065
MAE of Over-Predicted Cases	11.3	13.6	19.4	30.7	61.2
MAE of Under-Predicted Cases	49.5	34.8	28.9	26.4	28.3
RMSE of Over-Predicted Cases	25.3	23.6	29.5	44.3	81.7
RMSE of Under-Predicted Cases	74.8	55.0	47.4	44.5	46.3

Table 16 Evaluation of LQR over its quantiles

	Quantile				
	0.05	0.25	0.5	0.75	0.95
% of Over-Predicted Cases	0.044	0.232	0.487	0.730	0.946
% of Under-Predicted Cases	0.956	0.768	0.513	0.270	0.054
MAE of Over-Predicted Cases	13.5	17.9	25.0	41.8	89.5
MAE of Under-Predicted Cases	67.5	47.9	43.2	39.8	41.1
RMSE of Over-Predicted Cases	23.4	28.4	37.2	56.8	110.0
RMSE of Under-Predicted Cases	97.6	74.6	68.7	64.0	66.6

Table 17 Important/significant variables from QRF and LQR

Model	Features
QRF	the maximum of all RVU measures for the case, the sum of all RVU measures for the case, the mean of all RVU measures for the case, surgery room, indicator as inpatient or outpatient, indicator of robotic procedure, the number of cases scheduled on the day of surgery, number of panels in the surgical case, the number of times patient underwent surgeries, the number of days since patient's last surgery
LQR	the maximum of all RVU measures for the case, the sum of all RVU measures for the case, the mean of all RVU measures for the case, surgery room, indicator as inpatient or outpatient, indicator of robotic procedure, the number of times patient underwent surgeries, the number of cases scheduled on the day of surgery, gender, age, weight, the number of days since the patient's last surgery, race, the number of times patient underwent chemotherapy, height, the primary surgeon, the number of times patient had a CT scan

The ten most important variables for QRF are presented. Variable importance for QRF are computed using the prediction accuracy from out-of-bag sample data permuted for each predictor variable. More details can be found in [32] under the section titled 'importance'. The significant variables from LQR are based on the coefficient p-values and a significance level of 0.05

Table 18 Percent of patients that fall within the quantiles of QRF, LQR, L-Norm, and L-Norm-PS

	Quantile			
	0.25	0.5	0.75	0.95
QRF	0.25	0.502	0.751	0.935
LQR	0.232	0.487	0.730	0.946
L-Norm	0.363	0.501	0.625	0.986
L-Norm-PS	0.538	0.649	0.780	0.952

This table presents the percent of patients that were over predicted by each model for the specified quantiles. Taking 1 minus each value will give approximately the percent of patients that were under predicted

Table 19 MAE (RMSE) of Patients that Fall Within the Quantiles of QRF, LQR, L-Norm, and L-Norm-PS

	Quantile			
	0.25	0.5	0.75	0.95
QRF	12.6 (22.7)	18.6 (28.9)	30.4(44.1)	61.1 (81.7)
LQR	17.9 (28.4)	25.0 (37.2)	41.8 (56.8)	89.6 (110.0)
L-Norm	16.4 (18.3)	54.8 (57.8)	135 (133)	333 (355)
L-Norm-PS	35.1 (47.0)	65.8 (78.2)	118 (133)	280(399)

This table presents the MAE and RMSE for the patients that fell within (i.e., over predicted) the values corresponding to the specified quantiles. These should be compared by matching with the values in Table 18 first

Table 20 Sensitivity analysis of parameter T. The results show the weighted percentage improvement of the out-of-sample error when the QRF is used rather than the L-Norm, L-Norm-PS, LQR models in mean and percentiles for the Individualized Stochastic Optimization framework

		Number of Surgeries per Day				
	Model	2-3	4-5	6-7	8-9	10-11
$T \cdot 0.9$	L-Norm	21.0	45.2	48.7	53.9	60.4
	L-Norm-PS	25.1	42.7	39.9	48.2	57.7
	LQR	13.2	13.2	6.7	17.9	18.3
$T \cdot 0.95$	L-Norm-PS	28.4	51.3	56.2	60.4	66.3
	L-Norm-PS	37.3	48.2	48.2	54.7	64.4
	LQR	18.3	14.9	9.0	19.7	21.9
$T \cdot 1.05$	L-Norm	58.0	63.9	68.7	71.2	76.2
	L-Norm-PS	67.8	60.3	62.0	65.0	73.9
	LQR	38.6	19.4	12.5	23.2	27.5
$T \cdot 1.10$	L-Norm-PS	69.2	69.7	73.3	75.5	80.0
	L-Norm-PS	77.1	66.3	67.4	69.3	78.3
	LQR	45.6	23.2	13.3	24.8	29.0

Table 21 Sensitivity analysis of parameter T. The results show the weighted percentage improvement of the out-of-sample error when the QRF is used rather than the L-Norm, L-Norm-PS, LQR models in mean and percentiles for the Individualized Robust Optimization framework

		Number of Surgeries per Day				
	Model	2-3	4-5	6-7	8-9	10-11
$T \cdot 0.9$	L-Norm	29.7	59.0	75.8	77.0	77.5
	L-Norm-PS	47.6	61.0	71.8	74.4	74.8
	LQR	-0.7	4.9	1.7	3.0	4.7
$T \cdot 0.95$	L-Norm-PS	33.6	61.7	78.3	79.6	80.3
	L-Norm-PS	54.5	62.4	75.2	77.2	77.7
	LQR	-1.3	4.7	4.1	3.5	7.2
$T \cdot 1.05$	L-Norm	46.1	65.6	82.7	84.0	86.5
	L-Norm-PS	69.2	65.6	80.3	81.0	84.3
	LQR	2.0	5.8	9.7	4.3	18.1
$T \cdot 1.10$	L-Norm	52.8	67.6	84.6	85.7	88.5
	L-Norm-PS	75.8	66.6	82.1	82.9	86.7
	LQR	6.1	4.4	14.6	5.5	20.9

Table 22 Sensitivity analysis of parameter T. The results show the weighted percentage improvement of the out-of-sample error when the QRF is used rather than the L-Norm, L-Norm-PS, LQR models in mean and percentiles for the Individualized Distributionally Robust Optimization framework

		Number of Surgeries per Day				
	Model	2-3	4-5	6-7	8-9	10-11
$T \cdot 0.9$	L-Norm	44.1	63.5	71.0	71.8	67.8
	L-Norm-PS	41.3	59.0	66.6	66.7	63.4
	LQR	16.6	16.6	17.5	24.8	18.1
$T \cdot 0.95$	L-Norm-PS	49.3	66.9	74.4	74.5	69.8
	L-Norm-PS	46.8	62.4	69.9	69.5	65.0
	LQR	18.9	17.7	20.2	25.2	19.3
$T \cdot 1.05$	L-Norm	61.5	73.1	80.0	79.9	75.5
	L-Norm-PS	59.7	68.1	75.6	75.2	71.0
	LQR	27.8	22.9	24.2	29.3	22.4
$T \cdot 1.10$	L-Norm	67.6	75.5	82.5	82.2	77.8
	L-Norm-PS	66.4	70.0	78.3	77.8	73.3
	LQR	33.9	24.3	25.8	31.2	24.4

Table 23 Summary of median surgery durations by case size for sensitivity analysis

Case Size	2	3	4	5	6	7	8	9	10	11
Median Total Time of Training set	503	481.5	253	248	283	318	350	385	390	398
Median Total Time of Test set	494	486	242	237	298.5	328	349	363	415	398.5
Mean Total Time of Test set	484.1	447.3	319.1	260.9	303.8	344.5	374.3	369.9	439.6	403.2

Total times presented in minutes

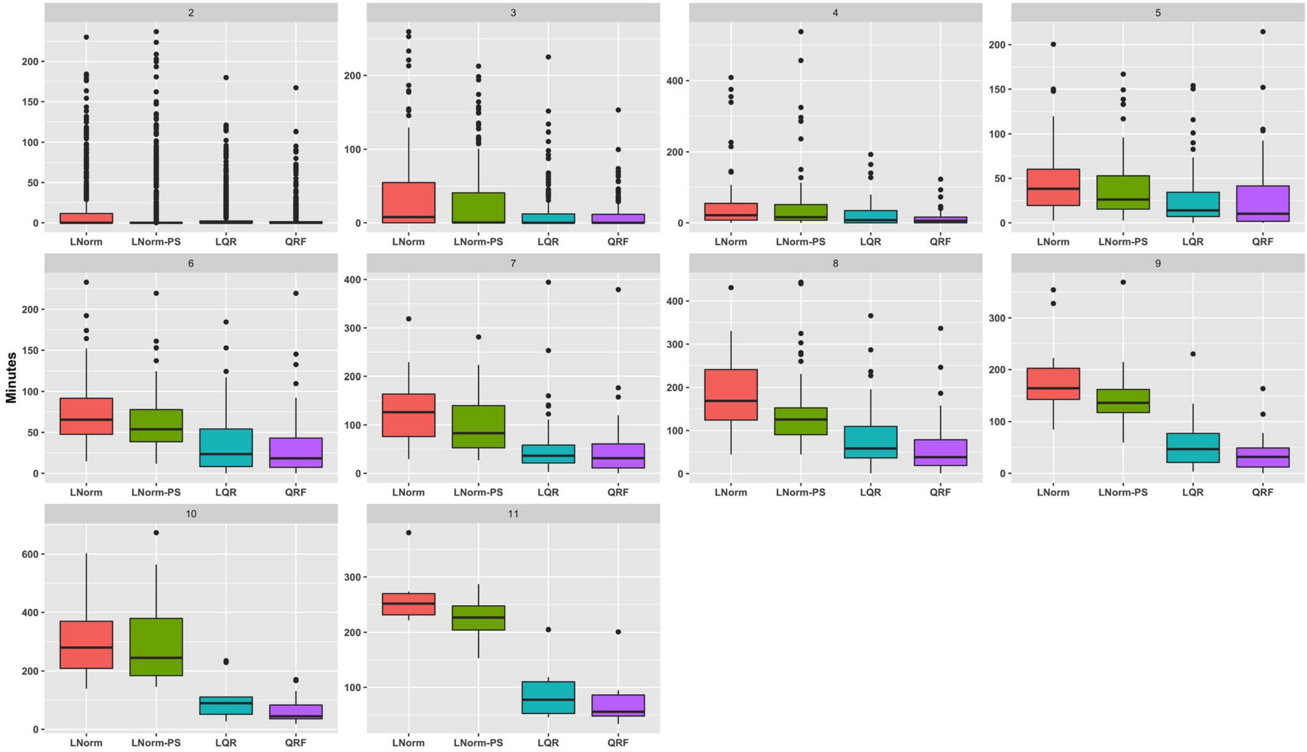


Fig. 5 Box-plots for the SAA method comparing the L-Norm, L-Norm-PS, LQR and QRF model's out-of-sample waiting time error for the different surgery regimes organized by the number of cases scheduled each day

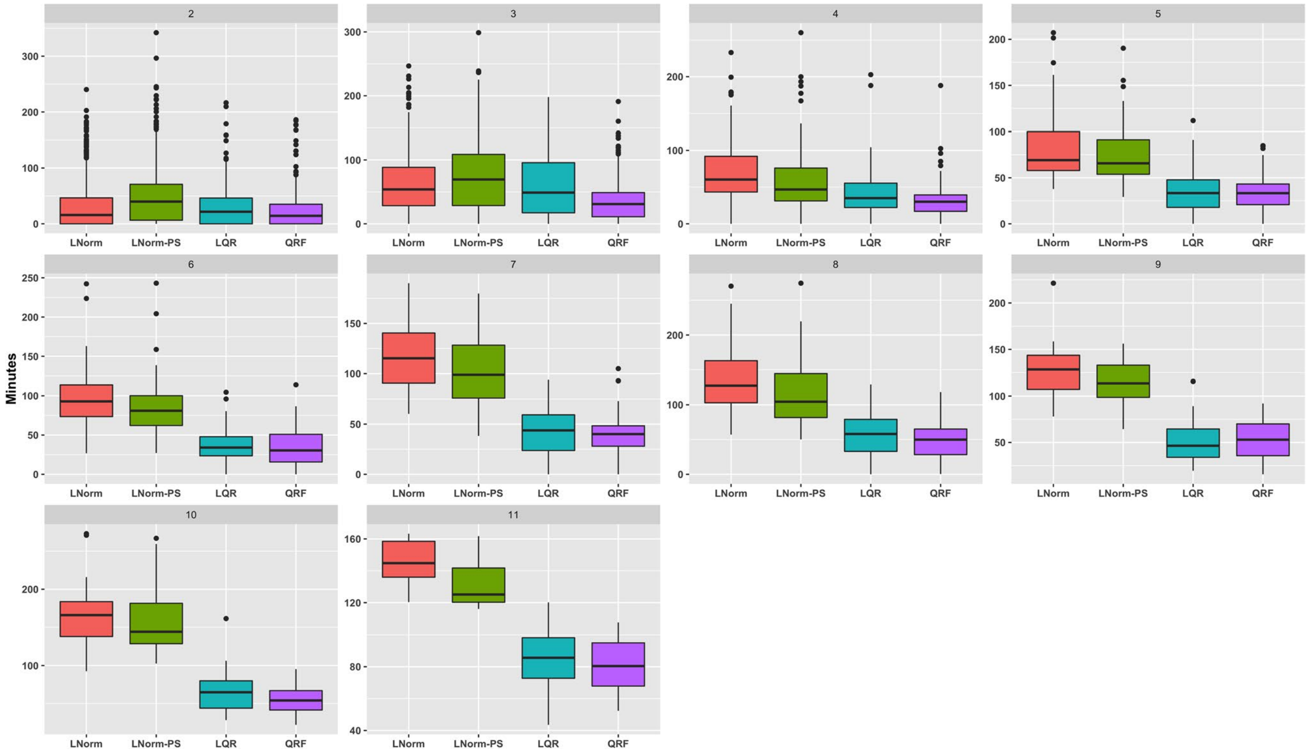


Fig. 6 Box-plots for the SAA method comparing the L-Norm, L-Norm-PS, LQR and QRF model's out-of-sample overtime error for the different surgery regimes organized by the number of cases scheduled each day

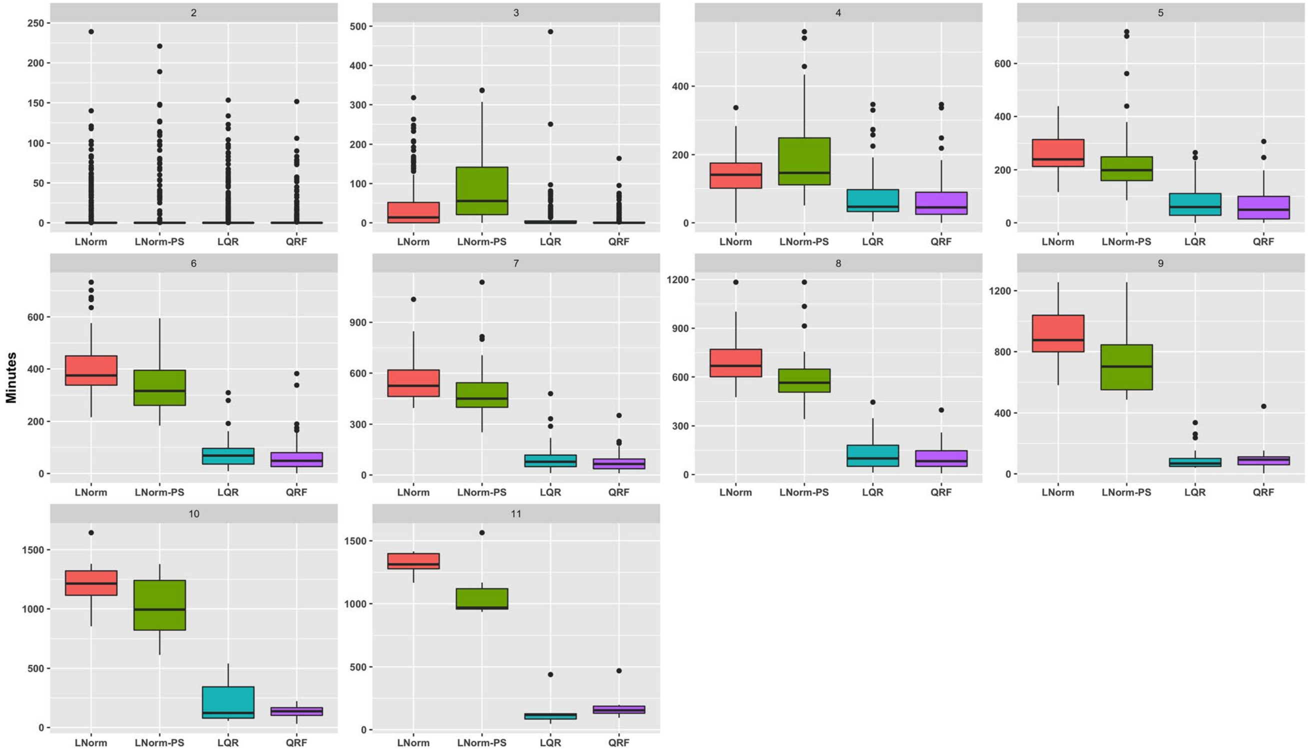


Fig. 7 Box-plots for the RO method comparing the L-Norm, L-Norm-PS, LQR and QRF model's out-of-sample waiting time error for the different surgery regimes organized by the number of cases scheduled each day

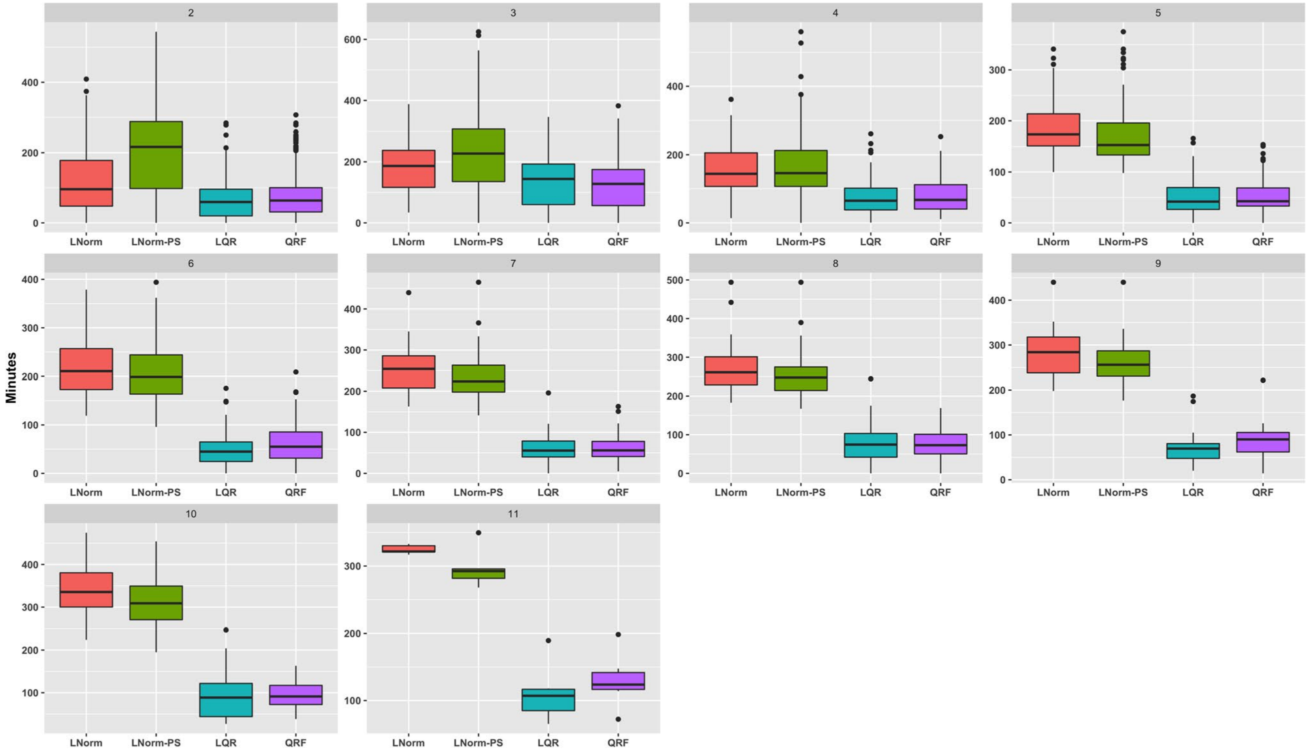


Fig. 8 Box-plots for the RO method comparing the L-Norm, L-Norm-PS, LQR and QRF models out-of-sample overtime error for the different surgery regimes organized by the number of cases scheduled each day

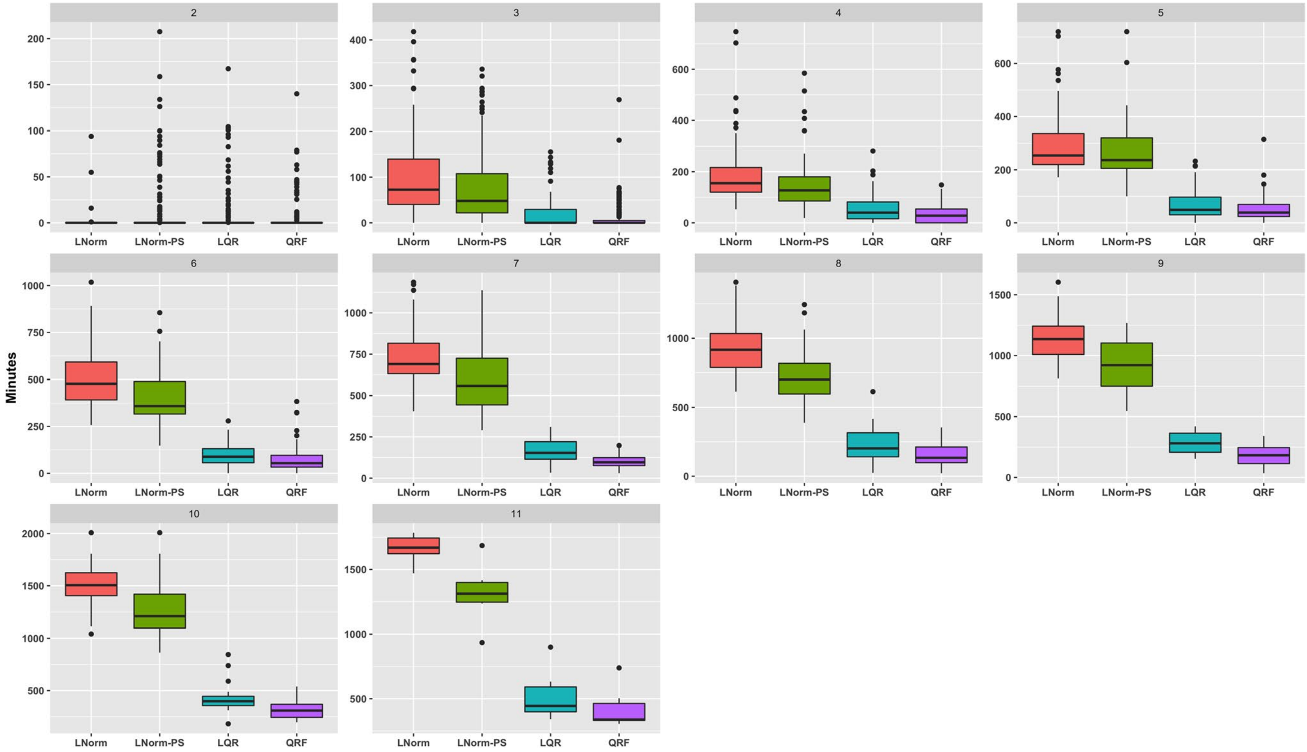


Fig. 9 Box-plots for the DRO method comparing the L-Norm, L-Norm-PS, LQR and QRF model's out-of-sample waiting time error for the different surgery regimes organized by the number of cases scheduled each day

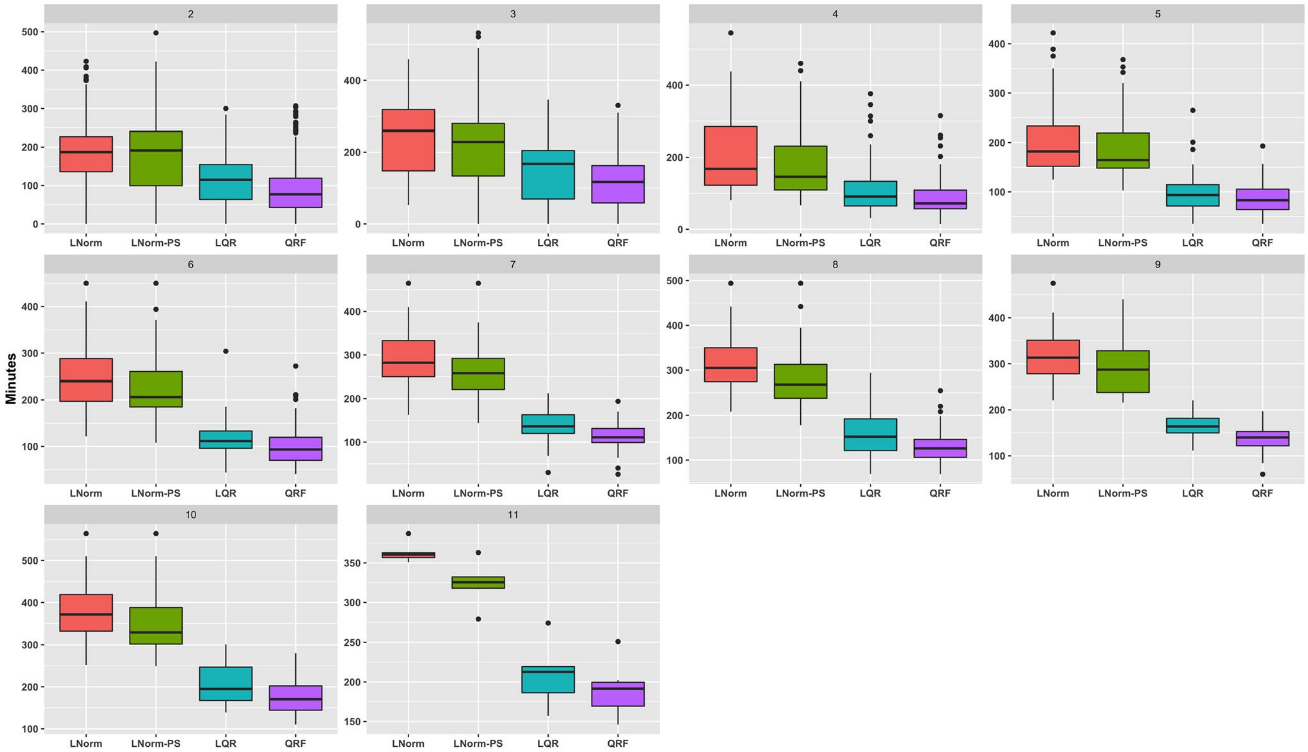


Fig. 10 Box-plots for the DRO method comparing the L-Norm, L-Norm-PS, LQR and QRF model's out-of-sample overtime error for the different surgery regimes organized by the number of cases scheduled each day

Funding Arlen Dean: National Science Foundation Graduate Research Fellowship (DGE1841052), Mark Van Oyen: National Science Foundation (CMMI1548201), and Henry Lam: National Science Foundation (CMMI1834710, CMMI1548201). Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors(s) and do not necessarily reflect the views of the National Science Foundation.

References

1. Ban GY, El Karoui N, Lim AE (2018) Machine learning and portfolio optimization. Manage Sci 64(3):1136–1154
2. Bandi C, Bertsimas D (2012) Tractable stochastic analysis in high dimensions via robust optimization. Math Program 134(1):23-70
3. Bartek MA, Saxena RC, Solomon S, Fong CT, Behara LD, Venigandla R, Velagapudi K, Lang JD, Nair BG (2019) Improving operating room efficiency: machine learning approach to predict case-time duration. J Am Coll Surg 229(4):346–354
4. Bastani H, Bayati M (2020) Online decision making with high-dimensional covariates. Oper Res 68(1):276–294
5. Ben-Tal A, El Ghaoui L, Nemirovski A (2009) Robust optimization. Princeton University Press
6. Ben-Tal A, Den Hertog D, De Waegenaere A, Melenberg B, Rennen G (2013) Robust solutions of optimization problems affected by uncertain probabilities. Manage Sci 59(2):341–357
7. Bertsimas D, Brown DB, Caramanis C (2011) Theory and applications of robust optimization. SIAM Rev 53(3):464–501
8. Bertsimas D, Gupta V, Kallus N (2018) Data-drivenrobust optimization. Math Program 167(2):235–292
9. Breiman L (2001) Random forests. Machine learning 45(1):5–32
10. Bühlmann P, Yu B et al (2002) Analyzing bagging. Ann Stat 30(4):927–961
11. Castaing J, Cohn A, Denton BT, Weizer A (2016) A stochastic programming approach to reduce patient wait times and overtime in an outpatient infusion center. IIE Transactions on Healthcare Systems Engineering 6(3):111–125
12. Davison AC, Hinkley DV (1997) Bootstrap methods and their application. 1, Cambridge university press
13. Delage E, Ye Y (2010) Distributionally robust optimization under moment uncertainty with application to data-driven problems. Oper Res 58(3):595–612
14. Denton B, Gupta D (2003) A sequential bounding approach for optimal appointment scheduling. IIE Trans 35(11):1003–1016
15. Denton B, Viapiano J, Vogl A (2007) Optimization of surgery sequencing and scheduling decisions under uncertainty. Health Care Manag Sci 10(1):13–24
16. Dexter F, Ledolter J (2005) Bayesian prediction bounds and comparisons of operating room times even for procedures with few or no historic data. Anesthesiology: The Journal of the American Society of Anesthesiologists 103(6):1259–1167
17. Dexter F, Dexter EU, Masursky D, Nussmeier NA (2008) Systematic review of general thoracic surgery articles to identify predictors of operating room case durations. Anesthesia & Analgesia 106(4):1232–1241
18. Dexter F, Epstein RH, Bayman EO, Ledolter J (2013) Estimating surgical case durations and making comparisons among

facilities: identifying facilities with lower anesthesia professional fees. Anesthesia & Analgesia 116(5):1103–1115

19. Efron B (2014) Estimation and accuracy after model selection. J Am Stat Assoc 109(507):991–1007
20. Erdogan SA, Denton B (2013) Dynamic appointment scheduling of a stochastic server with uncertain demand. INFORMS J Comput 25(1):116–132
21. Goldfarb D, Iyengar G (2003) Robust portfolio selection problems. Math Oper Res 28(1):1–38
22. Hong LJ, Huang Z, Lam H (2021) Learning-based robust optimization: Procedures and statistical guarantees. Manage Sci 67(6):3447–3467
23. Hyndman RJ, Athanasopoulos G (2018) Forecasting: principles and practice. OTexts
24. Jebali A, Diabat A (2015) A stochastic model for operating room planning under capacity constraints. Int J Prod Res 53(24):7252–7270
25. Jiang R, Shen S, Zhang Y (2017) Integer programming approaches for appointment scheduling with random no-shows and service durations. Oper Res 65(6):1638–1656
26. Kallus N, Udell M (2020) Dynamic assortment personalization in high dimensions. Oper Res 68(4):1020–103
27. Kayis E, Wang H, Patel M, Gonzalez T, Jain S, Ramamurthi R, Santos C, Singhal S, Suermondt J, Sylvester K (2012) Improving prediction of surgery duration using operational and temporal factors. In: AMIA Annual Symposium Proceedings, American Medical Informatics Association, vol 2012, p 456
28. Kleywegt AJ, Shapiro A, Homem-de Mello T (2002) The sample average approximation method for stochastic discrete optimization. SIAM J Optim 12(2):479–502
29. Koenker R, BassettJr G (1978) Regression quantiles. Econometrica: journal of the Econometric Society pp 33–50
30. Kong Q, Lee CY, Teo CP, Zheng Z (2013) Scheduling arrivals to a stochastic service delivery system using copositive cones. Oper Res 61(3):711–726
31. Lasserre JB (2010) Moments, positive polynomials and their applications, vol 1. World Scientific
32. Liaw A, Wiener M et al (2002) Classification and regression by randomforest. R news 2(3):18–22
33. Lin Y, Jeon Y (2006) Random forests and adaptive nearest neighbors. J Am Stat Assoc 101(474):578–590
34. Luangkesorn KL, Eren-Doğu Z (2016) Markov chain monte carlo methods for estimating surgery duration. J Stat Comput Simul 86(2):262–278
35. Mak HY, Rong Y, Zhang J (2014) Appointment scheduling with limited distributional information. Manage Sci 61(2):316–334
36. May JH, Strum DP, Vargas LG (2000) Fitting the lognormal distribution to surgical procedure times. Decis Sci 31(1):129–148
37. McManus ML, Long MC, Cooper A, Mandell J, Berwick DM, Pagano M, Litvak E (2003) Variability in surgical caseload and access to intensive care services. The Journal of the American Society of Anesthesiologists 98(6):1491–1496
38. Meinshausen N (2006) Quantile regression forests. Journal of Machine Learning Research 7(Jun):983–999
39. Meisami A (2018) Integrated machine learning and optimization frameworks with applications in operations management. PhD thesis, University of Michigan Ann Arbor
40. Pearce B, Hosseini N, Taaffe K, Huynh N, Harris S (2010) Modeling interruptions and patient flow in a preoperative hospital environment. In: Simulation Conference (WSC), Proceedings of the 2010 Winter, IEEE, pp 2261–2270
41. Robinson LW, Chen RR (2003) Scheduling doctors’ appointments: optimal and empirically-based heuristic policies. IIE Trans 35(3):295–307
42. Sabria F, Daganzo CF (1989) Approximate expressions for queueing systems with scheduled arrivals and established service order. Transp Sci 23(3):159–165
43. Serfling RJ (2009) Approximation theorems of mathematical statistics, vol 162. John Wiley & Sons
44. ShahabiKargar Z, Khanna S, Good N, Sattar A, Lind J, O’Dwyer J (2014) Predicting procedure duration to improve scheduling of elective surgery. In: Pacific Rim International Conference on Artificial Intelligence, Springer, pp 998–1009
45. Shapiro A, Dentcheva D, Ruszczyński A (2009) Lectures on stochastic programming: modeling and theory. SIAM
46. Sitompul D, Randhawa S (1990) Nurse scheduling models: a state-of-the-art review. J Soc Health Syst 2(1):62–72
47. Stepaniak PS, Heij C, De Vries G (2010) Modeling and prediction of surgical procedure times. Stat Neerl 64(1):1–18
48. Strömblad CT, Baxter-King RG, Meisami A, Yee SJ, Levine MR, Ostrovsky A, Stein D, Iasonos A, Weiser MR, Garcia-Aguilar J et al (2021) Effect of a predictive model on planned surgical duration accuracy, patient wait time, and use of presurgical resources: a randomized clinical trial. JAMA Surg 156(4):315–321
49. Strum DP, May JH, Vargas LG (2000) Modeling the uncertainty of surgical procedure timescomparison of log-normal and normal models. The Journal of the American Society of Anesthesiologists 92(4):1160–1167
50. Tulabandhula T, Rudin C (2014) Robust optimization using machine learning for uncertainty sets. arXiv preprint arXiv:1407.1097
51. Van DerVaart AW, Wellner JA (1996) Weak convergence. In: Weak convergence and empirical processes, Springer, pp 16–28
52. Wang PP (1993) Static and dynamic scheduling of customer arrivals to a single-server system. Naval Research Logistics (NRL) 40(3):345–360
53. Wiesemann W, Kuhn D, Sim M (2014) Distributionally robust convex optimization. Oper Res 62(6):1358–1376
54. Zangwill WI (1966) A deterministic multi-period production scheduling model with backlogging. Manage Sci 13(1):105–119
55. Zangwill WI (1969) A backlogging model and a multi-echelon model of a dynamic economic lot size production system–a network approach. Manage Sci 15(9):506–527
56. Zhou J, Dexter F (1998) Method to assist in the scheduling of add-on surgical cases-upper prediction bounds for surgical case durations based on the log-normal distribution. The Journal of the American Society of Anesthesiologists 89(5):1228–1232