MANAGEMENT SCIENCE
Vol. 00, No. 0, Xxxxx 0000, pp. 000–000
ISSN 0025-1909 | EISSN 1526-5501 | 00 | 0000 | 0001

INFORMS
DOI 10.1287/xxxx.0000.0000

Hospital-wide Patient Flow Optimization

Dimitris Bertsimas
Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA 02139, dbertsim@mit.edu

Jean Pauphilet
Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA 02139, jpauph@mit.edu

Our healthcare system operates under unbearable financial and operational stress. To improve quality of care and relieve clinicians and hospital staff, operations need to be coordinated and optimized across all services in real-time. We propose a multi-stage adaptive robust optimization approach combined with machine learning techniques to achieve this goal. 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. Based on historical data from a large academic medical center, we demonstrate that our optimization model can be solved in seconds for a 600-bed institution, reduces off-service placement by 33% on average, and boarding delays in the emergency department and post-anesthesia units by 30% and 19% respectively. We also illustrate the benefit from using adaptive linear decision rules instead of static assignment decisions. All together, holistic hospital optimization offers a unique opportunity to revitalize healthcare delivery with optimization and data at the core.

Key words: Hospital operations; flow management; machine learning; multi-stage 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 units to a global coordination scheme across the entire hospital. Indeed, Rutherford et al. (2017) identified five guiding principles, among which the utilization of advanced data analytics to "forecast patient demand patterns, and match capacity and demand" and "a system-wide approach to patient flow". These recommendations are easier said than done. In this paper, we propose a holistic optimization approach combined with machine learning techniques to achieve this goal: hospital-wide patient flow optimization. Based on historical data from a large academic hospital, we demonstrate that our approach can be implemented in a real-world environment and effectively reduces delays and patient misplacement.

1.1. Relevant literature

The operations research and management science (OR/MS) literature on healthcare systems is extensive and broad, both in terms of research questions and methods. Brandeau et al. (2004) showcase some applications of OR methods to healthcare environments, and notably to healthcare operations management. In an extensive survey, Hulshof et al. (2012) construct a taxonomy of the literature along three planning horizons (strategic, tactical and operational) and six care domains (ambulatory, emergency, surgical, inpatient, home and residential care). We refer to Lakshmi and Iyer (2013) for a more recent survey dedicated to applications of queueing methods in healthcare. Our intention here is not to be exhaustive but rather emphasize salient aspects of hospital operations and patient flow management which have received increased attention recently and motivate our work.

Armony et al. (2015) conducted one of the first data-based analysis of patient flows at a hospital level. As they acknowledge, "traditionally, hospital studies have focused on individual units, in isolation from the rest of the hospital; but this approach ignores interactions among units". They consider a sub-network of units composed of the emergency department (ED) and five inpatient units and compare historical patient flows (arrivals, waiting times, length of stays) with stochastic modeling.

A central performance metric in their exploratory data analysis is delays. Indeed, delays can be used as a measure of operational efficiency as well as quality of care. Empirical work suggests

that prolonged ED boarding time - the time needed for a patient in the ED to be admitted to an inpatient bed - is associated with negative health outcomes (Mathews et al. 2018, Chan et al. 2016b). Prolonged ED boarding time is usually due to unavailability of inpatient beds (Shi et al. 2016). Consequently, better understanding and modeling of discharge patterns are needed as well. Dai and Shi (2018) analyze and compare two service time models which have been recently proposed to capture non-stationarity in patient discharges: Shi et al. (2016) decompose service time onto two time scales to capture for inter- and intra-day non-stationarity. They separate the number of days a patient stays (LOS), which is primarily dictated by his/her medical condition, from the discharge (resp. admission) hour h_{dish} (resp. h_{adm}) which captures the discharge (resp. admission) process and its inefficiencies. Formally, $S = LOS + h_{dish} - h_{adm}$. In a different direction, Chan et al. (2016a) and Dong and Perry (2018) observe that discharge decisions are periodic events which can only occur once or twice a day and propose an inspection-delay service time model. Besides 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. Finally, Green (2008) surveys the potential for OR techniques in reducing hospital delays, with an emphasis on queueing models.

From a strategic and tactical point of view, adequate staffing levels (Green et al. 2006, Yankovic and Green 2011), refined ward dimensioning (De Bruin et al. 2010, Boulton et al. 2016), and optimized triage processes (Saghafian et al. 2012, 2014, Huang et al. 2015) have been proposed to reduce delays and improve overall operational efficiency. From an operational perspective, Chan et al. (2014) investigate speeding up the service rate in periods of congestion in the ED. He et al. (2019) propose an hybrid stochastic-robust optimization to dynamically allocation patients to physician in the ED so as to meet performance targets, such as the British four-hour target (Weber et al. 2011). A general insight of queueing theory is also that resource pooling might produce better performance. Due to heterogenity in patient needs however, Song et al. (2015) empirically find that pooling resources in the ED increases waiting time and overall length-of-stay. Indeed, even at a macroscopic scale, Kuntz et al. (2019) suggest that specialization of general hospitals might be more beneficial

than pooling. 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 unit designated for a different service than the service required by her condition. Using data from a large academic medical center, Song et al. (2019) find that off-service placement increases length-of-stay and readmission risk by 22.8% and 13.1% respectively. However, we should acknowledge the fact that all empirical studies on the matter do not corroborate their findings, as summarized in Table 1. Another related phenomenon is off-level placement, that is, when a patient needing an ICU is placed in a general care unit. Likewise off-service placement, evidence suggests that off-level placement is detrimental for the patient (see Table 1).

Table 1 Summary of the empirical evidence for the impact of off-service and off-level placement on resulting length-of-stay, mortality rate, and readmission rate. "No effect" indicates that the study found no significant effect. A blank indicates the study did not consider this outcome.

Question	Reference	Length-of-stay	Readmission rate	Mortality rate
Off-service	Song et al. (2019)	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
Off-level	Chan et al. (2018)	Increased	Increased	
	Kim et al. (2015)	Increased	Increased	No effect

As far as ED boarding is concerned, 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 and a Markov decision process model. Under admittedly simplifying assumptions, they prove that a threshold-based policy is optimal. Based on

such theoretical guarantees, they propose a heuristic policy which outperforms the generalized $c\mu$-rule widely used for routing in multi-class queueing systems (Cox and Smith 1991). On simulations calibrated with real-world hospital data, their policy reduces total cost by 14% and boarding time by 9%. However, their experiments includes only patients who were admitted to the ED for an initial diagnosis of either 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 and propose an approximate dynamic programming algorithm which effectively reduces the overflow proportion by 20% on a simulated hospitals with five services. However, as the authors acknowledge, their model is not realistic for it does not consider all units and cannot account for inter-unit transfers of inpatients.

Thompson et al. (2009) formulate a Markov decision process problem to allocate heterogeneous patients to inpatient beds dynamically. They have successfully implemented their solution in a 130-bed community hospital over an 18-day trial. Potentially, their solution could increase the hospital revenue by 1% and reduce boarding time by 50%. To do so, they divide the inpatient population into 12 clinical categories, assume each category is associated with a single primary unit and restrict the set of admissible policies accordingly. It remains unclear whether their approach could be generalized to larger institutions with a higher number of beds and a more diverse patient population. They also allow for inpatient reallocation, which might be hard to implement in practice. In a complementary direction, 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 have implemented a cloud-based solution for a large medical center which reduced ED request-to-bed-assignment times by 23%. However, their formulation is myopic and assigns patients to bed in an online fashion without forecasting of the future state of the hospital. Finally, (Meng et al. 2015) propose a distributionally robust optimization approach to optimize the scheduling of admissions for planned procedures, and protect this schedule against uncertainty in length-of-stay and unplanned admission. The problem we consider is complementary. They consider the tactical decision

of planning scheduled admissions for the next weeks, and use robust optimization as a framework to incorporate uncertain daily deviations from the schedule due ED patients. On the contrary, we consider the operational problem of dynamically admitting new patients (both scheduled and emergent), considering as uncertain future arrivals and requests. However, they only use static robust optimization while our work considers adaptive rules that, to the best of our knowledge, have not been applied in hospital operations yet.

Our present paper falls into this line of work but differs substantially in terms of scope and methodology: First, we adopt a holistic approach to optimize bed assignment decisions simultaneously for all inpatient units. To the best of our knowledge, no study has previously addressed the question in such breadth. Our approach not only accounts for admission of ED patients to inpatient beds but also for outside transfers, surgical patients boarding from the post-anesthesia care units (PACUs), and patient flows within inpatient units. Secondly, we build our analysis on data rather than stochastic modeling and distributional assumptions. Given the data rich environment that hospitals have become, we believe that queueing models are less meaningful, especially due to their stringent assumptions and the curse of dimensionality they suffer from.

1.2. Contributions and structure

Inspired by the words of Armony et al. (2015),

> "While theory has been the comfort zone of Operations Research [...], the situation dramatically differs when data is brought into the picture."

we propose a holistic optimization model to optimize bed assignment at a hospital level with data as the primitive. Our contributions can be summarized as follows:

1. We consider the entirety of the hospital and optimize patient flows at a system level, while previous work mostly focused on isolated units or a sub-network comprised of the ED and some inpatient wards.

2. We describe the location of each patient individually using integer decision variables, as opposed to stochastic queueing models which, at scale, rely on fluid model assumptions that dissolve individual movements into continuous flows. This distinction is relevant in practice because of tight capacity constraints which make each unit of capacity matter.

3. From a modeling perspective, we associate each patient with two locations, namely a physical location corresponding to the hospital unit she physically is, and a virtual location corresponding to the unit she *should be in* given her clinical need. Correspondingly, we can divide patient flows into two categories: Physical patient flows, which result from hospital management decisions to accept, place and discharge patients; Clinical flows, which are uncertain quantities. This perspective has the double advantage of being simple and dissociating the operational from the clinical decision-making process, the later being modeled as an uncertain quantity driven by each patient's condition rather than a decision variable.

4. In this framework, the optimal bed allocation decisions can be formulated as minimizing the mismatch between the physical flows (decisions) and the clinical ones (uncertainty). To the best of our knowledge, this formulation is novel, simple, and captures many operational deficiencies observed empirically such as boarding delays and off-service placement.

5. To account for uncertainty in the clinical trajectories, we integrate predictions obtained from machine learning techniques into an overall robust optimization formulation. While queueing models have been successfully used to model patient flows at a unit level, stochastic analysis of patients flows for hospital-wide bed assignment might be intractable due to non-stationarity of the arrival and departure processes, intricate network structure between the different units, soft pre-assignment rules of services to units, and high dimensions. In this work, we use outputs from machine learning models to build data-driven uncertainty sets for patients' clinical flows.

6. Finally, we demonstrate that our proposed formulation is tractable and leads to significant operational benefit. On data from a 600-bed medical center over 7 months, we solve the robust optimization problems in seconds and provide a bed assignment policy which reduces off-service placement by 33% on average, boarding delays in the emergency departments and post-anesthesia units by 30% and 19% respectively, while keeping overall occupation constant. We also demonstrate the additional benefit from using linear decision rules that allow for a more effective and flexible trade-off between waiting time and off-service placement.

Structure In Section 2, we provide a high-level description of our approach with some details on our partner hospital. Section 3 presents a nominal version of the holistic hospital optimization (H_2O) problem. In Section 4, we investigate the sources of uncertainty, detail the machine learning techniques used to anticipate future arrivals and departures, and provide a robust formulation for the H_2O problem. Finally, we assess the performance of our approach and the benefit from adaptive robust optimization on numerical experiments in Section 5.

2. Problem description and approach

In this section, we provide a high-level description of our model and approach. We first describe the main patient flows faced by a hospital on a daily basis and provide some characteristics on our partner hospital. We then discuss the single-stage problem of allocating beds to all pending requests and cover its multi-stage extension.

2.1. Patient flows at a large academic hospital

Figure 1 sketches the main patient flows into, in and out of the hospital, within a day. New requests for beds can either come from surgeries or scheduled admissions (OR), the emergency department (ED) or transfers from another institution (T). We refer to these units as *waiting units*. Current inpatients are dispersed into the different *inpatient units*, each of them corresponding to different services and levels of care. We consider here two levels of care, namely intensive care (IC) and general care (GC). Inpatient can either move to another inpatient unit - usually to change the level of care - or be discharged out of the hospital. For simplicity, we consider a single virtual *discharge unit* (D), although there are many potential discharge destinations such as home or rehabilitation clinics.

Note that, though stylized, this representation can capture almost all relevant patient flows within a day. For example, we do not explicitly allow patient flows from inpatient units to surgery - there is no edge from (IC)/(GC) to (OR). Yet, inpatients who are scheduled for surgery do fit in our model. If the inpatient needs a new bed after surgery, we can model it as a planned discharge followed by a readmission from (OR). On the other hand, if her procedure does not call for a change in bed, we can consider that the patient does not move.

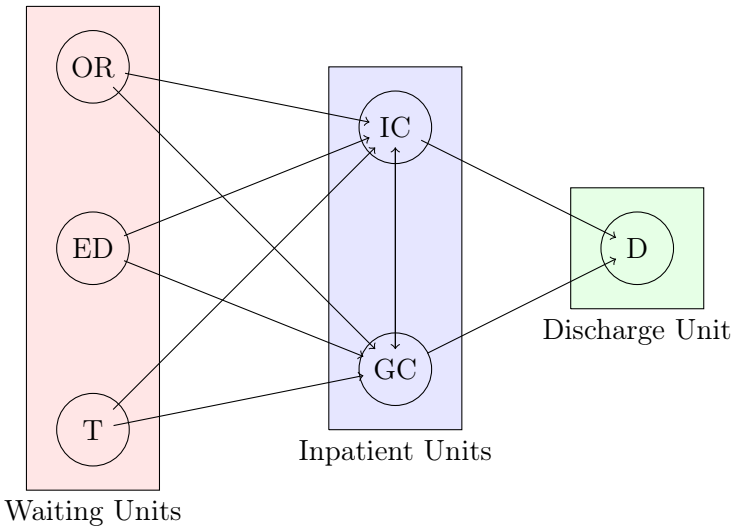

Figure 1 Schematic views of patient flows in a typical hospital.

We validated this flow description with our partner hospital and believe it to be fairly accurate and general. Our partner hospital is a large academic medical center with nearly 600 licensed beds. In 2018, it received 54,696 visits to the ED, admitted nearly 40,756 inpatients and performed 26,632 surgeries (51 % ambulatory). We excluded from our analysis Psychiatry and Obstetrics units because they manage a very distinct patient population and operate separately from the rest of the hospital. Finally, from a terminology perspective, 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 comprised of rooms and beds. Our partner hospital, for instance, has 15 services spread over 31 units, divided onto two campuses. Table 2 lists some of these units, alongside their main characteristics. The full list of units is given in Appendix EC.1, Table EC.1.

2.2. Single-stage problem: individual bed assignments

At each point in time, the hospital knows exactly which beds are available. Simultaneously, the hospital accumulates bed requests from patients in a waiting unit, as well as current inpatients needing a new bed. In short, for each patient, we know her current location in the hospital and her need. Consequently, we can associate a patient with a cost of being assigned to each unit and solve a matching problem to assign patients to units at minimal cost while satisfying capacity constraints.

For simplicity of presentation here, we restricted a "clinical need" to the combination of a hospital service and a level of care (general or intensive) and considered assigning patients to units. In practice

Table 2 Description of inpatients units on campus E at our partner hospital. The type is either GC (General Care) or IC (Intensive Care). The unit name follows the nomenclatura: Campus-Type Number.

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

however, we implement a more granular version of the model to assign patients to beds during the first stage, taking into account gender, isolation status due to infection, need for telemetry.

2.3. Multi-stage problem: optimal patient flows

The single-stage deterministic matching problem is inherently myopic and does not consider future capacity and requests. This is a major drawback and can lead to suboptimal decisions. For instance, it can lead to using all available beds to accommodate outside transfer requests while critical cardiac surgery patients are expected in a few hours - a situation we witnessed ourselves at our partner hospital. Therefore, we consider a multi-period extension.

We use a folding horizon formulation that optimizes bed assignment across the entire hospital from now on until 6 am. The rationale behind the folding horizon is that bed requests cannot be left pending 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 from postponing some bed assignment decisions indefinitely. Yet, we acknowledge the fact that some decisions such as admissions of outside transfers or OR (re)scheduling decisions impact the hospital on a weekly rather than daily time scale and that a longer or a rolling horizon might be better suited. We

shall incorporate these decisions and investigate the issue in further work. We use 6 am instead of midnight as the end of the optimization period for it usually corresponds to a low activity point and a transition time between night and the next day shifts. We divide the optimization horizon into two-hour time periods in order to continuously optimize bed assignments throughout the day while leaving time for the care teams to actually implement the assignment decisions.

The extension from single to multiple stages then follows by substituting knowledge of current capacity and bed requests by predictions on the future. For inpatients, Bertsimas et al. (2020) developed machine learning models to predict inpatients flows within and out of the hospital, with $80-90\%$ accuracy. We implemented a similar end-to-end solution that computes daily estimates of the probability of discharge and ICU need in the next 24 hours and integrated it into the EHR system of our partner hospital. These estimates are patient-specific risk scores. As a result, they take into account the particular patient mix of the hospital on a given day and can be aggregated at a unit or a hospital level. For the waiting units, we built predictive models to predict future bed requests from all the aforementioned sources: For the ED, we trained a regression model to predict the number of future bed requests for the whole optimization horizon. We then use historical data to anticipate how these requests will be distributed across the day. We adopted a similar approach for transfer patients. For surgical patients, we know the surgeries scheduled for the rest of the day. Using historical data, we associate each procedure code with a hospital service and a probability of needing an ICU. Hence, we convert the schedule into a sequence of need requests. We assume there is no uncertainty about bed requests from future surgeries. This is not overly restrictive since most unplanned surgeries are due to ED admissions so the resulting bed needs are already captured in the ED model. We include non-surgical scheduled admissions in this category.

The main difference between the single- and multi-stage problems is the level of granularity for the decision variables. For the single-stage model, we use a decision variable for each individual patient. For the multi-stage problem on the other hand, we consider aggregated patient flows. In our opinion, these variables correspond to reasonable approximations of the decisions to be taken in the future, while leading to a tractable optimization problem.

3. Mathematical formulation of the nominal problem

We now provide a formal formulation for the holistic hospital optimization problem (H_2O), following a similar outline: First, we consider the problem of assigning each patient to a unit for the first time period, taking into account available bed capacities and known clinical needs. Then, we consider the multi-stage problem of optimizing patients flows across the rest of the day. In this section, we will assume that future bed requests and discharges are fully known. We will relax this assumption in the next section.

3.1. Single-stage problem: individual unit assignments

Decision variables: Given information on current location and need, we decide on the location of each patient at time $t=1$. To do so, we introduce binary variables z^1_{ij}, indicating whether patient i is in unit j at time $t=1$, for each patient $i=1,\dots,I$ and for each unit $j=1,\dots,J$.

Constraints: The decision variables $\boldsymbol{z}^1$ must satisfy a set of constraints, concisely denoted $\boldsymbol{z}^1 \in \mathcal{Z}$, such as:

(Z1) Integrality: $z^1_{ij} \in \{0,1\},\ \forall i=1,\dots,I, j=1,\dots,J.$

(Z2) Single-location: $\sum_j z^1_{ij} = 1,\ \forall i=1,\dots,I.$

(Z3) Unit capacity: $\sum_i z^1_{ij} \le C^1_j,\ \forall j=1,\dots,J.$

(Z4) Forbidden moves: Given the patient's initial location, some moves might not be possible. For instance, as depicted on Figure 1, a patient currently in (ED) cannot move to the (OR), or a patient in the discharge unit cannot be readmitted to an inpatient ward. In short, for each patient, we have a number of forbidden locations $\mathcal{F}_i$, i.e., $z^1_{ij}=0,\ \forall j \in \mathcal{F}_i$.

Objective: In this framework, the overall unit assignment problem can be formulated as finding the physical locations $\boldsymbol{z}^1$ which are as close as possible to the clinical needs, namely solve

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

Rule for computing c_{ij}: The assignment cost should depend on the patient's current location and her need. We apply the following rule to compute c_{ij}. For each patient i, we define her *clinical need* as the combination of hospital service s and level of care ℓ that her condition requires. The level

of care $\ell \in \{0,1\}$ is a binary variable equal to one if the patient requires intensive care (IC), and zero, otherwise. If the patient has no particular need, namely if no bed request has been placed for the patient, then we set $c_{ij} = 0$ if patient i is in unit j at $t = 0$ and $c_{ij} = \infty$, otherwise, so that the patient does not move from her current location. Otherwise, c_{ij} captures the quality of the match in terms of hospital service and level of care. Each unit is pre-assigned a primary service as well as secondary services as described in Table 2. Correspondingly, we define a priority level p_{js} between each unit j and service s. For primary services, $p_{js} = 1$, for secondary services, $p_{js} = 2$, and so on. If (s, ℓ) is the clinical need of patient i, we decompose the cost c_{ij} into

$$c_{ij} = p_{js} - 1 \qquad \text{(Service match)}$$

$$+ \begin{cases} 3, & \text{if patient } i \text{ needs an ICU and unit } j \text{ is not an ICU,} \\ 2, & \text{if patient } i \text{ does not need an ICU and unit } j \text{ is an ICU,} \\ 0, & \text{otherwise.} \end{cases} \qquad \text{(Level of care match)}$$

Depending on the particular hospital, we might include a term to capture physical distance, especially if units are located in different buildings. We also can add a penalty depending on how long the patient has been waiting for a bed.

Problem size: A large academic hospital like our partner has about $J \approx 30$ units and $I \approx 600$ patients. This problem has at most $I \times J \approx 18,000$ binary variables. Besides the integrality constraints (Z1), the forbidden moves constraints (Z4) are imposed by design by defining variables z^1_{ij} only for $j \notin \mathcal{F}_i$, leaving only $I + J \approx 630$ constraints.

3.2. Multi-stage problem: optimal patient flows across units

We now introduce the multi-stage problem of optimizing patients flow across units over the entire day. We adopt a folding horizon approach, optimizing patient flows until 6 am and decomposing the day into two-hour time periods.

First attempt: The previous single-stage model could naturally be extended to a multi-stage setting. Denoting by $\boldsymbol{z}^t \in \{0,1\}^{I \times J}$ the vector of patients' physical locations at time $t = 1, \ldots, T$,

and $\boldsymbol{y}^t$ the vector of their clinical locations, one could try to minimize the distance between the physical and clinical locations, i.e., solve

$$\min_{\boldsymbol{z}\in\mathcal{Z}} \sum_t \sum_{i,j,} c(y_i^{t-1}) z_{ij}^t.$$

Drawbacks of this approach are twofold.

First, its tractability. At first sight, it involves at least $I \times J \times T \approx 180,000$ binary variables. Actually, some constraints, such as the forbidden moves constraints, are more naturally expressed in terms of flows rather than locations, so one would need to introduce $I \times J^2 \times T \approx 5,400,000$ variables to linearize the individual flows $z_{ij}^t z_{ij}^{t+1}$. Ten billion variables is far beyond what current solvers can handle and solve in a reasonable amount of time. Remember that we consider a multi-stage problem with 2-hour time periods, so to be practically useful and implementable, the overall optimization problem should be solved in a limited fraction of the 120 minutes available.

Second, its idealism. It is idealistic to assume one can decide in advance of the physical path in the hospital of each individual patient, whether there are currently here or expected to come. For planning purposes, relevant variables should be at a unit rather than individual level. For instance, at 8 am in the morning, "one expects x discharges from unit j between 4 pm and 6 pm" is a more realistic statement than "one expects patients x, y, z to be discharged between 4pm and 6pm" - and an equally useful one. Consequently, we will consider unit-level variables in the multi-stage setting.

Assumption: To further justify the shift from individual to aggregated decision variables, we assume that, over the entire time horizon, each patient should move at most once, both physically and clinically. Under this assumption, we can divide patients flows depending on the unit they originate while keeping some level of tracability on individual patients.

Decision variables: We introduce an integer variable, $f_{j,j'}^t$, indicating the number of patients who moved from unit j to unit j' during the period at time t, for each time $t = 0, \ldots, T$ and each pair of units $j, j' = 1, \ldots, J$. Note that because of the single-move assumption, these patients were in unit j during $[0, t)$.

Constraints: Flow variables have to satisfy the following constraints:

(F1) Integrality: $f^t_{j,j'} \in \mathbb{N}$.

(F2) Flow conservation: If j is an inpatient unit, $\sum_{t,j'} f^t_{j,j'} = z^0_j$, while if j is a waiting unit, $\sum_{t,j'} f^t_{j,j'} \geq z^0_j$ to account for future arrivals.

(F3) Forbidden moves: Those constraints are similar than the individual moves. We also add the constraints

$$f^t_{j,j'} = 0, \ \forall t = 0, \dots, T-1, j = 1, \dots, J,$$

to ensure patients who do not move during the day are all captured in $f^T_{j,j'}$.

(F4) Capacity constraints:

$$z^0_j + \sum_{s=0}^{t-1} \sum_{j'} f^s_{j,j'} - \sum_{s=0}^{t-1} \sum_{j'} f^s_{j,j'} \leq C^t_j, \ \forall j, \ \forall t = 0, \dots, T.$$

Objective: Similarly, we denote by $g^t_{j,j'}$ the number of patient who needed to be in unit j during $[0, t)$ and then need to be in unit j'. Assuming for the moment that these needs are perfectly known, an objective might be to minimize the number of patients from each unit j who requested unit j' and could not be assigned to it, i.e., the quantity

$$\left(\sum_{t=0}^{T-1} g^t_{j,j'} - \sum_{t=0}^{T-1} f^t_{j,j'} \right)_+,$$

where $(x)_+ := \max(x, 0)$ denotes the positive part of x. More precisely, we decompose the cost $c(\boldsymbol{f}, \boldsymbol{g})$ into the following quantities:

- For scheduled admissions ($j = OR$), 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

$$c_{OR,GC} \sum_{t=1}^{T} \left(\sum_{j' \in GC} \sum_{s=0}^{t-1} g^t_{j,j'} - \sum_{j' \in GC} \sum_{t=0}^{t-1} f^t_{j,j'} \right)_+ + c_{OR,IC} \sum_{t=1}^{T} \left(\sum_{j' \in IC} \sum_{s=0}^{t-1} g^t_{j,j'} - \sum_{j' \in IC} \sum_{t=0}^{t-1} f^t_{j,j'} \right)_+.$$

In anticipation of the following section, let us remark that in a robust approach, the aggregation of flows over units and time reduces the power of the adversary, hence leads to less conservative solutions.

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

$$c_{WU,GF}\left(\sum_{j\in\{ED,T\}}\sum_{j'\in GC}\sum_{t=0}^{T-1} g^t_{j,j'} - \sum_{j\in\{ED,T\}}\sum_{j'\in GC}\sum_{t=0}^{T-1} f^t_{j,j'}\right)_+,$$

and

$$c_{WU,IC}\left(\sum_{j\in\{ED,T\}}\sum_{j'\in IC}\sum_{t=0}^{T-1} g^t_{j,j'} - \sum_{j\in\{ED,T\}}\sum_{j'\in IC}\sum_{t=0}^{T-1} f^t_{j,j'}\right)_+.$$

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

$$c_{GC,IC}\left(\sum_{j\in GC}\sum_{j'\in IC}\sum_{t=0}^{T-1} g^t_{j,j'} - \sum_{j\in GC}\sum_{j'\in IC}\sum_{t=0}^{T-1} f^t_{j,j'}\right)_+.$$

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

$$f^t_{j,D} = g^t_{j,D},\ \forall j \in GC \cup IC,\ \forall t = 0, \dots, T. \tag{D}$$

All in all, the multi-stage problem can be written

$$\begin{aligned}
\min_{\boldsymbol{f}\in\mathcal{F}}\quad & c(\boldsymbol{f}, \boldsymbol{g}) \\
\text{s.t. } & z^0_j + \sum_{s=0}^{t-1}\sum_{j'} f^s_{j,j'} - \sum_{s=0}^{t-1}\sum_{j'} f^s_{j,j'} \le C^t_j,\ \forall j, t && \text{(F4)} \\
& f^t_{j,D} = g^t_{j,D},\ \forall j\ \forall j \in GC \cup IC,\ \forall t. && \text{(D)}
\end{aligned}$$

where $\mathcal{F}$ denotes the set of flows $\boldsymbol{f}$ satisfying the integrality, flow conservation and forbidden moves constraints (F1-3), and $c(\boldsymbol{f}, \boldsymbol{g})$ is a piece-wise linear convex function in both the decision variables $\boldsymbol{f}$ and the clinical needs $\boldsymbol{g}$.

Problem size: This problem has at most $T \times J^2 \approx 9,000$ integer variables. Besides the integrality constraints (F1) and the forbidden moves constraints (F3), there are $J + J \times T \approx 330$ constraints.

3.3. Final formulation: Holistic Hospital Operations (H_2O)

Finally, combining the first-stage and the multi-stage problem, we obtain the Holistic Hospital Optimization formulation (H_2O),

$$\min_{\boldsymbol{z}^1\in\mathcal{Z},\boldsymbol{f}\in\mathcal{F}} \sum_{i,j} c_{ij} z^1_{ij} + \lambda\, c(\boldsymbol{f},\boldsymbol{g}) \quad \text{s.t.} \quad f^0_{j,j'} = \sum_{i=1}^{I} z^1_{ij'} z^0_{ij},\ \forall j \neq j', \tag{C}$$

$$z^0_j + \sum_{s=0}^{t-1}\sum_{j'} f^s_{j,j'} - \sum_{s=0}^{t-1}\sum_{j'} f^s_{j,j'} \leq C^t_j,\ \forall j,t, \tag{F4}$$

$$f^t_{j,D} = g^t_{j,D},\ \forall j,t, \tag{D}$$

where $\lambda > 0$ controls the degree of foresight. The coupling constraints (C) ensure that individual and aggregated variables agree on the first time period. The binary vector z^0_{ij} encodes the physical location of patient i at $t = 0$. The above formulation contains all the fundamental components of the complete formulation we actually implemented, which considers capacity decisions and some individual trajectories, z^t_{ij} for $t = 1,\ldots,T$ as well. We discuss these extensions and some implementation considerations in Appendix EC.2.

Until now, we assumed clinical flows $\boldsymbol{g}$ were known, while in practice, they are not. Nonetheless, they should obey some constraints analogous to the ones satisfied by physical flows and can be partially predicted using machine learning techniques. In the following section we incorporate the physics of the problem and predictions from machine learning into the optimization formulation to account for this uncertainty.

4. Uncertainty on clinical trajectories

To make appropriate decisions in the presence of uncertainty, two aspects ought to be taken into account: the degree to which the uncertainty can be anticipated and the inherent noise in any estimate. In this section, we describe how we combine machine learning techniques and robust optimization methodology to capture both predictability and variability of clinical flows $\boldsymbol{g}$ and enrich the optimization problem (H_2O).

4.1. Predictability: Machine learning at the rescue

We first use historical data and machine learning techniques to construct predictive models for key clinical flows. In accordance with the single-move assumption and our modeling framework, we consider the patients based on their original unit separately. Table 3 summarizes the main out-of-sample predictive power of our methods.

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
ED	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

4.1.1. Inpatients We have rich information about current inpatients from their EHRs. Following the approach from Bertsimas et al. (2020), we built individual risk scores to predict, on a daily basis (each day at 6 a.m.) and for each patient, the probability to be discharged by the end of the day and the probability to be in an ICU. We summarize here the key steps of this predictive task.

Data: Patient-level EHR data about all inpatient admitted at 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, $\mathbb{P}(\text{discharge for patient } i, \text{ as of 6 a.m.})$, and the probability to be in an ICU by the end of the day, $\mathbb{P}(\text{ICU for patient } i, \text{ as of 6 a.m.})$.

Out-of-sample accuracy: On both prediction task, 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 intensive care midnight census with an MRE of 11.1% ($R^2 = 0.998$).

Implementation: The prediction models are integrated within the EHR system of the hospital and daily compute these predictions, every day at 6 am.

Limitation: For predicting the need for intensive care, the main limitation is that we do not yet have access to data about needs and 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 clinical flows: We dissociate the prediction of daily volume with the intra-day distribution. For daily volumes, we aggregate machine learning predictions at a hospital and unit level. For instance,

$$\sum_{i \text{ inpatient at t=0}} \mathbb{P}(\text{discharge for patient } i, \text{ as of } t=0) \text{ is an estimate of } \sum_j \sum_{t=0}^{T} g_{j,D}^t,$$

$$\sum_{i \text{ in unit } j \text{ at t=0}} \mathbb{P}(\text{discharge for patient } i, \text{ as of } t=0) \text{ is an estimate of } \sum_{t=0}^{T} g_{j,D}^t.$$

We now connect $\mathbb{P}(\text{discharge for patient } i, \text{ as of } t = 0)$ with the output of our model, $\mathbb{P}(\text{discharge for patient } i, \text{as of 6 a.m.})$, i.e., the probability that the patient will be discharged by

the end of the day as of 6 am. If it is h o'clock at the beginning of the time period ($t = 0$), then $\mathbb{P}(\text{discharge for patient } i, \text{ as of } t=0) = \mathbb{P}(\text{discharge for patient } i, \text{as of } h \text{ o'clock})$. We then simply estimate the right-hand side of the equality by $\alpha_h \mathbb{P}(\text{discharge for patient } i, \text{as of 6 a.m.})$, where the scaling factor α_h is calibrated empirically on the training set. Intuitively, $\alpha_h \mathbb{P}(\text{discharge for patient } i, \text{as of 6 a.m.})$ is the updated probability for patient i to be discharged conditioned on the fact that she is still at the hospital at h o'clock. Typically, α_h is equal to 1 before 8 am and close to 0 after 10 pm. Concerning intra-day distribution, we simply take the empirical distribution and compute average ratios β_t, the average fraction of discharges which occutred during $[t, t+1)$. As a result, one should expect

$$\sum_j g^t_{j,D} \Big/ \sum_j \sum_{t=0}^{T} g^t_{j,D} \text{ to be close to } \beta_t.$$

Note that β_t depends on the day of the week and the hour of the day at $t = 0$. Also, we model intra-day distribution of discharges for the entire hospital only, not for each unit. This decomposition scheme into discharge volume and intra-day distribution resembles the patient-level two-time-scale model of Shi et al. (2016), who decompose overall length-of-stay into days and hours. 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. Because of censorship issues, we believe that the empirical distribution of moves into and out of the intensive care units does not reflect the actual distribution of when those needs were expressed.

4.1.2. Scheduled surgeries and admissions Since we do not include OR scheduling decisions - beds are the only resource we consider in this paper - we assume there is no uncertainty about future surgeries. Indeed, most of unplanned surgeries are due to ED admissions and the bed needs emerging from the ED are captured in a dedicated model. We thus focus our attention to the scheduled cases.

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

Connection with clinical flows: To convert the schedule for surgeries into a sequence of future needs from OR patients for the rest of the day, we use data on past surgeries and we associate each procedure with a hospital service and an empirical probability of needing an ICU after surgery. If a current inpatient (who already has a bed in an inpatient unit) is scheduled for surgery, depending on the surgery type, she might either come back to her original bed after surgery (in which case, it makes no difference in our model to assume the patient will never leave her bed) or need a new bed assignment (in which case we consider the patient will be discharged when surgery starts and readmitted after surgery). Using decisions trees, we elicited a simple rule that explains 85% of the past reassignment decisions made by the hospital.

4.1.3. Emergent requests By nature, future bed requests from the ED involve patient not physically present in the hospital yet. So, little side information is available to us.

Data: All visits to the ED arrivals and their corresponding bed requests from November 2012 until July 2019. We excluded from the bed requests the patient admitted to the Clinical Decision Unit, a 5-bed section of the emergency room dedicated to overnight emergent patient and fully managed by ED staff. 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, number of future bed requests received until 6 am.

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.

Implementation: The linear prediction models are integrated within the IT system of the hospital and hourly computes a predicted number of requests until 6 am.

Connection with clinical flows: From the linear regression model, we compute an estimate of the total number of requests from the ED, namely $\sum_{j'} \sum_{t=0}^{T} g^t_{ED,j'}$. As for discharges, we use empirical data to estimate intra-day distribution of these requests along the day.

4.1.4. Outside transfers For outside transfers, we adopt the same approach as for the ED.

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.

Testing period: May 2019 - July 2019.

Prediction task: At each hour of the day, number of future bed requests received until 6 am.

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 1).

4.2. Variability: adaptive robust optimization at the rescue

Despite their accuracy, our predictive models cannot overcome the inherent variability in clinical flows, which we account for by adopting a robust optimization approach. We now characterize a so-called uncertainty set $\mathcal{G}$ and impose $\boldsymbol{g} \in \mathcal{G}$.

Physics: We first describe structural constraints which are similar to the ones satisfied by physical flows. Namely,

- Integrality on $g^t_{j,j'}$.
- Consistency with known needs at $t = 0$.
- Flow conservation, namely, if j is an inpatient unit, $\sum_{t,j'} g^t_{j,j'} = z^0_j$, while $\sum_{t,j'} g^t_{j,j'} \geq z^0_j$ for waiting units to account for future arrivals.
- Forbidden moves, such as $g^t_{ED,OR} = 0$ (imposed by design).

Predictions: We then include constraints based on the outputs from our machine learning models.

- Discharges. Given the estimate for the total number of discharges, $\hat{dis}$, we impose the following constraints

$$\lfloor (1-\varepsilon)\hat{dis} \rfloor \leq \sum_j \sum_{t=0}^{T} g_{j,D}^t \leq \lceil (1+\varepsilon)\hat{dis} \rceil, \quad \text{(overall volume)}$$

$$\lfloor \beta_t(1-\varepsilon)\hat{dis} \rfloor \leq \sum_j g_{jD}^t \leq \lceil \beta_t(1+\varepsilon)\hat{dis} \rceil. \quad \text{(intra-day distribution)}$$

We impose similar constraints for the discharge volume at a unit level.

- Intensive care. Given the estimate of the total number of intensive care patients, $\hat{icu}$, we impose the bounds

$$\lfloor (1-\varepsilon)\hat{icu} \rfloor \leq \sum_j \sum_{j' \in IC} \sum_{t=0}^{T} g_{j,j'}^t \leq \lceil (1+\varepsilon)\hat{icu} \rceil,$$

and as a unit level as well.

- ED/Transfer requests. We have an estimate on the total number of bed requests from the ED, $\hat{ed}$ and impose volume and intra-day constraints. Likewise for outside transfers.

- OR schedule. For the scheduled surgeries and admissions, the clinical flows are known and fixed.

Actually, we know the needs in terms of service and level of care, not in terms of target unit. So, we introduce extra variables, $h_{j,(s,\ell)}^t$, 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. We detail this extended formulation in Appendix EC.3.

4.3. Robust counterpart and affine adaptivity

Recall the nominal problem

$$\min_{\boldsymbol{z}^1 \in \mathcal{Z}, \boldsymbol{f} \in \mathcal{F}} \sum_{i,j} c_{ij} z_{ij}^1 + \lambda\, c(\boldsymbol{f}, \tilde{\boldsymbol{g}}) \quad \text{s.t.} \quad f_{j,j'}^0 = \sum_{i=1}^{I} z_{ij'}^1 z_{ij}^0, \ \forall j \neq j', \tag{C}$$

$$z_j^0 + \sum_{s=0}^{t-1} \sum_{j'} f_{j,j'}^s - \sum_{s=0}^{t-1} \sum_{j'} f_{j,j'}^s \leq C_j^t, \ \forall j,t, \tag{F4}$$

$$f_{jD}^t = \tilde{g}_{j,D}^t, \ \forall j,t. \tag{D}$$

We highlighted with red color and a tilde the uncertain quantities. In a robust approach, we impose all constraints to hold for any value of $\tilde{\boldsymbol{g}} \in \mathcal{G}$ and take the worst case cost $\max_{\tilde{\boldsymbol{g}}\in\mathcal{G}} c(\boldsymbol{f}, \tilde{\boldsymbol{g}})$. For the latter, $c(\boldsymbol{f}, \tilde{\boldsymbol{g}})$ is a piece-wise linear convex function. We can write its epigraph formulation and robustify each constraint independently. The main issue lies in the equality constraints (D). For a given solution, $f^t_{j,D}$ is fixed and such constraints cannot hold in general for all values of $\tilde{g}^t_{j,D}$. A simple, yet conservative, approach is to replace the equality by an inequality $\leq$ and take the robust counterpart. To containing conservatism, we adopt an affinely adaptive rule (Ben-Tal et al. 2004, Chen and Zhang 2009): We first eliminate the number of actual discharges, $f^t_{j,D}$, and replace it by the number of needed discharges, $\tilde{g}^t_{j,D}$. This substitution affects

- Flow conservation constraints (F3): for all inpatient units j

$$\sum_{t,j'} f^t_{j,j'} = \sum_{t,j'\neq D} f^t_{j,j'} + \sum_t \tilde{g}^t_{j,D} = z^0_j.$$

- Unit capacity constraints (F4): for all inpatients unit j

$$z^0_j + \sum_{s=0}^{t-1}\sum_{j'} f^s_{j',j} - \sum_{s=0}^{t-1}\sum_{j'\neq D} f^s_{j,j'} - \sum_{s=0}^{t-1} \tilde{g}^s_{j,D} \leq C_j,\ \forall t.$$

Then, we enforce a simple recourse policy: all variability in discharges from unit j will be directly reported on the number of patients in unit j who stays in unit j throughout the day, i.e., impose

$$f^T_{j,j} = \tilde{f}^T_{j,j} - \sum_t \tilde{g}^t_{j,D} \geq 0,$$

where $\tilde{f}^T_{j,j}$ is the new decision variable of interest. With this condition, the flow conservation policy is always satisfied. This strategy, which we called "affine with known recourse" policy or simply "affine" policy, is an affinely adaptive robust policy for the decision variables $\boldsymbol{f}$ depend affinely on the uncertainty $\tilde{\boldsymbol{g}}$. Yet, the affine rule is not considered as a decision variable but rather informed by intuition and imposed by design.

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 H_2O approach on data collected between January and July 2019. In order 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 Appendix EC.4 but highlight its main characteristics here:

- We consider each day independently. In other words, for each day, we initialize a new experiment with the actual state of the hospital on this day at 6 am, and not what this state would have been, had the optimization model run the previous day as well.
- We assume that the decisions we are making only impact bed assignments. As mentioned in introduction, there is strong evidence to suggest that is not the case and that patient assignment also impacts length-of-stay and health outcomes. However, we argue that since we are considering each day independently, this effect does not tangibly impact our counterfactual estimation.
- We assume that all bed assignment decisions made by (H_2O) 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 which arose and got solved between two optimization runs, we implement the decision from the hospital staff. This is notably the case for current inpatients for whom doctors often arrange inter-units transfers directly over the phone.

To assess performance, we compute:

- the number of off-service placements made throughout the day;
- the cumulative waiting time in the ED/post-OR for an inpatient, also referred to in the literature as boarding time. For each patient in a waiting unit at some point in the day, we count the number of time periods she waited for an assignment and then sum the waiting times over all patients on that day;

- the number of ED patients who waited more than 2 hours for a bed assignment.

For all metrics, either we compare absolute values (after normalization) between the historical policy and (H_2O) or we report the relative difference of (H_2O) with respect to the empirical decisions. In any case, the lower the better. To ensure, improvement in off-service placement is not due to the hospital admitting less patients, we also report the relative difference in terms of:

- overall throughput, defined as the difference in inpatient census between the end and the beginning of the day;
- peak census, defined as the maximum inpatient census during the day;
- number of outside transfer admissions.

5.2. Results

Computational time and scalability: Our historical data set consists of $N = 212$ days. Hence, we solved a total of $12 \times 212 = 2,544$ instances. The median solve time is 0.54 seconds with an interquartile range, the difference between the 75% and 25% percentiles, of 0.72 seconds. We report summary statistics in Table 4. In contrast with queueing models, our approach is very tractable and can be solved in less than a second for a 600-bed institution. As a result, one could explore near 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, 2 hours) needs to be long enough for the bed assignment policy to be computed *and implemented*.

Table 4 **Summary statistics of the solve time (in seconds) over $2,544$ instances of (H_2O).**

Percentile:	1st	25th	50th (median)	75th	99th	Average
	0.03	0.20	0.54	0.93	2.68	0.68

Impact on hospital performance: We now evaluate the operational benefit of the (H_2O) policy. Figure 2 illustrates the impact of (H_2O) on the distribution of boarding times, i.e., the time patients wait to be assigned to an inpatient bed, and patient misplacement. We make the following observations: First, overall (H_2O) leads to significant decrease in waiting times, both in the emergency

room and after surgery. Second, quality of the assigned bed, measured in terms of number of patient misplacement, also improves under (H_2O). Although there is a clear trade-off between quality of the bed assigned and time waited for the assignement, our simulations suggest that there is an opportunity for hospitals to improve on both metrics simultaneously compared to how they currently operate, by leveraging advanced prescriptive analytics. Thirdly, this is not only an improvement on average but a substantial distributional shift towards lower values for all metrics. Finally, the reduction in boarding time for patients in the PACUs (OR Wait) ought to be taken with a pinch of salt. Indeed, for the (H_2O) policy, we take into account bed requests from surgical patients as soon as they enter the PACU and assume that the assignment can be implemented in the 2 hours following the decision. However, after a surgery, a patient can often not be moved to an inpatient unit before she recovers from the anesthesia and is physically fit to be discharged from the PACU. This time-to-recovery generates delays for the historical policies, delays to which (H_2O) is oblivious. We are currently working with our partner hospital to keep a meticulous log of all the stages a patient goes through after surgery in order to properly assess the magnitude of these delays. Figure 3 summarizes the relative improvement of (H_2O) over the current bed assignment strategy over those four metrics.

Control for admission rate and throughput: We also compare the overall throughput and congestion of the hospital in Table 5. The results on these metrics are reassuring: The daily peak census at the hospital is largely unchanged (less than 1% increase), negating the possibility that (H_2O) would reduce boarding time and off-service placement by overcrowding units. In addition, overall throughput globally increases which demonstrates that (H_2O) does not admit less patients in order to achieve such improvement. However, we observe that admissions from outside transfers decrease substantially. This observation suggests that (H_2O) 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 of admitting external patients on the rest of the system when making their admission decisions. Our holistic approach, on the other hand, integrates outside admission decisions into an overall bed assignment decision tool and is able to quantitatively take into

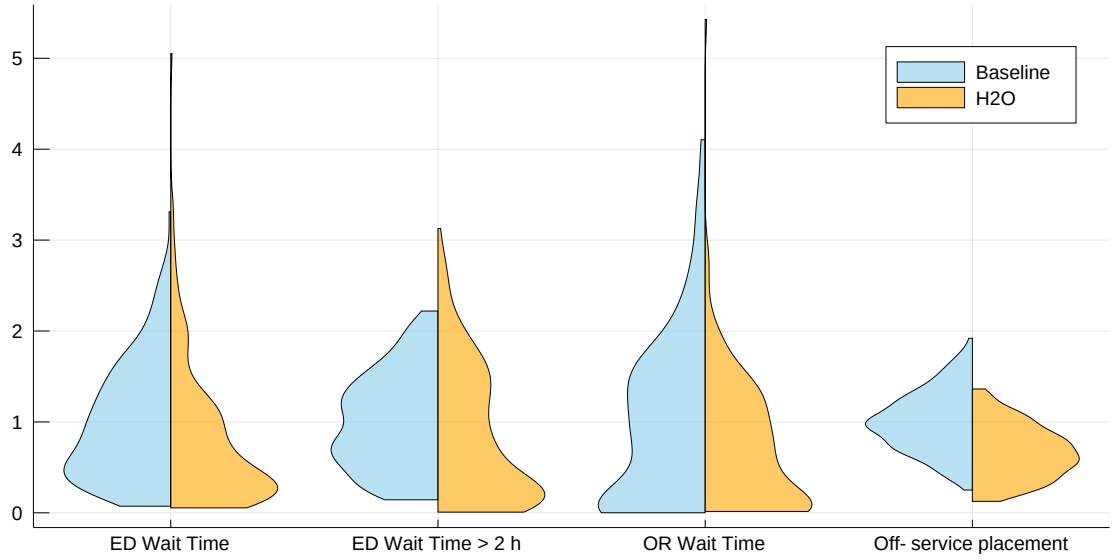

Figure 2 **Violin plots of the distribution of three waiting time-related metrics under the historical policy (in blue, left side) and (H_2O) (in yellow, right side). For each metric, the reported values are normalized so that median value for the historical policy equals one.**

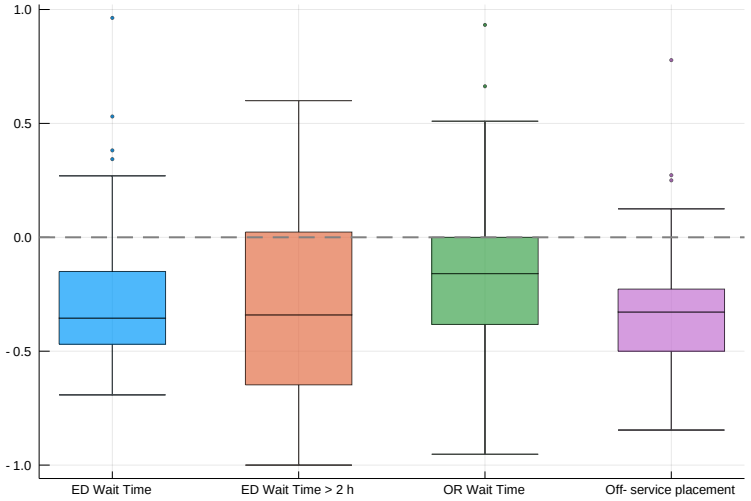

Figure 3 **Boxplot for the distribution of the relative difference between the (H_2O) policy over the historical bed assignment decisions in terms of four performance metrics: (from left to right) ED wait, number of ED patients waiting more than 2 hours, OR wait, number of off-service placements.**

account these effects. This constitutes a central contribution of the (H_2O) model. Furthermore, we do not believe that this negative impact on transfer admissions will persist in a real-life implementation. Indeed, each day in our simulation starts with a backlog of bed requests from the previous day. Since our methodology considers each day independently, the backlog of requests is inherited

from the historical policy. After (H_2O) is implemented, this backlog will be significantly reduced, thus creating more opportunities for outside transfer admission.

Table 5 Summary statistics on the relative difference between the (H_2O) policy over the historical bed assignment decisions in terms of peak census, overall throughput and admission from outside transfer.

Metric	10th percentile	25th percentile	Median	75th percentile	90th percentile
Peak census	0.2%	0.5%	0.9%	1.5%	2.0%
Throughput	-16.4%	14.3%	64.4%	200.0%	530.0%
TC admissions	-77.8%	-68.0%	-46.2%	0.0%	153.8%

Patient-level comparison of bed placement decisions: We now compare how the decisions made historically by the hospital differ from (H_2O)'s at a patient level (see Table 6). While the decisions made by (H_2O) jointly improve boarding times and service placements on average, we observe that not all patients benefit from the optimization approach. Approximately 38% of all patients experience better outcomes under (H_2O) while 20% experience worse outcomes, suggesting that (H_2O) is able to correctly identify and prioritize patients based on their individual impact on the overall system performance. Inequity implications of such optimized processes 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.

Trade-off between waiting time and off-service placement: As previously mentioned, there is an intuitive trade-off between time waited to be admitted and quality of the assigned bed. In our model, we can control this trade-off through the parameter λ that balances the first-stage and the multi-stage objectives. If $\lambda = 0$, the decision maker is myopic, waiting is not an option. As a result, the optimization model will design strategies with low waiting times but a potentially high number of off-placed patients. Conversely, as λ increases, the strategy becomes more forward-looking and improves placement at the expense of increased waiting times. We illustrate this trade-off by considering different values of λ, running our simulation engine on the first two weeks of May 2019,

Table 6 Comparison of individual bed assignement decisions from (H_2O) with the historical bed placements. A (+) (resp. (-)) indicates a strict improvement (deterioration). (=) indicates that both policies lead to the same outcome.

Overall impact of (H_2O)	Waiting time	Service Placement	Proportion of patients
Improvement	(+)	(+)	6.5%
	(+)	(=)	23.8%
	(=)	(+)	9.4%
Not comparable	(=)	(=)	35.1%
	(+)	(-)	3.0%
	(-)	(+)	3.6%
Deterioration	(-)	(=)	16.0%
	(=)	(-)	1.1%
	(-)	(-)	1.5%

and computing average waiting time in the emergency department and off-service placements for each value of λ. Results are graphically reported on Figure 4. Note that all performance metrics are reported as relative values compared to historical placements, hence a negative value indicates a reduction, i.e., an improvement. We also seize the opportunity to assess the improvement of our affinely adaptive robust policy presented in Section 4.1 (in blue circles) vs. the static robust approach (in pink diamonds). As presented on Figure 4, the static policy already provides a significant improvement over the historical placements at the hospital. Using affine decision rules further increases the benefit from optimization. In addition, the Pareto frontier for the affine policies displays less of an "all-or-nothing" shape, hence featuring a more flexible trade-off.

6. Concluding remarks

In this paper, we propose an optimization framework to achieve hospital-wide patient flow management. We believe our work constitutes a first step towards a major paradigm shift in hospital operations, research and practice. Our approach is comprised of two key elements: a holistic view to

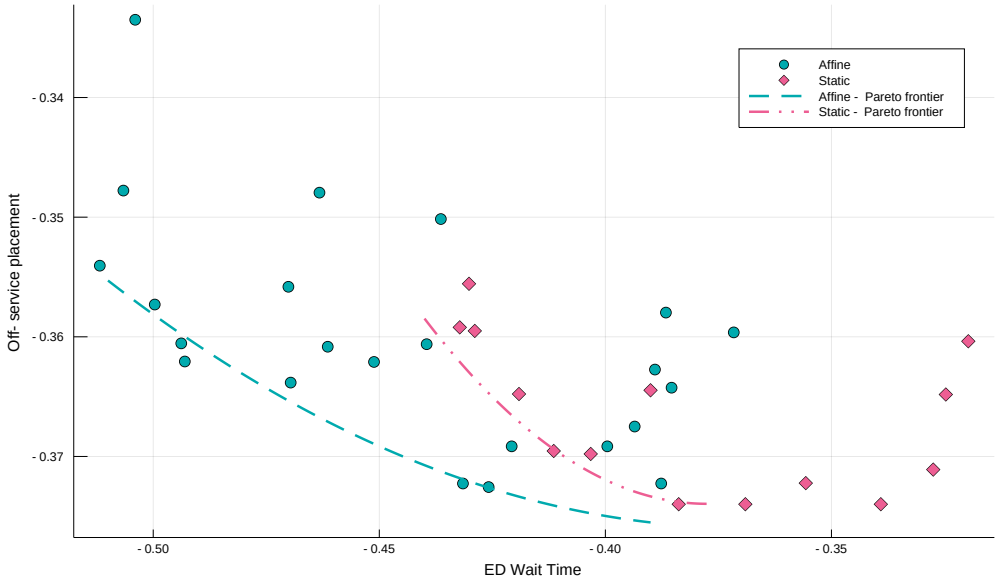

Figure 4 Trade-off waiting time in the emergency department vs. off-service placements. All quantities are relative values compared with historical placements. We compare the performance of the proposed robust affine policies (blue circles) with the static solution (pink diamonds).

account for the entirety of the hospital within a single model, and data to describe and anticipate future demand and supply of care. Numerically, we demonstrate that our framework is applicable to large institutions such as academic medical centers and provides tangible operational improvement.

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 Qual Saf*, 27(1):11–20, 2018.

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

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

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. *medRxiv preprint medRxiv:2020.05.12.20098848*, 2020. URL `https://www.medrxiv.org/content/early/2020/05/18/2020.05.12.20098848`.

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.

Margaret L Brandeau, François Sainfort, and William P Pierskalla. *Operations research and health care: a handbook of methods and applications*, volume 70. Springer Science & Business Media, 2004.

Carri W Chan, Galit B Yom-Tov, and Gabriel J Escobar. When to use speedup: An examination of service systems with returns. *Operations Research*, 62(2):462–482, 2014.

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.

David Roxbee Cox and Walter Smith. *Queues*, volume 2. CRC Press, 1991.

Jim G Dai and Pengyi Shi. Recent modeling and analytical advances in hospital inpatient flow management. *Available at SSRN*, 2018.

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

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. *Available at SSRN 3246641*, 2018.

Linda V Green. Using operations research to reduce delays for healthcare. In *State-of-the-Art Decision-Making Tools in the Information-Intensive Age*, pages 1–16. INFORMS, 2008.

Linda V Green, Joao Soares, James F Giglio, and Robert A Green. Using queueing theory to increase the effectiveness of emergency department provider staffing. *Academic Emergency Medicine*, 13(1):61–68, 2006.

Shuangchi He, Melvyn Sim, and Meilin Zhang. Data-driven patient scheduling in emergency departments: A hybrid robust-stochastic approach. *Management Science*, 2019.

Junfei Huang, Boaz Carmeli, and Avishai Mandelbaum. Control of patient flow in emergency departments, or multiclass queues with deadlines and feedback. *Operations Research*, 63(4):892–908, 2015.

Peter JH Hulshof, Nikky Kortbeek, Richard J Boucherie, Erwin W Hans, and Piet JM Bakker. Taxonomic classification of planning decisions in health care: a structured review of the state of the art in OR/MS. *Health Systems*, 1(2):129–175, 2012.

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.

Derya Kilinc, Soroush Saghafian, and Stephen Traub. Dynamic assignment of patients to primary and secondary inpatient units: Is patience a virtue? 2018.

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.

Ludwig Kuntz, Stefan Scholtes, and Sandra Sülz. Separate and concentrate: Accounting for patient complexity in general hospitals. *Management Science*, 65(6):2482–2501, 2019.

C Lakshmi and Sivakumar Appa Iyer. Application of queueing theory in health care: A literature review. *Operations Research for Health Care*, 2(1-2):25–39, 2013.

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.

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*. 2011.

Patricia A Rutherford, Lloyd P Provost, Uma R Kotagal, Katharine Luther, and Alex Anderson. Achieving hospital-wide patient flow. *IHI White Paper. Cambridge: Institute for Healthcare Improvement*, 2017.

Soroush Saghafian, Wallace J Hopp, Mark P Van Oyen, Jeffrey S Desmond, and Steven L Kronick. Patient streaming as a mechanism for improving responsiveness in emergency departments. *Operations Research*, 60(5):1080–1097, 2012.

Soroush Saghafian, Wallace J Hopp, Mark P Van Oyen, Jeffrey S Desmond, and Steven L Kronick. Complexity-augmented triage: A tool for improving patient safety and operational efficiency. *Manufacturing & Service Operations Management*, 16(3):329–345, 2014.

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 Yang. Capacity pooling in hospitals: The hidden consequences of off-service placement. *Available at SSRN 3186726*, 2019.

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.

Ellen J Weber, Suzanne Mason, Angela Carter, and Ruth L Hew. Emptying the corridors of shame: organizational lessons from england's 4-hour emergency throughput target. *Annals of emergency medicine*, 57(2):79–88, 2011.

Natalia Yankovic and Linda V Green. Identifying good nursing levels: A queuing approach. *Operations Research*, 59(4):942–955, 2011.

Supplementary material

This electronic companion provides the full Holistic Hospital Optimization (H_2O) formulation presented in Section 3, details on the description of the uncertainty set from Section 4 and describes the simulation procedure used for the numerical experiments in Section 5.

EC.1. Description of units in our partner hospital

Our partner hospital is comprised of two campuses (E and W). We provide a list and description of the units on campus E in Table 2. For completeness, we detail the composition of campus W here, in Table EC.1.

EC.2. Complete nominal formulation

In this section, we present the complete nominal formulation for the Holistic Hospital Optimization (H_2O) problem, which includes individual future trajectories and capacity decisions.

EC.2.1. Individual unit assignments

In practice, for current patient who expressed a need for a new bed, clinical staff might value having a plan. If the optimization problem decides not to assign a patient to a new bed immediately, can we elicit what it had in mind for this particular patient? When to move her and where. To do so, we introduce individual unit assignment variables which will track the location of the patients throughout the end of the optimization period. Since we only care about the patients who currently need a new bed, we shall introduce these variables just for them.

Decision variables: We introduce binary variables z_{ij}^t, indicating whether patient i is in unit j at time t, for each patient $i = 1, \dots, I$ and for each unit $j = 1, \dots, J$ who is currently in a waiting or inpatient unit.

In accordance with the single-move assumption, 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 and the time she moves, ℓ_{ij} and m_{it}

Table EC.1 Description of inpatients units on campus W at our partner hospital. The type is either GC (General Care) or IC (Intensive Care). The unit name follows the nomenclatura: Campus-Type Number.

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

Constraints: The decision variables z_{ij}^t, ℓ_{ij} and m_{it} must satisfy a set of constraints, concisely denoted $(\boldsymbol{z}, \boldsymbol{\ell}, \boldsymbol{m}) \in \mathcal{Z}$, namely:

- Integrality: $z_{ij}^t, \ell_{ij}, m_{it} \in \{0,1\}$.
- Single-location: $\sum_j z_{ij}^t = 1,\ \forall i,t$ and $\sum_j \ell_{ij} = 1,\ \forall i$.
- Single-move: $\sum_t m_{it} = 1,\ \forall i$.
- Logical constraints between $\boldsymbol{z}$ and $\boldsymbol{\ell}, \boldsymbol{m}$: For any patient i,

$$z_{ij}^t \leq \ell_{ij},\ \forall j,t$$
$$z_{ij}^t \leq \sum_{s=1}^{t} m_{is},\ \forall j,t$$

- Unit capacity: $\sum_i z_{ij}^t \leq C_j^t,\ \forall j$.
- Forbidden moves (imposed by design).

Objective: The objective is to minimize $\sum_{i,j,t} c_{ij}^t z_{ij}^t$, where the cost c_{ij}^t also captures waiting time, in addition to capturing the mismatch between actual and required service and level of care (and eventually physical distance). For instance, we add to c_{ij} defined in Section 3 a term $\max(WaitingTime_i + 2t - t_0, 0)$, where $WaitingTime_i$ is the time patient i already waited for an assignement, the factor 2 in the $2t$ term comes from the fact that we consider 2-hour time periods and t_0 is some threshold value below which we consider waiting to be "harmless". To reduce the problem size, we only consider in the sum the set of patients I_{need} who expressed a new need, i.e., patients in waiting units or inpatient who need to change bed at $t = 0$.

Note: In Section 3, we force patients who do not have particular care needs ($i \neq I_{need}$) to stay in their original unit by imposing an infinte cost to other locations ($c_{ij}^t = \infty$ if $z_{ij}^0 = 0$). Alternatively, we can enforce this requirement with

$$\sum_{i \notin I_{need}} \sum_{j: z_{ij}^0 = 0} \ell_{ij} \leq k,$$

with $k = 0$. Taking $k > 0$ allows for inpatient reallocation as in Thompson et al. (2009).

EC.2.2. Capacity constraints and decisions

In practice, unit capacity is another decision variable. Each unit j has a fixed number of licensed beds, $LicensedBeds_j$. However, it also 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 $j = 1, \dots, J$ and each time $t = 0, \dots, T$. With this notation, the actual unit capacity C_j^t would be equal to $C_j^t = LicensedBeds_j + v_j^t$. We consider such decision variables in our final formulation.

All beds in the same service or ward are not equivalent. Indeed, some patients - with viral infections for instance - require beds in private rooms. Depending on hospital policy, shared rooms can sometimes 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 breakdown 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 both genders $C_{j,ungendered}^t$. Then, the capacity constraints write as follows, for all units j, time t:

$$\sum_i z_{ij}^t \leq C_{j,male}^t + C_{j,female}^t + C_{j,ungendered}^t,$$

$$\sum_{i,\ i\ male} z_{ij}^t \leq C_{j,male}^t + C_{j,ungendered}^t,$$

$$\sum_{i,\ i\ female} z_{ij}^t \leq C_{j,female}^t + C_{j,ungendered}^t.$$

EC.3. Lifted formulation of the uncertainty set

In addition to $g_{j,j'}^t$, we introduce variables, $h_{j,(s,\ell)}^t$, 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. $h_{j,(s,\ell)}^t$ better correspond to how needs are expressed, whereas $g_{j,j'}^t$ capture how they materialize. We need constraints to ensure that these two description of clinical flows, from a unit and clinical need perspective, match on multiple aspects.

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

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

- 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 *clinical* flows, that is, 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 clinical needs it can serve, $\mathcal{N}(j')$, inducing the constraint

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

EC.4. Synthetic experiment procedure

We assess our approach on data collected between January and July 2019 at our partner hospital. We apply the following methodology:

For a given day, we start our experiment at $h = 6$ o'clock 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 robust H_2O formulation, obtaining a vector of first-stage variables $\boldsymbol{z}^1$. We virtually implement this vector and move all patients in or waiting for a bed as prescribed by $\boldsymbol{z}^1$.

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

For waiting units, we have the empirical lists of requests at h and $h+2$ o'clock. We construct our simulated list of bed requests by taking

- new requests available at $h+2$ o'clock which were not available at h o'clock,
- requests available at h o'clock which were not granted a bed by $\boldsymbol{z}^1$.

For current inpatients, we consider 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 her in the location prescribed by $\boldsymbol{z}^1$, otherwise we place her in her 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.
- 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.