Hospital-wide Inpatient Flow Optimization

Dimitris Bertsimas
Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA 02139, dbertsim@mit.edu
Jean Pauphilet
London Business School, London, NW1 4SA, jpauphilet@london.edu

An ideal that supports quality and delivery of care is to have hospital operations that are coordinated and optimized across all services in real-time. As a step toward this goal, we propose a multistage adaptive robust optimization approach combined with machine learning techniques. Informed by data and predictions, our framework unifies the bed assignment process across the entire hospital and accounts for present and future inpatient flows, discharges as well as bed requests – from the emergency department, scheduled surgeries and admissions, and outside transfers. We evaluate our approach through simulations calibrated on historical data from a large academic medical center. For the 600-bed institution, our optimization model was solved in seconds, reduced off-service placement by 24% on average, and boarding delays in the emergency department and post-anesthesia units by 35% and 18% respectively. We also illustrate the benefit from using adaptive linear decision rules instead of static assignment decisions.

Key words: Hospital operations; flow management; machine learning; multistage robust optimization

1. Introduction

A majority of hospitals in developed countries operate under increasing financial and operational stress. To improve the quality of care and alleviate the burden on clinicians and hospital staff, healthcare operations practitioners widely agree on the need to shift from isolated improvement in each individual unit to a global coordination scheme across the entire hospital (Rutherford et al. 2017). In this paper, we propose a system optimization approach combined with machine learning techniques to achieve hospital-wide inpatient flow optimization. Based on historical data from a large academic hospital, we conduct extensive simulations to assess the scalability of our approach and evaluate its potential impact on admission delays and patient misplacement.

1.1. Patient flows at a large academic hospital

In this work, we study the dynamic allocation of patients to inpatient beds, including admissions of scheduled and emergency patients. Our goal is to dynamically assign patients to beds while

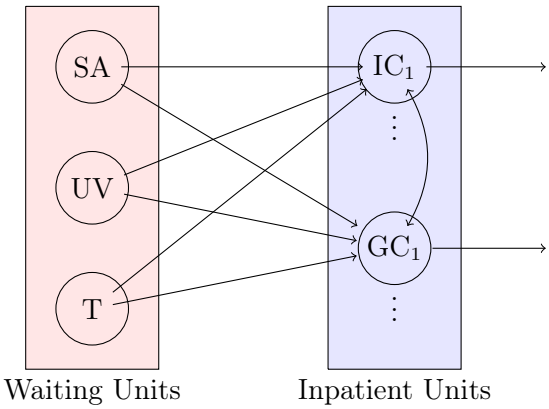


Figure 1 **Schematic views of patient flows in a typical hospital. We divide incoming flows of patients into scheduled admissions (SA), unscheduled visits (UV) and transfers (T). Once admitted, patients can flow between units corresponding to general care (GC) or intensive care (IC) units.**

accounting for future requests and discharges. We do so by formulating a robust multistage optimization problem, calibrating it with predictions from machine learning models, and considering affine decision rules to balance tractability and conservatism.

Figure 1 sketches the main patient flows in and out of a hospital within a day. New requests for beds can come from three main channels: scheduled admissions (SA), which include scheduled surgeries and other planned hospital stays; unscheduled visits (UV), i.e., emergency department (ED) visits as well as outpatient visits that unexpectedly require admission; and transfers from another institution (T). Current inpatients are allocated into the different inpatient units, each of them corresponding to different medical specialties and levels of care. We consider here two levels of care, namely intensive care (IC) and general care (GC). Each unit can serve different medical specialties or services. Typically, each unit is assigned to at least one "primary service". In some cases, they can also welcome patients from a "secondary" specialty, outside of their main expertise. Inpatients can either move to another inpatient unit –usually to change the level of care– or be discharged out of the hospital. Note that, though stylized, this representation can capture almost all relevant patient flows within a day, such as surgeries for currently admitted patients.

Our flow description reflects the practices at our (anonymous) partner institution, a large academic medical center with nearly 600 licensed beds. In 2018, it received almost 55,000 visits to the ED, admitted nearly 41,000 inpatients and performed 26,500 surgeries (c.50% ambulatory). We exclude from our analysis Psychiatry and Obstetrics units because they manage admissions separately from the rest of the hospital.

In this paper, a *service* refers to a medical specialty such as internal medicine, oncology, cardiology, or surgery. We refer to a *unit*, or equivalently a *ward* or a *floor*, as a physical space composed of rooms and beds. Our partner hospital, for instance, has 15 services spread over 31 units, divided

Table 1 Description of inpatients unit on campus A at our partner hospital. The type is either GC (general care) or IC (intensive care). The unit name follows the nomenclature: Campus-Type Number.

Name	Type	# Licensed beds	# Private rooms	Primary (secondary) services
A-GC 1	GC	36	8	Oncology (Internal Medicine)
A-GC 2	GC	43	8	Internal Medicine (Oncology)
A-GC 3	GC	24	8	Surgery (Orthopedics, Plastic Surgery)
A-GC 4	GC	28	18	Oncology
A-GC 5	GC	20	12	Orthopedics (Internal Medicine)
A-GC 6	GC	23	8	Internal Medicine
A-IC 1	IC	12	12	Surgery (Internal Medicine, Orthopedics)
	Total	186	74	

across two campuses. Table 1 lists some of these units, alongside their main characteristics. The full list of units is given in the electronic companion (Table EC.1).

1.2. Approach and contributions

Our approach can be summarized as follows:

1. **Tractable optimization model:** We develop a tractable discrete optimization model to find short-term bed allocations that satisfy patient-specific requirements, while accounting for future bed requests and availability for all inpatient units in the hospital in the coming hours and days. While no study is perfectly exhaustive, we believe our approach to this multifaceted issue is one of the most comprehensive of its kind. On this aspect, our model lies between patient-level models for bed assignment (Thompson et al. 2009, Thomas et al. 2013, Schäfer et al. 2019) that are often myopic and face much greater challenges to scale up to large institutions, and unit-level stochastic models (Kilinc et al. 2018, Dai and Shi 2019) that can inform the design of tactical admission rules but fail to capture individual patient requirements in the short term (e.g., gender, isolation status).

2. **Prediction-based robust optimization:** To account for uncertainty in future patient flows, we adopt a robust optimization approach, where the description of the uncertainty is based on machine learning predictions, instead of parametric assumptions employed in many queueing settings. Our robust description of uncertainty connects with the cautious attitude prevailing in healthcare management. Accordingly, we provide theoretical bounds on the probability of "disappointment", i.e., the probability that the decision maker will incur a higher cost than what the model estimated. Our result relies on new large deviation bounds for the sum of correlated binomial random variables that could be of independent interest. To mitigate the conservatism of our approach, we consider affinely adaptive rules, which, to the best of our knowledge, have not been studied for multistage optimization problems with discrete variables and uncertainty, let alone in a healthcare context.

3. **Numerical validation:** Finally, on extensive simulations, we verify that our proposed formulation is tractable and estimate its potential operational benefit. On data from a 600-bed medical

center over seven months, we solve the robust optimization problems in seconds and provide a bed assignment policy that reduces off-service placement by 24% on average, boarding delays in the ED and post-anesthesia units by 35% and 18% respectively, while keeping overall occupancy constant. We also illustrate the additional benefit from using linear decision rules, which partially adapt future decisions to the future realizations of patient arrivals/discharges. Because this adaptive model describes future decisions with this additional (and realistic) flexibility, it yields first-stage decisions that are more forward-looking, hence allowing for a more effective trade-off between waiting time and off-service placement.

Structure: Section 2 presents the relevant literature and positions our contribution. In Section 3, we describe our modeling strategy and assumptions, and derive a first version of the Hospital-wide Inpatient Flow Optimization (HIFO) problem. To account for uncertainty in future flows, we propose a prediction-based robust optimization approach, derive a theoretical connection with a risk-averse stochastic objective, and explore the use of affine decision rules in Section 4. Finally, we assess the performance of our approach and the benefit from adaptive robust optimization on numerical experiments in Section 5.

Notations: We reserve bold characters for vectors and uppercase characters for random variables. Hence x, $\boldsymbol{x}$, X, and $\boldsymbol{X}$ denote a scalar, a vector, a random variable, and a random vector respectively. For any integer n, we define $[n] := 1, \ldots, n$.

2. Literature review

We present relevant aspects of patient flow management which have received increased attention recently and provide motivation for our work.

2.1. Patient flow analysis and modeling

Armony et al. (2015) conducted one of the first data-based analyses of patient flows at a hospital level, emphasizing interactions between the emergency department (ED) and a subset of five inpatient wards. A central performance metric in their analysis is delays. Indeed, delays can be used as a measure of operational efficiency as well as quality of care. For instance, prolonged ED boarding time –the time needed for an ED patient to be admitted to an inpatient bed– is associated with negative health outcomes (Mathews et al. 2018, Chan et al. 2016b), and is usually due to lack of available inpatient beds (Shi et al. 2016). Consequently, better understanding and modeling of discharge patterns are needed. Dai and Shi (2021) analyze and compare two service time models which have been recently proposed to capture non-stationarity in patient discharges: the two-timescale model of Shi et al. (2016) and an inspection-delay service time model (Chan et al. 2016a, Dong and Perry 2020). Beyond the ED, Johnson et al. (2013), Long and Mathews (2018), Oliveira et al. (2018) empirically measure the negative consequences of prolonged Intensive Care Unit (ICU) boarding.

A general insight of queueing theory is that resource pooling might produce better performance. Due to heterogenity in patient needs however, Song et al. (2015) empirically find that pooling in the ED increases waiting time and overall length-of-stay. In an inpatient context, pooling resources leads to patient misplacement, also called off-service placement or patient overflow. Off-service placement occurs when an incoming patient is placed in a different service than the one required by their condition. Empirical studies have investigated the extent to which off-service placement increases length of stay and readmission risk (Alameda and Suárez 2009, Stowell et al. 2013, Liu et al. 2014, Stretch et al. 2018, Bai et al. 2018, Song et al. 2020). A related phenomenon is off-level placement, i.e., when a patient needing an ICU is placed in a general care unit. Empirical evidence suggests that off-level placement is also detrimental for the patient (Kim et al. 2015, Chan et al. 2018).

These empirical findings prominently demonstrate how operations impact on quality of care and have motivated many efforts, including ours, to improve healthcare operations.

2.2. Patient flow optimization

Here, we highlight the main angles through which patient flows can be and have been improved.

Bed capacity planning Kao and Tung (1981), De Bruin et al. (2010), Boulton et al. (2016) apply queueing theory tools to optimize the number of beds allocated to each ward. Kao and Tung (1981) develop a two-scale model where baseline capacity is designed to satisfy yearly overall demand, while monthly adjustments are considered to match month-in-month variability. Pinker and Tezcan (2013) extend the analysis and differentiate between shared and private rooms. Izady and Mohamed (2019) address the issue of off-service placement by dividing beds into two categories, specialty-specific and specialty-agnostic, while defining optimal functional groups of beds or clusters. Ward dimensioning decisions are revised on a yearly basis at best. In this work, we take the ward capacity as an input to the patient flow optimization problem, hence providing complementary benefits.

Scheduled admission optimization Scheduled admissions are responsible for most of the weekly variability in inpatient bed census (Helm and Van Oyen 2014). At a strategic level, Carnes et al. (2011) formulate the allocation of surgical blocks to surgeon as a deterministic integer optimization problem, so as to systematically reduce peak surgical census on Wednesdays. Bavafa et al. (2019) construct an optimal portfolio of elective procedures to maximize profit, while taking into account the utilization of surgical and recovery beds. At a tactical level, Bekker and Koeleman (2011), Helm and Van Oyen (2014) and Meng et al. (2015) optimize the scheduled admissions planning for the upcoming weeks, and incorporate uncertain daily deviations due to ED patients and evolving patient situations. To do so, Helm and Van Oyen (2014) propose a stochastic location process model for resource needs of elective and unplanned patients, and estimate the number of untreated patients using an Erlang loss model approximation. Meng et al. (2015) adopt a distributionally robust optimization approach to minimize bed shortfall directly. In a complementary direction, we consider the

operational problem of dynamically admitting new patients (both scheduled and emergent), while accounting for future scheduled and unscheduled flows. In terms of modeling, per Meng et al. (2015), we use a robust approach to capture uncertainty. Additionally, we study the impact of adaptive (namely affine) rules.

Admissions of emergent patients For ED admissions, there is a trade-off between waiting in the ED for the right bed to become available and immediately placing the patient in another service. Kilinc et al. (2018) explore this trade-off using a queueing framework. In a stylized two-ward setting, they prove that a threshold-based policy is optimal and propose a heuristic that reduces cost by 14% and boarding time by 9% on simulations. However, their experiment only includes patients who were admitted to the ED for chest pain or congestive heart failure, and is restricted to two inpatient units. In addition, the edge of their policy shrinks as the number of inpatient beds increases. Dai and Shi (2019) similarly formulate the overflow decision problem as a Markov decision process (MDP) and propose an approximate dynamic programming algorithm which reduces the overflow proportion by 20% on a simulated hospital with five services. However, as they acknowledge, their model does not consider all units and cannot account for inter-unit transfers of inpatients. Recently, Zhang et al. (2020) adopted a simulation-based approach to evaluate different overflow strategies, that is predefined guidelines or protocol, to divert patients from their primary service or unit to improve overall throughput. The scope of our study is wider: We do not restrict our analysis to isolated units or a sub-network of units, but instead consider admissions to all inpatient units, from different sources (ED, scheduled admissions, other hospitals), together with inter-unit transfers. Our paper also differs in methodology and intent. While Kilinc et al. (2018), Dai and Shi (2019) use stochastic queueing models and their fluid approximation to obtain structural properties on the optimal admission rule, we provide patient-level feasible placement recommendations using a discrete description of the decisions and the uncertainties.

Operational patient-bed assignment Thomas et al. (2013) automate the process of finding a feasible bed assignment for each ED patient to reduce bed assignment waiting time. Using mixed-integer optimization, they account for room type, gender or staff-to-patient ratio requirements, and obtain a 23% reduction in ED request-to-bed-assignment times. However, their formulation is myopic and assigns patients to beds without forecasting the future state of the hospital. This limitation is partially addressed by Schäfer et al. (2019), who include future bed requests, either scheduled or emergent. However, their resulting integer optimization formulation requires more than 12 hours to solve instances with 96 beds and ≥ 3 time steps, and needs to adopt a greedy look-ahead heuristic. On historical data from a German hospital, they achieve a 96% reduction in overflow and a 5% increase in utilization. Besides the fact that such improvement benefits might be unattainable for highly congested hospitals where utilization is already above 95%, their approach is

also not protected against prediction errors: future admissions are assumed equal to their predictions. Thompson et al. (2009) formulate an MDP problem to allocate heterogeneous patients to inpatient beds dynamically and have successfully implemented their solution in a 130-bed community hospital over an 18-day trial (+1% in revenues, −50% in boarding time). Even for a medium-size institution, they require many approximations to make their approach scalable: multi-dimensional expectations are approximated via sample average approximation, the policy space is reduced to a finite space with at most 48 policies, and the optimization horizon is reduced to two periods.

3. Problem description and model

In this section, we introduce our modeling framework for inpatient flow optimization, emphasizing the key modeling ingredients, assumptions, and flexibility. We describe the overall problem, its three competing horizons, and our aggregation strategy in Section 3.1, and then derive a mathematical formulation for each horizon in the subsequent sections.

3.1. A multi-horizon problem

When optimizing inpatient flows, hospital managers have to balance different objectives and time horizons. Throughout the day, they constantly decide when to admit patients and to which unit/bed. To do so, they should take into account the current state of the hospital (e.g., current demand and availability of beds) and also consider its evolution in the near future. For instance, it might be worth waiting for a bed in the right service to become available before admitting a patient. Alternatively, some beds should be kept empty in the morning in anticipation of future admissions later in the day. Such waiting and idling strategies are only justified in a multistage context. We formulate inpatient flow management as a multistage optimization problem with three different timescales: immediate, daily, and weekly horizon. A central aspect in our formulation is that we adjust the level of granularity of the decision variables to the time horizon: the immediate problem considers individual patient assignments, the daily timescale optimizes unit-level patient flow, and the weekly objective controls overall hospital occupancy. Table 2 summarizes our modeling strategy.

Our model balances practical relevance and numerical tractability. In the short term, patient-specific information is needed to find a feasible bed assignment. For instance, shared rooms should be same-gender; and patients with a viral infection require a private room. Regarding future flows, we consider aggregated statements on overall discharge and arrival volumes for each unit. In this regard, our approach can be seen as a compromise between patient-level models for bed assignment (Thompson et al. 2009, Thomas et al. 2013, Schäfer et al. 2019) that are often myopic and scale poorly in the size of the institution, and unit-level stochastic models (Kilinc et al. 2018, Dai and Shi 2019) that can inform the design of tactical admission rules but fail to capture individual patient requirements in the short term.

Table 2 The three optimization horizons and the characteristics of their respective optimization formulation. EoD stands for End of Day.

Horizon	Immediate	Daily	Weekly
Multistage strategy	✗	Folding horizon	Rolling horizon
Period length	2 hours	2 hours	1 day
# Periods	1	Between 1 and 12 (EoD)	7
Decision level	Patient	Unit	Hospital
Decision variable	Patient assignment z^1_{ij}	Patient flow $f^t_{jj'}$	Remaining capacity C_r

3.2. Immediate problem: individual patient placement

First, we consider current bed availability and requests - either by new patients or current inpatients needing a new bed - and seek to assign patients to beds at minimal cost. Formally, let I and J denote the number of patients and units respectively. For each patient $i \in [I]$, their current location is encoded through binary variables $z^0_{ij} \in \{0,1\}$, $j \in [J]$, where $z^0_{ij} = 1$, if patient i is in unit j, 0 otherwise. Similarly, we define a decision variable $\boldsymbol{z}^1 \in \{0,1\}^{I \times J}$ to describe the location of each patient at time $t = 1$. The objective is to find a patient-unit allocation that minimizes cost, i.e., solve some optimization problem of the form

$$\min_{\boldsymbol{z}^1 \in \mathcal{Z}} \sum_{i,j} c_{ij} z^1_{ij}.$$

In the optimization problem above, the set $\mathcal{Z}$ captures physical and operational constraints on $\boldsymbol{z}^1$, such as the fact that each patient can only be in one location, or that each unit has a fixed capacity.

The cost c_{ij} measures the disutility experienced from assigning patient i to unit j and depends on the patients' current location and need. Among others, c_{ij} captures both off-level and off-service placement. We can also incorporate physical distance and ambulance costs, or how long the patient has been waiting for a bed. Section EC.2.1 details how c_{ij}'s can be computed based on the organization of the hospital. Note that, for a large academic hospital like our partner, $J \approx 30$ units and $I \approx 600$ patients, so this problem involves $I \times J \approx 18,000$ binary variables and $I + J \approx 630$ constraints.

3.3. Daily problem: optimal patient flows

To incorporate near-term bed requests and availability, we extend the myopic single-stage problem described in the previous section with a multistage component. We consider two-hour time periods and optimize future patient flows until 6am. We later refer to 6am as "end of the day" (EoD) because it corresponds to a low activity point and a transition time between night and day shifts. Effectively, our formulation is a folding or receding horizon formulation: At 6:01am, we optimize over $T = 12$ time periods until 6am the next day. At 8:01am, we optimize over $T = 10$ time periods, and so on. The rationale behind a folding horizon is that bed requests cannot be left pending

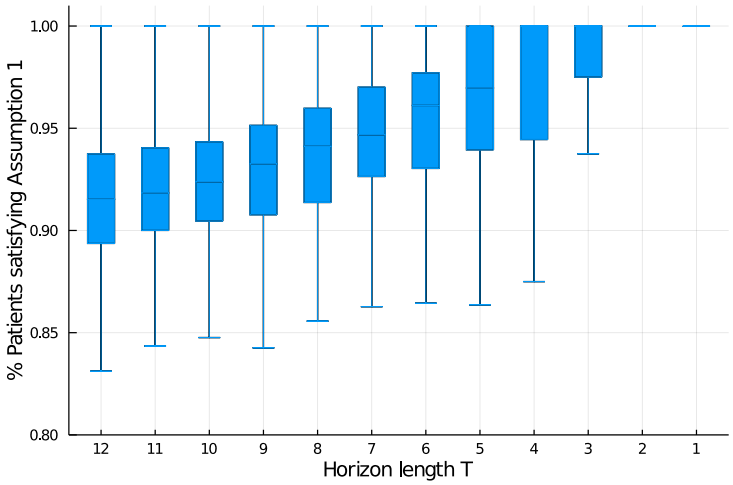


Figure 2 **Historical proportion of patients satisfying the single-move assumption (Assumption 1) among all patients that moved at least once, as a function of the horizon length.**

indefinitely, especially for surgical and emergent patients, and that a bed assignment will have to be made at some point. Using a folding horizon approach prevents postponing some bed assignment decisions indefinitely. Furthermore, we chose to divide the horizon into two-hour blocks because, if implemented in practice, we would need to leave enough time between two runs of the algorithm for the care teams to implement the assignment decisions.

As previously mentioned, the main difference between the immediate and daily problems is the level of granularity for the decision variables. For the multistage problem, we consider aggregated patient flows: let $f^t_{j,j'}$ indicate the number of patients who stayed in unit $j \in [J]$ until $t-1$ and moved from unit j to unit $j' \in [J]$ during period $(t-1, t]$. In other words, we do not allow patients to change units more than once until the end of the day, as stated in Assumption 1.

Assumption 1. *Each patient can move at most once between $t=0$ and the end of the day.*

This assumption is validated empirically as shown in Figure 2. At 6:01am, the vast majority ($>85\%$) of patients who will move at least once in the day will move only once. Naturally, throughout the day, the horizon shrinks and this proportion increases to 100%. The main reason for introducing Assumption 1 is to keep track of individual trajectories despite aggregation. Indeed, under this assumption, we can divide patients' flows depending on the unit they originated from, and connect aggregated flows to the individual patients involved. For instance, we have $f^1_{j,j'} = \sum_i z^1_{ij'} z^0_{ij}$. Similar to $\boldsymbol{z}^1$, $\boldsymbol{f}$ are integer-valued and satisfy flow-conservation, forbidden moves, and capacity constraints. The latter can be expressed as:

$$z^0_j + \sum_{\tau\in[t]} \sum_{j'} f^\tau_{j',j} - \sum_{\tau\in[t]} \sum_{j'} f^\tau_{j,j'} \le C^t_j, \ \forall j \in [J], \ t \in [T], \tag{1}$$

where C_j^t is the floor capacity at time t. Note that capacity might not be constant over time and might even be a decision variable (see Section EC.2.5).

To measure the quality of a patient flow routing $\boldsymbol{f}$, we compare it with what patients need from a medical standpoint. Let us consider $g_{j,j'}^t$, the number of patients who were in j until $t-1$ and needed to be moved from unit j to j' during $(t-1,t]$. In other words, $g_{j,j'}^t$ corresponds to the total demand for unit j' at time t from patients currently in unit j, while $f_{j,j'}^t$ corresponds to the satisfied demand. For the remainder of this section, we present our optimization formulation assuming that the $g_{j,j'}^t$ are known. In the next section, however, we present a methodology to estimate them from data under Assumption 1 and incorporate these predictions into the optimization problem.

A natural objective is to bring $\boldsymbol{f}$ as close as possible to $\boldsymbol{g}$. For example, the number of patients who requested to be moved from unit j to j' but did not move can be expressed as

$$\left(\sum_{t\in[T]} g_{j,j'}^t - \sum_{t\in[T]} f_{j,j'}^t \right)^+,$$

where $(x)^+ := \max(x,0)$ denotes the positive part of x. Using the same logic, we construct a cost function $c(\boldsymbol{f},\boldsymbol{g})$ to measure the difference between where patients need to be moved to and where they are effectively moved to. Section EC.2.4 details the exact cost formula we used for our numerical experiments. The key aspects are that $c(\boldsymbol{f},\boldsymbol{g})$ is piecewise linear, convex in $(\boldsymbol{f},\boldsymbol{g})$, and only penalizes off-level placement. All in all, the daily problem can be written as

$$\begin{aligned}
\min_{\boldsymbol{f}\in\mathcal{F}} \quad & c(\boldsymbol{f},\boldsymbol{g}) \\
\text{s.t. } & z_j^0 + \sum_{\tau\in[t]}\sum_{j'} f_{j',j}^\tau - \sum_{\tau\in[t]}\sum_{j'} f_{j,j'}^\tau \le C_j^t,\ \forall j\in[J], t\in[T], && (1)\\
& f_{j,DIS}^t = g_{j,DIS}^t,\ \forall j\ \in GC\cup IC,\ \forall t\in[T] && (2)
\end{aligned}$$

where $\mathcal{F}$ denotes the set of feasible flows $\boldsymbol{f}$ and the last constraints enforce that the number of patients discharged from unit j at time t is equal to the number of patients who need to be discharged from j at time t.

3.4. Weekly problem: Target hospital occupancy level

Due to its finite horizon nature, the previous optimization formulation might lead to undesirable end-game strategies and fill the hospital to capacity at the end of the day. To mitigate this effect, we add a long-term objective that accounts for bed needs in the coming week (seven days). Let C_r denote the number of beds available at the end of the day. C_r is a decision variable that depends on the previous decisions through the following constraint:

$$C_r \le \sum_{j\in[J]} \left(C_j^T - z_j^0 - \sum_{t\in[T]}\sum_{j'} f_{j',j}^t + \sum_{t\in[T]}\sum_{j'} f_{j,j'}^t \right).$$

We compute a target number of available beds for the next week C_w and minimize shortage with respect to this target: $\min_{C_r \geq 0} (C_w - C_r)^+$.

Although tactical planning of scheduled admission can help reduce intra-week admission volume (Carnes et al. 2011, Helm and Van Oyen 2014, Meng et al. 2015, Bavafa et al. 2019), scheduled admissions are not uniformly distributed over the week, with a peak weekly census around wednesdays. Similarly, visits to the ED and discharge volumes exhibit strong weekly seasonality patterns. Accordingly, C_w should vary depending on the day of the week and the known scheduled admissions. For each of the seven following days, we estimate the total number of admissions (scheduled and unscheduled) and discharges for the day. As a first approximation, these estimates could be taken as day-of-the-week averages. Technology permitting, more sophisticated predictions using weather forecasts or inpatient length-of-stay models (Bertsimas et al. 2021) could be used as well. The difference between expected admissions and discharge provide an estimate of the (potentially negative) number of beds needed on that day. We set C_w equal to the actualized value of these future needs, for some discount factor β: $C_w = \sum_{d\in[7]} \beta^d (Admissions_d - Discharges_d)$.

3.5. Final formulation: Hospital-wide Inpatient Flow Optimization (HIFO)

Finally, combining the three optimization horizons, we obtain the HIFO formulation:

$$\begin{aligned}
\min_{\boldsymbol{z}^1\in\mathcal{Z},\boldsymbol{f}\in\mathcal{F},C_r} \; & \sum_{i,j} c_{ij} z^1_{ij} \;+\; \lambda\; c(\boldsymbol{f},\boldsymbol{g}) \;+ \lambda_w \,(C_w - C_r)^+ \\
\text{s.t.} \quad & z^0_j + \sum_{\tau\in[t]} \sum_{j'} f^\tau_{j',j} - \sum_{\tau\in[t]} \sum_{j'} f^\tau_{j,j'} \leq C^t_j,\; \forall j,t, && (1)\\
& f^t_{j,DIS} = g^t_{j,DIS},\; \forall j,t, && (2)\\
& f^1_{j,j'} = \sum_{i\in[I]} z^1_{ij'} z^0_{ij},\; \forall j \neq j', && (3)\\
& C_r \leq \sum_{j\in[J]} \left(C^T_j - z^0_j - \sum_{t\in[T]} \sum_{j'} f^t_{j',j} + \sum_{t\in[T]} \sum_{j'} f^t_{j,j'} \right), && (4)
\end{aligned}$$

where $\lambda, \lambda_w > 0$ controls the degree of foresight. The coupling constraints (3)-(4) ensure that immediate, daily, and weekly variables agree. Among other extensions, our formulation can provide individual trajectories, z^t_{ij} for $t \in [T]$ for some patients as well (Section EC.2.2) and include additional variables to model surge capacity decisions (EC.2.5).

3.6. Limitations

From a modeling standpoint, the main assumption our model relies on is Assumption 1. Without this assumption, one would require a fixed cost for moving a patient from unit j to unit j' to avoid returning nonsensical solutions where patients are moved at each time period. Assumption 1 replaces such cost by a hard constraint. As previously discussed, it is a simplification because

patients may visit multiple units within a day. However, we alleviate this issue by implementing our policy in a close-loop fashion: At the beginning of each time period, the optimization model HIFO is solved. Patients who have requested a bed are admitted as prescribed by $\boldsymbol{z}^1$. At the next time period, the optimization model is resolved using the updated state of the hospital (i.e., data on bed requests, scheduled admissions and procedures, and bed availability is updated), and new decision variables $\boldsymbol{z}^1$ are computed and implemented. The longer-term decision variables $\boldsymbol{f}$ and C_r are never implemented but we include them in the optimization so that the here-and-now decisions are forward-looking and anticipate future flows. This implementation, which is referred to as 'close-loop' in the control literature, creates feedback and improves the adjustability and the sensitivity of the system to parameter errors, which will be crucial as we introduce uncertainty on future demand (also referred to as medical trajectories) $\boldsymbol{g}$. We refer to Dorf and Bishop (2008, chapter 3) for an introduction to control and feedback loops, and Ben-Tal et al. (2009, chapter 14.4) for a robust perspective on controllers. In addition, we need a close-loop implementation in practice because only the first-stage decisions $\boldsymbol{z}^1$ are patient-level, hence directly actionable for the hospital. Previous work has developed clustering techniques to predict more complex patient trajectories (Xu et al. 2016, Ranjan et al. 2017). Future work could investigate how to relax Assumption 1 in a tractable fashion to incorporate more complex patient dynamics and flows, especially for the weekly planning horizon.

A second limitation stems from the fact that our HIFO formulation requires the specification of costs associated with each decision. Some of these costs can be estimated from actual hospital expenditures or reimbursement costs, but most of them are implicit. For example, empirical evidence suggests that off-service placement can be detrimental to the patient, in terms of length of stay or readmission risk. Accordingly, we penalize off-service placement in our cost formula (c.f. EC.2.1), but precisely calibrating a penalty value remains an open challenge, which we solved through discussion with our partner hospital and iterative improvement of our model. Alternatively, Zhalechian et al. (2020) focus on a single patient outcome (e.g., hospital readmission) and propose an online learning approach to jointly learn and optimize for this particular objective. However, their approach applies to simple admission policies and for a single outcome that is directly observable (within reasonable time after the decision is made).

In our model, we do not incorporate the potential impact of the admission decisions recommended on the flow dynamics, i.e., the validity of the model itself. In our case, this impact is limited by the fact that, as discussed in our simulations in Section 5, a large fraction of the patient flows remain unchanged by the optimization. However, we must acknowledge that switching to an optimization-based approach might change the flow dynamics within the hospital and require some re-training

of the predictive models we will present in the next section, in addition to other sources of non-stationarity. Future work could investigate endogeneizing the impact of the decisions on the future flow dynamics directly.

4. Uncertainty on medical trajectories

In practice, the medical trajectories $\boldsymbol{g}$ are uncertain. To make appropriate decisions under uncertainty, two aspects ought to be taken into account: the degree to which the uncertainty can be anticipated (predictability) and the inherent noise in any estimate (variability). To do so, we propose to combine machine learning predictions with a robust optimization approach. Namely, let $\boldsymbol{G}$ denote the integer-value random vector, indexed by $[J]\times[J]\times[T]$, indicating future demand for each unit. In Section 4.1, we develop a collection of predictive models to estimate $\mathbb{E}[\boldsymbol{G}]=\hat{\boldsymbol{g}}$. Then, to account for prediction errors, we construct an uncertainty set $\mathcal{U}$ around the point-wise predictions $\hat{\boldsymbol{g}}$, as described in Section 4.2. Finally, we solve a robust version of HIFO . Specifically, let us denote $\mathcal{R}$ the set of recourse policies $\boldsymbol{f}:\mathcal{U}\to\mathcal{F}$ that are non-anticipative, i.e., $f^t_{j,j'}$ should only depend on $\tilde{g}^\tau_{u,u'}$, $u,u'\in[J]$, with $\tau\le t$. We consider the following problem:

$$
\begin{aligned}
\min_{\boldsymbol{z}^1\in\mathcal{Z},\boldsymbol{f}(\cdot)\in\mathcal{R},C_r(\cdot)} \quad & \boldsymbol{c}^\top\boldsymbol{z}^1 + \max_{\tilde{\boldsymbol{g}}\in\mathcal{U}} \lambda\, c(\boldsymbol{f}(\tilde{\boldsymbol{g}}),\tilde{\boldsymbol{g}}) + \lambda_w\left(C_w - C_r(\tilde{\boldsymbol{g}})\right)^+ \\
\text{s.t.} \quad & z^0_j + \sum_{\tau\in[t]}\sum_{j'} f^\tau_{j',j}(\tilde{\boldsymbol{g}}) - \sum_{\tau\in[t]}\sum_{j'} f^\tau_{j,j'}(\tilde{\boldsymbol{g}}) \le C^t_j,\ \forall j,t,\forall\tilde{\boldsymbol{g}}\in\mathcal{U},\\
& f^t_{j,DIS}(\tilde{\boldsymbol{g}}) = \tilde{g}^t_{j,DIS},\ \forall j,t,\forall\tilde{\boldsymbol{g}}\in\mathcal{U},\\
& f^1_{j,j'}(\tilde{\boldsymbol{g}}) = \sum_{i\in[I]} z^1_{ij'}z^0_{ij},\ \forall j\neq j',\forall\tilde{\boldsymbol{g}}\in\mathcal{U},\\
& C_r(\tilde{\boldsymbol{g}}) \le \sum_{j\in[J]}\left(C^T_j - z^0_j - \sum_{t\in[T]}\sum_{j'} f^t_{j',j}(\tilde{\boldsymbol{g}}) + \sum_{t\in[T]}\sum_{j'} f^t_{j,j'}(\tilde{\boldsymbol{g}})\right),\ \forall\tilde{\boldsymbol{g}}\in\mathcal{U},\\
& \sum_{t,j'} f^t_{j,j'}(\tilde{\boldsymbol{g}}) = z^0_j,\ \forall\tilde{\boldsymbol{g}}\in\mathcal{U}.
\end{aligned}
\tag{5}
$$

We discuss the benefits of a robust approach to a stochastic description of the uncertainty in Section 4.3 and theoretically connect the objective of our robust approach (i.e., worst-case cost) to a stochastic notion of value-at-risk. Section 4.4 presents our model with static decision rules, and explores the use of affine decision rules to alleviate conservatism.

4.1. Predictability: Machine learning to the rescue

We first use historical data and machine learning techniques to construct predictive models for key aggregate demand quantities. In accordance with the single-move assumption (Assumption 1) and our modeling framework, we consider the patients based on their unit of origin j and only predict the next unit they will request to be admitted to. We now review the prediction strategies we adopted

Table 3 Summary of out-of-sample performance for all prediction tasks, on their respective test set. We use optimal classification trees (OCT) (Bertsimas and Dunn 2017) for classification tasks and regularized regression (Lasso) (Tibshirani 1996) for regression tasks.

Patient category	Prediction task	Method	Metric	Value
Inpatients	Probability of discharge	OCT	AUC	0.810
	Daily discharges	OCT	Median relative error	6.0%
			R^2	0.847
	Probability of intensive care	OCT	AUC	0.973
	ICU census	OCT	Median relative error	11.1%
			R^2	0.998
Unscheduled visits	Bed requests	Lasso	Median absolute error	3.67
			Median relative error	14.0%
			R^2	0.910
Transfers	Bed requests	Lasso	Median absolute error	1.19
			Median relative error	58.1%
			R^2	0.805

for each unit of origin. Table 3 summarizes the main out-of-sample predictive power of our models. Details about the data, models, and training procedure are provided in EC.3.

For inpatient units, we separate the demand for beds in unit j' at time t, as the aggregation of individual patient needs. Mathematically, we decompose $\tilde{g}^t_{j,j'}$ into

$$\tilde{g}^t_{j,j'} = \sum_{i \text{ in unit } j \text{ at } t=0} \tilde{y}^t_{i,j'},$$

where $\tilde{y}^t_{i,j'} \in \{0,1\}$ indicates whether patient i will request a bed in unit j' at at time t. Following the approach from Bertsimas et al. (2021), we build individual risk scores to predict the quantities $\tilde{y}^t_{i,j'}$ using the rich information available from their electronic health records (EHRs). In particular, on a daily basis (each day at 6am) and for each inpatient, we predict the probability to be discharged by the end of the day and the probability to be in an ICU, i.e., the expected value of

$$\sum_{t\in[T]} \tilde{y}^t_{i,DIS} \quad \text{and} \quad \sum_{t\in[T], j' \text{ is an ICU}} \tilde{y}^t_{i,j'}.$$

By properly summing over inpatients i, we can estimate the expected number of discharges and ICU requests from any inpatient unit j as

$$\sum_{t\in[T]} \tilde{g}^t_{j,DIS} \quad \text{and} \quad \sum_{t\in[T], j' \text{ is an ICU}} \tilde{y}^t_{j,j'}.$$

Observe that, because we aggregate patient-level risk scores, our unit-level predictions take into account the patient mix of the unit in a very rich manner. A central operational challenge in practice is that these estimates cannot easily be updated throughout a day. Because they rely on patient-level

information available in the EHR system, data extraction and processing for all inpatients requires time and is performed on a fixed schedule (e.g., once a day). For instance, at 6am, we estimate the probability of discharge for each patient, $\mathbb{P}(\text{discharge for patient } i, \text{ as of 6am})$, which corresponds to the probability that patient i will be discharged between 6am and EoD. However, as the day goes by, we optimize patient flows from "now" $(t=0)$ onwards. So, if it is h o'clock at the beginning of the time period $(t=0)$, we are interested in

$$\mathbb{P}(\text{discharge for patient } i, \text{ as of } t=0) = \mathbb{P}(\text{discharge for patient } i, \text{ as of } h \text{ o'clock}).$$

However, there is often little new information available to update these probabilities compared with our initial estimate, $\mathbb{P}(\text{discharge for patient } i, \text{ as of 6am})$. We then simply estimate the updated probability by $\alpha_h \mathbb{P}(\text{discharge for patient } i, \text{ as of 6am})$, where the scaling factor α_h is calibrated empirically on the training set. Intuitively, $\alpha_h \mathbb{P}(\text{discharge for patient } i, \text{ as of 6am})$ is the updated probability for patient i to be discharged conditioned on the fact that they are still at the hospital at h o'clock. Typically, α_h is non-increasing, equal to 1 before 8am, and close to 0 after 10pm.

For scheduled admissions, we consider the latest schedule available at the beginning of each time period and assume that there is no uncertainty. Hence, we do not explicitly model the risk of cancellation in this version of the model. However, as soon as an admission is cancelled, it disappears from the schedule. In other words, our model adapts to the state of the schedule and cancellations. For surgeries, the schedule indicates the type of surgery performed, not the specific unit the patient should be admitted to afterwards. We automatically learn the mapping between the type of surgery and the required unit after surgery by training a decision tree model on historical data.

Unscheduled visits and transfers concern patients who, by definition, are not physically present in the hospital and have not requested a bed yet, so no (or little) patient-level information is available to predict future demand from these units. As a result, we build simple linear models using date/time-related variables such as month number, day of the week, time of the day, weekend or holiday indicator, and number of requests in previous time periods.

Note that, in our close-loop implementation, we resolve the robust optimization problem (5) at every time step (every two hours in our case). Accordingly, between two resolves, we not only update the status of the hospital for the immediate patient placement problem (e.g., number of patients in each unit, number and information about the pending bed requests) but also our predictions about future demand $\hat{\boldsymbol{g}}$. In other words, we update the state of the hospital, which includes beliefs about future arrivals and departures.

4.2. Variability: robust optimization to the rescue

Despite their accuracy, our predictive models cannot overcome the inherent variability in medical trajectories, which we account for by adopting a robust optimization approach. Based on the predictions we computed in the previous section, $\hat{\boldsymbol{g}}$, we construct a so-called uncertainty set $\mathcal{U}$, and restrict our attention to uncertain flows within $\mathcal{U}$, i.e., $\tilde{\boldsymbol{g}} \in \mathcal{U}$.

First, we impose structural constraints, similar to the ones satisfied by the physical flows $\boldsymbol{f}$: The variables $g^t_{j,j'}$ are integer; At $t=0$, $g^0_{j,j'}$ equals the number of patients currently in unit j and who need to go to unit j'; Flow conservation constraints, i.e., $\sum_{t,j'} g^t_{j,j'} \geq z^0_j$ for $j \in \{SA, UV, T\}$ (to allow for future arrivals), while $\sum_{t,j'} g^t_{j,j'} = z^0_j$ if j is an inpatient unit. Note that the last set of constraints impose that the future demand also satisfy Assumption 1. Constraints $\sum_{t,j'} g^t_{j,j'} \leq M z^0_j$ would translate the fact that each patient from unit j can generate no more than M requests on average.

We then include constraints based on the outputs of our machine learning models and bound the overall volume and intra-day distribution of future bed requests. For instance, for hospital discharges we impose

$$\sum_{j\in[J],t\in[T]} g^t_{j,DIS} \text{ close to } \sum_{j\in[J],t\in[T]} \hat{g}^t_{j,DIS}, \qquad \text{(overall volume)}$$

$$\sum_{j\in[J]} g^t_{j,DIS} \Big/ \sum_{j\in[J],t\in[T]} g^t_{j,DIS} \text{ close to } \hat{\beta}_t, \qquad \text{(intra-day distribution)}$$

where $\hat{\beta}_t$ corresponds to the average fraction of discharges which empirically occurred during $[t, t+1)$ (conditioned on the day of the week and hour of the day). This decomposition scheme into discharge volume and intra-day distribution resembles the patient-level two-timescale model of Shi et al. (2016), who decompose overall length of stay into days and hours. Similarly, we impose constraints on the overall volume of discharges from each inpatient unit j, on the overall number of ICU bed requests (from the entire hospital and each unit separately), and on the number of bed requests from UV or T. We impose additional constraints on the intra-day distribution of bed requests, and on the mix of requested units from UV/T. Namely, for $j \in \{UV, T\}$, we bound the ratios

$$\sum_{j'\in[J]} g^t_{j,j'} \Big/ \sum_{j'\in[J],t\in[T]} g^t_{j,j'} \quad \text{and} \quad \sum_{t\in[T]} g^t_{j,j'} \Big/ \sum_{j'\in[J],t\in[T]} g^t_{j,j'}.$$

However, we do not control the intra-day distribution of ICU bed requests, which are mostly driven by bed placement decisions, i.e., our decision variables. As previously mentioned, for scheduled admissions, we impose future requests to be equal to their expected values: $g^t_{SA,j'} = \hat{g}^t_{SA,j'}$.

Note that a robust approach of uncertainty allows for an easy control on overall demand levels directly $\tilde{g}^t_{j,j'}$ and even aggregates of them over units or time. In contrast, in a stochastic or distributionally robust approach, demands would need to be modeled as correlated random variables with the random variable $\boldsymbol{G}$ satisfying all these constraints above almost surely.

4.3. Discussion: Modeling uncertainty

From a modeling perspective, we would like to comment on the benefits of using a robust description of the uncertainty compared with stochastic models.

First, specifying distributions for inpatient flows is a complicated and largely open problem. Kim and Whitt (2014) discuss challenges and solutions to fit a non-homogeneous Poisson process for arrivals to the ED. By extension, it would be reasonable to consider independent Poisson random variables $G^t_{j,j'}$ with respective parameter $\hat{g}^t_{j,j'}$ to model the demands from $j \in \{UV, T\}$. For an inpatient unit j, however, Dai and Shi (2021) highlight the deficiencies of standard queueing models to capture the complex dynamics of patient flows. The most challenging aspect is that the random variables $G^t_{j,j'}$ are not independent since they depend on the requests from the same patients (those currently in unit j). The most reasonable assumption would be to decompose $G^t_{j,j'}$ as the sum of patient-level demands $G^t_{j,j'} = \sum_i Y^t_{i,j,j'}$ where the $\boldsymbol{Y}_i$'s are independent generalized Bernouilli random variables. Doing so, however, would break the benefit from aggregation and involve intractable multidimensional integration. In addition to tractability benefits, robust modeling theoretically relates to stochastic modeling and we can connect the worst-case cost of a policy $c^\star = \max_{\boldsymbol{g}\in\mathcal{U}} c(\boldsymbol{f}, \boldsymbol{g})$ with a value-at-risk statement. Formally, we restrict our attention to flows $\boldsymbol{f}$ that do not depend on $\tilde{\boldsymbol{g}}$. Under reasonable assumptions on the distribution of $\boldsymbol{G}$, we provide theoretical bounds on the quality of this pessimistic estimation in terms of disappointment, i.e., we elicit a threshold ε such that the probability that the actual cost $c(\boldsymbol{f}, \boldsymbol{G})$ exceeds $c^\star$, $\mathbb{P}\left(c(\boldsymbol{f}, \boldsymbol{G}) \geq c^\star\right)$, is bounded by ε. In other words, the robust objective $c^\star$ can be viewed as an upper-bound on the value-at-risk for $c(\boldsymbol{f}, \boldsymbol{G})$ at a level $1-\varepsilon$ (see Föllmer and Schied 2002, for an introduction to risk measures). We acknowledge the fact that minimizing value-at-risk is a risk-averse attitude that relates to the prevailing carefulness in healthcare and hospital managers. Yet, risk-neutral decision makers in other contexts might favor a stochastic modeling with the objective of minimizing the expected cost.

To do so, we leverage structural properties of the cost function and assume that $c(\boldsymbol{f}, \boldsymbol{g})$ can be decomposed as the sum of k pieces, where each piece k involves a distinct subset of units and time periods, i.e.,

$$c(\boldsymbol{f}, \boldsymbol{g}) = \sum_{k\in[K]} c_k \cdot \left(\sum_{(j,j',t)\in\mathcal{S}_k} g^t_{j,j'} - \sum_{(j,j',t)\in\mathcal{S}_k} f^t_{j,j'} \right)^+, \tag{6}$$

where $\mathcal{S}_k \subseteq [J]\times[J]\times[T]$ and $\forall k \neq k', \mathcal{S}_k \cap \mathcal{S}_{k'} = \emptyset$. Each piece penalizes the unsatisfied demand for a particular set of origin units j, destination units j' and time periods t. For instance, in the example described in Section EC.2.4, one piece specifically penalizes the unmet demand for intensive care from the ED ($j = UV$, $j' \in IC$, $t \in [T]$), another considers demand from outside transfers, and so on. Finally, we make distributional assumptions on the random vector $\boldsymbol{G}$: In accordance with the queueing literature, we first assume that the flows follow a Poisson distribution:

ASSUMPTION 2. *The random variables* $\sum_{(j,j',t)\in\mathcal{S}_k} G^t_{j,j'}$ *are independent Poisson random variables with respective rates* λ_k.

In practice, recall that we estimate $\hat{\boldsymbol{g}} = \mathbb{E}[\boldsymbol{G}]$ from data, so that the rates $\lambda_k = \sum_{(j,j',t)\in\mathcal{S}_k} \mathbb{E}[G^t_{j,j'}]$ can readily be estimated as well. Under this assumption, we provide the following guarantee:

PROPOSITION 1. *Fix* $\boldsymbol{f}\in\mathcal{F}$ *and let* $c^\star = \max_{\boldsymbol{g}\in\mathcal{U}} c(\boldsymbol{f},\boldsymbol{g})$. *Denote* $c_\infty = \max_k c_k$ *and* $\bar{c} = \sum_k \lambda_k c_k$. *We introduce a measure of the size the uncertainty set* $\mathcal{U}$ *in terms of relative cost deviation* $\Gamma_c := [c^\star - c(\boldsymbol{f},\hat{\boldsymbol{g}})]/\bar{c}$. *Under Assumption 2, we have*

$$\mathbb{P}(c(\boldsymbol{f},\boldsymbol{G}) \geq c^\star) \leq \exp\left(-\frac{\bar{c}}{c_\infty}\frac{\Gamma_c^2}{1+\Gamma_c}\right).$$

Proof of Proposition 1

$$\begin{aligned}\mathbb{P}\left(c(\boldsymbol{f},\boldsymbol{G}) \geq c^\star\right) &= \mathbb{P}\left(c(\boldsymbol{f},\boldsymbol{G}) - c(\boldsymbol{f},\hat{\boldsymbol{g}}) \geq c^\star - c(\boldsymbol{f},\hat{\boldsymbol{g}})\right)\\ &\leq \mathbb{P}\left(\sum_{k\in[K]} c_k \left|\sum_{(j,j',t)\in\mathcal{S}_k} G^t_{j,j'} - \sum_{(j,j',t)\in\mathcal{S}_k} \hat{g}^t_{j,j'}\right| \geq \Gamma_c\right).\end{aligned}$$

The result then follows from a large deviation bound for the sum of independent Poisson random variables (Lemma 1) that we derive in Section A.1. □

In Proposition 1, the quantity $\bar{c}/c_\infty$ corresponds to a cost-adjusted average number of patients to be treated - in particular, if the c_k's are identical, it simplifies to $\sum_k \lambda_k$. Hence, the right-hand side in Proposition 1 decays exponentially in the total number of patients. Note, however, that Proposition 1 is an *a posteriori* bound that can be computed for a specific solution $\boldsymbol{f}$. To approximate it uniformly over the set of all policies $\boldsymbol{f}$, further assumptions are needed. For instance, if $c(\boldsymbol{f},\boldsymbol{g})$ is linear and $\mathcal{U} = \{\boldsymbol{g} : \|\boldsymbol{g} - \hat{\boldsymbol{g}}\| \leq \Gamma\}$, then $c^\star \propto c(\boldsymbol{f},\hat{\boldsymbol{g}}) + c_\infty\Gamma$ and the exponential term decays in $-\Gamma^2$. Interestingly, Proposition 1 demonstrates that the worst-case approach relates to a stochastic modeling of the uncertainty –yet, with a risk-averse lens, since the worst-case cost relates to value-at-risk and not expected cost– despite the fact that the uncertainty sets considered are bounded while the random variables $\boldsymbol{G}$ are Poisson, hence potentially unbounded.

REMARK 1. While the Poisson assumption is pervasive in queueing theory, we acknowledge that the independence assumption is restrictive. In particular, the sum $\sum_{k\in[K]}\sum_{(j,j',t)\in\mathcal{S}_k} G^t_{j,j'}$ is bounded by the total number of patients, which negates independence. Alternatively, we assume that patients are independent, and that the flows $G^t_{j,j'}$ aggregate patient-level moves:

ASSUMPTION 3. *The random variables* $G^t_{j,j'}$ *can be decomposed into* $G^t_{j,j'} = \sum_{i\in I} Y^t_{i,j,j'}$ *where the vectors* $\boldsymbol{Z}_i \in \{0,1\}^{\mathcal{J}\times\mathcal{J}\times\mathcal{T}}$, $i = 1,\ldots,I$ *are independent and satisfy* $\sum_{(j,j',t)} Y^t_{i,j,j'} \leq 1$.

Note that we can still define a rate for $\sum_{(j,j',t)\in\mathcal{S}_k} G^t_{j,j'}$ as $\lambda_k := \mathbb{E}\left[\sum_{(j,j',t)\in\mathcal{S}_k} G^t_{j,j'}\right] = \sum_{i\in[I]}\sum_{(j,j',t)\in\mathcal{S}_k}\mathbb{E}[Y^t_{i,j,j'}]$. Under Assumption 3, we derive a similar probabilistic guarantee as Proposition 1 (see Section A.2 for statement and proof) that shares the same qualitative insights.

We should emphasize that, since these theoretical guarantees apply to the day-level flow decisions $\boldsymbol{f}$, they would correspond to the actual performance of the algorithm if implemented in an open-loop fashion. In practice, however, we adopt a close-loop implementation where, at each time step, we solve (5), implement $\boldsymbol{z}^1/\boldsymbol{f}^1$ only, and then resolve (5) at the next time step. We will evaluate the performance of the close-loop policy via simulations in Section 5 but deriving theoretical guarantees for close-loop policy constitutes an interesting and open research direction.

A second benefit of a robust optimization approach is that it requires no distributional assumptions, thus avoiding the risk of model misspecification. Distributionally robust optimization could be used to reconcile stochastic modeling with uncertainty (also called ambiguity in this context) on the true underlying distributions. However, such methods are less tractable than standard robust approaches, even when relaxing the integrality constraints on the variables $G^t_{j,j'}$. Also, the robust approach can be interpreted as a lower bound on the objective of a distributionally robust optimization problem: For any set of distributions $\mathcal{P}$, we have

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\boldsymbol{G}\sim\mathbb{P}}\left[c(\boldsymbol{f},\boldsymbol{G})\right] \geq \max_{\tilde{\boldsymbol{g}}\in\mathcal{U}} c(\boldsymbol{f},\tilde{\boldsymbol{g}}),$$

with $\mathcal{U} = \{\mathbb{E}_{\boldsymbol{G}\sim\mathbb{P}}[\boldsymbol{G}] : \mathbb{P}\in\mathcal{P}\}$, by convexity of c and Jensen's inequality. In other words, a distributionally robust problem is at least as conservative as a robust problem whose uncertainty set is the set of all possible average flows. Finally, as we will discuss in the following section and in the numerical experiments, we observe that using adaptive decision rules $\boldsymbol{f}$ has a first-order impact on the performance and is more important than the description of the uncertainty. In this regard, opting for a tractable model of uncertainty that can be solved efficiently and be generalized to affine decision rules without exploding computational time is important.

Finally, uncertainty on $\tilde{\boldsymbol{g}}$ partly comes from inherent stochasticity in the arrival and request process but can also come from deterministic biases in the estimation of future demand for beds due to censoring. For instance, for predicting the probability that a current inpatient will need an ICU bed, we do not have access to data about needs and only observe transfers to/out of the ICUs. These internal transfers are indeed critical for the patient and often arranged informally (e.g., over the phone) between physicians directly, without recording a formal bed request in the system. The robust modeling of uncertainty is better suited to such deterministic deviations between the actual vector of demand $\tilde{\boldsymbol{g}}$ and their estimates $\hat{\boldsymbol{g}}$ than a distributionally robust model of ambiguity.

4.4. Static and affine adaptive policies

As previously discussed, the future decision variables $\boldsymbol{f}$ are not readily implementable since they represent unit-level flows. Instead, one should solve (5) at each time step and implement the patient-level here-and-now decision variables $\boldsymbol{z}^1$. Accordingly, the decisions taken at time t will effectively depend on the realization of demand $\tilde{g}^\tau_{jj'}$ for $\tau < t$. To mimic this process as truthfully as possible, we impose some structure on how future decision variables $\boldsymbol{f}$ in (5) should depend (or *adapt*) to the state variables $\boldsymbol{g}$ and derive the robust equivalent of (5) for each level of adaptivity we impose.

First, we consider a static approximation and restrict our attention to flows $\boldsymbol{f}$ that do not depend on $\tilde{\boldsymbol{g}}$. In doing so, we observe that the equality constraints $f^t_{j,DIS} = \tilde{g}^t_{j,DIS}$, $j \in [J], t \in [T]$ cannot hold for all values of $\tilde{\boldsymbol{g}} \in \mathcal{U}$. As a result, we conservatively replace them by inequality constraints and solve the resulting optimization problem:

$$
\begin{aligned}
\min_{\boldsymbol{z}^1 \in \mathcal{Z}, \boldsymbol{f} \in \mathcal{F}, C_r} \quad & \boldsymbol{c}^\top \boldsymbol{z}^1 + \max_{\tilde{\boldsymbol{g}} \in \mathcal{U}} \lambda\, c(\boldsymbol{f}, \tilde{\boldsymbol{g}}) + \lambda_w \left(C_w - C_r\right)^+ \qquad (7)\\
\text{s.t.} \quad & z^0_j + \sum_{\tau \in [t]} \sum_{j'} f^\tau_{j',j} - \sum_{\tau \in [t]} \sum_{j'} f^\tau_{j,j'} \leq C^t_j,\ \forall j, t,\ \forall \tilde{\boldsymbol{g}} \in \mathcal{U},\\
& f^t_{j,DIS} \leq \tilde{g}^t_{j,DIS},\ \forall j, t,\ \forall \tilde{\boldsymbol{g}} \in \mathcal{U}\\
& f^1_{j,j'} = \sum_{i \in [I]} z^1_{ij'} z^0_{ij},\ \forall j \neq j',\ \forall \tilde{\boldsymbol{g}} \in \mathcal{U},\\
& C_r \leq \sum_{j \in [J]} \left(C^T_j - z^0_j - \sum_{t \in [T]} \sum_{j'} f^t_{j',j} + \sum_{t \in [T]} \sum_{j'} f^t_{j,j'} \right),\ \forall \tilde{\boldsymbol{g}} \in \mathcal{U},\\
& \sum_{t,j'} f^t_{j,j'} + \sum_{t} \tilde{g}^t_{j,DIS} \geq z^0_j,\ \forall j,\ \forall \tilde{\boldsymbol{g}} \in \mathcal{U}.
\end{aligned}
$$

Remember that $c(\boldsymbol{f}(\tilde{\boldsymbol{g}}), \tilde{\boldsymbol{g}})$ is a piecewise linear convex function. To bound its worst-case value, we can write $c(\cdot, \cdot)$ in epigraph formulation and robustify each constraint independently (Gorissen and den Hertog 2013). In addition to obvious tractability benefits, the static approximation is also theoretically appealing. Indeed, by solving (7), the decision maker has access to a pessimistic estimate of the future daily cost, $c^\star = \max_{\boldsymbol{g} \in \mathcal{U}} c(\boldsymbol{f}, \boldsymbol{g})$ that, as we proved in the previous section, can be interpreted as a bound on the value-at-risk for the random future cost $c(\boldsymbol{f}, \boldsymbol{G})$.

To mitigate the conservatism of the static approach, we consider affinely adaptive decision rules (Ben-Tal et al. 2004, Chen and Zhang 2009). Affine decision rules have been widely used in multistage robust optimization and have proved optimal for some highly structured problems (Bertsimas et al. 2010, Iancu et al. 2013). However, to the best of our knowledge, their performance for robust problems with discrete decision variables and uncertainty has not been investigated, neither theoretically nor numerically. First, let us observe that the equality constraints $f^t_{j,DIS}(\tilde{\boldsymbol{g}}) = \tilde{g}^t_{j,DIS}$, $\forall j, t, \forall (\tilde{\boldsymbol{g}}) \in \mathcal{U}$,

dictate how $f^t_{j,DIS}$ should depend on $\tilde{\boldsymbol{g}}$. As a result, we first replace the variables $f^t_{j,DIS}(\tilde{\boldsymbol{g}})$ in (5) by $\tilde{g}^t_{j,DIS}$, and then consider the following optimization problem:

$$\begin{aligned}
\min_{\boldsymbol{z}^1\in\mathcal{Z},\boldsymbol{f}(\cdot)\in\mathcal{R},C_r(\cdot)} \quad & \boldsymbol{c}^\top \boldsymbol{z}^1 + \max_{\tilde{\boldsymbol{g}}\in\mathcal{U}} \lambda\, c(\boldsymbol{f}(\tilde{\boldsymbol{g}}),\tilde{\boldsymbol{g}}) + \lambda_w \left(C_w - C_r(\tilde{\boldsymbol{g}})\right)^+ \\
\text{s.t.} \quad & z^0_j + \sum_{\tau\in[t]}\sum_{j'} f^\tau_{j',j}(\tilde{\boldsymbol{g}}) - \sum_{\tau\in[t]}\sum_{j'\neq DIS} f^\tau_{j,j'}(\tilde{\boldsymbol{g}}) - \sum_{\tau\in[t]} \tilde{\boldsymbol{g}}^\tau_{j,DIS} \le C^t_j,\ \forall j,t,\ \forall \tilde{\boldsymbol{g}}\in\mathcal{U}, \\
& f^1_{j,j'}(\tilde{\boldsymbol{g}}) = \sum_{i\in[I]} z^1_{ij'} z^0_{ij},\ \forall j\neq j',\ \forall \tilde{\boldsymbol{g}}\in\mathcal{U}, \\
& C_r(\tilde{\boldsymbol{g}}) \le \sum_{j\in[J]} \left(C^T_j - z^0_j - \sum_{t\in[T]}\sum_{j'} f^t_{j',j}(\tilde{\boldsymbol{g}}) + \sum_{t\in[T]}\sum_{j'\neq DIS} f^t_{j,j'}(\tilde{\boldsymbol{g}}) - \sum_{\tau\in[t]} \tilde{\boldsymbol{g}}^\tau_{j,DIS} \right),\ \forall \tilde{\boldsymbol{g}}\in\mathcal{U}.
\end{aligned}$$

Then, we narrow our search to affine policies, i.e., flows of the form $f^t_{j,j'}(\tilde{\boldsymbol{g}}) = \bar{f}^t_{j,j'} + \sum_{u,u',\tau\le t} \Phi^{t,\tau}_{j,j',u,u'} \tilde{g}^\tau_{u,u'}$, where the recourse matrix $\boldsymbol{\Phi}$ is itself a decision variable to be computed jointly with $\bar{\boldsymbol{f}}$. However, the resulting number of decision variables increases from J^2T to $J^4T^2 \approx 10^8$ and this simplifcation remains numerically intractable. Accordingly, we restrict our analysis to two special cases:

1. First, a case where all the variability in discharges from unit j are directly reported on the number of patients in unit j who stay in unit j throughout the day, i.e., impose the affine relationship $f^T_{j,j} = \bar{f}^T_{j,j} - \sum_{t\in[T]} \tilde{g}^t_{j,D} \ge 0$. Other flows $f^t_{j,j'}$ are considered nonadaptive. In this approximation, the recourse matrix $\boldsymbol{\Phi}$ is imposed by design and we shall refer to it as the "affine with known recourse" strategy. Also, in this policy, the future cost $c(\boldsymbol{f}(\tilde{\boldsymbol{g}}),\boldsymbol{G})$ only involves coordinates of $\boldsymbol{f}(\tilde{\boldsymbol{g}})$ that are nonadaptive so Proposition 1 still holds for this policy.

2. In the second approximation, which we call "restricted affine" policy, we assume that $f^t_{j,j'}$, $j\neq j'$, only depends on $\tilde{g}^t_{j,j'}$ while $f^T_{j,j}$ depends on $\tilde{g}^t_{j,DIS}$, $t\in[T]$ i.e.,

$$\begin{aligned}
f^t_{j,j'} &= \bar{f}^t_{j,j} + \Phi^t_{j,j'} \tilde{g}^t_{j,j'}, \text{for } j\neq j', \\
f^T_{j,j} &= \bar{f}^T_{j,j} + \sum_{t\in[T]} \Phi^t_{j,j} \tilde{g}^t_{j,DIS}.
\end{aligned}$$

Hence, in this model, the increase in the number of decision variables is linear in the number of initial decision variables (instead of quadratic).

5. Numerical experiments

We now apply our methodology to historical data from our partner hospital, and assess its performance.

5.1. Evaluation methodology

We apply the HIFO approach on data collected between January and August 2019. For our sandbox experiments to be as faithful as possible to real-world implementation, we developed a methodology based on historical rather than simulated data. We detail the precise experimental procedure in Section EC.4 but highlight its main assumptions here:

- We preserve the discharge patterns observed empirically. In other words, our simulation assumes that the decisions we are making (i.e., bed assignments) only impact bed assignments, and not length of stay or health outcomes. As previously mentioned, some empirical evidence suggests it is not the case. Since we find that our optimization approach improves bed assignment, it should reduce length of stay. As a result, keeping the empirical discharge patterns, as we do, leads to a conservative estimate of the potential improvement. We investigate the impact our method could have on length of stay via 212-day simulations in Section EC.4.4.
- To mitigate the aforementioned issue, we simulate the state of the hospital over small and non-overlapping periods (ranging from one to seven days): At the beginning of each period, we initialize the experiment with the historical state of the hospital at this point in time, and not what its state would have been had the optimization model run the previous period as well.
- We assume that all bed assignment decisions made by HIFO at the beginning of a time period are implemented and effective by the end of the time period, i.e., that delays between assignment and actual placement do not exceed two hours.
- Finally, our experiments correspond to a hybrid simulation where some allocation decisions are dictated by the optimization model and some are taken by hospital staff directly. Indeed, our optimization algorithm runs every two hours. For every bed request that arose and got solved between two optimization runs, we follow the decision from the hospital staff. This is notably the case for high-acuity patients or emergent surgeries from current inpatients.

We measure the quality of the bed assignment policy by the number of off-service placements made throughout the period and the waiting time of patients in the "scheduled admissions" and "unplanned visits" category, also referred to as boarding time. In reality, boarding delays recorded in the historical data may sometimes over-estimate actual delays (if a patient's movement is entered into the EHR system after it happened) or under-estimate them (if the movement is entered in the system preemptively to hold a bed for a patient). In addition, our simulations might not capture all the practical intricacies that generate delays between assignment and movement in practice, so the delays obtained in our simulations may deviate from what could be achieved if implemented in practice. To mitigate these issues, we measure historical and simulated delays in terms of number of optimization periods (i.e., 2-hour periods), hence offering a more conservative comparison of our optimization-based policy with the current state of practice. For these metrics, we either report

their normalized values[1] or the relative difference in performance of the optimization policy with the historical decisions made by the hospital on the same period. In any case, the lower the value the better. To ensure that improvement in off-service placement or delay is not due to the hospital admitting fewer patients or going above capacity, we also report peak census during the period: admission volume from unplanned visits and outside transfers.

5.2. Benchmarking the performance of HIFO

We first compare the behavior of the robust HIFO formulation, both static and affine (Section 4.4), with:

- A fluid model where all future demand $G^t_{j,j'}$ are taken equal to their average value and flows $\boldsymbol{f}$ are taken continuous and chosen so as to minimize the base-case cost $c(\boldsymbol{f}, \mathbb{E}[\boldsymbol{G}])$;
- A stochastic model where future demands from $j \in \{SA, UV, T\}$ are assumed to follow non-homogeneous Poisson distributions and demands from inpatient units j are decomposed as the sum of independent generalized Bernouilli random variables. $G^t_{j,j'}$ for inpatient units. Expected values are estimated via sample average approximation (SAA) with 100 scenarios, and we impose the constraints for all scenarios. In EC.4.2, we verify that imposing the constraints in expectation instead does not materially change the performance of the SAA model.

In this set of experiments, we consider a subset of $N = 92$ days from our historical data (March–May 2019) and perform simulations over one-day periods.

Computational time and scalability: In total, we solved $12 \times 92 = 1,104$ instances, for 11 values of λ and three values of λ_w. We report summary statistics in Table 4. In contrast to competitive approaches from the literature, our optimization model is very tractable and can be solved in seconds for a 600-bed institution. Using a robust approach to model prediction errors, in either its static or known affine recourse version, does not yield any noticeable solve time increase compared to the fluid approximation model and the stochastic (SAA) model. In addition, note that the SAA model requires additional computational time for sampling scenarios – 30.7 seconds on average to sample 100 scenarios in our simulations – and additional memory. The runtimes for the restricted affine policy are approximately 30–40% higher than for the other three methods, but remain below a minute for 99% of the instances. As a result, one could contemplate real-time implementation to make bed assignment decisions in an online fashion. Yet, one has to keep in mind that the time period width (here, two hours) needs to be long enough for the bed assignment policy to be computed *and implemented*.

[1] To preserve anonymity of our partner hospital and its actual performance, we divide the values of each metric by the median of the same metric on the historical data.

Table 4 Summary statistics of the solve time (in seconds) over $33 \times 1,104 = 36,432$ instances for each method.

Percentile:	1st	25th	50th (median)	75th	99th	Average
Fluid approximation	0.12	0.27	0.37	0.81	12.89	1.60
Sample average approximation	0.13	0.35	0.53	0.96	26.44	2.19
HIFO - Static	0.11	0.26	0.36	0.63	26.68	2.07
HIFO - Affine with known recourse	0.11	0.26	0.37	0.62	24.77	1.89
HIFO - Restricted affine	0.17	0.43	0.7	1.26	41.21	3.22

Impact on waiting time and off-service placement: There is an intuitive trade-off between time a patient waits to be admitted and the quality of the assigned bed. In our model, we can control this trade-off through the parameters λ and λ_w that balance the first-stage and the multistage objectives. We illustrate this trade-off by considering 11 (resp. three) different values of λ (resp. λ_w) and plotting the relative improvement compared with historical placements in terms of total waiting time (from both scheduled and unscheduled admissions) vs. off-service placements in Figure 3. We make a few observations: First, all four methods lie on the negative orthant, hence improve on both metrics compared to the current policy. Although there is a trade-off between these two performance measures, our simulations suggest that there is an opportunity for hospitals to improve on both simultaneously, by leveraging advanced prescriptive analytics. Second, the fluid approximation performs comparably with HIFO with static decisions on these average performance metrics although its objective is essentially a value-at-risk (according to Proposition 1). In our view, the good performance of robust approach in average terms is due to close-loop implementation, which alleviates conservatism by updating the problem input parameters and re-optimizing at each time period. Furthermore, using affine decision rules further and substantially increases the benefit from optimization. Unlike the static (robust or stochastic) methods that consider one decision variable $\boldsymbol{f}^t$ for future decision, adaptive models consider multiple plans for the future, which will later be chosen based on the realized level of arrivals and discharges. Since uncertainty directly affects the set of feasible decisions (through discharges and capacity constraints), capturing such flexibility in future decisions lead to more forward-looking first-stage decision. In particular, we observe that HIFO with affine decision rules tends to admit more patients early in the day (10 a.m.-4 p.m.) than the other policies, hence showing an anticipation of the daily discharges (see additional evidence in EC.4.2)

5.3. Impact of HIFO on hospital performance

We now evaluate in more depth the potential benefit of the HIFO policy for the hospital. In this set of experiments, we consider the whole $N = 236$ days from our historical data and perform simulations over $\{1, 3, 5, 7\}$-day periods. For simplicity, we only report results for HIFO with known affine recourse, for $\lambda = \lambda_w = 1$.

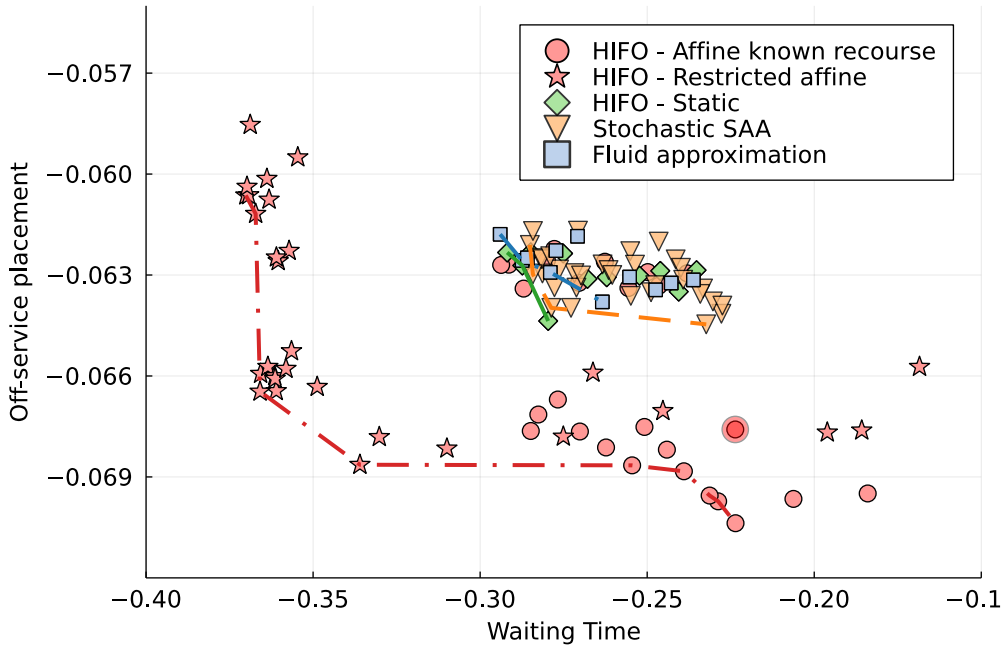


Figure 3 **Trade-off between waiting time and off-service placements. Each point represents the average relative difference between the considered policy and historical placements. A negative value indicates an improvement. We compare the performance of the proposed robust affine policies (red circles and stars) with the static solution (green diamonds) and their fluid approximation (blue squares). The highlighted point corresponds to the setting ($\lambda = 1$, $\lambda_w = 1$) used in the rest of the experiments.**

Quality of the bed assignment Figure 4 summarizes the relative improvement of HIFO over the current bed assignment strategy in terms of waiting time for (un)scheduled visits and off-service placement, for the different simulation lengths. On the one-day experiments, we recover the results observed in the previous section, namely the waiting time reduces by 20% and 11% on average for unscheduled and scheduled patients respectively, while off-service placement reduces by 6%. In addition, Figure 4 (and Figure EC.4) demonstrate that this is an improvement on average and also a substantial distributional shift towards lower values for all three metrics. As the simulation length increases from one to seven days, or as the optimization model is used on more consecutive days, the edge of optimization further increases and we observe an average reduction of 35%, 18%, and 24% on the same three metrics. Yet, the reduction in boarding time for patients in the post-anesthesia units (part of scheduled admissions) ought to be taken with a pinch of salt. Indeed, for patients in the post-anesthesia units (part of scheduled admissions) and for the HIFO policy, we take into account bed requests from surgical patients as soon as they enter the post-anesthesia care unit (PACU) and assume that the assignment can be implemented in the two hours following the decision. However, after a surgery, a patient can often not be moved to an inpatient unit before they recover from the

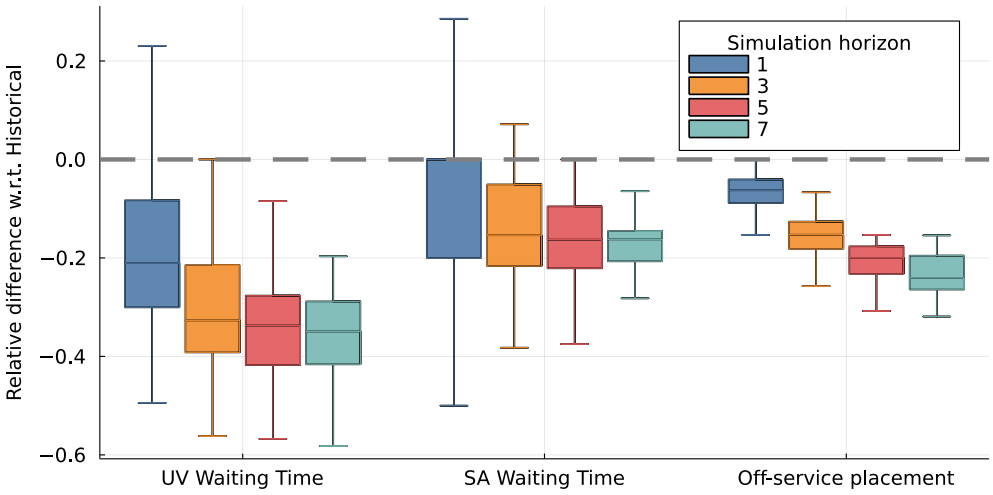


Figure 4 **Box plot for the distribution of the relative difference between the HIFO policy over the historical bed assignment decisions in terms of three performance metrics: (from left to right) waiting time from unscheduled visits, waiting time from scheduled admissions, and number of off-service placements. Results are reported for different simulation lengths (1, 3, 5, and 7 days)**

anesthesia and are physically fit to be discharged from the PACU. This time-to-recovery generates delays for the historical policies, delays to which HIFO is currently oblivious.

Control for admission volumes: We also compare the peak census and admission volumes in Table 5. Overall, the results of these metrics are reassuring: The daily peak census at the hospital is largely unchanged, negating the possibility that HIFO would reduce boarding time and off-service placement by overcrowding units. Admission volumes from scheduled admissions remain constant because there is no question about whether to admit patients from this category. However, for unscheduled visits, HIFO admits 12% more patients than current practice. These admissions correspond to patients who historically requested a bed, waited, and eventually exited the system without being admitted (the reason for exiting the queue is unavailable in the data). For these patients, HIFO assigns a bed faster. However, we observe that admissions from outside transfers decrease substantially on one-day simulations. This observation suggests that HIFO tends to favor patients physically present in the premises of the hospital over patients in another facility. Equivalently, it suggests that current hospital practice neglects the impact that admitting external patients has on the rest of the system when making their admission decisions. Our holistic approach integrates outside admission decisions into an overall bed assignment decision tool and is able to quantitatively take into account these effects. This constitutes a central contribution of the HIFO model. Furthermore, we observe that this negative impact vanishes as the simulation length increases. Since each simulation starts with a backlog of bed requests inherited from the historical policy, the observed trend supports the intuition that HIFO helps manage and reduce the backlog in the short term, thus creating opportunities for more outside transfer admissions in the long term.

Table 5 Median and inter-quartile range for the relative difference between the HIFO policy over the historical bed assignment decisions in terms of peak census and admission volume.

Metric	1-day simulation	3-day simulation	5-day simulation	7-day simulation
Peak census	0.91% (1.1%)	2.0% (1.31%)	2.64% (1.83%)	2.14% (1.63%)
TC volume	-12.92% (187.18%)	12.0% (70.63%)	11.11% (39.16%)	13.21% (31.31%)
UV volume	18.18% (21.67%)	12.75% (11.19%)	12.77% (9.25%)	12.38% (7.25%)
SA volume	0.0% (0.0%)	0.0% (2.08%)	0.0% (1.85%)	1.16% (1.52%)

5.4. Discussion on practical implications

In summary, our simulations illustrate that: a discrete optimization approach for patient-bed assignment is tractable, even for a hospital with 500+ beds; using adaptive decision rules has a first-order effect on the algorithm performance (more important than the framework used to describe the uncertainty); and an optimization-based patient-bed assignment process can simultaneously improve waiting times and off-service placement.

The implementation of a prescriptive bed-assignment model in practice might raise concerns related to fairness. We now compare how the decisions made historically by the hospital differ from HIFO's *at a patient level*. We consider all admissions of unscheduled visits and indicate whether their waiting time and service placement improved/deteriorated/remained the same when applying HIFO. Our analysis comprises nearly $7,000$ individual bed allocations. We then categorize the patients into four groups, depending on whether their bed allocation with HIFO is better, worse, identical, or not comparable. Figure 5 reports aggregate results for increasing simulation lengths. While the decisions made by HIFO improve boarding times and service placements on average, we observe that not all patients benefit from the optimization approach. On one-day experiments, approximately 38% of all patients experience better outcomes under HIFO while 20% experience worse outcomes, suggesting that HIFO is able to correctly identify and prioritize patients based on their individual impact on the overall system performance. In addition, the proportion of patients experiencing better placements increases as the simulation length increases, providing encouraging results on the potential benefit from implementing HIFO in practice. Inequity implications and fairness enforcement in healthcare have received attention in other contexts (Olsen 2011, Bertsimas et al. 2013, McCoy and Lee 2014), and constitute interesting and necessary future directions for our work.

When moving from simulation to actual implementation, one should expect the actual operational benefits to be less acute. This gap is partly due to the fact that the decision of assigning a patient to a bed is ultimately made by nurses and doctors, the output of the algorithm being a recommendation they can decide to ignore. Since our simulation setting assumes that all the HIFO recommendations are followed, our experiments might over-estimate the operational benefits. To

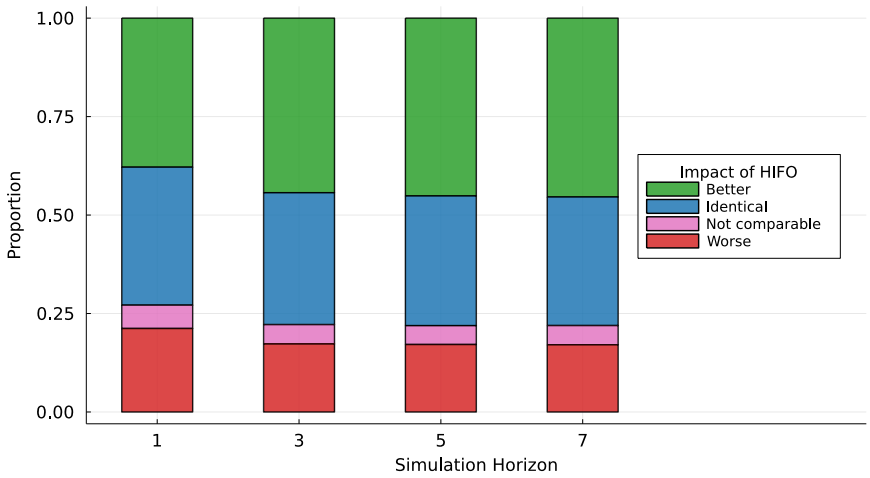


Figure 5 **Proportion of patients whose HIFO placement was better (green), worse (red), identical (blue), not comparable (pink) to their historical one, for different simulation lengths (1, 3, 5, and 7 days).**

quantify the impact non-compliance could have on our results, we conduct simulations over a one-day horizon, as in Section 5.2 but, for each patient, implement the recommendation of the algorithm with some probability $p \in [0,1]$ and implement the historical decision otherwise. The variable p (i.e., the level of compliance) creates a continuum between the historical policy ($p=0$) and our optimal policy ($p=1$). Among others, we observe in Figure 6 that implementing only 20% of the recommendations captures $\sim 70\%$ of the benefit in off-service placement while reducing waiting time by $\sim 7\%$. However, the impact on waiting time is almost linear in p. Modeling more rigorously the strategic behavior of users to the algorithmic recommendations and measuring in a real-life pilot the compliance and the ultimate operational impact are exciting potential future directions.

6. Concluding remarks

In this paper, we propose a robust discrete optimization framework to achieve hospital-wide patient flow management. Our approach comprises three key elements: a holistic view to account for the entirety of the hospital within a single model, an aggregation strategy to reconcile short- and medium-term objectives in a tractable fashion, and data to describe and anticipate future demand and supply of care. We use the framework of adaptive robust optimization to derive tractable optimization formulations, while alleviating conservatism inherent in a robust approach. Through extensive numerical simulations, we observe that our formulation can be solved within seconds for a 600-bed institution and could provide significant operational improvement. In addition to the implementation and evaluation of our algorithms in a real hospital environment, exciting future research directions include theoretically analyzing the performance of close-loop robust control policies, deriving tractable formulations for the corresponding adjustable *distributionally* robust optimization problem, and integrating fairness or behavioral considerations to increase adoption and impact.

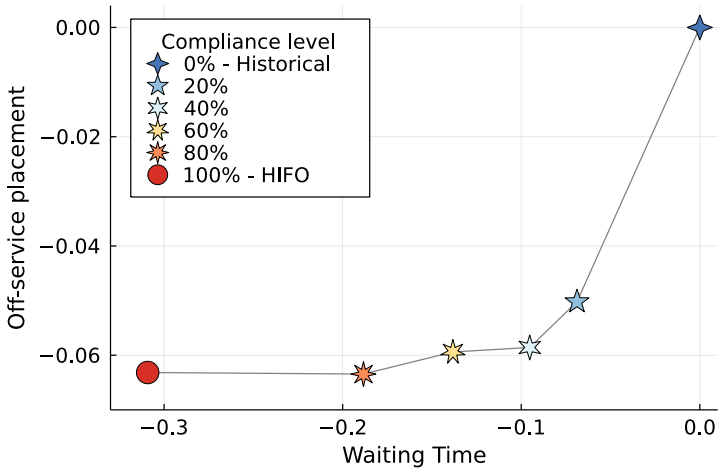


Figure 6 **Trade-off between waiting time and off-service placement for the HIFO policy with affine known recourse ($\lambda_w = \lambda = 1$) for varying compliance levels p.**

Acknowledgments

The authors thank the Associate Editor and three anonymous referees for their participation and constructive feedback during the review process, which greatly improved the quality of the manuscript.

Appendix A: Theoretical guarantees of the robust approach: lemmas and extension

In this section, we derive new tail bounds on the sum of independent Poisson random variables, used in the proof of Proposition 1. We also extend the result from Proposition 1 to a setting where we relax the independence assumption on the medical trajectories.

A.1. Tail bounds on the sum of Poisson random variables

LEMMA 1. *Let G_j, $j = 1, \dots, m$, be m independent Poisson random variables with respective rates λ_j. Fix $\boldsymbol{c} \in \mathbb{R}^m_+$ and denote $c_{\max} = \max_{j\in[m]} c_j$ and $\bar{c} = \sum_j \lambda_j c_j$. Then for any $\alpha \geq 0$,*

$$\mathbb{P}\left(\sum_{j\in[m]} c_j |G_j - \lambda_j| \geq \alpha \bar{c}\right) \leq \exp\left(-\frac{\bar{c}}{c_{\max}} \frac{\alpha^2}{1+\alpha}\right).$$

Proof of Lemma 1 Define $\rho(x) = e^x - x - 1$. For any Poisson random variable G with rate λ, $\mathbb{E}[e^{t[G-\lambda]}] = \exp(\lambda\rho(t))$. Fix $t \geq 0$. By a standard Chernoff bound technique, for any $\theta > 0$, we have

$$\mathbb{P}\left(\sum_{j\in[m]} c_j |G_j - \lambda_j| \geq t\right) \leq e^{-\theta t} \prod_{j\in[m]} \mathbb{E}\left[e^{\theta c_j G_j}\right] = \exp\left(-\theta t + \sum_{j\in[m]} \lambda_j \rho(c_j\theta)\right).$$

Since ρ is convex and $\rho(0) = 0$, for any $0 < x \leq y$, we have $\rho(x) \leq x/y\, \rho(y)$. Applying this inequality for $x = c_j\theta$ and $y = c_{\max}\theta$ yields

$$\mathbb{P}\left(\sum_{j\in[m]} c_j |G_j - \lambda_j| \geq t\right) \leq \exp\left(-\theta t + \frac{1}{c_{\max}} \sum_{j\in[m]} \lambda_j c_j \rho(c_{\max}\theta)\right) = \exp\left(-\theta t + \frac{\bar{c}}{c_{\max}} \rho(c_{\max}\theta)\right),$$

with $\bar{c} = \sum_j c_j \lambda_j$. The right-hand side is minimized at $\theta = \frac{1}{c_{\max}} \log(1 + t/\bar{c})$, which in turn leads to

$$\mathbb{P}\left(\sum_{j \in [m]} c_j |G_j - \lambda_j| \geq t \right) \leq \exp\left(\frac{\bar{c}}{c_{\max}} \left\{ \frac{t}{\bar{c}} - \left(1 + \frac{t}{\bar{c}}\right) \log\left(1 + \frac{t}{\bar{c}}\right) \right\} \right).$$

Since $\log(1+x) \geq 2x/(2+x)$ for $x \geq 0$,

$$\mathbb{P}\left(\sum_{j \in [m]} c_j |G_j - \lambda_j| \geq t \right) \leq \exp\left(-\frac{\bar{c}}{c_{\max}} \frac{(t/\bar{c})^2}{1 + t/\bar{c}} \right).$$

□

A.2. Probabilistic guarantees in case of dependent medical trajectories

We now relax the assumption that the flows are independent and Poisson distributed. Instead, we adopt a patient-level description: we assume that patients are independent and that flows are the aggregation of individual patient trajectories (Assumption 3). Note that under Assumption 3, we can still define a rate for $\sum_{(j,j',t) \in \mathcal{S}_k} G^t_{j,j'}$ as $\lambda_k := \mathbb{E}\left[\sum_{(j,j',t) \in \mathcal{S}_k} G^t_{j,j'} \right] = \sum_{i \in [I]} \sum_{(j,j',t) \in \mathcal{S}_k} \mathbb{E}[Z^t_{i,j,j'}]$. Among others, we have $\sum_{k \in [K]} \lambda_k \leq I$. Under this assumption, Proposition 1 translated into:

PROPOSITION 2. *Fix $\boldsymbol{f} \in \mathcal{F}$ and let $c^\star = \max_{\boldsymbol{g} \in \mathcal{U}} c(\boldsymbol{f}, \boldsymbol{g})$. Denote $c_\infty = \max_k c_k$ and $\tilde{c} = \sum_k \sqrt{\lambda_k} c_k$. We introduce a measure of the size the uncertainty set $\mathcal{U}$ in terms of relative cost deviation $\Gamma_c := [c^\star - c(\boldsymbol{f}, \hat{\boldsymbol{g}})] / \tilde{c}$. Under Assumption 3, for $\Gamma_c \geq 1$, we have*

$$\mathbb{P}(c(\boldsymbol{f}, \boldsymbol{G}) \geq c^\star) \leq \exp\left(-\frac{\tilde{c}^2}{2 I c_\infty} (\Gamma_c - 1)^2 \right).$$

Compared to Proposition 1, Proposition 2 qualitatively exhibits the same exponential dependency in $-\Gamma_c^2$. However, the rate involves $\tilde{c} = \sum_k \sqrt{\lambda_k} c_k \leq \sqrt{\bar{c} I}$, and is slower than the case where flows are assumed to be independent. The proof is similar to the proof of Proposition 1, except that we invoke a more technical tail bound derived from McDiarmid's inequality:

LEMMA 2. *Let $\boldsymbol{Z}_i \in \{0,1\}^m$, $i = 1, \ldots, n$, be n independent random vectors with $\sum_{j \in [m]} Z_{i,j} \leq 1$. Denote $\lambda_j := \sum_{i \in [n]} \mathbb{E}[Z_{i,j}]$. Fix $\boldsymbol{c} \in \mathbb{R}^m_+$ and denote $c_{\max} = \max_{j \in [m]} c_j$ and $\tilde{c} = \sum_j \sqrt{\lambda_j} c_j$. Then for any $\alpha \geq 1$,*

$$\mathbb{P}\left(\sum_{j \in [m]} c_j \left| \sum_{i \in [n]} (Z_{i,j} - p_{i,j}) \right| \geq \alpha \tilde{c} \right) \leq \exp\left(-\frac{\tilde{c}^2}{2 c_{\max} n} (\alpha - 1)^2 \right)$$

Proof of Lemma 2 Denote $h(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n) := \sum_{j \in [m]} c_j \left| \sum_{i \in [n]} (z_{i,j} - p_{i,j}) \right|$. The function h satisfies a so-called bounded difference property. Indeed, when we compare the difference in value of h for two n-tuples that differ only in their ith coordinate, since $\boldsymbol{z}_i$ and $\boldsymbol{z}'_i$ have at most one coordinate equal to 1, we obtain

$$|h(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_i, \ldots, \boldsymbol{z}_n) - h(\boldsymbol{z}_1, \ldots, \boldsymbol{z}'_i, \ldots, \boldsymbol{z}_n)| \leq 2 c_{\max}.$$

Hence, by McDiarmid's inequality (McDiarmid 1998, Theorem 3.1), for any $t \geq 0$,

$$\mathbb{P}\left(h(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n) - \mathbb{E}[h(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)] \geq t \right) \leq \exp\left(-\frac{t^2}{2 n c_{\max}} \right).$$

To conclude the proof, we need to bound $\mathbb{E}[h(\boldsymbol{Z}_1,\dots,\boldsymbol{Z}_n)] = \sum_{j\in[m]} c_j \mathbb{E}\left|\sum_{i\in[n]}(Z_{i,j}-p_{i,j})\right|$. By Cauchy-Schwarz inequality and the independence of the $\boldsymbol{Z}_i$'s, we obtain

$$\mathbb{E}\left|\sum_{i\in[n]}(Z_{i,j}-p_{i,j})\right| \le \sqrt{\sum_{i\in[n]} p_{i,j}(1-p_{i,j})} \le \sqrt{\sum_{i\in[n]} p_{i,j}} = \sqrt{\lambda_j}.$$

So $\mathbb{E}[h(\boldsymbol{Z}_1,\dots,\boldsymbol{Z}_n)] \le \sum_j c_j\sqrt{\lambda_j}$. Denoting $\tilde{c} = \sum_j c_j\sqrt{\lambda_j}$ and setting $t = (\alpha-1)\tilde{c}$, $\alpha \ge 0$, we obtain the result. □

References

César Alameda and Carmen Suárez. Clinical outcomes in medical outliers admitted to hospital with heart failure. *European Journal of Internal Medicine*, 20(8):764–767, 2009.

Mor Armony, Shlomo Israelit, Avishai Mandelbaum, Yariv N Marmor, Yulia Tseytlin, and Galit B Yom-Tov. On patient flow in hospitals: A data-based queueing-science perspective. *Stochastic Systems*, 5 (1):146–194, 2015.

Anthony D Bai, Siddhartha Srivastava, George A Tomlinson, Christopher A Smith, Chaim M Bell, and Sudeep S Gill. Mortality of hospitalised internal medicine patients bedspaced to non-internal medicine inpatient units: Retrospective cohort study. *BMJ Quality & Safety*, 27(1):11–20, 2018.

Hessam Bavafa, Charles M Leys, Lerzan Örmeci, and Sergei Savin. Managing portfolio of elective surgical procedures: A multidimensional inverse newsvendor problem. *Operations Research*, 67(6):1543–1563, 2019.

René Bekker and Paulien M Koeleman. Scheduling admissions and reducing variability in bed demand. *Health Care Management Science*, 14(3):237, 2011.

Aharon Ben-Tal, Alexander Goryashko, Elana Guslitzer, and Arkadi Nemirovski. Adjustable robust solutions of uncertain linear programs. *Mathematical Programming*, 99(2):351–376, 2004.

Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust Optimization*, volume 28. Princeton University Press, 2009.

Dimitris Bertsimas and Jack Dunn. Optimal classification trees. *Machine Learning*, 106(7):1039–1082, 2017.

Dimitris Bertsimas, Dan A Iancu, and Pablo A Parrilo. Optimality of affine policies in multistage robust optimization. *Mathematics of Operations Research*, 35(2):363–394, 2010.

Dimitris Bertsimas, Vivek F Farias, and Nikolaos Trichakis. Fairness, efficiency, and flexibility in organ allocation for kidney transplantation. *Operations Research*, 61(1):73–87, 2013.

Dimitris Bertsimas, Jean Pauphilet, Jennifer Stevens, and Manu Tandon. Predicting inpatient flow at a major hospital using interpretable analytics. *Manufacturing & Service Operations Management*, 2021.

John Boulton, Naveed Akhtar, Ashfaq Shuaib, and Paula Bourke. Waiting for a stroke bed: Planning stroke unit capacity using queuing theory. *International Journal of Healthcare Management*, 9(1):4–10, 2016.

Tim Carnes, Devon Price, Retsef Levi, Peter F Dunn, Bethany J Daily, and Sue Moss. An optimization framework for smoothing surgical bed census via strategic block scheduling. *Manufacturing Service Operation Management*, pages 488–494, 2011.

Carri W Chan, Jing Dong, and Linda V Green. Queues with time-varying arrivals and inspections with applications to hospital discharge policies. *Operations Research*, 65(2):469–495, 2016a.

Carri W Chan, Vivek F Farias, and Gabriel J Escobar. The impact of delays on service times in the intensive care unit. *Management Science*, 63(7):2049–2072, 2016b.

Carri W Chan, Linda V Green, Suparerk Lekwijit, Lijian Lu, and Gabriel Escobar. Assessing the impact of service level when customer needs are uncertain: An empirical investigation of hospital step-down units. *Management Science*, 65(2):751–775, 2018.

Xin Chen and Yuhan Zhang. Uncertain linear programs: Extended affinely adjustable robust counterparts. *Operations Research*, 57(6):1469–1482, 2009.

Jim G Dai and Pengyi Shi. Inpatient overflow: An approximate dynamic programming approach. *Manufacturing & Service Operations Management*, 2019.

Jim G Dai and Pengyi Shi. Recent modeling and analytical advances in hospital inpatient flow management. *Production and Operations Management*, 30(6):1838–1862, 2021.

Arnoud M De Bruin, René Bekker, Lillian Van Zanten, and GM Koole. Dimensioning hospital wards using the Erlang loss model. *Annals of Operations Research*, 178(1):23–43, 2010.

Jing Dong and Ohad Perry. Queueing models for patient-flow dynamics in inpatient wards. *Operations Research*, 68(1):250–275, 2020.

Richard C.. Dorf and Robert H Bishop. *Modern Control Systems*. Pearson Prentice Hall, 2008.

Hans Föllmer and Alexander Schied. Convex measures of risk and trading constraints. *Finance and Stochastics*, 6(4):429–447, 2002.

Bram L Gorissen and Dick den Hertog. Robust counterparts of inequalities containing sums of maxima of linear functions. *European Journal of Operational Research*, 227(1):30–43, 2013.

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

Dan A Iancu, Mayank Sharma, and Maxim Sviridenko. Supermodularity and affine policies in dynamic robust optimization. *Operations Research*, 61(4):941–956, 2013.

Navid Izady and Israa Mohamed. A clustered overflow configuration of inpatient beds in hospitals. *Manufacturing & Service Operations Management*, 2019.

Daniel W Johnson, Ulrich H Schmidt, Edward A Bittner, Benjamin Christensen, Retsef Levi, and Richard M Pino. Delay of transfer from the intensive care unit: A prospective observational study of incidence, causes, and financial impact. *Critical Care*, 17(4):R128, 2013.

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

Derya Kilinc, Soroush Saghafian, and Stephen Traub. Dynamic assignment of patients to primary and secondary inpatient units: Is patience a virtue? *HKS Working Paper No. RWP17-010*, 2018.

Song-Hee Kim and Ward Whitt. Are call center and hospital arrivals well modeled by nonhomogeneous poisson processes? *Manufacturing & Service Operations Management*, 16(3):464–480, 2014.

Song-Hee Kim, Carri W Chan, Marcelo Olivares, and Gabriel Escobar. Icu admission control: An empirical study of capacity allocation and its implication for patient outcomes. *Management Science*, 61(1): 19–38, 2015.

Jessica Liu, Joshua Griesman, Rosane Nisenbaum, and Chaim M Bell. Quality of care of hospitalized internal medicine patients bedspaced to non-internal medicine inpatient units. *PloS One*, 9(9):e106763, 2014.

Elisa F Long and Kusum S Mathews. The boarding patient: Effects of ICU and hospital occupancy surges on patient flow. *Production and Operations Management*, 27(12):2122–2143, 2018.

Kusum S Mathews, Matthew S Durst, Carmen Vargas-Torres, Ashley D Olson, Madhu Mazumdar, and Lynne D Richardson. Effect of emergency department and ICU occupancy on admission decisions and outcomes for critically ill patients. *Critical Care Medicine*, 46(5):720–727, 2018.

Jessica H McCoy and Hau L Lee. Using fairness models to improve equity in health delivery fleet management. *Production and Operations Management*, 23(6):965–977, 2014.

Colin McDiarmid. Concentration. In *Probabilistic methods for algorithmic discrete mathematics*, pages 195–248. Springer, 1998.

Fanwen Meng, Jin Qi, Meilin Zhang, James Ang, Singfat Chu, and Melvyn Sim. A robust optimization model for managing elective admission in a public hospital. *Operations Research*, 63(6):1452–1467, 2015.

Ester Góes Oliveira, Paulo Carlos Garcia, Clairton Marcos Citolino Filho, and Lilia de Souza Nogueira. The influence of delayed admission to intensive care unit on mortality and nursing workload: A cohort study. *Nursing in Critical Care*, 2018.

Jan Abel Olsen. Concepts of Equity and Fairness in Health and Health Care. In *The Oxford Handbook of Health Economics*. Oxford University Press, 04 2011.

Edieal Pinker and Tolga Tezcan. Determining the optimal configuration of hospital inpatient rooms in the presence of isolation patients. *Operations Research*, 61(6):1259–1276, 2013.

Chitta Ranjan, Kamran Paynabar, Jonathan E Helm, and Julian Pan. The impact of estimation: A new method for clustering and trajectory estimation in patient flow modeling. *Production and Operations Management*, 26(10):1893–1914, 2017.

Patricia A Rutherford, Lloyd P Provost, Uma R Kotagal, Katharine Luther, and Alex Anderson. Achieving hospital-wide patient flow. Technical report, Institute for Healthcare Improvement, 2017.

Fabian Schäfer, Manuel Walther, Alexander Hübner, and Heinrich Kuhn. Operational patient-bed assignment problem in large hospital settings including overflow and uncertainty management. *Flexible Services and Manufacturing Journal*, 31(4):1012–1041, 2019.

Pengyi Shi, Mabel C Chou, Jim G Dai, Ding Ding, and Joe Sim. Models and insights for hospital inpatient operations: Time-dependent ED boarding time. *Management Science*, 62(1):1–28, 2016.

Hummy Song, Anita L Tucker, and Karen L Murrell. The diseconomies of queue pooling: An empirical investigation of emergency department length of stay. *Management Science*, 61(12):3032–3053, 2015.

Hummy Song, Anita L Tucker, Ryan Graue, Sarah Moravick, and Julius J Yang. Capacity pooling in hospitals: The hidden consequences of off-service placement. *Management Science*, 66(9):3825–3842, 2020.

Andrew Stowell, Pierre-Geraud Claret, Mustapha Sebbane, Xavier Bobbia, Charlotte Boyard, Romain Genre Grandpierre, Alexandre Moreau, and Jean-Emmanuel de La Coussaye. Hospital out-lying through lack of beds and its impact on care and patient outcome. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine*, 21(1):17, 2013.

Robert Stretch, Nicolàs Della Penna, Leo A Celi, and Bruce E Landon. Effect of boarding on mortality in ICUs. *Critical Care Medicine*, 46(4):525–531, 2018.

Bex George Thomas, Srinivas Bollapragada, Kunter Akbay, David Toledano, Peter Katlic, Onur Dulgeroglu, and Dan Yang. Automated bed assignments in a complex and dynamic hospital environment. *Interfaces*, 43(5):435–448, 2013.

Steven Thompson, Manuel Nunez, Robert Garfinkel, and Matthew D Dean. Or practice—efficient short-term allocation and reallocation of patients to floors of a hospital during demand surges. *Operations Research*, 57(2):261–273, 2009.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Hongteng Xu, Weichang Wu, Shamim Nemati, and Hongyuan Zha. Patient flow prediction via discriminative learning of mutually-correcting processes. *IEEE Transactions on Knowledge and Data Engineering*, 29 (1):157–171, 2016.

Mohammad Zhalechian, Esmaeil Keyvanshokooh, Cong Shi, and Mark P Van Oyen. Personalized hospital admission control: a contextual learning approach. *Available at SSRN 3653433*, 2020.

Hui Zhang, Thomas J Best, Anton Chivu, and David O Meltzer. Simulation-based optimization to improve hospital patient assignment to physicians and clinical units. *Health Care Management Science*, 23(1): 117–141, 2020.

Electronic Companion

This electronic companion (EC) provides the full Hospital-wide Inpatient Flow Optimization HIFO formulation presented in Section 3. It also gives details on the uncertainty set from Section 4, and describes the simulation procedure used for the numerical experiments in Section 5 alongside additional numerical results.

EC.1. Description of units in our partner hospital

Our partner hospital comprises two campuses (A and B). We provide a list and description of the units on campus A in Table 1. For completeness, we detail the composition of campus B here in Table EC.1.

Table EC.1 Description of inpatient units on campus B at our partner hospital. The type is either GC (general care) or IC (intensive care). The unit name follows the nomenclature: Campus-Type Number.

Name	Type	# Licensed beds	# Private rooms	Primary (secondary) services
B-GC 1	GC	36	16	Surgery, Trauma (Orthopedics)
B-GC 2	GC	36	16	Orthopedics, Surgery
B-GC 3	GC	28	7	General Medicine, Surgery
B-GC 4	GC	30	8	Neurology, Neural Surgery
B-GC 5	GC	32	10	General Medicine
B-GC 6	GC	29	10	Cardiology (Cardiac Surgery)
B-GC 7	GC	30	8	Cardiology (Cardiac Surgery)
B-GC 8	GC	14	4	Neurology, Neural Surgery
B-GC 9	GC	28	2	General Medicine
B-GC 10	GC	20	20	Cardiac Surgery, Thoracic Surgery
B-GC 11	GC	34	6	General Medicine
B-IC 1	IC	10	10	Surgery (Orthopedics)
B-IC 2	IC	7	7	Cardiac Surgery
B-IC 3	IC	8	8	Cardiac Surgery
B-IC 4	IC	8	8	Cardiac Surgery (Cardiology)
B-IC 5	IC	8	8	Neurology, Neural Surgery
B-IC 6	IC	8	8	General Medicine
B-IC 7	IC	8	8	General Medicine
B-IC 8	IC	8	8	Neurology, Neural Surgery
B-IC 9	IC	8	4	Vascular Surgery

EC.2. Detailed optimization formulation and extensions

In this section, we provide some practical details about the optimization formulation introduced in Section 3 and describe extensions of the model that are relevant to practice. In this section, we present the complete nominal formulation for the HIFO problem, which includes individual future trajectories and capacity decisions.

EC.2.1. Immediate problem: cost calculation

The cost of assigning patient i to unit j, c_{ij}, should depend on the patient's current location and their medical need.

Each unit or ward is pre-assigned a primary service as well as secondary services, as described in Table 1. Correspondingly, we define a priority level p_{js} between each unit j and service s. For primary services, $p_{js}=1$, for secondary services, $p_{js}=2$. Otherwise, we have $p_{js}=\infty$. Note that, in some large hospitals, some units might provide a third level of medical specialty ($p_{js}=3$) –typically, those would correspond to medical specialties that are not unit j's primary or secondary specialties. These patients are often off-placed in unit j so the nursing staff in j are now familiar with this new patient population.

For each patient i, we describe their medical need as the combination of hospital service/medical specialty s and level of care ℓ that their condition requires. The level of care $\ell=1$ if the patient requires an intensive care bed, 0 otherwise. With this notation, we decompose the cost c_{ij} into

$$\begin{aligned} c_{ij} = p_{js} - 1 & \qquad \text{(Service match)} \\ + \begin{cases} 6, & \text{if } \ell=1 \text{ and } j \notin IC, \\ 3, & \text{if } \ell=0 \text{ and } j \notin GC, \\ 0, & \text{otherwise.} \end{cases} & \qquad \text{(Level of care match)} \end{aligned}$$

Depending on the particular hospital, we might include a term to capture physical distance between patient i and unit j, especially if units are spread across different buildings.

EC.2.2. Immediate problem: future individual assignments

In practice, for patients who need a new bed, medical staff might value having a plan for where to place the patients immediately (z^1_{ij}) and when to move them in the future and where. To do so, we introduce individual unit assignment variables z^t_{ij} which will track the location of the patients through to the end of the day. $z^t_{ij}=1$ if patient i is in unit j at time t, 0 otherwise. Since we only care about the patients who currently need a new bed, we shall introduce these variables just for them.

In accordance with the single-move assumption (Assumption 1), each patient should move (at most) once during the entire time horizon. Therefore, we also introduce binary variables to encode for the final location of the patient, ℓ_{ij}, and the time they move, m^t_i. All things considered, the decision variables z^t_{ij}, ℓ_{ij} and m^t_i must satisfy a set of constraints, concisely denoted $(\boldsymbol{z},\boldsymbol{\ell},\boldsymbol{m})\in\mathcal{Z}$, namely:

$$\begin{aligned} z^t_{ij}, \ell_{ij}, m^t_i \in \{0,1\},\ \forall i\in[I], j\in[J], t\in[T], \\ \sum_{t\in[T]} m^t_i \leq 1,\ \sum_{j\in[J]} \ell_{ij} = 1,\ \forall i\in[I], \end{aligned}$$

$$z_{ij}^t \le \ell_{ij},\ \forall j \in [J], t \in [T],$$
$$\sum_{\tau \in [t]} m_i^\tau \le 1 - z_{ij}^0 z_{ij}^t,\ \forall j \in [J], t \in [T-1],$$
$$\sum_{i \in I} z_{ij}^t \le C_j^t,\ \forall j \in [J], t \in [T],$$
$$z_{ij}^t = 0, \forall i \in [I], t \in [T], j \in \mathcal{F}_i.$$

The first set of constraints ensures that $\boldsymbol{z}, \boldsymbol{\ell}, \boldsymbol{m}$ are binary. The second set of constraints ensure that each patient i moves at most once. The third constraints capture the logic $\ell_{i,j} = 0 \implies z_{i,j}^t = 0$. The third constraint ensures that if the patient has not moved, i.e., $z_{ij}^0 z_{ij}^t = 1$, then $m_i^\tau = 0$ for $\tau \le t$. The fourth constraints are capacity restrictions. Finally, the last set of constraints involves a patient-specific set $\mathcal{F}_i$ describing the units that patient i cannot go to. For instance, if patient i is initially in unit j_0, i.e., $z_{ij_0}^0 = 1$, and did not request a new bed, then $\mathcal{F}_i = [J] \setminus \{j_0\}$. Note that we could relax this requirement to allow for inpatient reallocation as in Thompson et al. (2009).

Similar to Section 3, the objective is to minimize cost. For simplicity, we decompose cost into an assignment quality term, $c_{ij}\ell_{ij}$, with c_{ij} defined as in the previous section, and a waiting time cost, $c_{it}m_i^t$. The waiting time cost has a piece-wise linear structure, $c_{it} = a_i(PastWait_i + 2t - \bar{\tau}_i)^+$, i.e., we linearly penalize waiting time (by a marginal cost of a_i) whenever the cumulative waiting time for patient i (the sum of the time they have been waiting up until $t = 0$, denoted $PastWait_i$, plus the time they will wait if they are assigned at time t, $2t$ because we consider 2-hour time steps) exceeds a target threshold of $\bar{\tau}_i$. The target threshold $\bar{\tau}_i$ depends on the unit in which the patients are (we use 0 hours for scheduled admissions, 4 hours for unscheduled visits, and 6 hours for transfers). The marginal cost of waiting a_i depends on (a) the unit the patients are waiting in (among SA, UV, T), (b) the level of care they are waiting for, and (c), for SA and UV patients, the time they have been waiting so far, $PastWait_i$.

Remark EC.1. If one does not to want to introduce these additional variables to describe the future trajectories of patients currently waiting for a bed, the unit assignment costs c_{ij} defined EC.2.1 could be modified to also prioritize patients based on the time they have waited so far, e.g., by considering $c_{ij} + a_i(WaitTime_i - \bar{\tau}_i)^+$ instead.

EC.2.3. Daily problem: feasible set

The flow variables $f_{j,j'}^t$ ought to satisfy the following constraints:

(F1) Integrality: $f_{j,j'}^t \in \mathbb{N}$.

(F2) Flow conservation: If j is an inpatient unit, we have $\sum_{t,j'} f_{j,j'}^t = z_j^0$. Alternatively, for $j \in \{SA, US, T\}$, the total number of patients moved from j (for instance, from the emergency department) throughout the day, $\sum_{t,j'} f_{j,j'}^t$ can be greater than the number of patients currently waiting to be moved at $t = 0$, z_j^0, because of future arrivals. Hence, we have $\sum_{t,j'} f_{j,j'}^t \ge z_j^0$.

(F3) Forbidden moves: Those constraints are similar to those on the individual moves. We also add the constraints $f^t_{j,j}=0,\ \forall t\in[T-1], j\in[J]$ to ensure that patients who do not move during the day are all captured in $f^T_{j,j'}$.

(F4) Capacity constraints:

$$z^0_j+\sum_{\tau\in[T]}\sum_{j'} f^\tau_{j',j}-\sum_{\tau\in[T]}\sum_{j'} f^\tau_{j,j'}\le C^t_j,\ \forall j\in[J],\forall t\in[T]. \tag{1}$$

EC.2.4. Daily problem: cost formula

Let $(x)^+:=\max(x,0)$ denote the positive part of x. We define a cost function $c(\boldsymbol{f},\boldsymbol{g})$ between inpatient flows between units, $\boldsymbol{f}$, and the total number of patients requesting to be moved (i.e., the demand) between units, $\boldsymbol{g}$, as the sum of the following quantities:

- For scheduled admissions ($j=SA$), we emphasize delays as well as mismatch and try to satisfy demand for general and intensive care throughout the day. We add to $c(\boldsymbol{f},\boldsymbol{g})$ the terms

$$c_{SA,GC}\sum_{t=1}^{T}\left(\sum_{j'\in GC}\sum_{\tau\in[t]} g^\tau_{SA,j'}-\sum_{j'\in GC}\sum_{\tau\in[t]} f^\tau_{SA,j'}\right)^+ + c_{SA,IC}\sum_{t=1}^{T}\left(\sum_{j'\in IC}\sum_{\tau\in[t]} g^\tau_{j,j'}-\sum_{j'\in IC}\sum_{\tau\in[t]} f^\tau_{SA,j'}\right)^+.$$

In a robust approach, the aggregation of flows over units and time reduces the power of the adversary, and so leads to less conservative solutions.

- For $j\in\{UV,T\}$, we only consider the overall demand for general and intensive care respectively, i.e., augment the objective with the terms

$$c_{UV/T,GF}\left(\sum_{j\in\{UV,T\}}\sum_{j'\in GC}\sum_{t\in[T]} g^t_{j,j'}-\sum_{j\in\{UV,T\}}\sum_{j'\in GC}\sum_{t\in[T]} f^t_{j,j'}\right)^+$$

and

$$c_{UV/T,IC}\left(\sum_{j\in\{UV,T\}}\sum_{j'\in IC}\sum_{t\in[T]} g^t_{j,j'}-\sum_{j\in\{UV,T\}}\sum_{j'\in IC}\sum_{t\in[T]} f^t_{j,j'}\right)^+.$$

- For inpatient units, we try to satisfy demand for intensive care across the entire time horizon and consider

$$c_{GC,IC}\left(\sum_{j\in GC}\sum_{j'\in IC}\sum_{t\in[T]} g^t_{j,j'}-\sum_{j\in GC}\sum_{j'\in IC}\sum_{t\in[T]} f^t_{j,j'}\right)^+.$$

- Finally, we only allow for discharges that are needed through constraints of the form

$$f^t_{j,DIS}=g^t_{j,DIS},\ \forall j\in GC\cup IC,\ \forall t\in[T]. \tag{2}$$

EC.2.5. Capacity constraints and decisions

In practice, unit capacities C_j^t depend on hospital managers' decisions. First, the total number of beds effectively available in a unit on a given day depends on the number of physical beds in this ward, but also on the number of nurses staffed in this ward on this day. Due to their strategic/tactical nature, we consider them as inputs to our optimization problem and, in our numerical experiments, set C_j^t equal to the effective capacity of unit j on that day.

Furthermore, the hospital might have an extra operational lever to increase capacity: In addition to licensed beds, $LicensedBeds_j$, each unit j has the possibility to accommodate more patients than the number of licensed beds, by using outpatient stretchers or placing some patients in the corridors. We refer to these options as extra or virtual beds respectively. Though undesirable, situations where virtual beds are used are not uncommon, especially in the middle of the day when newly admitted and soon discharged patients overlap. Hence, we could introduce decision variables v_j^t to indicate the number of virtual beds used for each unit $j \in [J]$ and each time $t \in [T]$, and associate some cost with using virtual beds. With this notation, the actual unit capacity C_j^t would be equal to $C_j^t = LicensedBeds_j + v_j^t$.

Also, all beds in the same service or ward are not equivalent. Indeed, some patients –with viral infections for instance– require beds in private rooms; shared rooms can usually be occupied by same-sex patients only. This heterogenity in resources can also be captured in our optimization model by considering assignment of patients to beds $z_{i,b}^t$ directly, at the expense of a higher number of decision variables. Alternatively, we can break down the bed capacity C_j^t into beds that can be assigned to male only $C_{j,male}^t$, female only $C_{j,female}^t$, or to all genders $C_{j,all}^t$. Then, the capacity constraints write as follows, for all units j, time t:

$$\begin{aligned}
\sum_i z_{i,j}^t &\leq C_{j,male}^t + C_{j,female}^t + C_{j,all}^t, \\
\sum_{i,\ i\ male} z_{i,j}^t &\leq C_{j,male}^t + C_{j,all}^t, \\
\sum_{i,\ i\ female} z_{i,j}^t &\leq C_{j,female}^t + C_{j,all}^t.
\end{aligned}$$

EC.3. Calibration of the machine learning models to predict future flows

In this section, we detail the data and machine learning models used to predict the different patient flows. We also present an extended formulation for the uncertainty set $\mathcal{U}$.

EC.3.1. Predicting inpatient flows

We have rich information about current inpatients from their electronic health records (EHRs). Following the approach from Bertsimas et al. (2021), we build individual risk scores to predict, on a daily basis (each day at 6am) and for each patient, the probability to be discharged by the end

of the day, and the probability to be in an intensive care unit (ICU). We summarize here the key steps of this predictive task.

Data: Patient-level EHR data about all inpatients admitted to the hospital between January 2017 and July 2019.

Training period: January 2017–April 2018.

Testing period: May 2018 –July 2018.

Prediction task: Probability to be discharged by the end of the day, $\mathbb{P}(\text{discharge for patient } i, \text{as of 6am})$, and the probability to be in an ICU by the end of the day, $\mathbb{P}(\text{ICU for patient } i, \text{as of 6am})$.

Out-of-sample accuracy: On both prediction tasks, we reach an Area Under the receiver operating Curve (AUC) of 0.810 and 0.973 respectively. Consequently, we predict the number of daily discharges with a median relative error (MRE) of 6.0% ($R^2 = 0.847$) and the ICU midnight census with an MRE of 11.1% ($R^2 = 0.998$).

Integration: The prediction models should be integrated within the hospital's EHR system and predictions computed daily at 6am.

Limitation: For predicting the need for intensive care, the main limitation is that we do not yet have access to data about needs. We only observe transfers to/out of the ICUs, which is a censored version of the quantity of interest. To circumvent this difficulty, we did not include logistical or operational covariates, such as the overall or unit census. Also, in the next section, we will allow for some variability around these predictions, which will alleviate the issue.

Connection with demand variables $\boldsymbol{g}$: We dissociate the prediction of daily volume with the intra-day distribution as described in Section 4.2. For daily volumes, we aggregate machine learning predictions at a hospital and unit level. Concerning intra-day distribution, we simply take the empirical distribution and compute average ratios β_t, the average fraction of discharges which occurred during $[t, t+1)$. As a result, one should expect

$$\sum_j g^t_{j,DIS} \Big/ \sum_j \sum_{t\in[T]} g^t_{j,DIS} \text{ to be close to } \beta_t.$$

Note that β_t depends on the day of the week and the hour of the day at $t = 0$. In contrast with overall discharge volume, we model intra-day distribution of discharges for the entire hospital only and not for each unit. Regarding the need for intensive care, we use a similar approach to predict overall volume. However, we do not try to control the intra-day distribution, which is mostly driven by bed placement decisions, i.e., our decision variables.

EC.3.2. Scheduled admissions

For scheduled admissions, we consider the latest schedule available at the beginning of each time period, and assume there is no uncertainty. Hence, we do not explicitly model the risk of cancellation in this version of the model. However, as soon as an admission is cancelled, it disappears from the schedule. In other words, our model adapts to the state of the schedule and cancellations. For surgeries, the schedule indicates the type of surgery performed, not the specific unit the patient should be admitted to afterwards. We automate this translation in the following way:

Data: All surgical cases from January 2012 until July 2019.

Connection with demand variables $\boldsymbol{g}$: To convert the schedule for surgeries into a sequence of future needs from surgical patients for the rest of the day, we use data on past surgeries and we associate each Current Procedural Terminology (CPT) procedure code with a hospital service and an empirical probability of needing an ICU after surgery.

Double counting of inpatients scheduled for surgery: Some of the scheduled surgeries involve patients who have already been admitted and who occupy a bed in an inpatient unit. For these patients, however, we have already developed models to predict their probability of discharge and needing an ICU in the previous section. Accordingly, there is a risk of counting these patients twice, as contributing both to demand from current inpatients and SA. Based on our discussion with our partner hospital, we adopt the following strategy to prevent double counting: Using historical data and decisions trees, we elicited a simple rule[2] that predicts whether a current inpatient scheduled for surgery will request a new bed with 85% accuracy. If we anticipate that a current inpatient will need a new bed assignment, then we consider that they will be discharged when surgery starts and will be readmitted after surgery as an SA. Otherwise, we assume that the patient will come back to their original bed after surgery and pretend that they will never leave their bed (which, as far as bed assignment is concerned, is safe to assume).

EC.3.3. Emergency department (ED) and unscheduled visits

By nature, future bed requests from the ED involve patients not physically present in the hospital yet, so no patient-level information is available to predict these flows. Other unscheduled visits can correspond to patients who have an existing relationship with the hospital. For instance, some outpatients can experience complications and need to be admitted as an inpatient after their procedure. In these cases, one could use information about the outpatient schedule and the patient to predict the risk of inpatient admission. For simplicity, and to limit the number of models to train and maintain, we do not adopt this approach here. Instead we propose to aggregate all bed requests coming from the ED and other unscheduled visits and consider all of them as "ED requests".

[2] Rule: If the patient is currently in an ICU, no reassignment is needed, otherwise a reassignment is needed if the empirical probability of needing an ICU after this CPT code is above 35%.

Data: All visits to the ED arrivals and their corresponding bed requests from November 2012 until July 2019. We did not consider the Clinical Decision Unit, a five-bed observation unit of the emergency room dedicated to overnight emergent patient and fully managed by ED staff, as an inpatient unit. We constructed features based on the date (month number, day of the week, weekend or holiday indicator), the hour of the day and previous workload (requests received on the same time period yesterday, same day last week, on average last week, on average on the same day of the week last month).

Training period: November 2012–May 2018.

Testing period: July 2018–July 2019.

Prediction task: At each hour of the day, the number of future bed requests received until the end of dat (EoD).

Out-of-sample accuracy: Linear regression model with ℓ_1-regularization achieves a median relative error of 14.0% ($R^2 = 0.907$). Other predictive methods such as decision trees achieved comparable but not significantly higher accuracy so we implemented the linear model.

Integration: The linear prediction models should be integrated within the hospital's IT system and hourly compute a predicted number of requests until 6am (EoD).

Connection with demand variables $\boldsymbol{g}$: From the linear regression model, we compute an estimate of the total number of requests from the ED, namely $\sum_{j'} \sum_{t\in[T]} g^t_{ED,j'}$. We force the distribution of these requests across hours of the day and across hospital units to match the distribution observed empirically in our training data. We only consider the last year of our training data to estimate the distribution of requests across units (or at least service/levels of care) since the hospital structure has changed between the beginning of our dataset (2012) and 2018-2019.

EC.3.4. Transfers

For outside transfers, we adopt the same approach as for unscheduled visits.

Data: All transfer requests received from November 2018 until July 2019. We excluded patients transferred to the emergency room for there is usually no decision (the patients are immediately accepted to the ED) and, if they eventually request an inpatient bed, it will be accounted for by the ED model. Similar to the ED model, we constructed features based on the date (month number, day of the week, weekend or holiday indicator), the hour of the day and previous workload (requests received on the same time period yesterday, same day last week, on average last week, on average on the same day of the week last month).

Training period: November 2018–April 2019. Note that, unfortunately, we could not collect this data prior to November 2018. To keep the size of the training data reasonable, we extended the training period in April 2019, hence overlapping with the simulation period used for validation..

Testing period: May 2019–July 2019.

Prediction task: At each hour of the day, number of future bed requests received until the EoD.

Out-of-sample accuracy: Linear regression model with ℓ_1-regularization achieves a median absolute error of 1.19 requests. The median relative error is fairly high (58.1%), though, since the total number of requests is sometimes relatively low (close to one).

EC.3.5. Lifted formulation of the uncertainty set

In Section 4.2, we constructed an uncertainty set to describe deviations between the actual bed requests $\tilde{g}^t_{j,j'}$ and their estimates $g^t_{j,j'}$. However, in some cases, especially for larger institutions, it might be more convenient to describe/predict bed requests for a medical specialty and/or level of care instead of target unit, i.e., consider uncertain vectors $h^t_{j,(s,\ell)}$ instead of $g^t_{j,j'}$. Again, in a robust framework, we can easily introduce such additional variables and add simple (linear) constraints, to ensure that they relate to the $g^t_{j,j'}$s appropriately.

We introduce variables, $h^t_{j,(s,\ell)}$, which encode for the number of patients who were in unit j during $[0,t)$ and then need to be in service s at level of care ℓ. Here, $\ell \in \{0,1\}$ equals one if the patient needs intensive care (IC), zero otherwise. $h^t_{j,(s,\ell)}$ better corresponds to how needs are expressed, whereas $g^t_{j,j'}$ captures how they materialize. We need constraints to ensure that these two descriptions of medical trajectories, from a unit and medical need perspective, match on multiple aspects.

- Discharges. For simplicity, we created a virtual unit for discharges alongside a virtual service and impose

$$g^t_{j,DIS} = h^t_{j,(DIS,0)} + h^t_{j,(DIS,1)}, \ \forall j,t.$$

- Level of care. Each unit is either an intensive care (IC) or general care (GC) unit. Consequently, the following holds, for any unit j and any time t,

$$\sum_{j' \in IC} g^t_{j,j'} = \sum_s g^t_{j,(s,1)},$$
$$\sum_{j' \in GC} g^t_{j,j'} = \sum_s g^t_{j,(s,0)}.$$

- Total flows. Namely, $\sum_{j'} g^t_{j,j'} = \sum_{(s,\ell)} h^t_{j,(s,\ell)}$ for all unit j and time t.

- Pre-assignment unit-service. As previously mentioned, each unit can serve a list of primary services and vice versa. Since we are considering *medical* trajectories, i.e., what should happen in an ideal world, we do not consider secondary or tertiary services. Consequently, for each need (s,ℓ), we have a list of designated primary units $\mathcal{U}(s,\ell)$. So, for this particular need, we should have for all origin j and time t,

$$h^t_{j,(s,\ell)} \leq \sum_{j' \in \mathcal{U}(s,\ell)} g^t_{j,j'}.$$

Symmetrically, each floor j' is associated with a list of medical needs it can serve, $\mathcal{N}(j')$, inducing the constraint

$$g^t_{j,j'} \leq \sum_{(s,\ell)\in\mathcal{N}(j')} h^t_{j,(s,\ell)}, \forall j, t.$$

EC.4. Supplement to the numerical experiments

We assess our approach on data collected between January 1 and August 24, 2019 at our partner hospital.

EC.4.1. Simulation procedure

We apply the following methodology: We specify a start date and a simulation length (in days).

We start our experiment at $h = 6$ o'clock on the start date and download all the available data about the hospital at that time, namely the list of all pending bed requests, current inpatients, and scheduled surgeries and admissions. We also run the predictive models from Section 4.1. We then construct and solve the optimization model, i.e., robust HIFO or its fluid approximation, obtaining a vector of first-stage variables $\boldsymbol{z}^1$. We virtually implement this vector, i.e., we update the location of all patients as prescribed by $\boldsymbol{z}^1$.

For the next time period, at h o'clock +2 hours, we similarly download all the data available at that time. However, this empirical data needs to be adapted to reflect the previous assignment decisions $\boldsymbol{z}^1$, and not the empirical decisions. We use the following update rules:

For waiting units (Scheduled Admissions, Unscheduled Visits, Transfers), we have the empirical lists of requests at h and $h+2$ o'clock, and the prescribed assignments $\boldsymbol{z}^1$. We construct the list of pending of bed requests at $h+2$ o'clock by combining:

- new requests, i.e., requests that were pending at $h+2$ o'clock but unavailable at h o'clock;
- pending requests, i.e., requests that were pending at h o'clock but were not granted a bed according $\boldsymbol{z}^1$.

For current inpatients, we consider three different cases:

- *Case 1: Inpatient at $h+2$ o'clock who was an inpatient at h o'clock.* If this patient needed a bed at h o'clock, then we place them in the location prescribed by $\boldsymbol{z}^1$, otherwise we place them in their empirical location at $h+2$ o'clock. Note that, in the latter case, if the location of the patient between h and $h+2$ changed, we apply this change, although it was not prescribed by the optimization. By doing so, our simulation procedures accounts for unplanned surgery or unplanned ICU admissions from the current inpatient population.
- *Case 2: Inpatient at $h+2$ o'clock who was not an inpatient at h o'clock.* This patient was historically admitted between h and $h+2$ o'clock. If, in addition, the patient was in a waiting unit at h o'clock, then we follow $\boldsymbol{z}^1$. On the contrary, if the patient was not in a waiting unit at h o'clock, then we apply the empirical decision.

• *Case 3: Inpatient at h o'clock who is no longer an inpatient at $h+2$ o'clock, i.e., discharged patient.* We follow the empirical decision and discharge the patient from the hospital.

We repeat these steps until reaching the predefined simulation length (e.g., one or three day days).

In Section 5.2, we run the different optimization formulations for $\lambda \in \{0,1,2,3,4,5,6,7,8,9,10\}$ and $\lambda_w \in \{0,1,2\}$.

EC.4.2. Additional comparisons of adaptive HIFO with the static policies

Figure 3 demonstrates a strong improvement from using optimization over the historical policy, and a further improvement from using adaptive decision rules when modeling future decisions in this multi-stage optimization problem under uncertainty.

First, we observe that, in our problem, uncertainty affects the feasibility of the decision variables $\boldsymbol{f}^t$, primarily through discharges and capacity constraints. Hence, there are two possible implementations for the SAA approach. We can either impose the constraints for all scenarios ('feasible almost surely'), or, less conservatively, we can impose them in expectation ('feasible on average'). We reported the performance of the former implementation in Figure 3. However, Figure EC.1 compares the two alternatives and displays no material difference in performance.

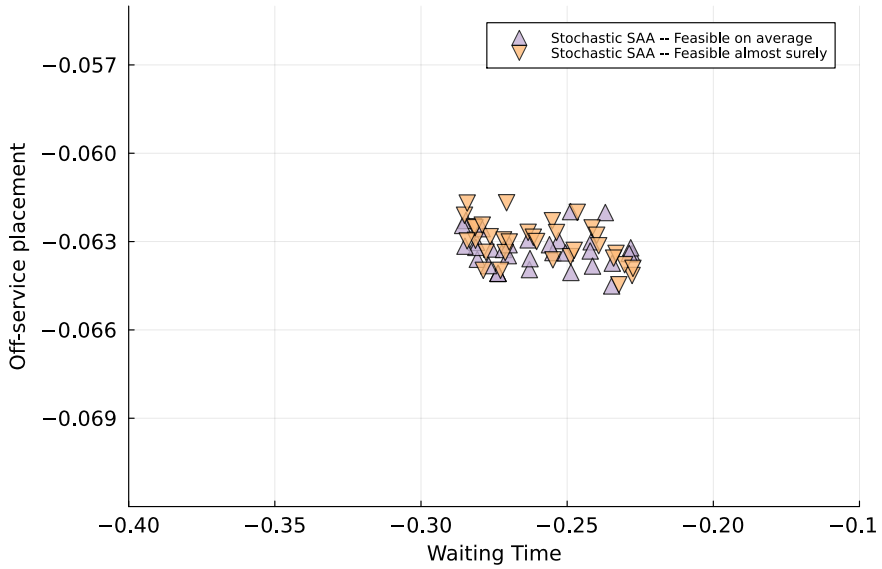


Figure EC.1 **Trade-off between waiting time and off-service placements for two alternative implementation of the SAA approaches.**

To further analyze why HIFO with affine decision rules uncovers even more efficient policies than static policies (Stochastic - SAA, Fluid approximation, HIFO with static decision rules), we display in Figure EC.2 the distribution of the hour of admission, for each class of policy. We observe that HIFO with affine decision rules admits more patients in the beginning of the day (between 10 a.m.

and 4 p.m.), which shows that, despite variability in future daily discharges, the affinely adaptive model is less conservative and better anticipates future discharges when making early admission decisions. Note that, for each policy class, the numbers displayed in Figure EC.2 are averaged over all policies within the class, as displayed in Figure 3. Accordingly, we observe more variance in the admission patterns for adaptive HIFO since the range of policies obtained is more varied than for the static approaches.

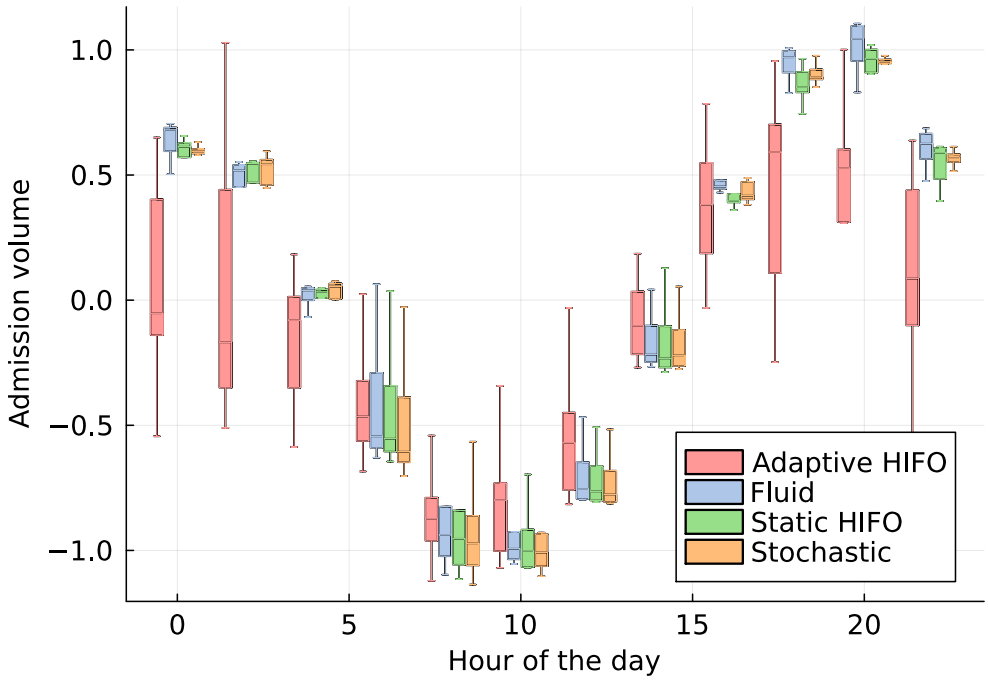


Figure EC.2 **Distributions of hour of admission for different optimization-based policies. Reported values are centered and scaled according the the mean and standard deviation of the empirical policy.**

We also look at individual patient decisions and identify the characteristics that explain most of the difference between the decisions made by the adaptive HIFO vs. the other three policies[3]. We measure the difference in decisions in terms of absolute difference in waiting time (centered and normalized according to the mean and standard deviation of the historical policy vs. affine HIFO) and predict it using information about the time (hour, day of the week) and the type of request (source channel, whether the patient will need an ICU). The resulting tree is displayed in Figure EC.3. Overall, we observe more discrepant decisions for transfer patients and on weekends.

[3] For this comparison, we consider one policy per policy class, namely the one with $\lambda = \lambda_w = 1$.

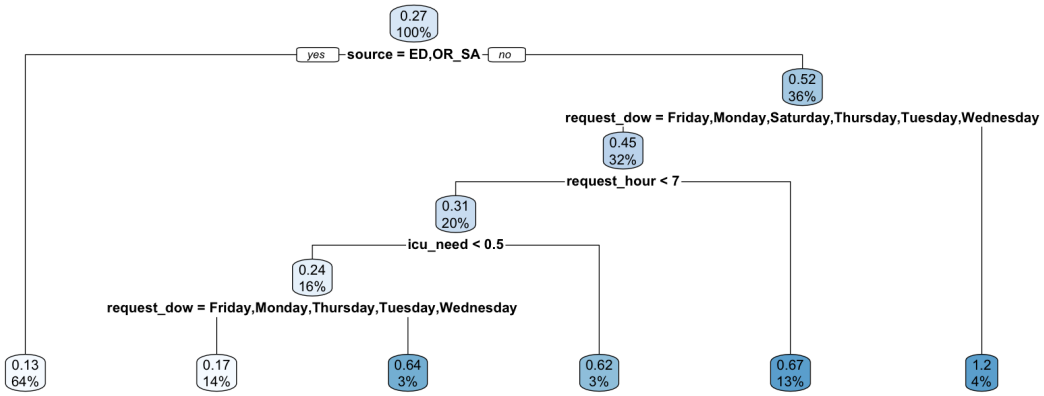


Figure EC.3 **Decision tree predicting the (standardized) absolute difference in waiting time of the static policies vs. adaptive HIFO.**

EC.4.3. Additional results for the affine with known recourse policy on $\{1,3,5,7\}$-day simulations

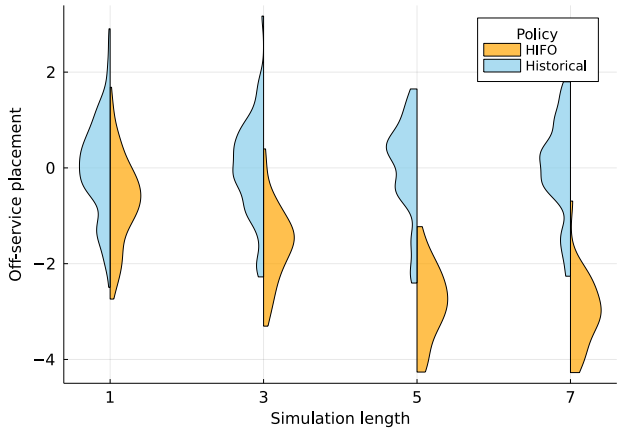
Figure EC.4 compares the distribution in the total number of off-service placements (Panel a), total waiting time of patients from UV (Panel b) and SA (Panel c), historically and using HIFO with known affine recourse, for different simulation lengths. Note that, for each simulation length and each metric, the values are normalized by the historical mean. We observe that, under HIFO, the distributions of these metrics are all shifted towards lower values, and that the improvement from HIFO increases as the simulation length increases.

To assess the robustness of our simulation strategy, we assess the impact of the choice of the starting day on the seven-day simulation results. We conduct simulations on all the possible seven-day periods in our dataset (229 possibilities) and report results in Figure EC.5, aggregated by which day of the week the first day of the period is. We observe consistent results in all metrics of interests, with waiting time from unscheduled visits being the most variable.

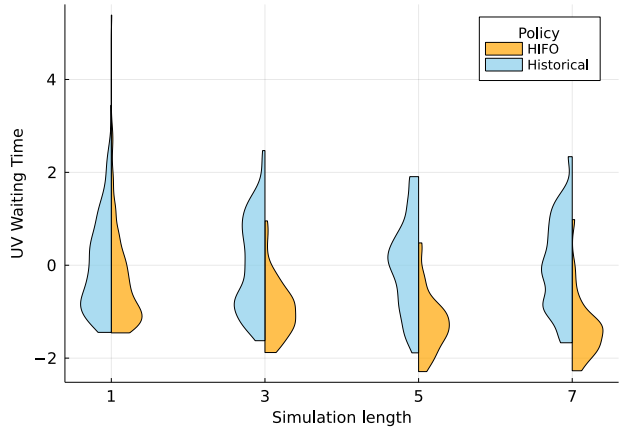
Finally, in Section 5.4, we discussed the fairness implications of implenting HIFO in practice. While Figure 5 reports the proportion of patients experiencing better, worse, identical, or not comparable outcomes for increasing simulation lengths, Table EC.2 reports detailed results for the one-day experiments.

EC.4.4. Quantification of the impact of HIFO on length of stay

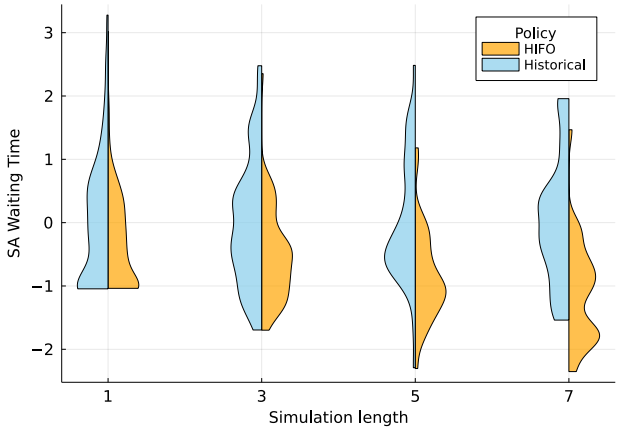
As demonstrated in Section 5, our robust optimization policy HIFO can effectively reduce the number of off-service placements. In turn, some empirical evidence suggests that off-service placement can have a negative (i.e., increasing) impact on length of stay. However, as summarized in Table EC.3, empirical studies are not unanimous on the matter. Accordingly, in our simulations, we



(a) Off-service placement



(b) UV Waiting Time



(c) SA Waiting Time

Figure EC.4 **Distributions of three performance metrics under the historical (blue left side) and HIFO (orange right side) bed assignment policy, for different simulation lengths (1, 3, 5, and 7 days). Reported values are normalized so that the historical policy achieves a mean of 1 on each metric.**

Table EC.2 **Comparison of individual bed assignment decisions from HIFO with the historical bed placements, on one-day simulations. A (+) (resp. (-)) indicates a strict improvement (deterioration). (=) indicates that both policies lead to the same outcome.**

Overall impact of HIFO	Waiting time	Service placement	Proportion of patients
Better	(+)	(+)	6.1%
	(+)	(=)	22.0%
	(=)	(+)	9.7%
Identical	(=)	(=)	35.0%
Not comparable	(+)	(-)	0.7%
	(-)	(+)	5.3%
Worse	(-)	(=)	20.9%
	(=)	(-)	0.1%
	(-)	(-)	0.2%

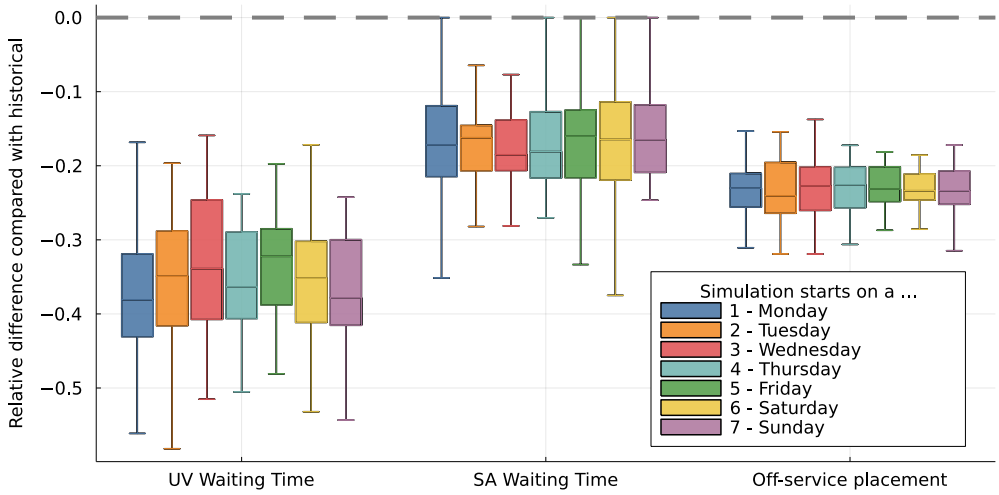


Figure EC.5 7-day simulations results depending on which day of the week the first day of the simulation period is.

adopted the conservative assumption that service placement has no impact on length of stay. We challenge this assumption in this section and try to quantify the impact of HIFO on length of stay reduction.

Table EC.3 Summary of the empirical evidence for the impact of off-service placement on resulting length of stay, mortality rate, and readmission rate. "No effect" indicates that the study found no significant effect. A blank indicates that the study did not consider this outcome.

Reference	Length of stay	Readmission rate	Mortality rate
Song et al. (2020)	Increased	Increased	No effect
Bai et al. (2018)			Increased
Stretch et al. (2018)			Increased
Stowell et al. (2013)	Increased	Increased	No effect
Liu et al. (2014)	No effect	No effect	No effect
Alameda and Suárez (2009)	Increased	No effect	No effect

As in Song et al. (2020), we posit that off-service placement has a multiplicative effect on length of stay. In the absence of rich patient-level information allowing rigorous econometric analysis, we conduct a stratified analysis and divide our patient population into six categories based on medical specialty (general medicine, surgery and surgical specialties, cardiology, oncology, neurology, other specialties). Let L_0 (resp. L_1) denote the average length of stay for in-service (resp. off-service) patients. Within each group, we estimate the mean and standard error of L_i, $i=0,1$, and assume that $L_i \sim \mathcal{N}(\mu_i, \sigma_i)$. Instead of using μ_1/μ_0, we adopt a conservative approach and use the 95th percentile of L_1/L_0 for each group, defined as

$$\max\left\{t \;:\; \mathbb{P}\left(\frac{L_1}{L_0} \geq t\right) \geq 0.95\right\} = \max\left\{t : \mathbb{P}\left(\mathcal{N}(0,1) \geq \frac{t\mu_0 - \mu_1}{\sqrt{t^2\sigma_0^2 + \sigma_1^2}}\right) \geq 0.95\right\},$$

which we compute by exhaustive search over a discretization grid of t values. Table EC.4 reports our conservative ratios of length of stay alongside the relative size of each group, i.e., the fraction of admissions belonging to each group. In particular, we find in our data that off-service placement increases length of stay by 9% on (weighted) average with heterogeneity across medical specialties (range: -23%; $+21\%$).

Table EC.4 Empirical estimates of the effect of off-service placement on hospital length of stay (LOS). We report the 95th percentile of the LOS ratio L_1/L_0 alongside the relative size of each service group in our partner hospital.

Group	Relative size	LOS inflation factor
General medicine	40%	1.10
Surgery and surgical specialties	28%	1.11
Cardiology	15%	1.21
Oncology	7%	0.77
Neurology	5%	0.86
Other specialties	5%	1.21

We then conduct simulation on the entire data set ($N = 212$ days). For patients whose service placement is different (but not necessarily better) under the historical and the HIFO policy, we compare the historical length-of-stay distribution with the expected new distribution based on the inflation factors from Table EC.4. In Figure EC.6, we do observe a minor global reduction in length of stay under HIFO, which could be due to the fact that most patients are admitted in the right medical specialty, that the impact of off-service placement on length of stay is limited, and that a large portion of patients are placed in the same service under both policies. Beyond length of stay, off-service placement might lead to increased mortality and readmission risk and put additional pressure on the medical staff, two issues that are hard to capture but should not be undervalued.

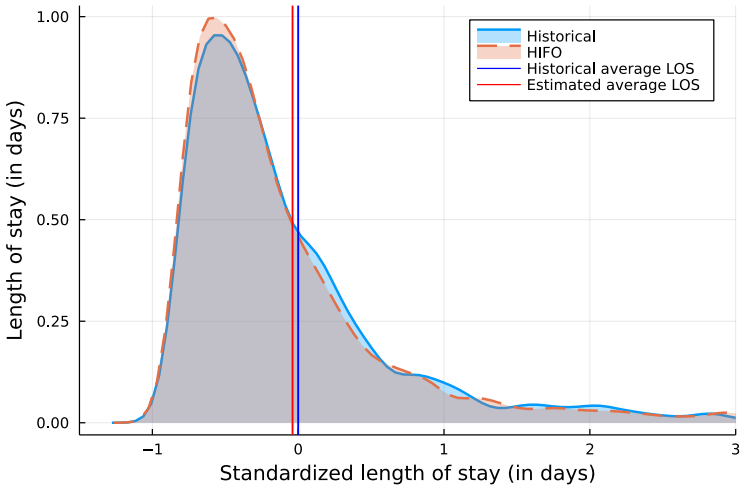


Figure EC.6 **Empirical length-of-stay distribution and estimated length-of-stay distribution under HIFO for the patients whose service placement differs in the two scenarios.**