Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Timing It Right: Balancing Inpatient Congestion vs. Readmission Risk at Discharge

Pengyi Shi, Jonathan E. Helm, Jivan Deglise-Hawkinson, Julian Pan

To cite this article:
Pengyi Shi, Jonathan E. Helm, Jivan Deglise-Hawkinson, Julian Pan (2021) Timing It Right: Balancing Inpatient Congestion vs. Readmission Risk at Discharge. Operations Research

Published online in Articles in Advance 04 Mar 2021

. https://doi.org/10.1287/opre.2020.2044

http://pubsonline.informs.org/journal/opre

OPERATIONS RESEARCH
Articles in Advance, pp. 1–24
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Methods

Timing It Right: Balancing Inpatient Congestion vs. Readmission Risk at Discharge

Pengyi Shi,[a] Jonathan E. Helm,[b] Jivan Deglise-Hawkinson,[c] Julian Pan[c]

[a] Krannert School of Management, Purdue University, West Lafayette, Indiana 47907; [b] Kelley School of Business, Indiana University, Bloomington, Indiana 47405; [c] Lean Care Solutions Corporation Pte. Ltd., Singapore 139959, Singapore

Contact: shi178@purdue.edu, https://orcid.org/0000-0003-0905-7858 (PS); helmj@indiana.edu, https://orcid.org/0000-0001-5577-5530 (JEH); jivan@leancaresolutions.com (JD-H); jp@leancaresolutions.com (JP)

Received: January 9, 2019
Revised: August 25, 2019; November 27, 2019
Accepted: December 30, 2019
Published Online in Articles in Advance: March 4, 2021

Subject Classifications: hospitals: healthcare applications; dynamic programming; optimal control: dynamic programming
Area of Review: Policy Modeling and Public-Sector Operations Research

https://doi.org/10.1287/opre.2020.2044

Abstract. When to discharge a patient plays an important role in hospital patient flow management and the quality of care and patient outcomes. In this work, we develop and implement a data-integrated decision support framework to aid hospitals in managing the delicate balance between readmission risk at discharge and ward congestion. We formulate a large-scale Markov decision process (MDP) that integrates a personalized readmission prediction model to dynamically prescribe both how many and which patients to discharge on each day. Because of patient heterogeneity and the fact that length of stay is not memoryless, the MDP has the curse of dimensionality. We leverage structural properties and an analytical solution for a special cost setting to transform the MDP into a univariate optimization; this leads to a novel, efficient dynamic heuristic. Furthermore, for our decision framework to be implementable in practice, we build a unified prediction model that integrates several statistical methods and provides key inputs to the decision framework; existing off-the-shelf readmission prediction models alone could not adequately parametrize our decision support. Through extensive counterfactual analyses, we demonstrate the value of our discharge decision tool over our partner hospital's historical discharge behavior. We also obtain generalizable insights by applying the tool to a broad range of hospital types through a high-fidelity simulation. Last, we showcase an implementation of our tool at our partner hospital to demonstrate broader applicability through our framework's *plug-and-play* design for integration with general hospital data systems and workflows.

Supplemental Material: The online appendices are available at https://doi.org/10.1287/opre.2020.2044.

Keywords: readmission risk • inpatient flow management • state-dependent discharge • large-scale Markov decision process (MDP) • approximation algorithms • tool implementation

1. Introduction

> A hospitalist makes many decisions that influence the cost of an inpatient stay . . . but probably none has more impact than "Should this patient go home today or tomorrow?" —Cover story for *American College of Physicians Hospitalist* (Colwell 2014)

This paper highlights the key tradeoff in making discharge decisions: Under the Affordable Care Act, it is still in hospitals' financial interest to discharge patients as soon as possible but also to facilitate post-discharge care and prevent 30-day readmissions. Rather than just lowering length of stay (LOS), hospitals now aim to *optimize it at the intersection of quality and cost.* Balancing this tradeoff has broad implications for patient flow, inpatient unit congestion, quality of care, and postdischarge risk, impacting all care providers from small community hospitals to major teaching hospitals.

Frequent overloading of inpatient units contributes to emergency department (ED) overcrowding (Proudlove et al. 2003), denial of intensive care unit (ICU) admission (Kim et al. 2014), cancellation of elective surgeries (Helm et al. 2011), and higher risk of mortality (Kuntz et al. 2014), among other consequences. When inpatient units become congested, doctors frequently discharge existing patients early (Kc and Terwiesch 2012, 2017; Berry Jaeker and Tucker 2017). This practice alleviates overcrowding in the ward by shifting the burden to the early discharge patients, who may experience increased risk of readmission, mortality, and other adverse outcomes (Kc and Terwiesch 2009, 2012). By contrast, when occupancy levels are low, hospitals may keep patients longer (Anderson et al. 2011), which can have a positive impact on patient outcomes (Bartel et al. 2020). The balancing act between individual discharge risk and ward congestion has grown into a major stress point in the face of recent pressures to reduce readmissions (Kocher and Adashi 2011) while limiting LOS, as indicated in the paper referenced

previously and others (Frenz 2014, Frakt 2016). Hospitals manage this tradeoff through ad hoc practices that lack analytical decision support. The operations management literature has made significant strides in this area, although new research is needed to support the development of practical tools that can be implemented as part of a hospital's workflow.

In this paper, we develop a data-integrated decision support framework for managing the tradeoff between readmission risk and inpatient crowding. The decision support optimizes *who and how many* to discharge each day based on a personalized trajectory that predicts how readmission risk evolves over the course of each patient's stay in the hospital. We focus on readmissions because it is the target area of improvement in our partner hospital, although the decision framework is easily adapted to other types of adverse events such as mortality and to a wide range of hospitals. Figure 1 illustrates the two main components of our decision support framework:

- To integrate with daily hospital operations, the discharge decision support needs to account for patient heterogeneity and history-dependent health status progression, which necessitates our development of a nontraditional modeling framework and optimization approach. To handle the curse of dimensionality, we develop a simple, efficient, and robust heuristic by integrating general structural properties with analytical solutions to a special cost problem.
- Implementing the decision support in practice requires personalized prediction of each patient's risk trajectory from data, that is, how the probability and timing of readmission evolves as a function of LOS. To the best of our knowledge, no off-the-shelf tool could be directly applied to our decision framework because of three challenges that have yet to be addressed in a unified method (see Section 1.1). We develop a prediction model that combines several statistical methods and allows a seamless integration of our decision support tool with existing hospital data systems.

A final product from this work is tested and implemented in a partner hospital in the state of Indiana, demonstrating the practical value of our decision framework and providing a showcase for other hospitals in need of a similar tool. Through the plug-and-play design of our framework, such a discharge decision tool can be easily adapted to a broad range of hospitals and hospital information technology (IT) systems in a (nearly) automated manner with some input from a trained analyst. We summarize our contributions in further detail in Section 1.2.

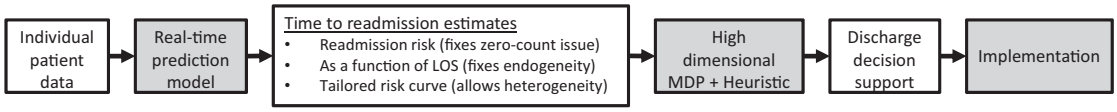
1.1. Challenges in Discharge Management

In this section, we first elaborate on how challenges associated with discharge management impact the day-to-day operations and patient outcomes at our partner hospital. Through this, we illustrate the broad challenges facing the hospital industry and demonstrate the real need for developing analytical decision support for inpatient discharge. We then discuss why new research is needed to develop an *implementable* tool to support daily discharge decisions.

1.1.1 Discharge Optimization. Discharge planners face complex decisions on how many patients and whom to discharge on a given day. Currently, hospitals engage in adaptive discharge practices in a reactive and ad hoc manner. For example, when our partner hospital becomes overcrowded, a communication is sent to all physicians asking them to discharge as many patients as possible to free up beds, which is a practice found in other hospitals we have spoken with as well. This unstructured approach may end up discharging too many or too few patients or discharging a suboptimal group of patients. This is highlighted by a recent empirical paper that demonstrates that individual physicians lack a system perspective and as a result react to occupancy crises poorly (Adepoju et al. 2019).

In our study, we show that discharge decisions must be far more nuanced to properly balance hospital and patient needs; in fact, adaptive discharge may be activated not only when occupancy is high but also when occupancy is low: keeping the right set of patients longer to reduce readmissions. These decisions need to account for the risk of each individual patient in the hospital unit and individual patient's risk evolution over future days in conjunction with expected future patient arrivals and current and future occupancy levels. The inpatient arrival day-of-week phenomenon and diverse patient characteristics further complicate discharge decisions. These subtleties necessitate a sophisticated decision support from data-integrated analytical models that prescribes discharge recommendations based on the dynamic patient profiles and system status. The complexities in developing such a decision support led our

Figure 1. Conceptual Diagram of Decision Support Development for Discharge Management



partner hospital and its data analytics contractor (Lean Care Solutions) to approach us, with the goal of more effectively leveraging discharge timing as a powerful tool in the readmission battle.

1.1.2. An Implementable Tool. The discharge optimization problem connects to the service rate control literature (see more review in Section 2). However, for the discharge decision support to be implementable, we find that the optimal discharge (service) rate—output from the conventional service rate control works—is not sufficient to inform the hospital as to how many and whom to discharge for a given mix and volume of patients currently in the hospital unit. Moreover, conventional models usually use memoryless service time distributions and other stylized assumptions that abstract away information that is crucial to day-to-day discharge management in practice. As a result, new modeling efforts are needed to develop an implementable decision framework and an appropriate solution algorithm.

In addition to the theoretical challenges, another barrier for the tool to be implementable is the need for a sufficiently accurate input to the prediction of patient readmission risk evolution as a function of LOS. Most existing readmission prediction tools treat LOS as an exogenous variable. Directly applying these tools by varying LOS suffers from endogeneity (sicker patients tend to stay longer and have a higher readmission risk), which often leads to the incorrect conclusion that extending LOS for an individual patient results in higher readmission risk. In addition, the discharge decision framework requires not only the prediction of the readmission probability as an input but also the prediction of the readmission timing, where we find more challenges when applying the classical Cox proportional hazard model to predict readmission timing (see details in Section 6). As such, we need to develop a unified prediction method that can work with different hospital data systems and can provide adequate parameterization of our decision support system.

1.2. Overview and Main Contributions

The main goal of this paper is to develop a novel, implementable decision support tool to allow hospitals to use discharge timing as a powerful lever to reduce readmissions and improve patient outcomes. This paper makes both technical and practical contributions to the literature.

1. *Discharge decision framework.* In Section 3, we build a large-scale Markov decision process (MDP) based on a patient flow model with reentries. This MDP deviates from traditional service rate control models and accounts for personalized patient risk trajectory and history-dependent state. It dynamically optimizes the number of patients and specifically who to discharge each day by balancing the tradeoff between the individual-level cost (readmission risk) and the system-level cost (ward congestion).

2. *Analytical results and heuristic.* Because of the patient heterogeneity and the fact that patient LOS is no longer exogeneous or memoryless, the formulated MDP has a high-dimensional state and action space. To overcome the curse of dimensionality, in Section 4, we prove structural properties of the MDP showing that discharge decisions should depend on *marginal risk* among all future days, contradicting the belief in the medical community that current absolute risk should be the criterion for discharge. We develop both a strong dominance and a more general weak dominance criterion to rank patients in terms of discharge desirability. In Section 5, we leverage a special case of the MDP that can be solved efficiently as a linear quadratic stochastic control problem. We develop a novel algorithm that combines the closed-form solutions of this special MDP with the patient ranking to transform the original MDP into a heuristic univariate optimization, significantly reducing the computational complexity.

In addition to being computationally appealing, our algorithm provides a simple and easily interpretable method for implementation, suggesting how many patients to discharge, with the who being determined directly from the ranking. Furthermore, our algorithm integrates with the complex and data-rich environment in hospitals by being flexible enough to incorporate *personalized* risk trajectories for all patients currently in the hospital ward and a nonstationary arrival process to account for day-of-week variability. Given the simplicity and flexibility, this algorithm can be easily implemented in hospitals beyond our partner hospital.

3. *Risk prediction and implementation.* We worked closely with our partner hospital to test and implement our discharge decision support framework. As a first step, in Section 6, we develop a unified prediction model that combines several statistical methodologies to overcome the challenges mentioned in Section 1.1. In particular, it is necessary to use an instrumental variable (IV) approach to correct the estimation bias caused by endogeneity—patient severity correlates with both LOS and readmission risk. Although there is no theoretical guarantee for the IV approach, this prediction model provides a reasonably accurate prediction. Most important, our prediction model allows a direct integration of our discharge decision support with the hospital's IT infrastructure and provider workflow.

Figure 2(a) shows a snapshot of the main portal of the implemented tool from our work. The tool displays (1) patients currently in the hospital unit

(represented by each block), ranked with different color codes in terms of their discharge desirability from the dominance criterion (see Section 4), (2) discharge risk curve for future possible LOS of each patient (with past LOS and generally recommended LOS), and (3) postdischarge readmission timing risk curve. See Figure 2(b) for features (1) and (2), enlarged in Online Appendix F.

4. *Tool value and broader insights.* During the development of our decision support framework, we provided performance measures and analysis for each component (optimization and prediction) individually for completeness and robustness. The overall value of our integrated decision support tool is captured in Section 7, where we measure LOS, net change in readmissions, and positive catch rate (proportion of actual readmissions our tool suggests intervening on) through careful counterfactual analyses using data from our partner hospital. To properly measure the holistic value of our system, we use a trace-based counterfactual that demonstrates Pareto dominance of our heuristic policy over the historical discharge policy in occupancy and readmissions. Importantly, our policy properly catches more than 50% of actual readmissions, suggesting extending their LOS. We generalize these results to other hospitals through high-fidelity simulation analyses in Section 7.3, showing which types of hospitals gain the most benefit from our dynamic discharge.

Remark 1. In practice, discharge decisions are complicated and involve a variety of factors. We emphasize that our tool is meant to provide analytical support for discharge decisions. Doctors can still use their own discretion in discharging patients according to the individual patient's condition and needs. Our tool is flexible to accommodate such deviations and can update the recommendations after incorporating the actual decisions; a more detailed discussion is presented in Section 8.

2. Literature Review

We review three streams of literature relevant to our paper.

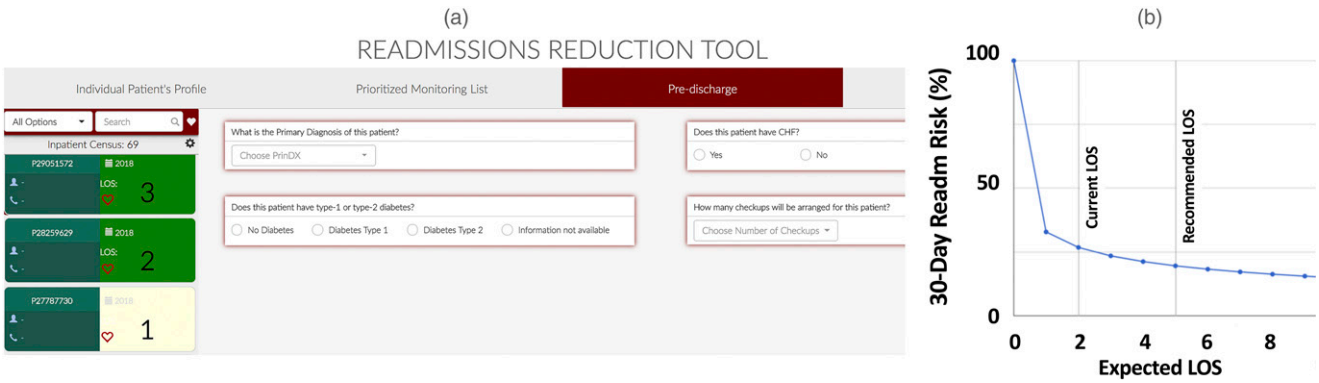
2.1. Empirical Evidence

A rich body of empirical research has provided evidence that hospitals tend to use discharge decisions to reduce inpatient unit congestion (Kc and Terwiesch (2012, 2017; Berry Jaeker and Tucker 2017). Long and Mathews (2017) show that ICU occupancy impacts the less essential *boarding time* but not the medically necessary LOS. Although early discharge can alleviate congestion and increase the chance of admission for future patients, it compromises patient outcomes. Using a large data set on patients with congestive heart failure, Oh et al. (2017) find inpatient stays that are shorter than the Centers for Medicare and Medicaid Services–suggested LOS are likely to exhibit a 1.1% greater risk of readmission. Kc and Terwiesch (2009, 2012) find that patients discharged early exhibit increased risk of readmission, mortality, and other adverse outcomes. The medical literature has discovered similar findings between LOS and patient outcomes (Heggestad 2002, Kuo and Goodwin 2011). By contrast, Bartel et al. (2020) show that keeping a patient one extra day can reduce mortality risk by nearly 6%. Oh et al. (2017) and Carey (2015) suggest that keeping patients longer can significantly reduce hospital costs. Our paper is largely motivated by these empirical studies, although it is important to differentiate our work from them because our focus in Section 6 is on dynamic, personalized *prediction* of readmission risk evolution over a patient's hospital stay, which is an important component of our integrated discharge optimization tool.

2.2. State-Dependent Discharge Optimization

Our modeling and discharge decision analysis connects with the literature on optimal service rate control.

Figure 2. Screenshot of the Discharge Decision Support Web Portal Implemented in Our Partner Hospital



Notes. The first plot (a) shows the screenshot of the main portal. The second plot (b) shows the individual risk prediction tab.

Within this area, several papers specifically study discharges in the hospital. Berk and Moinzadeh (1998) provide an early paper to study the tradeoff between discharge risk and inpatient occupancy. The authors model patient care for a homogeneous population in two stages, where stage 1 is *not dischargable* and stage 2 is a less critical stage in which early discharge can be exercised. They focus on steady-state performance analysis under two fixed policies (with and without early discharge), which is different from our focus on decision support. Crawford et al. (2014) develop a simulation study on the impact of inpatient discharge policies on ED congestion and readmission, where they evaluate the performance of three fixed discharge policies. Armony and Yom-Tov (2018) study discharge management for hematology patients who have both risk of infection and risk of mortality after chemotherapy. They leverage fluid approximations and perform steady-state analysis to identify discharge thresholds that minimize the combined patients' infection and mortality risks under capacity constraints.

Chan et al. (2012) consider the scenario where a new patient arrives to a full ICU, and doctors must decide which patient to discharge to free a bed. Our discrete time model is more natural for the inpatient discharge setting because discharges are usually processed once a day during rounds. Hence, we determine both *which patients* and *how many* patients to discharge, considering patient risk trajectories and current and future occupancy. We also track how long a patient has stayed, linking this to a LOS-dependent risk from our prediction model. Ouyang et al. (2020) consider the joint decision of ICU admission and discharge decisions. The decision maker decides whether to admit an arriving patient to the ICU or to the general ward and also who to discharge early if a patient needs to be admitted to a full ICU. An important insight the authors provide is that the optimal decisions depend not only on the expected ICU benefit of a patient but also how long the patient will stay to get this benefit. This is similar to our finding that the discharge *desirability* of a patient depends on the magnitude of risk reduction in future days and not just the absolute risk level. Bavafa et al. (2019) study the joint problem of coordinating elective case mix and discharge policies. They find that coordination has benefits over a siloed approach when costs of either the operating theater and/or inpatient beds are sufficiently high. Atlaeddini et al. (2019) use a nonparametric method to predict the impact of LOS on readmission risk and demonstrate how this could be used to support discharge planning via a simple, static optimization model without modeling the readmission process.

George and Harrison (2001), Ata and Shneorson (2006), Bekker and Boxma (2007), Chan et al. (2014), and Ingolfsson et al. (2018) study optimal control of queuing systems with state-dependent service rates. Huang and Gurvich (2018) and Braverman et al. (2020) develop new frameworks to identify asymptotic optimal control in single-server queues with abandonment. Such service rate control works generally provide an optimal rate but not which customer(s) to discharge. Furthermore, customers are usually assumed to be homogeneous, lacking rich, personalized profiles.

2.3. Readmission Prediction

The closest works relating to our readmission prediction model are those of Bardhan et al. (2014), Bartel et al. (*2019*), and Helm et al. (2016). The two-stage concept we use comes from Bardhan et al. (2014), but the authors take LOS as an exogenous variable, as in Helm et al. (2016). We use a similar IV strategy as Bartel et al. (2020) to correct for endogeneity, but that paper does not have the prediction for timing. Having the readmission density timing is not only for the completeness of the patient flow model in our discharge optimization but also because (1) having more detailed knowledge of when a patient is at risk enables the hospital to better target postdischarge follow-ups (as requested by our partner hospital), and (2) the medical literature has shown that the timing of readmission is correlated with readmission intensity (e.g., resource use, burden on staff; Skolarus et al. 2015).

3. Modeling Framework for Discharge Decision Optimization

In this section, we formulate an infinite-horizon, average-cost MDP for optimal discharge decisions based on the predicted risk trajectories from Section 6.1. This discharge decision framework is designed for use with patients whose medical LOS falls into a normal range (e.g., 1–15 days), which compromises most inpatients. We are not suggesting that our model should control the discharge of patients having excessively long LOSs or complicated reasons for remaining in the hospital, who should be handled on a case-by-case basis based on the doctor's discretion.

3.1. Patient Flow Model

Figure 3 depicts the patient flow model. New patients arrive to a hospital ward with N beds according to a time-nonhomogeneous Poisson process with a periodic arrival rate function $\lambda(t)$. At this point, we assume that the period is one day, with

$$\Lambda = \int_0^1 \lambda(s)ds = \int_t^{t+1} \lambda(s)ds \tag{1}$$

denoting the exogenous daily arrival rate. We study the impact of the day-of-week arrival variability in Section 7.3. Patients are admitted following the first-come,

first-serve discipline. After admission, each patient's LOS depends on the discharge action and is no longer an exogenous variable as in conventional queuing models. Once a patient is discharged, he or she is either cured or will be readmitted with probability that depends on their risk class and LOS at discharge.

We assume that each patient belongs to a discharge risk class m with probability p_m, where there are M possible classes; that is, $m = 1, \ldots, M$, and $\sum_{m=1}^{M} p_m = 1$. Let $r(m; j)$ be the readmission probability for a class m patient with an LOS of j, and let $q(t; m, j)$ be the probability that this class m patient will be readmitted on day t after discharge. (We will remove this classification later to account for individual risk curves of each patient in the hospital when a discharge decision must be made; see Remark 2 and Section 5.1.3.) We have

$$\sum_{t=1}^{T} q(t; m, j) = r(m, j), \tag{2}$$

where T denotes the maximum time we count a patient visit as a readmission (90 days in this paper). A class m patient may become a different class $\tilde{m}$ patient on readmission, with probability $\delta_{m,\tilde{m}}$ ($\sum_{\tilde{m}=1}^{M} \delta_{m,\tilde{m}} = 1$), because prior readmission is found to be an important factor in our prediction analysis. Correspondingly, we define one more quantity for later use, $q_{\tilde{m}}(t; m, j) = \delta_{m,\tilde{m}} \cdot q(t; m, j)$.

Here, $r(m; j)$ and $q(t; m, j)$ depend not only on class m but also on LOS (index j) before being discharged, where LOS is no longer an exogeneous variable but depends on the discharge decision to be specified in the next section. As key model inputs, these risk trajectories $r(m; j)$ and $q(t; m, j)$ are estimated from our prediction model developed in Section 6. Figure 4(a) shows the *personalized* predictions for the 90-day readmission probability, as a function of LOS, for 50 random patients from our data. Figure 4(b) shows the aggregate risk trajectories $\{r(m; j)\}$ for $M = 3$ classes using the k-means method to group patients based on the personalized risk curves. Figure 4(c) shows the estimated density curves $q(t; m, j)$ for $t = 7, 14, 21, 30, 60$ days after discharge for class $m = 2$.

Remark 2 (Personalized Trajectory)**.** For analytical convenience, we assume in the modeling framework the patient classifications and the aggregated risk trajectories. However, our eventual optimization algorithm developed in Section 5.1 does *not* rely on the classification and can directly take in the personalized risk trajectory from the prediction. We remove the risk classification and incorporate personalized patient risk in the numerical study and implementation in Section 7.

3.2. Infinite-Horizon Average Cost Problem

We formulate the discharge decision as a discrete-time, infinite-horizon average cost MDP. We specify the system state, action, transition dynamics, cost, and objective function.

3.2.1. System State. The system state is captured by the following $M \times (J + 1)$-dimensional vector:

$$X(t) = \left(X^0(t), X^1(t), \ldots, X^J(t)\right)',$$

where $X^j(t) = (X^{1,j}(t), X^{2,j}(t), \ldots, X^{M,j}(t))'$, and $X^{m,j}(t)$ denotes the number of class m patients who have spent j days in the system, $j = 0, \ldots, J$. Note that $X^{m,j}(t)$ includes both patients in service and waiting for a bed. Here waiting may take various forms other than being physically waiting, for example, patients who have finished treatment but have to remain in the ED (known as *ED boarding*) before being admitted to the inpatient hospital and patients who have received surgery waiting in the recovery room, among others. Hospitals are able to track the patient demand (and thus $X^{m,j}(t)$) through bed management systems that record times of bed requests and bed assignments. We assume that a patient begins the treatment and recovery process immediately on arrival because patients still receive care even if not immediately placed in a bed.

3.2.2. Action. Each day the decision maker observes the system state at a decision epoch (e.g., at the time of rounds) and determines the number of patients to discharge. For notational convenience, we assume that this observation occurs at time zero of each day. Mathematically, let $\{X(t), t \geq 0\}$ denote the system state, which is a continuous-time stochastic process.

Figure 3. Patient Flow Model of the Hospital Ward

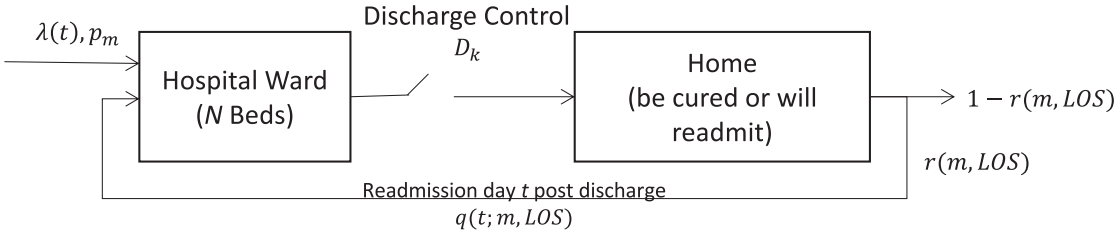
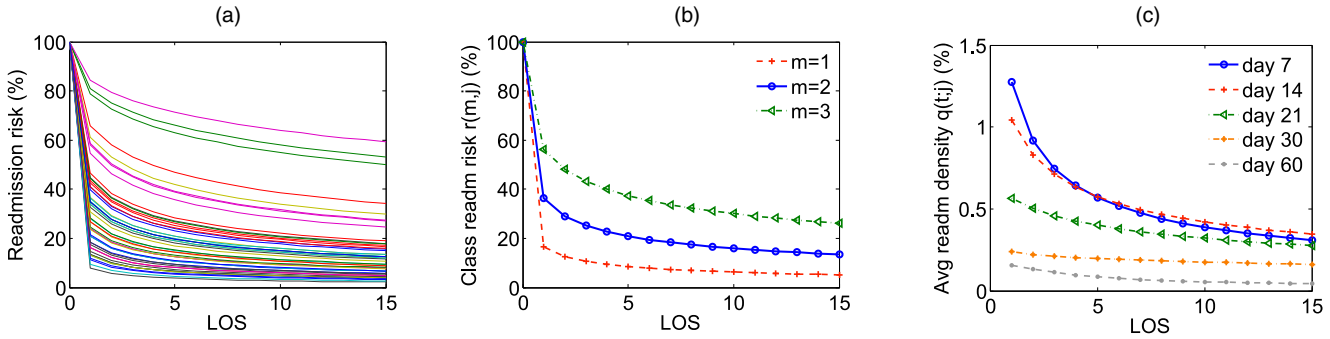


Figure 4. Predicted Readmission Risk Trajectory against LOS



Notes. The first and second plots show the 90-day cumulative probability as a function of LOS (50 random patients versus average curves for $M = 3$ classes from the clustering). The third plot shows the readmission density on different days (7–60 days) as a function of LOS for class $m = 2$ under $M = 3$ classes.

Let $X(k-)$ and $X(k)$ denote the preaction and postaction state at decision epoch k (day k). Unless otherwise specified, we use $X_k = X(k-), k = 0, 1, \ldots,$ to denote the preaction state. At decision epoch k, we take discharge action $D_k = (D_k^0, D_k^1, \ldots, D_k^J)'$, where $D_k^j = (D_k^{1,j}, D_k^{2,j}, \ldots, D_k^{M,j})'$, and $D_k^{m,j}$ represents the number of discharges of class m patients who have spent j days in the system.

As Berk and Moinzadeh (1998) point out, a patient may progress through a critical stage and then a stable stage, and the patient can only be discharged in the stable stage. To capture this feature, we impose a minimum LOS requirement on the discharge actions. That is, for patients belonging to class m, we can only discharge them when their LOS reaches a class-dependent threshold $\underline{L}_m \in [0, J]$. Incorporating this minimum LOS requirement, a feasible discharge action satisfies:

$$D_k^{m,j} \in \left[0, X_k^{m,j}\right], \ \forall m, j \geq \underline{L}_m; \tag{3}$$

$$D_k^{m,j} = 0, \qquad \forall m, j < \underline{L}_m. \tag{4}$$

3.2.3. Transition Dynamics. Let $A_k^{m,0}$ denote the number of new arrivals belonging to class m in period k (between decision epochs k and $k+1$), and let $A_k^{'m,0}$ denote the number of readmissions belonging to class m. For each m, the state evolution is

$$X_{k+1}^{m,0} = A_k^{m,0} + A_k^{'m,0}; \tag{5}$$

$$X_{k+1}^{m,j} = X_k^{m,j-1} - D_k^{m,j-1}, \quad j = 1, \ldots, J. \tag{6}$$

Equation (5) captures the arrivals to the hospital ward in period k. Equation (6) says that patients who have stayed $j-1$ days in period k become patients who have stayed j days in period $k+1$, except for those who are discharged, $D_k^{m,j-1}$.

The total number of exogenous arrivals $\sum_{m=1}^{M} A_k^{m,0}$ follows a Poisson distribution with mean Λ. By Poisson splitting, $A_k^{m,0}$ is distributed as $Poiss(\Lambda_m)$ with $\Lambda_m = p_m \Lambda$. The readmission arrival stream $A_k^{'m,0}$, by contrast, depends on past discharge actions; that is,

$$A_k^{'m,0} = \sum_{t=1}^{T} \sum_{j=1}^{J} \sum_{\tilde{m}=1}^{M} Bin\left(D_{k-t}^{\tilde{m},j}, q_m(t; \tilde{m}, j)\right),$$

where $Bin(\cdot, \cdot)$ denotes a binomial random variable, and $q_m(t; \tilde{m}, j)$ and T are given in Section 3.1.

3.2.4. One-Period Cost. The current period cost function is composed of the ward occupancy *congestion cost* $c_h(X_k)$ and the *discharge cost* $c_d(D_k)$, which depends on the expected number of readmissions given the patients being discharged. One reasonable form for $c_h(X_k)$ and $c_d(D_k)$ follows:

$$c_h(X_k) = C \cdot (S_k - N)^+, \tag{7}$$

where C is the unit holding cost, $S_k = \sum_{m=1}^{M} \sum_{j=0}^{J} X_k^{m,j}$ is the total patient census, and $(S_k - N)^+$ captures the *overage*, that is, the number of patients who cannot be accommodated in a ward bed at epoch k; that is,

$$\begin{aligned} c_d(D_k) &= \sum_{m=1}^{M} \sum_{j=0}^{J} \sum_{t=0}^{T} R_t \mathbb{E}\left[Bin\left(D_k^{m,j}, q(t; m, j)\right)\right] \\ &= \sum_{m=1}^{M} \sum_{j=0}^{J} D_k^{m,j} \sum_{t=0}^{T} R_t q(t; m, j), \end{aligned} \tag{8}$$

where $q(t; m, j)$ is given in (2), and R_t denotes the corresponding penalty cost. We allow the penalty cost R_t to depend on the timing of readmission because early readmitters are found to require more intensive care than late readmitters (Skolarus et al. 2015). In the case where $R_t = R$ for all t, we have

$$c_d(D_k) = \sum_{m=1}^{M} \sum_{j=0}^{J} D_k^{m,j} R \cdot r(m, j)$$

from (2). The overage cost from (7) is commonly adopted in modeling inpatient ward congestion (Samiedaluie et al. 2017, Armony et al. 2018), which captures the undesirable consequences for blocked admissions to inpatient beds, for example, patients boarding in the ED and cancellation of elective surgeries. Unused inpatient bed capacity is often regarded as a sunk cost. In Section D.1 of Online Appendix D, we also consider other forms of $c_h(X_k)$ such as $c_h(X_k) = C \cdot S_k$ and $C \cdot S_k^2$ that capture the effect of the total patient census and ward occupancy.

3.2.5. Objective. Let $\Pi = \{D_k^{m,j}\}$ be the set of admission policies such that $D_k^{m,j}$ satisfies (3) and (4). Let $\{X_k^\pi\}$ be the resulting state under a policy $\pi \in \Pi$. We can now write the objective function for the infinite-horizon long-run average cost problem for policy $\pi \in \Pi$:

$$Z^\pi = \limsup_{\tau\to\infty} \frac{1}{\tau} \sum_{t=0}^{\tau-1} \mathbb{E}\left[c_h\left(X_k^\pi\right) + c_d(D_k)\right]. \tag{9}$$

We end this section with two remarks on the cost parameters and the readmission arrivals $\{A_k^{'m,0}\}$.

Remark 3 (Cost Parameters)**.** The cost parameters C and R_t may be difficult to estimate in practice. In our analysis, we use them primarily as *tuning parameters* to reflect the tradeoff between discharging too few versus too many patients. In the numerical study and the implementation, we derive efficient frontiers for the decision makers to identify an operating regime to achieve their desired performance measures, eliminating reliance on the cost parameters themselves.

Remark 4 (Readmission Arrival)**.** Because the readmission window is long and the readmission probability is not large (with an average of around 10%–20%), each discharged patient's contribution to the readmission arrival rate on any given day is small. This leads to a smoothing effect on the readmission arrival rate across discharge policies as long as the day-to-day discharge actions do not fluctuate too much (Greenberg 1989; see Section D.2 of Online Appendix D for some numerical evidence). To maintain the Markov property of the MDP and to obtain insightful structural properties for developing solution algorithms, we consider the readmission arrivals $\{A_k^{'m,0}\}$ as exogenous variables for Theorem 1 and technical results in Sections 4 and 5. However, in the numerical study in Section 7 and implementation, we relax the assumption on the readmission arrivals being exogenous. We simulate patient readmissions according to the estimated readmission timing distributions, which depend on the discharge actions and individual patient characteristics. We show that although the solution algorithm is developed under the exogenous assumption, it provides significant improvement over current practice. To implement the algorithm in the relaxed setting, we estimate the readmission arrival distributions from a static policy developed in Section 3.3, which explicitly characterizes the impact of discharge on readmission and has an analytical solution.

3.3. Bellman Equation and Challenges in Solving the MDP

Denote the optimal solution to (9) as

$$\gamma^* = \inf_{\pi\in\Pi} Z^\pi. \tag{10}$$

Theorem 1 proves the existence of an average cost optimal stationary policy. Its proof is detailed in Section B.1 of Online Appendix B.

Theorem 1. *For the average cost optimality equation defined by Equations* (9) *and* (10)*, there exists an average cost optimal stationary policy.*

Let A_k denote the vector of random arrivals from each class, including both new and readmission arrivals (see Remark 4). For a given state $x = (x^0, x^1, \ldots, x^J)$ with $x^j = \{x^{1,j}, \ldots, x^{M,j}\}$, the Bellman equation is given by

$$V(x) = \min_{D\in\Pi}\ c_h(x) + c_d(D) - \gamma^* + \mathbb{E}_{A_k} V\left(A_k, x^0 - D^0, x^1 - D^1, \ldots, x^{J-1} - D^{J-1}\right), \forall x \in \mathcal{S}, \tag{11}$$

where $D = (D^0, \ldots, D^J)$ is the action vector with $D^j = (D^{1,j}, \ldots, D^{M,j})'$, $\mathcal{S}$ is the state space, γ^* is the optimal long-run average cost, and $V(\cdot)$ is the (relative) value function.

Because we have to track both the patient class (index m) and how long the patient has spent in the hospital (index j) in the state space and the action space, solving the Bellman equation (11) has a curse of dimensionality. If we cap the number of patients of each type (class, LOS) to be $\bar{S}$, both the state space and action space are of size $\bar{S}^{M\cdot(J+1)}$. In a simple two-class setting where patients are kept at most three days and $\bar{S} = 30$, the state space is of size $30^{4\times2} = 6.56\times10^{11}$. Conventional MDP techniques such as value or policy iteration become computationally challenging, if not entirely infeasible, even in this simple setting. In Sections 4 and 5, by identifying useful structural properties and leveraging solutions from a special cost setting, we heuristically convert this MDP to a univariate optimization problem and develop an efficient dynamic algorithm.

3.3.1. A Benchmark Solution: Static Discharge Thresholds. When the action space is constrained to static-threshold policies (i.e., a patient is not discharged until his or her risk level drops below a preset threshold), the MDP becomes tractable via steady-state analysis, which we specify later. The resulting static policy provides a benchmark solution that we can compare with policies from the dynamic algorithm developed under the full action space. In addition, this steady-state analysis also generates useful insights into the interplay between system congestion and discharge risk.

For the static optimization, there is a one-to-one relationship between the risk level and the LOS. Thus, it is equivalent to optimize the thresholds on LOS for each class m, denoted as l_m. For a given policy π with thresholds for LOS, $(l_1, \ldots, l_M)$, the distribution on number of discharges from class m on day k, $D_k^m = X_k^{m,l_m}$, is stationary under π. Let $\mathbb{E}[D^m]$ denote the steady-state expectation of D_k^m, and let $\mathbb{E}[Q]$ denote the steady-state expected queue length under policy π. We show in Section B.6 of Online Appendix B that minimizing the long-run average cost is equivalent to minimizing

$$\sum_{m=1}^{M}\sum_{t=1}^{T} R_t \cdot q_{\tilde{m}}(t; m, l_m) \cdot \mathbb{E}[D^m] + C \cdot \mathbb{E}[Q]\,. \tag{12}$$

Here $\mathbb{E}[D^m]$ can be found by solving the following set of flow-balance equations:

$$\begin{aligned}\mathbb{E}[D^m] &= \mathbb{E}\big[A_k^{m,0}\big] + \mathbb{E}\big[A_k'^{\,m,0}\big] \\ &= \Lambda_m + \sum_{\tilde{m}=1}^{M}\sum_{t=1}^{T} q_m(t; \tilde{m}, l_{\tilde{m}}) \cdot \mathbb{E}\big[D^{\tilde{m}}\big], \\ &\quad m = 1, \ldots, M.\end{aligned} \tag{13}$$

If patients do not change classes on readmission, we have $\mathbb{E}[D^m] = \Lambda_m/(1 - r(m, l_m))$ by (2). For $\mathbb{E}[Q]$, we discharge all $D_k^m = X_k^{m,l_m}$ patients on each day k. Here X_k^{m,l_m} is composed of new arrivals $A_{k-l_m}^m \sim Poiss(\Lambda_m)$ and readmissions $A_{k-l_m}'^{\,m}$, which we approximate as a Poisson for tractability. Then $\mathbb{E}[Q] \approx \mathbb{E}[(\sum_{m=1}^{M}\sum_{j=0}^{l_m-1} Poiss(\Lambda_m + \mathbb{E}[A_k'^{\,m,0}]) - N)^+]$.

For ease of exposition, we present the results on the optimal occupancy in steady state for a single class of patients; the results and insights extend to the multiclass problem. We drop the index on class m and use ℓ to denote the discharge threshold, with a slight abuse of notation, to define the associated discharge risk as $q(t;\ell)$ and $r(\ell) = \sum_{t=1}^{T} q(t;\ell)$. Let $\tilde{\Lambda} = \Lambda/(1 - r(\ell)) = \Lambda + \Lambda r(\ell)/(1 - r(\ell))$ denote the total arrival rate including readmissions.

Proposition 1. *Under a normal approximation for the Poisson distribution with mean $\tilde{\Lambda}$, the optimal discharge threshold l solves the following equation:*

$$\begin{aligned}&\sum_{t=1}^{T} R_t \left|\frac{\partial q(t;\ell)}{\partial \ell}\right| \cdot \tilde{\Lambda} + \sum_{t=1}^{T} R_t \cdot q(t;\ell) \left|\frac{\partial \tilde{\Lambda}}{\partial \ell}\right| \\ &\qquad = C(1 - \Phi(\alpha))\frac{\partial B}{\partial \ell},\end{aligned} \tag{14}$$

where $\Phi(\cdot)$ is the cumulative distribution function of standard normal, $B = \ell \cdot \tilde{\Lambda}$ is the system offered load, and $\alpha = \frac{N-B}{\sqrt{B}}$.

The proof is given Section B.6 of in Online Appendix B. Equation (14) characterizes the optimal discharge threshold and, consequently, the optimal offered load B under the given cost parameters.

We make two observations from Proposition 1. First, the equation characterizes the optimal threshold in terms the marginal increase in readmission risk versus marginal increase in congestion. That is, the left-hand side depends on the marginal change in the discharge risk $\partial q(t;\ell)/\partial \ell$; the right-hand side is the probability of exceeding capacity $(1 - \Phi(\alpha))$ times the marginal increase in system workload $\partial B/\partial \ell$, that is, the marginal for system congestion. To see the latter, note that

$$\begin{aligned}1 - \Phi(\alpha) &= \mathbb{P}(Z > \alpha) = \mathbb{P}\left(Z > \frac{N - B}{\sqrt{B}}\right) \\ &= \mathbb{P}\Big(B + Z\sqrt{B} > N\Big),\end{aligned}$$

where $B + Z\sqrt{B}$ approximates the total number of patients (or workload) in the system. This equation reveals the interplay between the two key tradeoffs we capture in the decision framework: balancing readmission risk and ward congestion.

Second, the optimal threshold explicitly depends on R_t and $\partial q(t;\ell)/\partial \ell$, the time-dependent cost parameters, and how ℓ affects the readmission density $q(t;\ell)$ for different t. Figure 4(c) shows that extending LOS mostly impacts the readmission density for t before 21 days; for t larger than 60 days, the density has negligible changes when changing LOS. Thus, if one puts more weight on readmissions before 21 days, it is more beneficial to extend the patient's stay, and vice versa.

4. Who to Discharge: Structural Properties on Optimal Actions

In this section, we analyze the structure of the optimal solution to the Bellman equation and establish two results: (1) a ranking of patients to discharge and (2) a threshold discharge policy that follows the ranking when a strong dominance property holds. That is, the

optimal policy will discharge all patients of a higher rank before discharging any patients of a lower rank in terms of strong dominance. The structural results developed here provide insights into one of the two key research questions: who to discharge; we answer the how many to discharge question in Section 5.

Definition 1 (Strong Dominance)**.** Define patient type via (class, LOS). Type (m_1, t_1) strongly dominates type (m_2, t_2), or $(m_1, t_1) \succ (m_2, t_2)$, if and only if

$$\left|\frac{\partial r(m_1, t_1 + t)}{\partial t}\right| \le \left|\frac{\partial r(m_2, t_2 + t)}{\partial t}\right|, \quad \forall t \ge 0. \tag{15}$$

The strong dominance says that the (absolute) marginal change in the readmission risk between today and *any* future day for type (m_1, t_1) is *smaller* than for type (m_2, t_2). At a high level, one trajectory strongly dominates another if the absolute value of the slope in the LOS dimension is smaller. Because the readmission risk always decreases in the LOS from our prediction, the derivative is negative, and this definition is equivalent to $\frac{\partial r(m_1,t_1+t)}{\partial t} \ge \frac{\partial r(m_2,t_2+t)}{\partial t}, \forall t \ge 0$.

Strong dominance always holds within a class. That is, for any $t_1 > t_0$, $(m, t_1) \succ (m, t_0)$, the marginal change in the readmission risk for a patient with a longer LOS is always smaller than that for a patient with a shorter LOS because $r(m, \cdot)$ is decreasing and convex (see the proof of convexity in Section B.5 of Online Appendix B). Our prediction results also show that this strong dominance property is satisfied when we group patients into M from 3 to 10 classes and use the corresponding aggregated risk trajectories (Figure 4(b) shows aggregated curves for $M = 3$). However, the strong dominance does not necessarily hold for any two patients when we use completely personalized risk trajectories in the numerical study and implementation. In this case, although the optimality of the structural properties we prove in the rest of this section is not guaranteed, we follow the same spirit and develop a weak dominance criterion that still allows us to rank patients and apply a similar threshold discharge policy (see Section 5.1.3).

We start analyzing the structural of the optimal actions by first proving the following proposition, which demonstrates that keeping the strong-dominant patient longer provides less benefit (smaller discharge cost reduction) than keeping the dominated patient longer. Let $\mathbf{e}_{(m,t)}$ denote the unit vector with one in the position corresponding to type (m, t) and zero elsewhere. Adding this vector to D_k indicates adding a single discharge of patient type (m, t) in the action.

Proposition 2. *For $c_d(D)$ of the form (8) and $R_t = R$, if $(m_1, t_1) \succ (m_2, t_2)$, then in any epoch k, the following holds for any future epoch, $k' > k$: $c_d(D_k + \mathbf{e}_{(m_1,t_1)}) - c_d(D_{k'} + \mathbf{e}_{(m_1,t_1+k'-k)}) \le c_d(D_k + \mathbf{e}_{(m_2,t_2)}) - c_d(D_{k'} + \mathbf{e}_{(m_2,t_2+k'-k)})$.*

The proof is in Section B.2 of Online Appendix B. Next, we leverage Proposition 2 to prove a theorem that allows us to rank patients and discharge them in strict order of their ranking. We specify the proof for the simpler case where the two interchanged patients have not reached their maximum LOS and leave the other more tedious case to Section B.3 of Online Appendix B. The unit vector $\mathbf{e}_{(m,t)}$ multiplying D_k or X_k gives the number of (m, t) patients in the corresponding action or state; for example, $D_k \cdot \mathbf{e}_{(m,t)} = D_k^{m,t}$.

Theorem 2. *For $c_d(D)$ of the form (8) and $R_t = R$, consider two patient types $(m_1, t_1) \succ (m_2, t_2)$. Then the optimal discharge action $D_k^* \cdot \mathbf{e}_{(m_2,t_2)} > 0$ only if $D_k^* \cdot \mathbf{e}_{(m_1,t_1)} = X_k \cdot \mathbf{e}_{(m_1,t_1)}$; that is, we would discharge patient type (m_2, t_2) only if we have discharged all type (m_1, t_1) patients.*

Proof. We prove the theorem via an interchange argument on two patients of type (m_1, t_1) and (m_2, t_2), respectively. Suppose at period k that $D_k^* \cdot \mathbf{e}_{(m_2,t_2)} = 1$ and that $D_k^* \cdot \mathbf{e}_{(m_1,t_1)} = X_k \cdot \mathbf{e}_{(m_1,t_1)} - 1$. If $D_k^* \cdot \mathbf{e}_{(m_2,t_2)} = n$ and $D_k^* \cdot \mathbf{e}_{(m_1,t_1)} = X_k \cdot \mathbf{e}_{(m_1,t_1)} - n$, we can repeat the interchange argument iteratively to achieve the same result because the discharge cost function is linear. Suppose that this one patient of class (m_1, t_1) who was not discharged in period k is discharged at a later time k'. Call this policy π_1, with value function $V_k^{\pi_1}(X_k)$. Now consider a second policy π_2 that switches the discharge timing of the type (m_1, t_1) patient and the type (m_2, t_2) patient that we track in the interchange argument. All other actions remain the same. First, consider the case where $t_1 + k' - k \le J$, that is, that k' is not beyond type (m_2, t_2) patient's maximum LOS. Let D_k and $D_{k'}$ be the discharge actions excluding the two switched patients. Then

$$\begin{aligned} V^{\pi_2} - V^{\pi_1} &= c_d\big(D_k + \mathbf{e}_{(m_1,t_1)}\big) + c_d\big(D_{k'} + \mathbf{e}_{(m_2,t_2+k'-k)}\big) \\ &\quad - c_d\big(D_k + \mathbf{e}_{(m_2,t_2)}\big) - c_d\big(D_{k'} + \mathbf{e}_{(m_1,t_1+k'-k)}\big) \\ &= c_d\big(D_k + \mathbf{e}_{(m_1,t_1)}\big) - c_d\big(D_{k'} + \mathbf{e}_{(m_1,t_1+k'-k)}\big) \\ &\quad - \big(c_d\big(D_k + \mathbf{e}_{(m_2,t_2)}\big) - c_d\big(D_{k'} + \mathbf{e}_{(m_2,t_2+k'-k)}\big)\big) \\ &\le 0\,. \end{aligned}$$

The first equality follows because the occupancies are the same under both policies under all sample paths, so the occupancy costs cancel out, and so do the discharges costs for all patients except the two interchanged patients because our discharge cost is linear. The inequality follows from Proposition 2, which shows that the first two terms in the second line are smaller

than the second two. The proof for the second case where $t_1 + k' - k > J$ is given in Section B.3 of Online Appendix B. Because π_2 produces a smaller cost than π_1, π_1 cannot be the optimal policy, contradicting the assumptions made at the beginning. □

Theorem 2 implicitly provides a ranking of patient types. In particular, this discharge ranking depends on marginal risk and not absolute risk, as has been the prevailing approach according to our discussions with hospitals. We now formalize the ranking, letting $[i]$ be the ith-ranked patient type, which means that $[j] \succ [i], \forall j < i$, [1] being the highest rank (most desirable to discharge). We prove the univariate threshold structure of the optimal policy.

Corollary 1. *The optimal discharge policy D_k^* for the optimization defined by (11) is of threshold form, with univariate threshold $\tilde{D}^*(X)$, where*

$$D_k^* \cdot \mathbf{e}_{[i]} = \min\left(X_k^{[i]}, \left(\tilde{D}(X_k) - \sum_{j<i} X_k^{[j]}\right)^+\right). \tag{16}$$

Proof. We prove this by contradiction. Suppose that policy D_k is not of the form in (16). Then it must be the case that there exists $D_k \cdot \mathbf{e}_{[i]} > 0$ and $D_k \cdot \mathbf{e}_{[j]} < X_k^{[j]}$ for some $j < i$. However, this cannot be optimal by Theorem 2. □

Remark 5. For ease of exposition, in the rest of this paper, we let $R_t = R$. The analytical results from this section extend to the more tedious time-varying case R_t if we redefine the strong-dominance criterion for $(m_1, t_1) \succ (m_2, t_2)$ as $\left|\frac{\partial \tilde{R}(m_1, t_1+t)}{\partial t}\right| \le \left|\frac{\partial \tilde{R}(m_2, t_2+t)}{\partial t}\right| \forall t \ge 0$, where $\tilde{R}(m, j) = \sum_{t=0}^{T} R_t q(m; j, t)$.

5. How Many to Discharge: Dynamic Discharge Decision Support

In the preceding section, we answered the question of whom to discharge. To solve the final discharge optimization, however, we still need to answer the question of how many to discharge from the high-dimensional MDP. To overcome this challenge, we leverage a linear-quadratic solution from a special cost setting to approximate the cost-to-go in the original MDP. Along with the structure of the optimal policy obtained in Section 4, we transform the original MDP into a univariate optimization problem, which significantly reduces the computational complexity and makes the solution tractable. In Section 5.1, we first present this dynamic algorithm. We then show its adaption to the realistic, complex hospital environment by relaxing the analytical assumptions in Section 3. In Section 5.2, we demonstrate the near-optimal performance of this dynamic algorithm in small-scale problems where the conventional value iteration is still feasible.

5.1. Dynamic Discharge Algorithm

The key to the univariate transformation of our dynamic algorithm relies on two properties: (1) the structural properties proved in Section 4 map the high-dimensional action space into an equivalent univariate action space, and (2) analytical solutions from a special quadratic cost setting allow us to approximate the cost-to-go with a quadratic function of the total occupancy, which only depends on the univariate action in (1), maintaining the univariate action space. We first analyze this special cost setting as a linear-quadratic stochastic control in Section 5.1.1. Then we show how to leverage its solutions to approximate the cost-to-go in the original MDP and transform it into the univariate optimization in Section 5.1.2. In Section 5.1.3, we discuss how to adapt the dynamic algorithm when we incorporate personalized risk trajectories (where strong dominance may no longer hold) and nonstationary arrivals. This demonstrates the flexibility and robustness of our dynamic algorithm in more complex hospital environments.

5.1.1. Linear-Quadratic Special Case. In this section, we consider a special finite-horizon version of (9) with $T + 1$ periods and a quadratic structure for both $c_h(\cdot)$ and $c_d(\cdot)$. Recall that $S_k = \sum_{m,j} X_k^{m,j}$ denotes the total number of patients in the system. Let $\tilde{c}_h(X_k) = C \cdot S_k^2$ and $\tilde{c}_d(D_k) = R \cdot \sum_{m,j} r(m, j)(D_k^{m,j})^2$, where we use $\tilde{c}_h$ and $\tilde{c}_d$ to emphasize that they are for the special setting. Furthermore, we assume that (1) $X_k^{m,J}$ is an absorbing state—that is, once reaching day J (the maximum LOS in the original problem), the patients will stay in that state—and (2) the action space is unconstrained—that is, $\Pi_{un} = \{D_k^{m,j} \in \mathbb{R}\}$. Given an initial state X_0, the optimization problem can be written as

$$\begin{aligned} V_{k,LQ}^{\pi}(X_k) &= \sum_{t=k}^{T} \mathbb{E}\left[C \cdot S_t^2 + R \cdot \sum_{m,j} r(m,j)\left(D_t^{m,j}\right)^2\right] \\ &\quad + \mathbb{E}\left[C \cdot S_{T+1}^2\right], \\ V_{LQ}^{*}(X_0) &= \min_{\pi \in \Pi_{un}} V_{0,LQ}^{\pi}(X_0). \end{aligned} \tag{17}$$

Proposition 3 shows that in this finite-horizon setting, the optimal discharge action is linear in S_k, and the value function is a quadratic function of S_k.

Proposition 3. *The optimal discharge decision of problem (17) is given by*

$$D_{k,LQ}^{m,j} = a_k^{m,j} S_k + b_k^{m,j}, \tag{18}$$

and the value function is given by

$$V^{\pi}_{k,LQ}(X_k) = \alpha_k S_k^2 + \theta_k S_k + \kappa_k\,. \tag{19}$$

Here $a_k^{m,j}$ *denotes the* $(j-1)M + m^{th}$ *entry of vector* $\mathbf{U_k^{-1}}$ $(\alpha_{k+1},\ldots,\alpha_{k+1})'$, *and* $b_k^{m,j}$ *denotes the* $(j-1)M + m^{th}$ *entry of vector* $\mathbf{U_k^{-1}B_k}$. *The matrix* $\mathbf{U_k}$ *and vector* $\mathbf{B_k}$ *are given by*

$$\mathbf{U_k} = \begin{pmatrix} \alpha_{k+1}+R_0 & \alpha_{k+1}\ldots & \alpha_{k+1} \\ \alpha_{k+1} & \alpha_{k+1}+R_1\ldots & \alpha_{k+1} \\ & \cdots & \\ \alpha_{k+1} & \alpha_{k+1}\ldots & \alpha_{k+1}+R_J \end{pmatrix},$$

$$\mathbf{B_k} = (\alpha_{k+1}\mathbb{E}[A_k] + \theta_{k+1}/2)\cdot\mathbf{1},$$

where $R_j = R\cdot(r(1,j),\ldots,r(M,j))'$, $\mathbf{1} = (1,\ldots,1)'$ *denotes the ones vector,* A_k *denotes the total arrivals from all classes including readmissions, and constants* α_k, θ_k, *and* κ_k *can be recursively calculated using* a_k, b_k, α_{k+1}, θ_{k+1}, and κ_{k+1}.

The proof in Section B.4 of Online Appendix B uses an induction argument where we also provide specifications of the constants α_k, θ_k, and κ_k. In addition, $D^{m,j}_{k,LQ}$ in Equation (18) connects to the linear decision rule used in stochastic optimization (Chen et al. 2008). Next, we use this linear decision rule as an approximation for the optimal actions (in future periods) in the original MDP and the quadratic function $V^{\pi}_{k,LQ}(X_k)$ as an approximation for the cost-to-go. We refer to these approximations as the *linear-quadratic approximation*. This is also the main purpose for analyzing this special case: to provide an analytical building block for tackling the original, general-cost MDP.

5.1.2. Transformation into Univariate Optimization. ***5.1.2.1. Action Space Reduction.*** If the strong-dominance property holds, Corollary 1 allows us to reduce the action space to a univariate decision $\tilde{D} = \tilde{D}(X_k)$, the total number of discharges on day k; there is a one-to-one mapping from $\tilde{D}$ to the discharge action $D_k^{m,j}(\tilde{D})$ for each class m and LOS j from (16). When strong dominance does not hold, we develop in Section 5.1.3 a more general criterion that still allows us to rank patients and maintain a one-to-one mapping from $\tilde{D}$ to $D_k^{m,j}(\tilde{D})$ such that we can still focus on solving the univariate decision $\tilde{D}$.

5.1.2.2. Cost-to-Go Approximation. To approximate the cost-to-go function (the last term in (11)), we start from the state in period k and expand period $k+1$ with actions approximated by the linear decision rule (18), $a^j_{k+1}S_{k+1} + b^j_{k+1}$, where $a^{m,j}_{k+1}$ and $b^{m,j}_{k+1}$ are the linear coefficients for the number of class (m,j) patients to discharge given total occupancy of S_{k+1} in period $k+1$. We then use the quadratic approximation for the future cost (period $k+2$ and beyond), $V^*_{LQ}(S_{k+2})$, given by (17). The optimal total discharge is therefore

$$\begin{aligned}\tilde{D}^*(X_k) = \operatorname*{argmin}_{0\le\tilde{D}\le S_k}\Bigg\{ & c_h(X_k) + \sum_{m,j} R\cdot r(m,j) D_k^{m,j}(\tilde{D}) \\ & + \mathbb{E}[c_h(X_{k+1})] \\ & + \mathbb{E}\left[\sum_m R\cdot r(m,J)\left(X_k^{m,J-1} - D_k^{m,J-1}(\tilde{D})\right)\right] \\ & + \sum_m\sum_{j\neq J} R\cdot r(m,j)\left(a^{m,j}_{k+1}S_{k+1} + b^{m,j}_{k+1}\right) \\ & + \mathbb{E}\left[V^*_{LQ}(S_{k+2})\right]\Bigg]\Bigg\}. \end{aligned} \tag{20}$$

The first two terms capture the congestion cost and discharge cost for the current period k. The third term captures the congestion cost in period $k+1$, whereas the fourth and fifth terms capture the discharge cost in period $k+1$. The fourth term is the discharge cost for patients who have reached their maximum LOS in period $k+1$, and the fifth term is the discharge cost for all other patients approximated by the optimal actions (18). The final term is the quadratic approximation for future cost in period $k+2$ and beyond.

5.1.2.3. Univariate Optimization. Equation (20) is a univariate optimization when $c_h(X_{k+1})$ only depends on S_{k+1}, for example, when c_h follows (7) or $c_h(X_{k+1}) = C\cdot S_{k+1}$ or $C\cdot S^2_{k+1}$. To see this, notice that $S_{k+1} = S_k - \tilde{D} + A_k$, and $S_{k+2} = S_{k+1} - \sum_m(X_k^{m,J-1} - D_k^{m,J-1}(\tilde{D})) - \sum_m \sum_{j\neq J}(a^{m,j}_{k+1}S_{k+1} + b^{m,j}_{k+1}) + A_{k+1}$; both only depend on the univariate decision $\tilde{D}$ given the realization of the *total* arrivals in k and $k+1$ (A_k and A_{k+1}), which are not class specific. Furthermore, $V^*_{LQ}(S_{k+2})$ only depends on the total occupancy S_{k+2}. In other words, the linear-quadratic solutions from Proposition 3 we use for approximating the future decisions and value functions only depend on the total number of discharges, which allows us to calculate the distribution of S_{k+1} and S_{k+2} with the univariate decision variable $\tilde{D}$. Combining this with the action space reduction results in a univariate optimization problem in $\tilde{D}$.

5.1.2.4. Tuning Parameters. If the number of patients reaching the maximum LOS, $X_k^{m,J-1} - D_k^{m,J-1}(\tilde{D})$, is small, which we expect to be the case given that J is chosen as an upper bound on LOS, then $S_{k+2} \approx (1 - \sum_m \sum_{j\neq J} a^{m,j}_{k+1})\, S_{k+1} - \sum_m \sum_{j\neq J} b^{m,j}_{k+1} + A_{k+1}$ is linear in S_{k+1} and A_{k+1}. Recall from Proposition 3 that $V^*_{LQ}(S_{k+2}) = \alpha_0 S^2_{k+2} + \theta_0 S_{k+2} + \kappa_0$. Thus, $\mathbb{E}_{A_{k+1}}[V^*_{LQ}(S_{k+2})]$ can be written as a quadratic function in S_{k+1}. Using the linear structure of $(a^j_{k+1}S_{k+1} + b^j_{k+1})$ in S_{k+1} and

taking out $c_h(X_k)$ because it does not depend on D_k, we can further simplify (20) as

$$\tilde{D}^*(X_k) = \underset{0\le\tilde{D}\le S_k}{\operatorname{argmin}} \left\{ \sum_{m,j} R\cdot r(m,j) D_k^{m,j}(\tilde{D}). \right. \\ \left. + \mathbb{E}_{A_k}[c_h(X_{k+1})] + \mathbb{E}_{A_k}\left[\tilde{\alpha} S_{k+1}^2 + \tilde{\beta} S_{k+1} + \tilde{\kappa}\right] \right\}, \tag{21}$$

where $\tilde{\alpha}$ and $\tilde{\beta}$ can be treated as tuning parameters. In implementation of the dynamic algorithm, we fine-tune $\tilde{\alpha}$ and $\tilde{\beta}$ to achieve a better performance for different cost structures.

The approximation for cost-to-go based on $V_{LQ}^*(S_{k+2})$ works well in most of our tested settings when the holding-cost structures are close to liner or quadratic, for example, $c_h(X_k) = C\cdot S_k$ or $C\cdot S_k^2$, or as in (7), where the queue length $(S_k - N)^+$ is a piecewise linear function and can be well approximated by quadratic functions in S_k.

5.1.3. Adaption to the Realistic Hospital Environment. ***5.1.3.1. Nonstationary Arrival.*** When the arrival rate $\mathbb{E}[A_k]$ is not stationary, for example, exhibiting the day-of-week phenomenon, we can evaluate the two expectation terms in (21) using the corresponding arrival rates and tune the parameters $\tilde{\alpha}$ and $\tilde{\beta}$ separately for different day of week.

5.1.3.2. Personalized Risk Trajectory. As mentioned in the Introduction, our readmission risk prediction is able to produce personalized risk trajectory based on individual patient profile. Notice that the last two (expectation) terms in the optimization in (21) do *not* depend on class-specific information. As long as we have a ranking that maps $\tilde{D}$ to specific individual patients to discharge, that is, $D_k^i(\tilde{D})$ for each patient i currently in the hospital unit, then we are able to eliminate the dependence on patient classification and incorporate personalized risk trajectory $r(i,j)$. That is, we simply replace $\sum_{m,j} R\cdot r(m,j) D_k^{m,j}(\tilde{D})$ with $\sum_i R\cdot r(i,j) D_k^i(\tilde{D})$ in (21).

5.1.3.3. Weak Dominance. If the strong-dominance property in (15) is a complete order (i.e., it is satisfied for any pair of patients), we can simply extend Corollary 1 to define the mapping D_k^i. However, our prediction results suggest that this is not always the case when using individual risk trajectories (even though we have this property when using the aggregated trajectories for up to $M = 10$). The structure of the optimal actions becomes much more nuanced when strong dominance is violated.

To address this challenge, we again leverage the linear-quadratic approximations from Proposition 3 to approximate future decision rules and value functions. Combining with a decomposition heuristic specified in Section C.1 of Online Appendix C, we identify determinants that drive the discharge quantities for each patient type (m,j) and obtain a more general weak-dominance criterion in terms of the following score:

$$\omega(m,j) = \psi_1^{m,j}\left(R_{m,j} - R_{m,j+1}\right) + \psi_2^{m,j}\left(R_{m,j} - R_{m,j+2}\right) \\ + \cdots + \psi_{J-j}^{m,j}\left(R_{m,j} - R_{m,J}\right), \quad \sum_{t=1}^{J-j} \psi_t^{m,j} = 1, \tag{22}$$

where $R_{m,j} = R\cdot r(m,j)$, and $\psi_t^{m,j}$ relates to the linear coefficients $a_k^{m,j}$ and $b_k^{m,j}$ (see Section C.1 of Online Appendix C). To incorporate the individual risk trajectory, we can substitute the terms $R_{m,j} - R_{m,j+t}$ with $R_{i,j} - R_{i,j+t}$ for each patient i. The strong dominance is a special case of the weak dominance.

To interpret (22), the parameter $\psi_t^{m,j}$ can be seen as a reflection of the probability (proportion) that a type (m,j) patient is discharged t days from today (under the linear-quadratic approximation). Thus, this score is a weighted average of the marginal improvements in readmission risk over the patient's remaining LOS trajectory, where the weight is driven by how likely the patient is to be discharged on a certain day. Tuning the weight parameters in (22) for different cost structures is not as straightforward as what we did for (21). Motivated by the preceding interpretation, in the counterfactual study and implementation, we leverage the static discharge thresholds l_m^* from Section 3.3 and set $\psi_t^{m,j} = 1$ for $t = l_m^* - j$ and zero otherwise. Our numerical results suggest that this modified weak dominance performs well in a variety of settings (see Section C.2 of Online Appendix C for some examples).

5.2. Performance of the Dynamic Algorithm

In this section, we demonstrate the performance of the algorithm by (1) comparing the actions solved from the dynamic algorithm with those from value iteration in small-scale MDPs and show that the dynamic algorithm is able to achieve near-optimal performance and (2) comparing the dynamic algorithm with several heuristics, including the static threshold policy and a one-step improvement on the static threshold policy, in a simulation setting for a suite of different parameters. In the interest of space, we demonstrate one set of results for the small-scale MDP here and leave other experiments to Online Appendix C: (1) additional numerical results in the small-scale MDP, including when the strong dominance no longer holds, and (2) comparison with additional heuristics.

In the small-scale MDP, we consider two classes of patients with maximum LOS $J = 3$ days. The risk trajectory for each class follows $r(1, j) = \{1, 0.3, 0.2, 0.15\}$ and $r(2, j) = \{1, 0.3, 0.08, 0.06\}$, satisfying the strong-dominance criterion. We set $C = 1$, $R = 6$, and $N = 12, 20$. Figure 5 shows the gap between the optimal actions obtained from value iteration versus the actions obtained from our dynamic heuristic for each class. For comparison, we also plot the gap for a myopic policy, where we ignore the cost-to-go and set $\tilde{\alpha} = \tilde{\beta} = 0$ in (21). From all the states, the maximum gap between our dynamic heuristic and the optimal action is one (Figure 5(a)) or two in only a few states (Figure 5(b)), whereas the myopic policy performs poorly. This demonstrates the importance of a more nuanced approach to discharge management.

6. Patient Risk Trajectory Prediction

For the development and implementation of the complete decision support system, a critical component is the patient risk trajectory prediction, which provides input to the discharge optimization framework. As mentioned in the Introduction, we need to address the endogeneity issue and provide a prediction in both the probability and timing of readmission. When applying the classical Cox proportional hazard model to predict readmission timing, we find two additional challenges. First, the prediction curves were being heavily distorted by the fact that many patients are never readmitted in the data, that is, the excess zero count issue (Bardhan et al. 2014). Second, the Cox model produces a time-to-readmission curve that has the same baseline hazard function for all patients: the postdischarge readmission risk peaks at the same time for all patients. However, our data indicate that some patients exhibit early readmitter behavior, whereas others exhibit late readmitter behavior (see Figure 6 from our partner hospital).

To the best of our knowledge, existing off-the-shelf models alone cannot provide sufficient input. In this section, we develop a new prediction approach to generate the risk trajectory inputs for our optimization framework. We described the development of this tool in Section 6.1 and then validate and report the performance of the tool in Section 6.2.

6.1. Prediction Model

Our approach, which addresses the aforementioned challenges, is summarized in Figure 7. In stage 1, we use a cure model to address excess zero count. This stage predicts the overall probability that a patient is eventually going to be cured or readmitted (Yu 2008, Bardhan et al. 2014). For patients who are not cured (will be readmitted), we move to stage 2, where we use a mixture model to capture patient heterogeneity with clustering and estimate different time-to-readmission curves for each cluster. We use an IV method, replacing LOS with a predicted LOS in stages 1 and 2, as a preprocessing stage to correct for endogeneity.

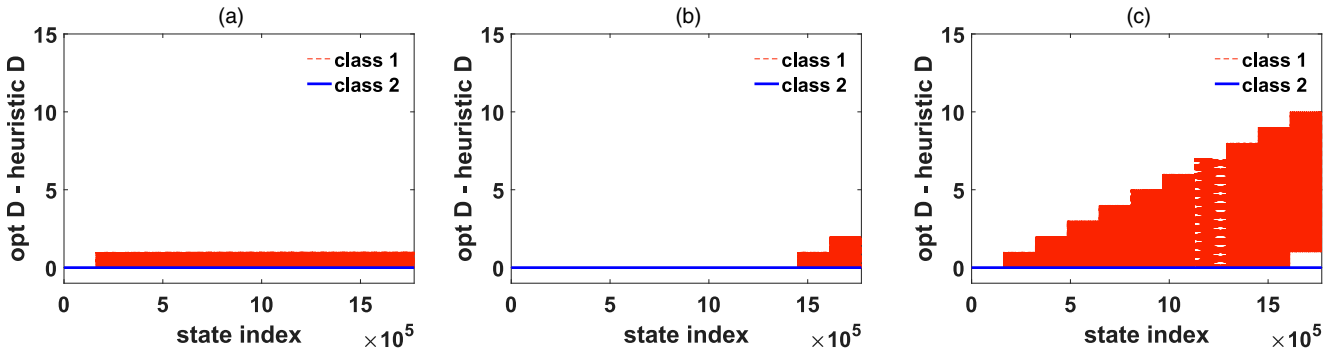
Mathematically, for stages 1 and 2, let C_i denote patient i's cure status, where

$$C_i = \begin{cases} 1, & \text{cured,} \\ 0, & \text{uncured.} \end{cases}$$

Conditioning on $C_i = 0$, a patient belongs to one of W clusters with probability π_w for $w = 1, \ldots, W$ (e.g., early-readmission versus late-readmission cluster), and $\{\pi_w\}$ values are population membership probabilities with $\sum_{w=1}^{W} \pi_w = 1$. Each cluster w has an associated time-to-readmission proportional hazard rate function $h^w(t; i)$. For each patient i, we use Z_i to denote which cluster patient i belongs to, with Z_i being drawn from a multinomial distribution with mixing weights $\{\pi_w\}$.

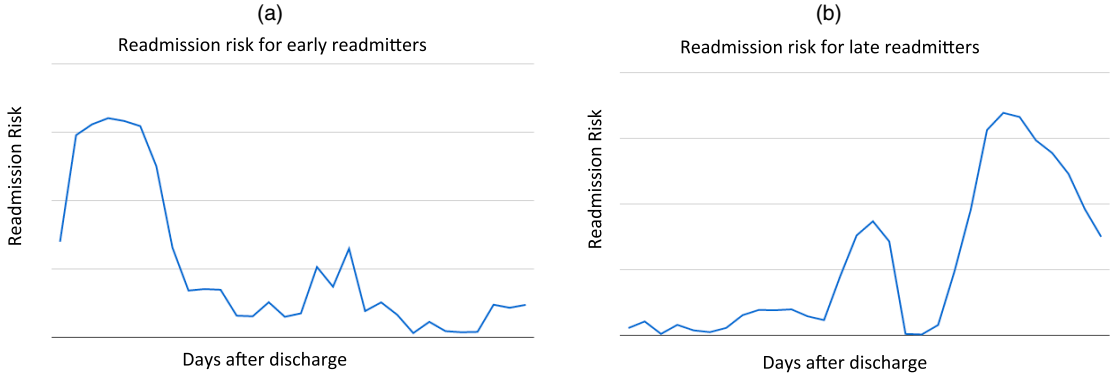
Both the cured status C_i and the membership variable Z_i are latent variables, that is, unobservable. In particular, the cured status C_i is partially observable (Yu 2008). That is, if a patient is readmitted, we know that he or she must have been uncured; however, if this patient is not readmitted, we do not know whether he or she is cured or uncured, but the readmission time is censored. To perform parameter estimation with

Figure 5. Small-Scale MDP Validation



Notes. We set $N = 12$ or 20, $\Lambda_1 = \Lambda_2 = 2$, $C = 1$, and $R = 6$. In plot (a), $N = 12$. In plot (b), $N = 20$. In plot (c), $N = 12$ with myopic policy.

Figure 6. Plot of Readmission Risk for Sampled Early and Late Readmitters from Data



these latent variables, we develop an expectation–maximization (EM) framework. The basic idea is to first find surrogates for the cured probability and membership probability. Then we iteratively update the estimation of these surrogates in the E-step of the EM algorithm and update the estimation of the parameters in the M-step by maximizing the (approximate) log-likelihood function. We leave the details of the EM algorithm to Section A.2 of Online Appendix A.

Next, we specify (1) the parametric models for stages 1 and 2, (2) readmission probability calculation, and (3) the IV method for the preprosessing stage.

6.1.1. Parametric Models. In the first stage, we assume that the probability of a patient being cured follows a logit model. For patient i with features Y_i, his or her cure probability $\theta_{0i} = \mathbb{P}(C_i = 1)$ follows

$$\log\left(\frac{\theta_{0i}}{1-\theta_{0i}}\right) = Y_i\xi, \tag{23}$$

where ξ is the set of coefficients associated with the individual patient's characteristics and risk factors, denoted as Y_i:

$$Y_i = \left(\log(LOS_i), Y_i^e\right).$$

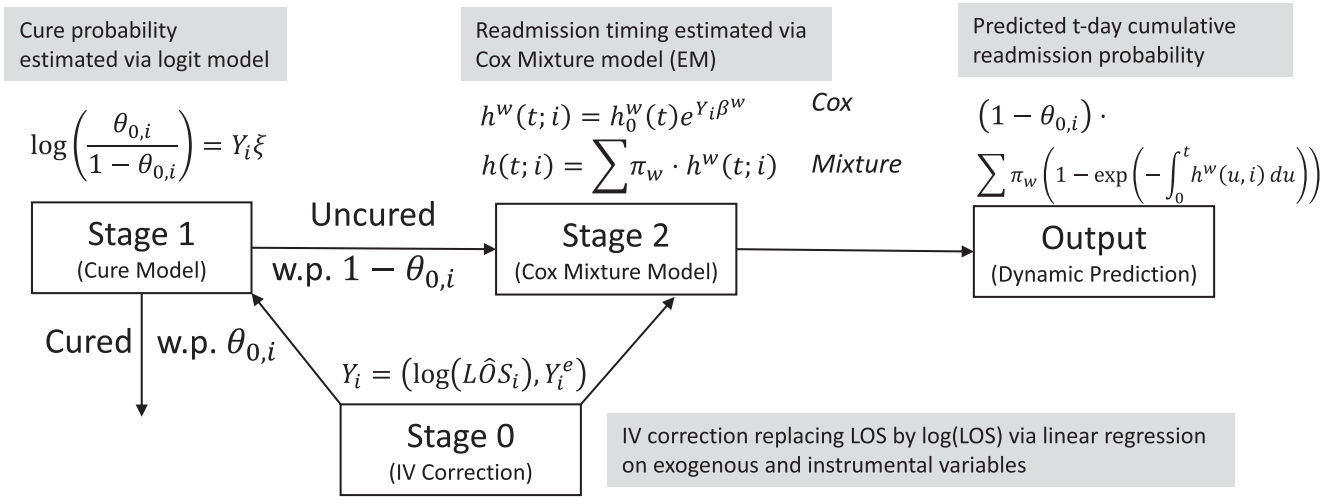
tHere $Y_i^e = (Y_{i,1}^e, \ldots, Y_{i,K-1}^e)$ denotes $K-1$ exogenous variables such as patient age, gender, and medical specialties. During the model training phase (for parameter estimation), we replace the actual $\log(LOS_i)$ observed from data with $\log(\hat{LOS}_i)$ predicted from a linear regression in stage 0, the preprocessing stage.

For the second stage, we use the Cox model. For patient i in cluster w, the hazard rate is

$$h^w(t;i) = h_0^w(t)e^{Y_i\beta^w}, \quad w = 1, \ldots, W. \tag{24}$$

The baseline hazard rate function $h_0^w(t)$ follows the Weibull form (see Section A.1 of Online Appendix A) with cluster-dependent parameters λ^w and k^w. The coefficients $\beta^w = \{\beta_1^w, \ldots, \beta_K^w\}$ are also cluster dependent, and Y_i is the same as the first stage, but β^w can be different from ξ.

Figure 7. Conceptual Diagram of Prediction Model Development



6.1.2. Readmission Probability. For a patient i' (could be either in or not in the training data), we calculate his or her readmission timing by first obtaining the cure probability $\hat{\theta}_{0i'}$ from the logit model (23). Then the t^*-day cumulative readmission probability for this patient i' equals $(1-\hat{\theta}_{0i'})H(t^*;i')$, where

$$H(t;i') = \sum_{w=1}^{W} \pi_w \left(1 - \exp\left(-\int_0^t h^w(u;i')du\right)\right). \quad (25)$$

The probability of being readmitted on a particular day t equals

$$(1-\hat{\theta}_{0i'})(H(t;i') - H(t-1;i')).$$

6.1.3. Correcting the Endogeneity of LOS with IVs. Patient severity may be positively correlated with both LOS and the dependent variables Y_i because sicker patients tend to stay longer and are also more likely to be readmitted. Thus, treating LOS as an exogenous variable can lead to the incorrect conclusion that longer LOS results in a higher readmission risk. To address this issue, we follow the IV technique developed by Bartel et al. (2020), replacing $\log(LOS_i)$ with $\log(\hat{LOS}_i)$ predicted by a linear regression on exogenous features Y_i^e (which are included in the first- and second-stage models) and additional IVs, IV_i, which only appear in stage 0, the preprocessing stage. The linear regression on $\log(LOS_i)$ follows:

$$\log(LOS_i) = \phi Y_i^e + \zeta IV_i\,. \quad (26)$$

After obtaining the estimates for ϕ, we replace the variable $\log(LOS_i)$ with the predicted value $\log(\hat{LOS}_i)$ in Y_i and estimate the parameters in the Cox cure model.

In their econometric study, Bartel et al. (2020) propose using admission day-of-week indicators as the IV in the regression (26). The rationale is that physicians prefer not to keep patients over the weekend. Thus, a patient who would have stayed for four days in the hospital may get discharged early if he or she is admitted on Tuesday because of such operational considerations. We adopt the same IV proposed by Bartel et al. (2020) in our implementation. Our prediction results show that this IV is able to help us correct the estimation bias caused by endogeneity and achieves satisfactory prediction performance on various metrics such as area under the curve (AUC; see more details in Section 6.2). We conclude with the following remark.

Remark 6. We do not intend to claim causal relationship or obtain causal inference for the effect of LOS on readmission such as Bartel et al. (2020). The LOS can be seen as a variable reflecting the aggregate effect of other drivers for patient recovery in the hospital; we choose LOS because it is the most intuitive variable for hospital to control, and it directly affects the workload. Our primary goal is to obtain a reasonably accurate prediction on how individual patient readmission risk evolves as a function of LOS, providing a necessary input to our decision framework.

There is no theoretical guarantee when using IV in conjunction with the nonlinear Cox survival model except when the hazard rate function is linear in Y_i (MacKenzie et al. 2014, Zheng et al. 2017), although there are reported empirical successes in correcting bias with IVs (Atiyat 2011, Tian 2016). We tried using the control function method for applying IV in nonlinear models (Petrin and Train 2010, Arıkan et al. 2017); the AUC is not as good as directly applying the IV. We would also like to point out that Bartel et al. (2020) focus on nondeferrable conditions such as heart attack patients. Hence, patients admitted on weekdays and weekends have similar conditions, and their IV is likely to be valid. However, because our model framework is not limited to a certain type of patient, the exclusion criterion for the IV may no longer hold. Although this IV helps to correct the estimation bias in our prediction analysis and is sufficient to generate an appropriate input for our discharge optimization, readers should be cautious, particularly when applying this IV to settings where causality is important.

We also emphasize that it is not our goal in this paper to fully explore all possible methods to achieve the best prediction capability or fully address the endogeneity issue. Rather, we aim to provide one method that works well and use it to showcase in a real hospital setting that our proof-of-concept is implementable, that is, using discharge optimization to balance congestion versus readmission risk. To this end, we show in Section 6.2 that the prediction tool demonstrates reasonably accurate performance. More important, in our trace-based counterfactual analysis with hospital data (Section 7), we demonstrate that combining the prediction tool and the discharge optimization, significant gains can be achieved in LOS, net readmissions, and positive catch rate. Thus, we believe that our prediction model serves its purpose in achieving the primary goal of this paper. We leave to future research further improvement of the prediction model and using more advanced methods to address the endogeneity issue in survival analysis.

6.2. Model Validation and Implementation

6.2.1. Data Description. The data set used to estimate and validate our prediction model is from our partner hospital in the state of Indiana, spanning January 2010 to September 2017. We exclude planned readmissions, expired patients, patients under the age of 18 (including newborns), and obstetric and gynecology patients.

The final data set included $n = 25{,}601$ patients. Of the 200 available features, we exclude the binary predictors that are recorded for less than 5% of the population and perform feature selections using cross-validation. The final considered features include patient demographics (age, gender, race, weight, and height), psychosocial data (e.g., ZIP code and marital status), diagnostic information (e.g., International Classification of Diseases code and major diagnostic categories), postdischarge dispositions, and indicators of secondary illnesses (e.g., depression, diabetes, smoking habits). Also, dates of admission and discharge for each patient visit are used to construct possible IVs.

6.2.2. Model Selection and Validation. To train and fine-tune our prediction model, we use a bootstrapping algorithm. Our main performance metric for prediction capability is the AUC. For bootstrapping, we use a subsampling method to randomly sample patients from the original data set 50 times, generating 50 different training data sets. Each time, we sample $m \approx 0.623n$ patients from the original data set without replacement (De Bin et al. 2016). From each training data set i, we estimate parameters $\Psi_i = \{\xi, \{\pi_w\}, \{\lambda^w\}, \{k^w\}, \{\beta^w\}\}$.

We use the average estimates $\overline{\Psi} = (1/50)\sum_{i=1}^{50} \Psi_i$ as the final estimated parameters for a given model configuration. We compare different configurations and select the one that performs the best with respect to $AUC(\overline{\Psi})$, which is obtained by verifying the prediction performance for all patients in the testing data (i.e., the original data set as for the bootstrapping method); see more details in Section A.3 of Online Appendix A. In the model selection process, we also consider $\overline{AUC} = (1/50)\sum_{i=1}^{50} AUC_i$ and several other performance metrics reported later. Compared with cross-validation typically used in the literature, bootstrapping allows for both a larger training and testing sample size, which is important in our setting because the number of readmitted patients is small compared with the total patient population (Austin and Steyerberg 2017).

6.2.3. Prediction Performance. The estimated coefficients $\overline{\Psi}$ of selected main features for our final model for implementation are reported in Section A.4 of Online Appendix A. Table 1 compares a few performance metrics on the prediction accuracy of our final model. For the 90-day readmission risk, $AUC(\overline{\Psi})$ is 69.4% on the testing data set, and $\overline{AUC}$ is 68.2% (±1.1% for the 95% confidence interval) from the 50 training data sets. Other performance metrics are calculated in a similar way. AUC, accuracy, and *F*-score are metrics for binary classifications, which we report for two different cutoff times (30 and 90 days). The *C*-statistic is a concordance metric for the continuous event time (Uno et al. 2011). It reflects the proportion of pairs of subjects whose observed time-to-failure event (possibly censored) and predicted risk scores agree among all possible pairs. Thus, only one number is reported. The AUC performance is comparable with those reported in the literature (Bayati et al. 2014, Min et al. 2019).

In addition, we show supporting evidence that this prediction tool has improved our partner hospital's prediction metrics since the pilot implementation (see details in Online Appendix F). More important, given that the goal of this paper is to reduce readmissions through our decision framework, instead of simply predicting the readmissions, we demonstrate in Section 7 that when we combine the prediction tool and the discharge optimization framework, significant gains can be achieved in LOS, net readmissions, and positive catch rate.

7. Improving over Practice: Counterfactual and Simulation Analysis

In this section, we develop a case study based on data from our partner hospital. We begin with a counterfactual analysis on the hospital's historical practice to demonstrate how our dynamic algorithm could have improved the hospital's performance (see Sections 7.1 and 7.2). Our results demonstrate (1) Pareto dominance of the dynamic policy over historical practice, reducing both the readmission rate and the proportion of early readmitters, (2) a high positive catch rate, properly identifying and intervening on patients who were readmitted in the data, and (3) occupancy smoothing, an unintended additional benefit. This has served as an important step in our implementation process to demonstrate to the hospital

Table 1. Predicting Capacity Performance of the Final Prediction Tool

	AUC		Accuracy		*F*-Score		*C*-Statistic	
Day	Testing	Training	Testing	Training	Testing	Training	Testing	Training
30	69.2%	66.2% ± 1.0%	87.8%	87.7% ± 0.1%	93.4%	93.4% ± 0.1%	67.0%	66.9% ± 0.6%
90	69.4%	68.2% ± 1.1%	78.3%	77.5% ± 0.5%	87.4%	86.8% ± 0.4%		

Note. In each panel, the first number is gained on the testing data set, whereas the second number is averaged from the 50 testing data sets (the number following the ± sign is the half-width of the 95% confidence interval).

the potential value of our tool and to explain the logic for when our recommendations differed from historical practice. In Section 7.3, we develop a data-driven discrete event simulation to generate insights for the broader application of our method in a wide range of hospitals through sensitivity analyses.

7.1. Trace-Driven Counterfactual Analysis Based on Historical Data

To gain buy-in from management for a pilot implementation, we designed a counterfactual to compare the dynamic policy (based on the dynamic algorithm) with historical practice in our partner hospital. To create a realistic comparison with historical discharge behavior, we use a trace-driven approach (Sherman and Browne 1973) in which the system inputs are generated from observations in the data instead of using parametric assumptions. We then use an additive/subtractive counterfactual of readmission events to evaluate the impact of changing historical discharge decisions along several dimensions.

7.1.1. Avoiding/Adding Readmission Events. Our dynamic discharge policy may either extend or shorten the actual LOS observed in the data. If a patient had a readmission event in the data and our model recommends extending his or her LOS, we avoid the readmission with a probability that is proportional to the risk reduction from extending the LOS. Specifically, we calculate the ratio between the predicted risk at our recommended discharge date versus the predicted risk at the actual discharge date. If a uniformly generated random variable exceeds this ratio, we avoid the readmission. To explain this approach, imagine that there is a random draw between 0% and 100% for each patient that produces the readmission event. If the readmission risk is 20% for this patient using the historical LOS and we observe a readmission event in the data, then we know the realized outcome of the random draw lies anywhere between 0% and 20%. Suppose that by extending the LOS, we reduce the readmission risk to 15%. Then we can avoid this readmission event if the realized random draw lies between 15% and 20% with a probability 5/20 = 25%. Similarly, if a patient did not have a readmission event and our model recommends shortening his or her LOS, we generate a new readmission event with a probability proportional to the risk increase. We use Monte Carlo simulation with 50 replications to generate the sequence of the random draws for each patient.

7.1.2. Personalized Risk Curves. In the implementation of the dynamic heuristic on real data, we incorporate personalized risk trajectories directly from our prediction model, removing the reliance on patient classes, as in the modeling framework in Section 3. We follow the modified weak-dominance rule, developed in Section 5.1.3, to rank all patients who are currently in the hospital when a discharge decision needs to be made.

7.1.3. Minimum LOS. To estimate the minimum LOS $\underline{L}_m$, we cluster patients into $M = 3$ classes with the k-means method based on their predicted curves (Figure 4(b)), roughly corresponding to the low-, medium-, and high-risk groups. We further divide each class into 10 subclasses and estimate the 10th percentile as a proxy for the minimum LOS. The estimated minimum LOS is one day for all subclasses in the low- and medium-risk groups and is two days for all subclasses in the high-risk group.

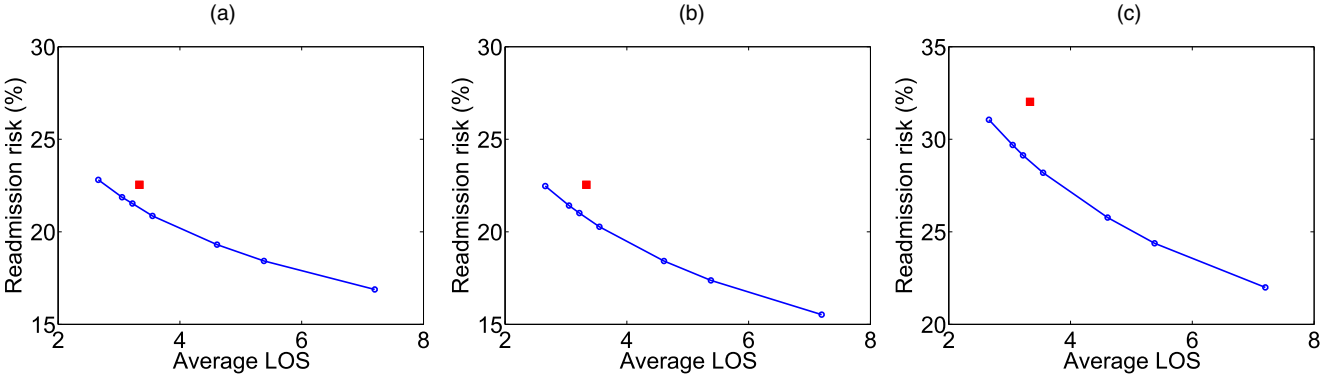
7.1.4. Tuning Parameters. Estimating the cost parameter C is challenging. Also, in our data, there is no reliable estimate of the capacity N. Instead, we use these two as tuning parameters and plot the efficient frontiers with respect to different performance metrics by varying N and the ratio between C and R. This allows hospital managers to choose a parameter regime to achieve their desired performance metrics. During the actual implementation, we use a default setting based on the management team's feedback on their preferred target point of the efficient frontier. In the tool delivered to the hospital, we also maintain an interactive tab where the efficient frontier is shown, and hospital managers can change the parameter settings to adjust their occupancy and readmission targets (see more details in Section 8, where we describe the pilot implementation).

7.2. Value of the Tool

7.2.1. Summary of Results. Figure 10 plots the readmission risk against the average LOS (which also implies the average occupancy by Little's law). In this figure, the solid line corresponds to the performance of dynamic policies from a series of experiments where we vary R from 0.01 to 120, with $N = 40$ and $C = 1$; the dotted line corresponds to the performance of the hospital's historical practice. We report three types of readmission risk in each of the subplots:

- *Realized readmissions.* This is the historical number of readmissions (adjusted by the number of avoided or added readmissions for the dynamic policy) divided by the total number of patients.
- *Predicted readmissions.* To account for other sample paths than the realized one (in the data), we compare the predicted readmission risk using the historical versus the dynamic policies.
- *High and medium risk.* We compare the predicted readmission risk for the historical versus dynamic policies for only high- and medium-risk patients.

Figure 8. Efficiency Frontier



Notes. We set $N = 40$ and $C = 1$ and change R from 0.01 to 120 to get the solid line in each plot (using the dynamic heuristic). The dot corresponds to the performance of the historical discharge policy. The 95% confidence intervals are tight, and we omit them in the plots. Plot (a) shows the realized readmission; plot (b) shows the predicted readmission; plot (c) shows the predicted readmission for high- and medium-risk patients.

From the figure, the dynamic policy can either reduce the readmission risk while maintaining a similar average LOS or maintain the same readmission risk with shorter LOS (lower occupancy); that is, our dynamic heuristic exhibits Pareto dominance over the historical one. For example, the dynamic policy can significantly reduce readmission risk for medium- and high-risk groups (from 32% to 28%) when extending the LOS slightly (from 3.33 to 3.55 days). Table 2 reports the average LOS and the average added and avoided number of readmissions, along with the 95% confidence interval.

7.2.2. Positive Catch. A *positive catch* is defined as extending the LOS for at least one day for a patient who was actually readmitted in the historical data. This performance metric measures the ability of our tool to properly identify and intervene on at-risk patients. It is also appealing to our industry partner (Lean Care Solutions) because they find that such metrics could be easily explained to hospital management. Using a policy that maintains a similar average LOS as the historical one, we have a positive catch rate above 50%. Positive catch rates for other ratios of R and C are shown in Table 2.

7.2.3. Impact on Readmission Timing. Under the average LOS of 3.55 days ($R/C = 3$), of the total number of avoided readmissions, approximately 25% of them were readmissions within 14 days, and 40% of them were within 30 days. Given that 14 and 30 days correspond to 15% and 33% of the total readmission window (90 days), we can see that our recommended discharge policies are able to reduce more early readmitters: this is beneficial to hospitals because the early readmitters usually require more intensive care during their readmission visits. By further extending the LOS, the percentage of avoided readmissions within 14 and 30 days can be increased to up to 30% and 45%, respectively. These observations are consistent with our conjecture that increasing LOS has a more significant effect on reducing early readmits (Figure 4(c)).

We also perform a similar additive/subtractive counterfactual to examine whether extending LOS would change the timing of readmissions for patients who received the extension intervention but did not avoid the readmission (from the random draw). To do so, we use the conditional probability of readmission timing before day t, conditioning on the event that the patient is readmitted. Under the average LOS of 3.55 days ($R/C = 3$), 3% patients were shifted from before $t = 14$ days to after 14 days, in addition to the avoided readmissions.

7.2.4. Occupancy Smoothing. Harrison et al. (2005) hypothesized that discharge policies could be effective in smoothing hospital occupancies, which has numerous benefits beyond the objectives of our study, such as reducing cancellations of elective surgery,

Table 2. Summary of Statistics from Dynamic Policies under Different Cost Parameters

R/C	0.01	0.1	1	3	40	120
Average LOS	2.67	3.05	3.22	3.55	4.61	7.20
No. readmission avoided	398 ± 4.7	514 ± 5.6	554 ± 5.5	658 ± 6.0	935 ± 6.8	1,481 ± 8.7
No. readmission added	513 ± 5.5	379 ± 5.5	333 ± 4.6	258 ± 4.3	127 ± 2.6	24 ± 1.2
Positive catch	37%	45%	49%	54%	68%	84%

Note. Each number following the ± sign in the second and third rows denotes the half-width of the 95% confidence interval of the corresponding entry.

boarding time in the emergency department, off-servicing of patients, and stress on hospital staff, among others (Kc and Terwiesch 2017, Dai and Shi 2019). Although occupancy smoothing is not explicitly incorporated in our objective function of the MDP, Figure 9 shows that the dynamic policy does produce a much smoother occupancy curve than the historical one, an unintended benefit. The *peakedness* of the occupancy, defined as the sum of squared differences between the daily and overall mean occupancy (normalized by the overall mean), decreases from 1.44 under the historical practice to less than 0.25 under the dynamic policy for settings in Figure 9.

7.3. Broader Insights for Different Hospital Environments and Operational Characteristics

In this section, we develop a high-fidelity, data-driven discrete event simulation to study the operational characteristics of hospitals that influence the efficacy of the dynamic discharge policy, where we vary hospital unit size and utilization, shape of the risk curves, and variability in the arrival process. We summarize generalizable insights into which types of hospitals benefit most from the dynamic policy through comparison with the static threshold policy and an empirical policy that mimics historical behavior. The rationale behind these insights is explained through two operating regimes: occupancy-driven versus quality-driven regimes, which we detail in Online Appendix E.

7.3.1. Simulation Design and Policy Description. The simulation platform is similar to that in Section 7.1, except that we assume that the patient arrivals follow a time-nonhomogeneous Poisson process, with the arrival rates depending on both time of day and day of week. The exogenous daily arrival rate is estimated to be $\Lambda = 6.5$ patients per day. For each arriving patient, we randomly sample his or her risk trajectory from the predicted risk curves of all patients in the data set.

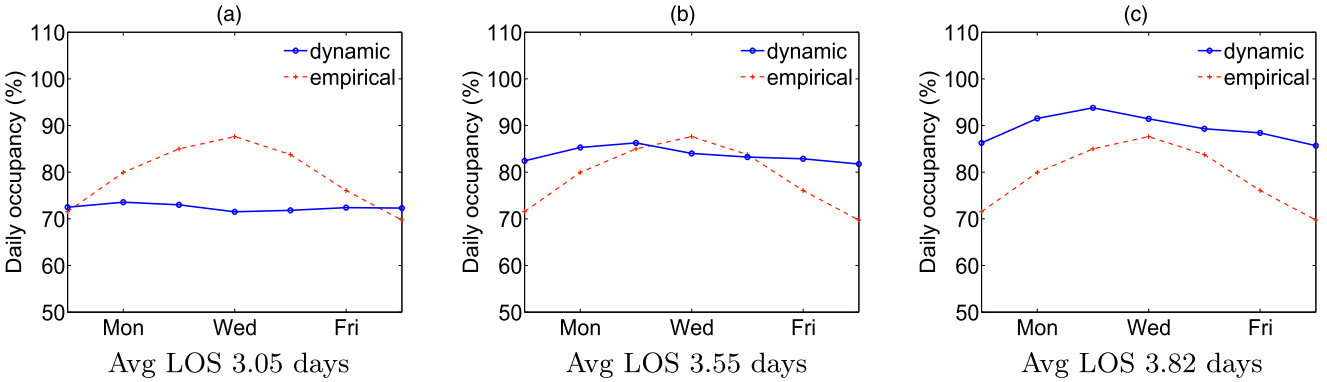
We compare the performance of the dynamic policy against two benchmark policies: (1) an empirical policy based on historical discharge behavior and (2) the optimized static threshold policy from Section 3.3. For these two policies, we group patients into $M = 3$ classes as in Figure 4(b), roughly corresponding to the low-, medium-, and high-risk groups. The empirical policy is similar to the static policy, where we estimate from data a fixed set of discharge risk thresholds: 10%, 20%, and 35% for the low-, medium-, and high-risk patients, respectively. The corresponding thresholds of LOS for each class are 3, 4, and 5 days; the average LOS of 3.69 days is close to the historical average (3.35 days).

For the baseline, we select the cost parameters $C = 1$, $R = 3$, and capacity $N = 40$, under which setting the empirical policy has the smallest performance gap with the static and dynamic policies. In this way, we give the empirical policy the benefit of the doubt by assuming that the hospital is aiming to optimize under the given system conditions.

7.3.2. Main Insights for Hospital Operating Characteristics. Figure 10 plots the performance improvement from the dynamic and static policies over the empirical policy (i.e., performance gap) under a variety of settings.

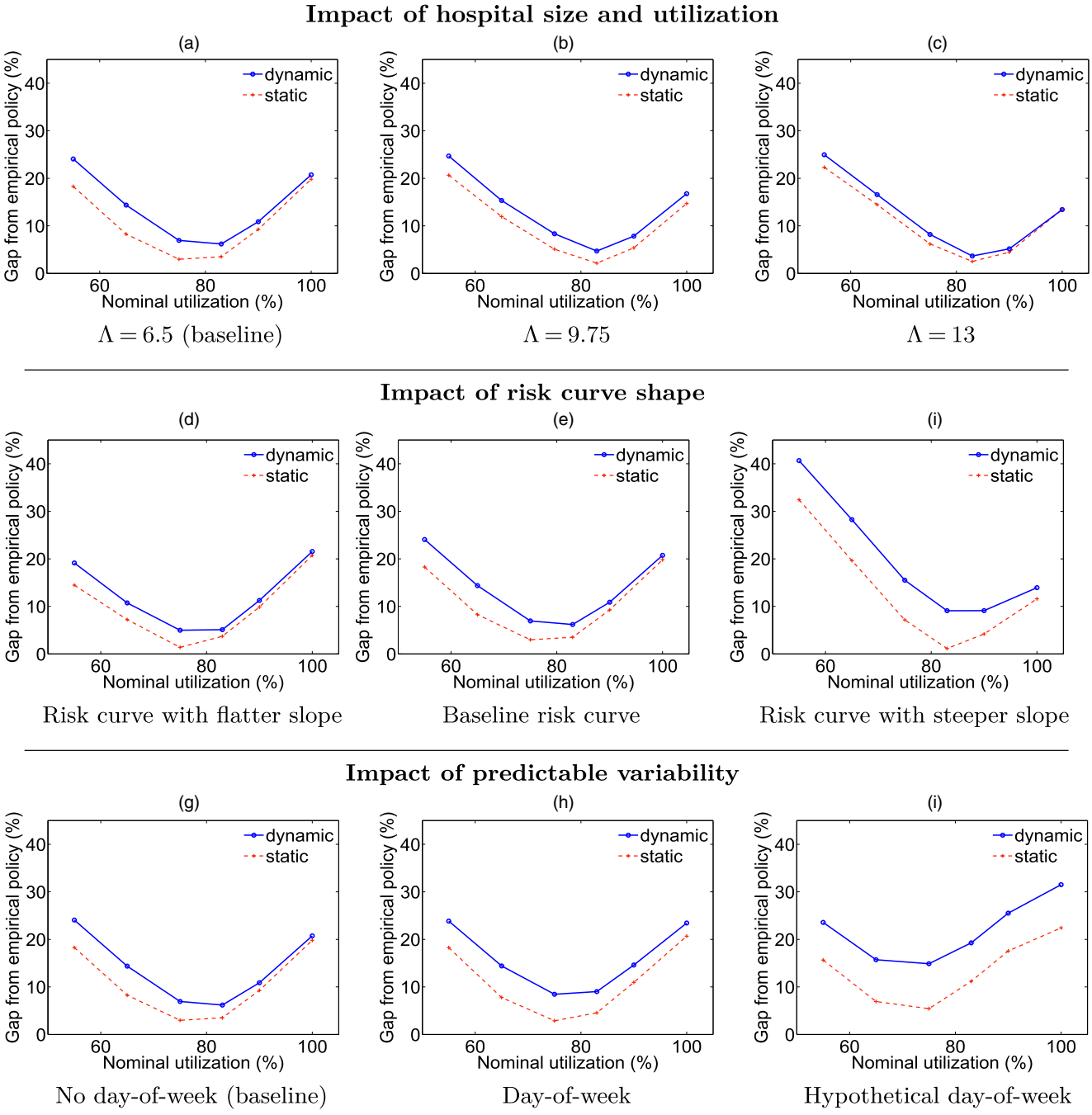
7.3.2.1. Size and Utilization. Figure 10, (a)–(c), shows the performance under different utilizations when we increase the system size (reflected by the higher arrival rate while maintaining the same utilization). The performance of both dynamic and static policies exhibits a U-shaped pattern as a function of utilization, with the largest gains occurring at low and high occupancies. The more interesting finding is that the

Figure 9. Daily Occupancy Level on Different Days of a Week



Notes. We set $C = 1$ and $N = 40$ and change R to get different average LOSs under the dynamic policy (reported in the title of each subplot). The historical average LOS is 3.35 days.

Figure 10. Performance Gap under Different Sensitivity Analysis Settings



Notes. In the baseline, we set $\Lambda = 6.52$, $N = 40$, $C = 1$, and $R = 3$. To measure the utilization, we calculate the offered load using the empirical policy because it does not react directly to occupancy levels. We then adjust the arrival rates to achieve different levels of utilization.

gap between the static and dynamic policies converges as system utilization increases, and the gap converges faster as hospital size increases. We explain this phenomenon through the two operating regimes identified in Online Appendix E. Thus, in smaller hospitals or single wards and in hospitals with lower utilizations (common in community hospitals), the dynamic policy is needed to obtain the greatest benefit. In larger and/or more highly used hospitals, such as urban and teaching hospitals, the simpler static policy may be sufficient. This insight may explain why the dynamic policy works well in our partner hospital, which is a small community hospital.

7.3.2.2. Risk Curves. Figure 10, (d)–(f), shows the performance improvement for different shapes of the

risk curve, moving from flatter to steeper slopes. As the slope increases, the gap between the dynamic and static policies increases significantly. The performance of the dynamic policy always improves as the slope increases, whereas the static policy does not exhibit this monotone behavior. Linking this to practice, the dynamic policy is more useful when patients recover faster. An example of this environment could be specialized elective surgery hospitals or wards because these patients often have a more rapid recovery process than patients with complicated medical conditions.

7.3.2.3. Day-of-Week Variability. Figure 10, (g)–(i), shows the performance improvement when the arrival process exhibits an increasing day-of-week arrival variability. As the variability in the arrival process increases, the gap between dynamic and static increases significantly. The dynamic policy always performs better as variability increases, whereas the static policy does not necessarily improve. Hence, the dynamic policy is increasingly valuable along multiple dimensions as the variability in the arrival process increases, providing both better performance and greater occupancy smoothing. This again indicates that the dynamic policy may be more useful in elective surgical hospitals or wards because surgical scheduling generates much of the variability in the arrival process.

8. Conclusion

> Through constant interaction with multiple doctors, case management, and hospital executives, especially chief medical officers, from community hospitals to academic hospitals, we have identified the need for discharge optimization as a valuable addition to the readmissions reduction offering of our tool. —Chief executive officer of Lean Care Solutions

In this paper, we develop a practical tool that integrates personalized readmission risk prediction into inpatient discharge planning. We test and implement this tool through collaborations with a data analytics company and a local partner hospital. Based on extensive counterfactual and simulation analyses, we demonstrate the value of this tool compared with the hospital's historical discharge behavior and identify hospital characteristics that would benefit the most from our discharge optimization. We show that by increasing the average LOS moderately, the readmission risk can be reduced significantly; for example, when the average LOS increases from 3.33 days under the historical practice to 3.55 days under the dynamic policy, the corresponding readmission risk for medium- and high-risk groups decreases from 32% to 28%. The purpose of our tool is not to promote early discharge (unless necessary when system is highly congested); rather, it provides analytical support for a hospital to balance the benefits of shortening or extending LOS. We conclude with a brief discussion of our ongoing implementation efforts, including introduction of the user interface, integration into current workflow, and future research opportunities.

8.1. Implementation Efforts

8.1.1. User Interface. Figure 2(a) shows the discharge tool's user interface. On the left side of the predischarge module, every inpatient is represented by a rectangle, which shows admission date and current LOS. These rectangles are ranked based on the ranking criterion we developed earlier in this main paper. In addition to having the patients ranked in terms of readiness for discharge, the boxes are displayed in three colors: green, yellow, and red. Green indicates that the patient can be discharged under a conservative setting that puts more focus on the readmission cost. Yellow indicates that the patient can be discharged under the baseline cost setting; see more discussion on cost choices below. Red indicates that the patient should not be discharged except in extreme cases (e.g., mass casualty events).

Clicking on one of the patient rectangles reveals additional information, such as principal diagnostic and relevant psychosocial data in the center panel. By clicking on "Submit" in the lower right-hand corner, a window pops up with the discharge risk versus LOS curve shown in Figure 2(b). The vertical bars show the current LOS and the recommended discharge date from our optimization tool. Enlarged figures are available in Online Appendix F. At the hospital's request, the tool also provides a postdischarge tab on prediction of the time to readmission when discharge is initiated (see a snapshot in Online Appendix F).

8.1.2. Integration into Workflow. The discharge optimization tool is intended to be used daily during the time when discharges are being evaluated and processed, typically before morning rounds. When considering the discharge, the staff will check the tool for a patient's discharge suitability/ranking and the other analytics features described previously that support the discharge decision. We emphasize that this tool is intended for decision support and can adapt to deviation from recommendations based on clinicians' assessment and medical judgment. These deviations are logged in the tool, where clinicians can also provide notes about their discharge decision. The dynamic data gathered provide live feedback to our algorithms, which can help to make further adjustments for the decision support. In the implementation, the discharge recommendations are provided based on a default setting for the parameters C and R, which were chosen by soliciting the management

team's input on their target LOS/readmission rate combination. Two sets of R/C ratios are selected to produce the color codes mentioned previously: $R/C = 3$ as the baseline version where the resulting average LOS (from our simulation analysis) roughly matches the historical average of 3.33 days and $R/C = 40$ as a more conservative setting to match the hospital's target to reduce its readmission rate below 20%. In addition, we provide a pop-up box in the tool where a gatekeeper can change the default setting to allow for dynamic adjustment of management goals in the event, for example, that management has additional information or strategy that would require a temporary or more permanent change in policy. In this pop-up box, we demonstrate a tradeoff curve, where we vary the tuning parameters and demonstrate the tradeoff between average LOS (occupancy) and average readmission rate.

More details of the implementation are documented in Online Appendix F. Thus far, the pilot has been well received and is considered a success by our industry partners. Moving forward, Lean Care Solutions' chief executive officer indicates that "[Our tool] has received very positive interest from another leading academic hospital on the east coast, praising the applicability of the real world decisions that need to be made on the ground and the simplicity of information that the tool communicates to the users. We are looking to roll out this module after more testing as part of the Readmissions Reduction Tool offering to all our current and future clients."

8.2. Future Research

This paper can be extended in a few directions. We are working with the hospitals to explore the possibility of collecting more time-varying covariates that may better reflect patient condition changes in the hospital, which could help improve both the risk prediction and discharge decision optimization. Newly emerging machine learning tools such as deep neural networks could further improve the accuracy of the risk prediction. There are significant avenues of future work in developing these tools to improve the prediction performance and quantifying how much such improvement can increase the value of our discharge decision support. In addition, we focus on readmission risk as the main outcome metric, whereas future work could extend our framework to other patient outcomes. From the analytical side, we propose a weak-dominance ranking criterion when strong dominance is not met but not as a main focus of this paper. A more thorough study could take a deeper look into this setting along with establishing possible performance bounds, which likely would require new methodology and approximation methods. In addition, future works may consider jointly optimizing discharge decisions along with other decisions such as admission control or patient diversion (including off-service placements) to further reduce ward congestion.

References

Adepoju T, Tucker A, Jin H, Manasseh C (2019) Hospital boarding crises: The impact of urgent versus prevention responses on length of stay. Working paper, Boston University, Boston.

Anderson D, Price C, Golden B, Jank W, Wasil E (2011) Examining the discharge practices of surgeons at a large medical center. *Health Care Management Sci.* 14(4):338–347.

Arıkan M, Ata B, Friedewald JJ, Parker RP (2017) Enhancing kidney supply through geographic sharing in the united states. *Production Oper. Management* 27(12):2103–2121.

Armony M, Yom-Tov G (2018) Optimizing discharge decisions in a hematology ward. Working paper, New York University, New York.

Armony M, Chan CW, Zhu B (2018) Critical care capacity management: Understanding the role of a step down unit. *Production Oper. Management* 27(5):859–883.

Ata B, Shneorson S (2006) Dynamic control of an $M/M/1$ service system with adjustable arrival and service rates. *Management Sci.* 52(11):1778–1791.

Atiyat M (2011) Instrumental variable modeling in a survival analysis framework. PhD thesis, University of Pennsylvania, Philadelphia.

Atlaeddini A, Helm JE, Shi P, Faruqui S (2019) An integrated framework for reducing hospital readmissions using risk trajectories characterization and discharge timing optimization. *IISE Trans. Healthcare Systems Engrg.* 9(2):1–14.

Austin PC, SteyerbergEW (2017) Events per variable (EPV) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models. *Statist. Methods Medicine Res.* 26(2):796–808.

Bardhan I, Jeong-ha O, Zheng Z, Kirksey K (2014) Predictive analytics for readmission of patients with congestive heart failure. *Inform. Systems Res.* 26(1):19–39.

Bartel AP, Chan CW, Kim SHH (2020) Should hospitals keep their patients longer? The role of inpatient care in reducing post-discharge mortality. *Management Sci.* 66(6):2326–2346.

Bavafa H, Ormeci L, Savin S, Virudachalam V (2019) Surgical case-mix and discharge decisions: Does within-hospital coordination matter? Working paper, University of Wisconsin Madison, Madison, WI.

Bayati M, Braverman M, Gillam M, Mack KM, Ruiz G, Smith MSS, Horvitz E (2014) Data-driven decisions for reducing readmissions for heart failure: General methodology and case study. *PLoS One* 9(10):e109264.

Bekker R, Boxma OJ (2007) An $M/G/1$ queue with adaptable service speed. *Stochastic Models* 23(3):373–396.

Berk E, Moinzadeh K (1998) The impact of discharge decisions on healthcare quality. *Management Sci.* 44(3):400–415.

Berry Jaeker JA, Tucker AL (2017) Past the point of speeding up: The negative effects of workload saturation on efficiency and patient severity. *Management Sci.* 63(4):1042–1062.

Braverman A, Gurvich I, Huang J (2020) On the Taylor expansion of value functions. *Oper. Res.* 68(2):631–654.

Carey K (2015) Measuring the hospital length of stay/readmission cost trade-off under a bundled payment mechanism. *Health Econom.* 24(7):790–802.

Chan CW, Yom-Tov G, Escobar G (2014) When to use speedup: An examination of service systems with returns. *Oper. Res.* 62(2): 462–482.

Chan CW, Farias VF, Bambos N, Escobar GJ (2012) Optimizing intensive care unit discharge decisions with patient readmissions. *Oper. Res.* 60(6):1323–1341.

Chen X, Sim M, Sun P, Zhang J (2008) A linear decision-based approximation approach to stochastic programming. *Oper. Res.* 56(2):344–357.

Colwell J (2014) Length of stay: Timing it right. Strategies for achieving efficient, high-quality care. Accessed June 22, 2018, http://www.acphospitalist.org/archives/2014/10/los.htm.

Crawford EA, Parikh PJ, Kong N, Thakar CV (2014) Analyzing discharge strategies during acute care: A discrete-event simulation study. *Medical Decision Making* 34(2):231–241.

Dai J, Shi P (2019) Inpatient overflow: An approximate dynamic programming approach. *Manufacturing Service Oper. Management* 21(4):894–911.

De Bin R, Janitza S, Sauerbrei W, Boulesteix A-L (2016) Subsampling vs. bootstrapping in resampling-based model selection for multivariable regression. *Biometrics* 72(1):272–280.

Frakt A (2016) The hidden financial incentives behind your shorter hospital stay. *New York Times* (January 5), https://www.nytimes.com/2016/01/05/upshot/the-hidden-financial-incentives-behind-your-shorter-hospital-stay.html.

Frenz DA (2014) Not too long, not too short, just right. *Today's Hospitalist*. Accessed January 3, 2019, https://www.todayshospitalist.com/not-too-long-not-too-short-just-right/.

George JM, Harrison JM (2001) Dynamic control of a queue with adjustable service rate. *Oper. Res.* 49(5):720–731.

Greenberg BS (1989) *M/G/*1 queueing systems with returning customers. *J. Appl. Probabilities* 26(1):152–163.

Harrison GW, Shafer A, Mackay M (2005) Modelling variability in hospital bed occupancy. *Health Care Management Sci.* 8(4):325–334.

Heggestad T (2002) Do hospital length of stay and staffing ratio affect elderly patients' risk of readmission? A nation-wide study of norwegian hospitals. *Health Services Res.* 37(3):647–665.

Helm JE, AhmadBeygi S, Van Oyen MP (2011) Design and analysis of hospital admission control for operational effectiveness. *Production Oper. Management* 20(3):359–374.

Helm JE, Alaeddini A, Stauffer JM, Bretthauer KM, Skolarus TA (2016) Reducing hospital readmissions by integrating empirical prediction with resource optimization. *Production Oper. Management* 25(2):233–257.

Huang J, Gurvich I (2018) Beyond heavy-traffic regimes: Universal bounds and controls for the single-server queue. *Oper. Res.* 66(4):1168–1188.

Ingolfsson A, Almehdawe E, Pedram A, Tran M (2018) Comparison of fluid approximations for service systems with state-dependent service rates and return probabilities. Working paper.

Kc DS, Terwiesch C (2009) Impact of workload on service time and patient safety: An econometric analysis of hospital operations. *Management Sci.* 55(9):1486–1498.

Kc DS, Terwiesch C (2012) An econometric analysis of patient flows in the cardiac intensive care unit. *Manufacturing Service Oper. Management* 14(1):50–65.

Kc DS, Terwiesch C (2017) Benefits of surgical smoothing and spare capacity: An econometric analysis of patient flow. *Production Oper. Management* 26(9):1663–1684.

Kim S-H, Chan CW, Olivares M, Escobar G (2014) Icu admission control: An empirical study of capacity allocation and its implication for patient outcomes. *Management Sci.* 61(1):19–38.

Kocher RP, Adashi EY (2011) Hospital readmissions and the affordable care act: Paying for coordinated quality care. *JAMA* 306(16):1794–1795.

Kuntz L, Mennicken R, Scholtes S (2014) Stress on the ward: Evidence of safety tipping points in hospitals. *Management Sci.* 61(4):754–771.

Kuo Y-F, Goodwin JS (2011) Association of hospitalist care with medical utilization after discharge: Evidence of cost shift from a cohort study. *Ann. Internal Medicine* 155(3):152–159.

Long EF, Mathews KS (2017) The boarding patient: Effects of ICU and hospital occupancy surges on patient flow. *Production Oper. Management* 27(12):2122–2143.

MacKenzie TA, Tosteson TD, Morden NE, Stukel TA, O'Malley AJ (2014) Using instrumental variables to estimate a Cox's proportional hazards regression subject to additive confounding. *Health Services Outcomes Res. Methodology* 14(1–2): 54–68.

Min X, Yu B, Wang F (2019) Predictive modeling of the hospital readmission risk from patients' claims data using machine learning: A case study on COPD. *Sci. Rep.* 9(1):2362.

Oh JC, Zheng ZE, Bardhan IR (2017) Sooner or later? Health information technology, length of stay and readmission risk. *Production Oper. Management* 27(11):2038–2053.

Ouyang H, Argon NT, Ziya S (2020) Allocation of intensive care unit beds in periods of high demands. *Oper. Res.* 68(2):591–608.

Petrin A, Train K (2010) A control function approach to endogeneity in consumer choice models. *J. Marketing Res.* 47(1):3–13.

Proudlove NC, Gordon K, Boaden R (2003) Can good bed management solve the overcrowding in accident and emergency departments? *BMJ* 20(2):149–155.

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

Sherman SW, Browne JC (1973) Trace driven modeling: Review and overview. Morris MF, Roth PF, Kiviat PJ, eds. *Proc. 1st Sympos. Simulation Comput. Systems* (IEEE Press, New York), 200–207.

Skolarus TA, Jacobs BL, Schroeck FR, He C, Helfand AM, Helm J, Hu M, et al. (2015) Understanding hospital readmission intensity after radical cystectomy. *J. Urology* 193(5):1500–1506.

Tian Z (2016) Time dependent covariate in instrumental variable analysis. PhD thesis, McGill University Libraries, Montreal.

Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ (2011) On the *C*-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statist. Medicine* 30(10): 1105–1117.

Yu B (2008) A frailty mixture cure model with application to hospital readmission CATA. *Biomedical J.* 50(3):386–394.

Zheng C, Dai R, Hari PN, Zhang M-J (2017) Instrumental variable with competing risk model. *Statist. Medicine* 36(8):1240–1255.

Pengyi Shi joined Krannert School of Management, Purdue University as an assistant professor in January 2014. Her research interests include data-driven modeling and decision making in healthcare operations. She collaborated with practitioners and faculty members from different healthcare organizations. Her research won first place for INFORMS Pierskalla Best Paper Award in 2018 and second place for POMS CHOM Best Paper Award in 2019 and 2020.

Jonathan Helm is currently the Grant Thornton Associate Professor of Operations and Decision Technologies at Kelley School of Business, Indiana University. He was a National Science Foundation fellow and winner of the Pierskalla Award for best healthcare paper. His research aims to improve the delivery of healthcare at the system level, the organizational level, and the individual patient level.

Jivan Deglise-Hawkinson is an operations research senior analyst in the Operations Research and Advanced Analytics Department at American Airlines. He was previously the chief data scientist at Lean Care Solutions Corporation. His interests include data-driven forecasting methodologies and leveraging them to create optimized decision support tools.

Julian Pan is the chief executive officer and cofounder of Lean Care Solutions, a healthcare startup providing analytics and operational decision support products to hospitals globally.