A prescriptive optimization approach to identification of minimal barriers for surgical patients

Taghi Khaniyev[*‡], Martin S. Copenhaver[♮♯], Kyan C. Safavi[†§], Ana Cecilia Zenteno Langle[†], Keren S. Starobinski[‡], Bethany Daily[†], Peter Dunn[†§], *and* Retsef Levi[‡]

[*]Department of Industrial Engineering, Bilkent University, Ankara, Turkey
[♮]Johns Hopkins Hospital, Baltimore MD
[♯]School of Medicine, Johns Hopkins University, Baltimore MD
[†]Massachusetts General Hospital, Boston MA
[§]Harvard Medical School, Boston MA
[‡]Sloan School of Management, Massachusetts Institute of Technology, Cambridge MA
Contact: taghi.khaniyev@bilkent.edu.tr

Problem definition: Ensuring timely patient discharges is central to managing a hospital's patient flow; however, discharges are dependent on the coordination of multiple care teams and thus are highly decentralized in nature. Therefore, central capacity teams often lack transparency into how to prioritize scarce resources (e.g., imaging, post-hospital facility placement) for patients who need them for timely discharge. Our goal is to identify a small subset of such barriers for hospitalized surgical patients by balancing two aims: a high likelihood that the patient will be discharged from the hospital in the next 24 hours if these barriers are resolved; and a high likelihood that these barriers will indeed be resolved.

Methodology/results: We leverage a machine learning model that predicts the likelihood of a patient's discharge in the next 24 hours and train another set of machine learning models that predict, for each open barrier of a patient, the likelihood that the barrier will be resolved prior to discharge. We then combine these predictive models with a mixed-integer prescriptive optimization model to identify a subset of issues called minimal barriers. We empirically demonstrate the efficacy of the proposed formulation and solution methodology in identifying a small number of minimal barriers using real data from a large academic medical center. We also study the influence of variability in discharge prediction estimates on the prescribed barriers; we introduce a robust optimization variant which incorporates this directly, and we demonstrate that this is capable of overcoming several shortcomings of the nominal approach.

Managerial implications: The modeling framework we propose can significantly enhance the ability of capacity management teams to identify significant barriers to care progression and understand which barriers are most important to address (based on their relationship with patient discharge) while taking into account the heterogeneity of barriers and their resolution.

1. Introduction

Large hospitals often operate at or near maximum capacity, thereby experiencing significant congestion and delays in patient care. Ensuring smooth patient flow in such settings requires central

capacity management teams to maintain a delicate balance between hospital admissions (inflow) and discharges (outflow). The inflow of admissions of new patients is a relatively transparent process that can be centrally tracked. Elective admissions (e.g., elective surgeries or chemotherapy administration) are known in advance, and even unscheduled admissions through the emergency department (ED) or hospital-to-hospital transfers require a submission of a bed request to a central admitting department. In contrast, the process of discharging patients from the hospital is usually highly decentralized and requires coordination of resources across numerous role groups, making it opaque to capacity planners. Furthermore, frequent delays in the discharge process lead to congestion in critical areas like the ED, intensive care units, and the operating rooms (Copenhaver et al. 2019).

Ensuring patients' timely discharge is a multifactorial challenge. First, there is a high degree of heterogeneity across patients due to their clinical characteristics as well as non-clinical factors such as the availability of social support. Second, the discharge management of care is often spread across different groups of care providers (e.g., physicians and nursing teams) and other professionals (e.g., case managers and social workers), who in turn may have a different perspective on each patient's discharge readiness. Moreover, as hospitals cope with increasing congestion, capacity managers are asked to help prioritize scarce resources (such as imaging or lab tests) in order to expedite patient flow. However, it is time-consuming (and challenging) to verify whether a specific pending test is truly the last thing a patient needs prior to discharge. Thus, the resulting lack of central transparency in patients' discharge readiness can lead to the suboptimal allocation of resources that are essential to patient flow and the discharge of patients. These problems are further compounded by scale—a large hospital system typically has to manage hundreds of patient admissions and discharges each day. While many hospital systems are investing in new technological or personnel strategies to address capacity challenges, such as "capacity physicians" (Safavi et al. 2022) or capacity command centers (Franklin et al. 2022, Åhlin et al. 2023, Franklin et al. 2023), these approaches still face challenges in reliably identifying high-yield patient barriers.

To address this challenge in a systematic, data-driven way, we build upon the work described in Safavi et al. (2019), which introduced the terminology of *barriers to discharge* to dynamically represent the list of issues (e.g., "abnormal lab results", "issues with mobility", or "patient awaiting placement at a facility") that could potentially stand in the way of discharging a patient. This framework uses clinical and administrative barriers to discharge as building blocks to construct a neural network model that can accurately predict how likely a patient is to be discharged within the next 24 hours. It outputs, for each patient, the likelihood of discharge together with a list of open barriers at the moment of prediction. This is important because the concept of barriers is central to how care teams approach the day-to-day management of inpatients, making the model's output

easily interpretable and actionable. The model, however, does not give any guidance on which of these barriers should be prioritized for those patients who are closer to discharge. Moreover, retrospective analysis shows that not all barriers of a patient need to be resolved before discharge. Conceptually, this suggests that there exists a subset of *minimal barriers* whose collective resolution constitutes the minimal effort to enable a patient's discharge. In this work, we combine machine learning models with a practically tractable integer programming approach to identify patients' minimal barriers. This methodology uncovers targeted, data-driven interventions which may be the most effective to expedite patient flow by balancing the likelihood that barriers will be resolved with the possible change in discharge likelihood associated with their collective resolution.

1.1. Approach—Discharge Prediction, Resolution Likelihood, and Minimal Barriers

Our approach is three-fold. First, we change the input and output representation of the model by Safavi et al. (2019) in a way that allows us to provide clinically interpretable model predictions. Specifically, for the input representation, we propose a framework to dynamically represent a patient's clinical and administrative characteristics throughout her hospitalization in terms of which barriers to discharge are active and resolved. If we consider a patient's data as a combination of *static* features (such as age and sex) which do not change during the hospitalization, and dynamic features (such a blood pressure and mobility status), new barriers can be *triggered* (e.g., patient being placed on intravenous narcotics), and remain *open* until they are subsequently *resolved* (the patient being taken off of intravenous narcotics). A binary representation of this trigger-resolution framework, combined with the mathematical optimization machinery, allows us to incorporate sequential and logic-based dependencies among barriers in our prescriptive optimization model. For the output representation, using the discharge likelihoods predicted by the model, we categorize patients by their discharge readiness: those highly likely to be discharged, those unlikely to be discharged, and those "maybe" patients for whom discharge could potentially be facilitated by resolving specific barriers.

Second, we use this input representation to train another set of machine learning models predicting, for each existing barrier of a patient, the likelihood it will be resolved prior to discharge. This step is critical given the complex, personalized nature of barriers which is influenced by several attributes, such as frequency (how often a barrier is triggered in the population), overall resolution likelihood (how often that barrier is resolved amongst those for whom it is triggered), and the workload burden (the amount of work required to resolve the barrier prior to discharge). Intuitively, this step assesses whether the respective barrier is necessary for the patient's discharge.

Finally, these predictive models are subsequently combined within a mixed integer optimization model to answer the following question: among a patient's numerous barriers, does there exist a

small subset of those barriers which balances the likelihood that such barriers are resolved prior to discharge *and* the increased likelihood of discharge for that patient *given that those barriers are resolved*? We call this the *minimal barriers* (MIN) identification problem. Intuitively, among a patient's currently open barriers, the MIN should be defined as those which are both necessary and sufficient for discharge. That is, they are likely to be resolved prior to discharge, and a patient's discharge is likely to occur given that these barriers are resolved. We hypothesize that "maybe" patients will account for most patients with non-empty MIN sets. We also study the influence of estimation error (in discharge predictions) on the minimal barrier set and propose a robust optimization approach to directly incorporate variability in such estimates.

The central modeling challenge posed by such a problem is that this is not merely about prediction, as the outcome (a patient's discharge) is dependent on actions (the resolution of barriers) which may or may not take place. Understanding and addressing this question is a critical step in achieving the broader aim of developing a transparent, centralized, scalable, and effective discharge process for hospitalized patients. Although the framework outlined above is more broadly applicable, in this work we limit our focus to surgical inpatients, i.e., patients who are cared for by a surgical service and require hospitalization following their surgical procedure. One would need to create additional clinical barriers that are relevant for the patient population of interest by working with subject-matter experts from that clinical domain.

1.2. Contributions

Our main contributions can be summarized as follows:

1. We combine a discharge prediction model and barrier resolution likelihood models in a prescriptive mixed integer optimization model that identifies which barriers of a patient, if resolved, are most effective to increase discharge readiness. By jointly optimizing for both discharge likelihood and barrier resolution likelihood, we are able to select a small set of realistic and impactful barriers. We also develop a scalable iterative approach to solve this problem that finds optimal solutions for each patient in less than a second on average.
2. We present a detailed empirical assessment using real patient data to illustrate the effectiveness of the approach in identifying actionable patients and their respective MIN for surgical inpatients at a large academic medical center. In doing so, we discuss how to calibrate our approach to identify a practically reasonable number of patients and barriers. We also show that our personalized barrier resolution likelihood approach is superior to a naïve approach that does not incorporate such information. This analysis supports the original hypothesis that "maybe" patients are those most likely to be identified by the MIN approach.

3. Finally, we demonstrate an application of our approach to the specific problem of identifying patients to prioritize for post-hospital facility placement. We also study how estimation error in the discharge prediction model (namely, variance in prediction estimates) can contribute to changes in the minimal barrier set, and we propose a robust optimization model which can be used to overcome challenges of the nominal approach.

Implications for Practice: This study has several practical implications for hospital capacity management. As hospitals increasingly focus on creating centralized management to direct patient flow (such as capacity command centers), the modeling and algorithmic framework and tools we have created significantly enhance the ability to prioritize patients and specific actions per patient:

1. *Identify a small subset of barriers*: by incorporating the likelihoods that a patient's barriers will be resolved prior to discharge, we are able to identify a small subset of barriers that both require resolution and for which resolution is likely to result in discharge. We show that our personalization approach is advantageous (relative to an appropriate baseline), in contrast to current practice which often uses a one-size-fits-all approach and frequently ignores barrier resolution likelihoods entirely. Importantly, we show that our approach can provide nontrivial insights about which barriers are frequent versus those which are identified in the minimal barrier set. This is especially important as certain high frequency barriers (such as imaging, which is often identified as a barrier, e.g., Safavi et al. 2022) are rarely prescribed.
2. *Enable targeted interventions*: this prescriptive framework allows hospitals to pursue targeted, barrier-specific interventions. We demonstrate this for the use case of post-hospital facility (skilled nursing facility, rehabilitation hospital, or long-term acute care) placement. Even though many hospitals are investing in partnerships with post-acute facilities (or creation of their own facilities), they often lack insight into the operational workflow needed to identify which patients (across an entire hospital) should be prioritized for placement. Our approach addresses this challenge directly, and we estimate the potential benefit of our approach for this use case. While our proposal requires prospective clinical validation, we believe it forms a useful basis for numerous targeted interventions which can address the worsening hospital congestion that capacity management teams face.

1.3. Related Literature

Our research is closely related to existing research in healthcare operations, interpretable machine learning, causal inference (particularly treatment personalization), and combinatorial multi-armed bandits. We briefly summarize the connections to such work and how our work differs.

Healthcare operations: Inpatient capacity management and patient flow planning has been studied from a variety of perspectives. Conforti et al. (2011) and Gartner and Kolisch (2014) look at the problem from a hospital resource allocation perspective and employ mathematical models to maximize patient flow and contribution margins, respectively, while ensuring each patient goes through their respective clinical pathway. From inpatient bed capacity management perspective, some studies such as Kim et al. (2014), McCoy et al. (2018) focus on predicting total discharge volume on a given day while others such as Liu et al. (2010), Azari et al. (2012), Morton et al. (2014) focus on predicting length-of-stay (LOS) for each individual patient. There has also been work to understand discharge readiness as it progresses through a patient's hospitalization, both manually (De Grood et al. 2016) and using machine learning (Levin et al. 2012, Barnes et al. 2016, Safavi et al. 2019). More recently, Na et al. (2023) use machine learning to predict multiple patient outcomes including 24-hour and 48-hour discharge likelihoods. After deploying a webtool that communicates the patient outcome predictions and integrating it into the daily patient review processes at the study hospitals, they report a significant reduction in the average LOS of patients, which highlights the practical potential of such ML-driven models when implemented properly. In the present study, we extend our previous work (Safavi et al. 2019) in which we developed a neural network model to predict discharge likelihoods for surgical inpatients. We use this underlying model as one input in our novel prescriptive approach while also incorporating additional information about how likely a specific barrier is to be resolved by the time of discharge.

Interpretable machine learning: Machine learning (ML) algorithms, able to process substantial electronic clinical and operational data, offer an opportunity to create systematic, scalable, and accurate predictions of which patients will be discharged (Temple et al. 2015, Karhade et al. 2018, Van Walraven and Forster 2017). These algorithms, however, must be combined with an interpretable clinical representation to create an output that helps guide the clinical discharge processes. Rajkomar et al. (2018) proposed an automatized patient representation which analyze electronic health records (EHRs) and construct relevant features in an unsupervised way. They demonstrate that deep learning methods using this representation are capable of accurately predicting multiple medical events such as readmission, prolonged length of stay from multiple centers without site-specific data harmonization. In contrast, Bertsimas et al. (2020b) construct an expertise-driven patient representation on top of EHR data and apply an optimal classification tree machine learning approach to predict several aspects of patient flow such as mortality and discharge within 24 hours; they argue that the black-box nature of deep learning models impedes adoption from doctors and caregivers which are not engaged in the modeling process.

There has been a variety of work to explain black-box models (e.g., Ribeiro et al. 2016). However, these frameworks typically account only for which features contribute to the current prediction

score, and as such, they do not answer the decision-making problem of "what interventions should be made (i.e., barriers resolved) such that a desired outcome is achieved?" Fischetti and Jo (2018) and Anderson et al. (2020) tackle this problem by answering the question "what is the minimal subset of features that needs to be modified in order to change the currently predicted class?" They present MIP formulations for high-dimensional piecewise linear functions that correspond to trained neural networks. Anderson et al. (2020) demonstrate the effectiveness of their approach in a well-known benchmark image classification problem. From methodological point of view, the work of Anderson et al. (2020) is the closest to the framework we are presenting in this study. Their approach, however, lacks a few key characteristics to be applicable to the problem of identifying MIN. First, they assume changing each feature is equally costly/likely to happen. In the context of identifying MIN, the likelihood of each barrier's resolution is an important parameter to take into account. Ignoring this aspect has the potential to output a solution that includes barriers which, if resolved, could significantly improve the patient's score, but is very unlikely to be resolved. Secondly, they also assume that all features can be modified independently of one another, whereas the way we design our feature space ties the values of pairs of trigger and resolution features to each other according to precedence relationships and does not allow making changes to the static features. This is an important characteristic that needs to be taken into account in order to generate feasible solutions in our setting.

Personalized Treatments: In the causal inference literature, the *Potential Outcomes Framework* (Rubin 2005) is often used to quantify the causal effect of a treatment for a given patient by comparing the potential outcomes. The standard framework was developed for a binary treatment (i.e., treatment or control), and has been adapted for multiple discrete or continuous treatments since. Recent approaches such as Shalit et al. (2017) and Bica et al. (2020) use machine learning to estimate personalized potential outcome function from historical observational data containing both contextual features and a treatment feature. In our setting, resolution of an open barrier could be thought of as a "treatment." A major challenge with directly adapting such techniques here is that there are typically multiple (potentially tens of) open barriers for a given patient and more than one of them could require resolution prior to discharge (note, though, that there are some recent literature on multi-arm personalized treatment problems, e.g., Bertsimas et al. 2016, 2020a).

Combinatorial Multi-armed Bandits: Lastly, the class of *combinatorial multi-armed bandit* problems (CMAB) from reinforcement and online learning problems (Chen et al. 2016) share some similarities with our approach. In CMAB, the objective in each round is to choose a subset of base arms (collectively called a *super arm*) in a way that maximizes the expected gain. It is assumed

that each arm's properties are only partially known and may become better understood in the subsequent rounds only by choosing that arm. A variant of the problem is called Contextual CMAB (CC-MAB) which incorporates side information associated with each arm in each round (Chen et al. 2018). The exploitation phase of the CC-MAB problem resembles our problem very closely in that it allows personalization (via context features) and combinatorial selection of multiple arms. The key difference, however, is that they assume that the reward function (or the expected outcome) is a known submodular function of the played arms, whereas in our case, the expected outcome function is not known explicitly and must be learned from historical data.

1.4. Paper Structure

The remainder of the paper is structured as follows: we begin Section 2 by describing the discharge prediction model adopted from Safavi et al. (2019) followed by the changes we made to the model to make it amenable for the subsequent prescriptive optimization model. In Section 2.2, we formally define the notion of barriers and describe the trigger-resolution framework to dynamically represent barriers. Next, in Section 2.3, we introduce the predictive models for barrier resolution likelihood, and subsequently incorporate the predictive models into a prescriptive optimization model in Section 2.4 to identify the minimal barriers. Section 3 includes detailed empirical results when the prescriptive approach is applied to data from a large academic medical center. Section 4 focuses on a model use case and extension. In Section 5, we draw some final conclusions.

2. Modeling Approach

In earlier work, Safavi et al. (2019) trained machine learning models to reliably predict whether a given patient will be discharged from the hospital within the next 24 hours. Such a predictive model takes as its input a variety of clinical and administrative data extracted from a hospital's EHR and provides a prediction score (in $[0, 1]$) as its output which can be interpreted as the likelihood that the patient will be discharged within 24 hours. They tested three commonly used machine learning algorithms (logistic regression, classification trees, and neural networks) with their respective hyperparameters optimized and ultimately selected a feed-forward neural network (with a single hidden layer with 100 nodes) which outperformed the others across all evaluation metrics.

Data elements included in their study were structured, including either numerical values (e.g., laboratory results) or multiple choice of predetermined text phrases (e.g., diet descriptions within diet orders). A summary of data included is shown in Table SM1 in the Supplemental Material. Inclusion of a data source was determined based on (i) whether the clinical care teams currently use it to evaluate a patient's discharge readiness, (ii) the reliability of the data, and (iii) the consistency of usage by different services. From the data included, they extracted two types of features: *static*

and *dynamic*. The static features, $\mathbf{s}$, refer to the elements that do not change throughout the same hospitalization (e.g., age, sex), while dynamic features are those that can change (e.g., whether patient is receiving intravenous narcotics). The first dynamic feature they used denoted the current length of stay, d, expressed as the number of days elapsed since admission. To further represent a patient's dynamic status, they introduced the concept of barriers. A barrier could be specific to a type of surgical procedure or generalizable across multiple types of surgical procedures. A list of 149 potential barriers were identified in Safavi et al. (2019) by a multidisciplinary team that included physicians, nurses, case managers, physical therapists, and nutritionists.

We note that Safavi et al. (2019) also uses the term milestone to represent a patient's dynamic status. In the subsequent sections, we opted for using a single term, *barrier*, for simplicity since not having achieved a milestone can be interpreted as a barrier. We used the same data sources (consequently, the same list of barriers) as in Safavi et al. (2019), but needed to make a number of changes to the feature representation and the discharge prediction model they proposed in order to make it amenable for a prescriptive optimization approach. We next describe these changes.

2.1. Changes Made to the Discharge Prediction Model

Data: To conduct our study, we use data from a 1019-bed academic medical center where more than 36,000 surgical procedures are performed annually. We include all adult inpatients (aged 18 years or older) who had a surgical procedure completed during their hospitalization, were cared for by a surgical team on an inpatient floor during their hospitalization, and were discharged between May 1, 2016, to September 30, 2019. Overall during the study period, there were 52,724 unique hospitalization episodes with average length of stay of 4.45 days. To mimic how a prescriptive model might work in practice (with clinical databases updated once daily shortly after midnight), we capture data at a daily level. In particular, for a given patient on a given day, we use all data collected prior to that day (i.e., until 23:59 on the day prior). This time was used because the clinical and administrative database is updated every 24 hours, starting at 23:59. Our approach reflects a common limitation of real-world implementation, namely, that daily updates are the most common (and feasible) in practice. Therefore, we do not evaluate more frequent data (nor model) updates.

Therefore, our central unit of observation is a patient-day pair, and our outcome of interest is whether the patient was discharged within 24 hours (i.e., on that calendar day). The predictive models discussed subsequently, predict, for each patient, in each day they are in the the hospital. During the study period, the average number of patients on which the models predicted was 188.2, amounting to overall 234,649 patient-day pairs. The average number of discharges per day was 42.3. At the outset, we temporally split the data intro three sets: training, validation, and test. The

training set includes all observations of patients admitted on or before December 31, 2018 (78.9% of patients and 78.2% of patient-days); similarly, the validation set includes all observations of patients admitted between January 1, 2019 and April 30, 2019 (9.5% of patients and 10.0% of patient-days); and finally, the test set includes patients admitted between May 1, 2019 and September 30, 2019 (11.6% of patients and 11.8% of patient-days). All results presented later are based on test set (out-of-sample) performance using models built from the training and validation data. This study was conducted with appropriate approval from the Mass General Brigham Institutional Review Board (protocol `#2011P001124`).

Model Input and Performance: For the purposes of this study, we also trained the same three models (logistic regression, classification tress, and neural networks) as in Safavi et al. (2019) using a modified feature vector representation which transforms dynamic data into features using a trigger-resolution framework where each barrier is represented by a pair of binary features. Dynamic representation of barriers in the feature vector is detailed in Section 2.2. Performance of the models (shown in the Supplementary Material Table SM2) were similar to the ones reported in Safavi et al. (2019). Given its superior discriminative performance (as measured by area under the receiver-operator characteristic curve, or AUC), the neural network model was implemented in practice at our institution and therefore we use this model structure as the basis of the discharge prediction part of our approach.

Model Output: Next, to better reflect clinical and operational ambiguity in discharge predictions, we convert the continuous prediction score into three discrete predictions (as opposed to two); in particular, instead of a binary Yes or No prediction, a third Maybe category is also used, and the groups are defined as follows (given discharge prediction score s):

$$\text{predict} \begin{cases} \text{No}, & s \in [0, S_-) \\ \text{Maybe}, & s \in [S_-, S_+] \\ \text{Yes}, & s \in (S_+, 1] \end{cases}$$

where S_- and S_+ fixed thresholds are defined so that the No and Yes categories have high negative and positive predictive power (NPV/PPV), respectively. Precisely,

- S_- was chosen so that 95% of the (training sample) predictions that fall under No category were not discharged within 24 hours (NPV=95%); and
- S_+ was chosen so that 75% of the predictions that fall under Yes category were discharged within 24 hours (PPV=75%).

These percentages were based on clinical input about tradeoffs of making the respective predictions. Based on the training and validation data, we set the thresholds as $S_- = 0.19$ and $S_+ = 0.59$. The performance of these choices on the test data is shown in Table 1.

Table 1 Test sample discharge prediction categories and comparison of actual discharge times

Prediction category	# (%) of Observations	Leave within 24 hours	Leave in 24-48 hours	Leave after 48 hours
No	16373 (59.3%)	823 (5.0%)	2017 (12.3%)	13533 (82.7%)
Maybe	7980 (28.9%)	2933 (36.8%)	2167 (27.1%)	2880 (36.1%)
Yes	3248 (11.8%)	2367 (72.9%)	548 (16.9%)	333 (10.2%)

Note that the Yes and No categories make up of approximately 71% of all predictions. Moreover, these two correspond informally to "leave within 24 hours" and "leave after 48 hours," respectively. On the other hand, most of the patients who leave between 24-48 hours fall under the Maybe category. Combining this with the insight from Safavi et al. (2019) that "false positive patients without a clinical reason to stay are typically discharged within 24 to 48 hours", the Maybe category roughly corresponds to the potentially actionable patients for the purposes of saving bed-days. **Therefore, we hypothesized that the Maybe patients would be the most prevalent patients to have specific MIN identified**; we perform a detailed assessment of this hypothesis in Section 3.

Calibration: A critical notion for the meaningful intreptation of clinical prediction models in practice is that of model *calibration* (Huang et al. 2020, Steyerberg 2019). For a predictive model $h : F \to [0,1]$, where F is the feature space and the observed outcome data is a binary $y \in \{0,1\}$, model calibration requires assessing how well $h(\mathbf{x})$ estimates $\mathbb{P}(y=1|\mathbf{x})$, i.e., the probability that $y=1$ given feature vector $\mathbf{x} \in F$. There are numerous approaches to assessing model calibration. For our purposes, we follow one of the common approaches based on the *Brier score* for h (Huang et al. 2020). For a set of data $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^n \subseteq F \times \{0,1\}$, it is defined as

$$\text{Brier}_D(h) := \frac{1}{n} \sum_i \left(y_i - h(\mathbf{x}_i)\right)^2 .$$

Lower Brier scores typically correspond to better calibration for a given prediction task.

In order to assess whether a given model h is properly calibrated on a data set D, we apply the Spiegelhalter Z-test (Huang et al. 2020). In particular, a model is defined as calibrated per a Spiegelhalter test (at the 0.05 level) if $\text{Spieg}_D(h) \leq 1.96$, where

$$\text{Spieg}_D(h) := \left| \frac{\sum_i (y_i - h_i)(1-2h_i)}{\sqrt{\sum_i (1-2h_i)^2 h_i (1-h_i)}} \right|,$$

and $h_i := h(\mathbf{x}_i)$. We chose to apply this definition because it is easy to use in settings where many models are being tested and where it is not feasible to visually inspect usual calibration plots; indeed, this is important in the subsequent section on barrier resolution models. (Note that Spiegelhalter tests are sometimes interpreted via p values instead. In particular, a p value can

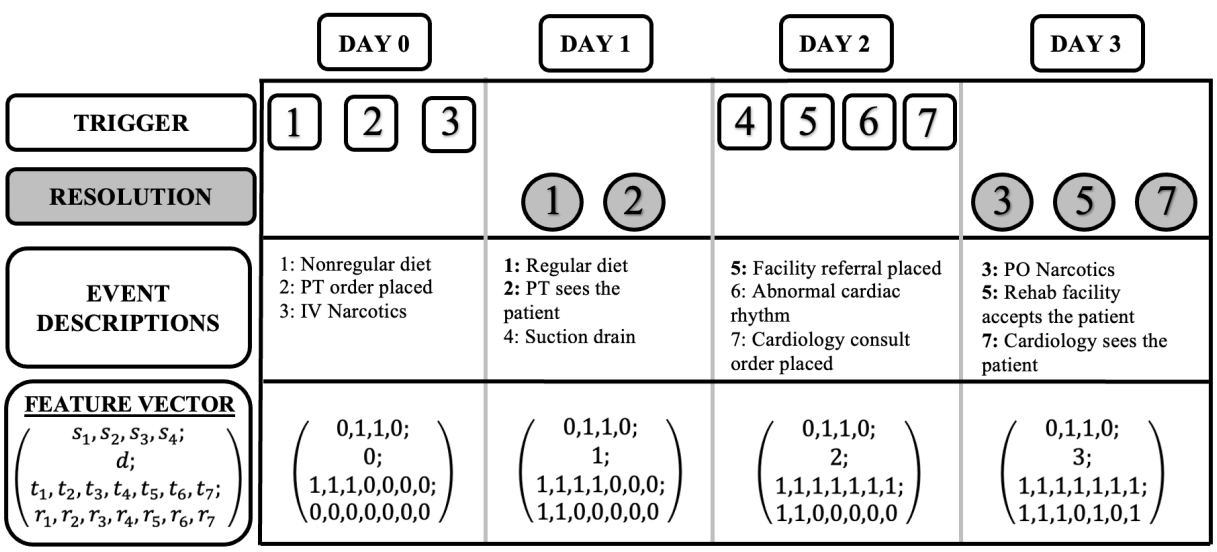
be found by computing $p = 2 \cdot \Phi\left(-\text{Spieg}_D(h)\right)$, where Φ is the cumulative density function of a standard normal; therefore, calibration corresponds precisely with failing to the reject the null at a significance level of 0.05.)

Given this definition, we verified that the discharge prediction model (as defined above) is properly calibrated on the training set. The Brier scores of a Platt-scaled version of the model on the training, validation, and testing sets were 0.109, 0.111, and 0.113, respectively; a calibration plot is shown in Figure SM1. Platt scaling was used given that the discharge estimates are not well calibrated otherwise, as they consistently overestimate the likelihood of discharge. For further details, see the Supplemental Material.

2.2. Dynamic Representation of Barriers

We represent the dynamic status of a patient's barriers by using a trigger-resolution framework where each barrier is encoded by a pair of binary features: (1) a trigger feature that marks whether the barrier has been activated during the hospitalization; and (2) a resolution feature that signals whether the barrier has been resolved. Figure 1 provides an illustrative example of how the trigger-resolution framework could be used to capture the dynamic state of a patient from the moment the patient is admitted and until discharge.

Figure 1 **Example timeline of a total knee replacement patient's hospitalization**



Each vertical block in Figure 1 represents a single day during the hospitalization. The first horizontal block illustrates the events' trigger and resolution timeline. Each type of event (e.g., diet, physical therapy, medication, etc.) is marked with a different number, when it is triggered (white rectangles), and when it is resolved (gray circles). The second horizontal block lists the event

descriptions denoted by the rectangles and circles in the block above. Finally, the third horizontal block illustrates the feature vector representation of the patient for each day during the hospitalization. Entire feature vector $\mathbf{x}_i = (\mathbf{s}_i, d_i, \mathbf{t}_i, \mathbf{r}_i)$ of observation i is composed of four components as described above: *static*, *days since admission*, *trigger*, and *resolution* features, respectively. In this specific example, the static features are illustrated by the first four binary values $(s_1, s_2, s_3, s_4) = (0, 1, 1, 0)$. For the purposes of illustration, let us assume that these features represent, respectively, "whether the patient's sex is *Female*", "whether the patient's age is over 65", "whether the patient had a knee replacement surgery", and "whether the patient is on a Cardiology service floor". Observe that these four features do not change throughout the hospitalization, as expected. The number of days since admission is denoted by d on the second line of the feature vector and is incremented by 1 each day. Finally, the third and fourth lines in the feature vector indicate whether each barrier (seven of them illustrated in this example) is triggered and resolved, respectively.

According to Figure 1, right after the surgery, the patient was unable to have a regular (oral) diet, had a need to be seen by a physical therapist (PT), and was on intravenous (IV) narcotics. On the next day, two of the barriers from the previous day (i.e., diet and PT consult) were resolved, but a new one was introduced as a suction drain was placed. On Day 2, several other barriers were introduced. The patient was assessed to need a post acute care facility after discharge, hence a referral to facility was placed awaiting for acceptance. Moreover, patient appeared to have some heart issues indicated by an abnormal cardiac rhythm. As a result, a consult order to Cardiology was placed. On the third and last day, patient was taken off of the IV narcotics and started on oral (PO) narcotics for the pain. A rehab facility accepted the patient and a cardiology specialist saw the patient.

It can be seen from the barriers representation on the Day 3 that the patient was discharged without two of the triggered barriers resolved: "Suction drain" and "Abnormal cardiac rhythm." This may seem like an anomaly, but in fact it is not unusual with surgical patients. Depending on which cohort a patient belongs to (type of surgery, age, post acute care need, etc.), the resolution of a barrier may not be necessary for discharge. In this example, the patient is discharged to a rehab facility where the patient will continue to receive professional medical care, which means as long as it did not constitute a major immediate problem for the patient (which may be inferred from a cardiologist signing off after the consult), a cardiac rhythm issue could be kept under control at the next care facility. Similarly, the suction drain can be removed at a later time, if, for example, the patient is discharged to a rehab facility. Thus, in itself, the fact the suction was not removed does not mean necessarily that the patient could not be discharged. In fact, our analysis shows that, on average, only 43.6% of the barriers which get triggered at some point during the hospitalization of the patient get resolved by the time of discharge. This highlights the importance of identifying the

barriers that are critical and necessary for patient's discharge. Identifying this subset of necessary and sufficient barriers among many barriers that might be triggered is key to enabling a proactive approach to facilitating discharges.

Table 2 illustrates the heterogeneity in barrier occurrences and resolutions for a subset of ten barriers. Specifically, observe that some barriers such as "pending physical therapy (PT) discharge disposition recommendation," "irregular diet," and "on intravenous narcotics" occur in many patients, while others mostly occur only in a smaller fraction (for example, only 1.3% of patients have an "on heparin" barrier during their hospitalization). There is also heterogeneity in the percentage of patients for which the respective barrier is resolved prior to discharge. The data in Table 2 imply that not all barriers are equally likely to be resolved prior to discharge. Moreover, the importance of a barrier's resolution for a patient's discharge is not identical across different patients, specifically, for some the barrier might be critical to allow discharge, whereas for others it might not prevent the discharge. Taking this heterogeneity into account, in the next subsection, we train a new set of predictive models to predict, for a given patient, how likely an open barrier of the patient is to be resolved prior to discharge.

Table 2 Ten barriers and their respective out-of-sample summary statistics. "Patient level" indicates the percentage of patients for whom the barrier was triggered (and subsequently resolved prior to discharge), while "Patient-day level" provides the percentage of hospital days for which the barrier was activated and then the percentage where it was resolved (after activated).

Barrier	**Patient level**		**Patient-day level**		**Resolution model**	
	% triggered	% resolved	% open	% resolved	AUC	Brier score
Pending Physical Therapy's discharge recommendation	96.0	46.3	59.5	45.4	0.91	0.12
On oral narcotics	69.2	24.9	49.7	21.2	0.72	0.15
Irregular diet	58.8	58.2	44.4	40.6	0.87	0.14
Urinary catheter	42.7	84.9	25.8	78.4	0.79	0.13
On intravenous narcotics	37.2	83.8	13.5	85.9	0.85	0.10
On oxygen device	34.8	93.6	22.4	83.7	0.77	0.12
Low urine output	34.7	87.0	17.1	83.9	0.89	0.08
X-ray study	33.9	80.6	20.7	75.3	0.70	0.17
Awaiting post-hospital facility acceptance	15.6	78.8	18.5	82.2	0.80	0.11
On heparin	1.3	94.9	0.5	92.5	0.58	0.08

2.3. Barrier Resolution Prediction Models

The feature vector described in Section 2.2 is designed to represent a patient's current status, including their barriers. Therefore, by extracting the feature vector for each (historical) patient at their time of discharge, we can construct a binary outcome vector where each element represents whether the patient was discharged with the respective barrier resolved. The availability of this

historical data allows us to train a model for each barrier to predict whether an open barrier of a patient will be resolved by the time of discharge. This requires training a resolution prediction model for each of the 149 barriers included in the model (a list of the barriers can be found in the Supplemental Material).

At this stage, our goal for each barrier was to build a model that predicted, for any observation (patient-day) with that barrier open, whether that barrier is resolved by the time of discharge (given the patient's current feature vector). Our primary objective in doing so was to create models with high discriminative power with respect to the resolution outcome (i.e., high AUC) while ensuring that the models were appropriately calibrated on the training and validation sets. In particular, for each of the 149 barriers we performed the following:

1. On the subset of the training data with that barrier open, we train a sequence of ℓ_1-regularized logistic regression models. In particular, we use `glmnet` (Friedman et al. 2010) to compute the regularization path for regularization parameters $\lambda_1 > \lambda_2 > \cdots > \lambda_{\min} > 0$ (we use the default automated grid choice and remove any ℓ_2 regularization). Using the parameter $\lambda_{1\text{se}}$ corresponding to the parameter for the model within one standard error of that with the lowest cross-validation error (as performed only based on the training set), we exclude any models with $\lambda < \lambda_{1\text{se}}$. (The $\lambda_{1\text{se}}$ choice is based on standard practice recommendations, *cf.* Hastie et al. 2009.) This results in a set of regularized logistic regression models $M_{\lambda_1}, \ldots, M_{\lambda_{1\text{se}}}$. We assume that we have K such models (i.e., $|\{\lambda_1, \ldots, \lambda_{1\text{se}}\}| = K$).
2. We also train K ordinary logistic regression models $U_{\lambda_1}, \ldots, U_{\lambda_{1\text{se}}}$ defined as follows: for each model M_λ on the regularization path, we identify the set of nonzero variables for that model and train an ordinary (unregularized) logistic regression model (again with outcome being whether the barrier is resolved) but only on the subset of features for which the corresponding coefficient in M_λ is nonzero. (The motivation for training such unregularized models as well in this setting is that regularization is known to potentially adversely affect model calibration.)
3. For the set of $2K$ logistic regression models, we exclude any models with validation AUC below 0.5; we also exclude any models which fail to calibrate per a Spiegelhalter test (either on the training data or on the validation data). Among the remaining models, we pick the one that maximizes $\text{AUC}_{\text{val}} - \text{Brier}_{\text{val}} - \max\{\text{AUC}_{\text{train}} - \text{AUC}_{\text{val}}, 0\}$, where AUC_D denotes the AUC on data set D. The motivation for this choice is that we simultaneously want to maximize validation AUC while minimizing Brier score and controlling the imbalance between the training and validation AUCs.

Based on this process, there were 16 barriers with validation-calibrated models. There were 25 barriers for which there are no calibrated logistic regression models with validation AUC above 0.5; for such barriers, we use the baseline resolution percentage (in the training and validation

sets combined) for all observations. There were also 8 barriers which were never present in the training set, never present in the validation, or never resolved in both; resolution likelihoods were set uniformly to 0 for these barriers. Therefore, for each observation, we can obtain the personalized barrier resolution likelihood vector, denoted $\mathbf{L} \in [0,1]^{149}$, by concatenating the prediction scores of all resolution prediction models (L_j represents the resolution likelihood for barrier j; the dependence on the observation is omitted for readability). Table 2 includes the AUC and Brier score for a sample of 10 of these barriers. Further discussion of the relationship between the number of observations with a given barrier and out-of-sample AUC is included in the Supplemental Material. In initial model-building, we also evaluated a multi-output neural network model (to predict joint resolution of all 149 barriers), but this approach performed significantly worse than simple models on the single-barrier-resolution prediction task. Therefore, we restrict our attention to the barrier-by-barrier approach described above. However, the creation of reliable barrier resolution prediction models which capture the dependency structure (and account for low prevalence of many barriers and low resolution frequencies for many barriers) should be considered in future work.

Having trained the discharge prediction model and the barrier resolution prediction models, we next describe how these predictive models can be leveraged to formulate and solve the MIN problem, which in turn can be used to inform decisions on which barriers to target for each patient.

2.4. Prescriptive Model — Identifying Minimal Barriers

As was highlighted in Section 2.2, not all barriers of a patient are resolved prior to discharge. Moreover, among the barriers that are resolved before discharge, there may be some whose resolutions were not necessary for discharge. Conceptually, this suggests that there exists a subset of barriers (we call them *minimal barriers*, or MIN) whose collective resolution constitutes the minimal effort to enable a patient's discharge. Our goal is to identify such a set of barriers.

To define our objective, we start with an idealized model. In particular, for a subset S of a patient's current barriers, let $\mathbb{P}_{\text{res}}(S)$ denote the (idealized) probability that all the barriers in S are resolved prior to discharge, and let $\mathbb{P}_{\text{dc}}(S)$ denote the (idealized) probability that the patient is discharged within 24 hours *given* that the S barriers are resolved. We aim to solve the problem of finding S that maximizes

$$\max_{\text{barriers } S} \mathbb{P}_{\text{res}}(S) \cdot \mathbb{P}_{\text{dc}}(S),$$

the product of the probability of the barriers being resolved and the probability of discharge given that resolution. An optimal barrier set S^* to this problem balances the likelihood of those barriers being resolved with the contribution that such barrier resolution makes to a patient's discharge likelihood. The advantage of such an objective is that it implicitly accounts for the fact that some barriers are unlikely to be resolved even if they contribute significantly to discharge likelihood,

while other barriers might be likely to be resolved but make little or no difference to a patient's progression towards discharge. To solve this idealized problem, we need a way of estimating the two central quantities $\mathbb{P}_{\text{res}}$ and $\mathbb{P}_{\text{dc}}$. The latter can be estimated using the discharge prediction model described in Section 2.1 and the former can be estimated using the barrier resolution prediction models described in Section 2.3.

We now translate this idealized problem of identifying a patient's *minimal barriers* into an optimization problem where the decisions on the open barriers ("whether the barrier should be selected to resolve") are reflected as changes in the current feature vector $\mathbf{x}^c$ of a patient to a modified vector $\mathbf{x}$ (in this section, we suppress the index for observation i to make the exposition simpler). Consequently, we try to decide which features to modify, by allowable changes, so that the product of the following two objective functions is maximized:

- $J_1(\mathbf{x};\mathbf{x}^c)$: the discharge likelihood within 24 hours if all selected barriers are resolved; this approximates the idealized $\mathbb{P}_{\text{dc}}$.
- $J_2(\mathbf{x};\mathbf{x}^c)$: the product (among the selected barriers S represented in changing $\mathbf{x}^c$ to $\mathbf{x}$) of each of the likelihoods that the given barrier will be resolved before discharge; this approximates $\mathbb{P}_{\text{res}}(S)$, the probability of the barriers in S being resolved, assuming all barriers resolve independently of one another.

The functions J_1 and J_2 are computed using the predictive models described in Sections 2.1 and 2.3, respectively, where $\mathbf{x}$ and $\mathbf{x}^c$ are the modified feature vector and the original feature vector of the patient, respectively.

Using these two objectives, we define the composite objective function J as the product $J = J_1 \cdot J_2$ and we call this the **overall likelihood given a subset of barriers**. This is directly analogous to the idealized objective we discussed at the beginning of Section 2.4, with J_1 estimating $\mathbb{P}_{\text{dc}}$ and J_2 estimating $\mathbb{P}_{\text{res}}$. Instead of optimizing J directly, we adopt a constraint-based multiobjective optimization perspective. In particular, we compute the Pareto frontier for J_1 and J_2 by solving instances of maximizing J_2 subject to (iteratively updated) lower bounds on J_1. We proceed by describing the structure of the two different function J_1 and J_2.

Structure of J_2: Notice that the second objective function J_2 can be calculated by multiplying the personalized resolution likelihoods $\mathbf{L}$ of the selected barriers readily available from the barrier resolution prediction models described in Section 2.3. To linearize J_2, we take the natural logarithm of the personalized likelihoods $\log(L_j)$ so that maximizing the sum of log-likelihoods of the selected barriers would be equivalent to maximizing the likelihood that all selected barriers will be resolved, which we refer to as their *collective resolution likelihood*. Addition of a new barrier to the list of selected barriers always reduces the collective resolution likelihood. Therefore, this objective function implicitly favors reducing the number of selected barriers.

The advantage over simply minimizing the number of selected barriers is that this way we are able to account for the differences in the resolution likelihoods of barriers. It is intuitive to conclude that one would prefer selecting a list of two barriers both very likely to be resolved by the time of discharge rather than a single barrier which will almost surely not be resolved. In this way, J_2 can be generally thought of as a measure of **cost** (higher J_2 corresponding to lower cost), and in other settings we expect that J_2 might be more directly defined using actual costs.

Structure of J_1: On the other hand, optimizing with the first objective function, J_1, is a more challenging task for two reasons: it is implicitly expressed through a trained neural network; and the underlying decision space is discrete (i.e., binary decisions for whether each barrier is resolved) and enumeration is not practically feasible. To address this, we make use of the work of Anderson et al. (2020) who proposed and evaluated tractable mixed integer program formulations to find the minimal subsets of features that leads to a desired change in neural network predictions. To apply their framework, we address two specific modifications:

- Not all features are modifiable; only the resolution features of open barriers can be changed from 0 to 1. In other words, the only allowed action is to "resolve an existing barrier."
- The objective is not to minimize the number of selected barriers that can raise prediction score to a predefined level, but instead to maximize the function $\log(J_2)$ subject to the prediction score being at least some threshold z_0.

In the Supplemental Material, we include a detailed exposition showing how these modifications can be made to yield a linear MIP formulation of the subproblem

$$[\mathrm{SP}(z_0;\mathbf{x}^c)] \qquad \max_{\text{feasible } \mathbf{x}} \ \log J_2(\mathbf{x};\mathbf{x}^c) \quad \text{s.t.} \quad J_1(\mathbf{x};\mathbf{x}^c) \geq z_0,$$

where $z_0 \in [0,1]$ is a fixed threshold and $\mathbf{x}^c$ is a patient's current feature vector. (The underlying decision variables are binary variables capturing whether each barrier is resolved.) Note that the key to this reformulation is using the underlying neural network structure of the model for predicting discharges. The resulting formulation has $2h + p_b$ variables and $4h + 2p_b + 1$ constraints, where h is the dimension of the hidden layer in the neural network (in our case, 100) and p_b is the number of barriers (in our case, 149).

2.4.1. An Iterative Algorithm to Optimize J: Given the formulation of SP as a MIP which is amenable to commercial solvers, we are now ready to proceed with an algorithm for maximizing J for a given observation $\mathbf{x}^c$. As noted earlier, we do so by solving $\mathrm{SP}(z_0;\mathbf{x}^c)$ across a spectrum of thresholds $z_0 \in [0,1]$. The corresponding algorithm is shown in pseudocode in Algorithm 1, which we refer to as the *Barriers-to-Discharge* (or BtD) algorithm. The optimal solution $\mathbf{x}^*$ to a given feature vector $\mathbf{x}^c$ identifies a set of **prescribed (or minimal) barriers**, namely, the

MIN set. The algorithm follows the usual approach used when computing Pareto frontiers, namely, instead of solving $\mathrm{SP}(z_0; \mathbf{x}^c)$ for all choices of $z_0 \in \{0, \epsilon, 2\epsilon, \ldots, 1\}$, we only solve it for a (typically smaller) set of values.

In the remainder of the paper, we make use of two important definitions:

Definition 1 (Feasible patient). Given a patient observation with corresponding feature vector $\mathbf{x}^c$, we say that the observation is **feasible** if there exists some set of open barriers such that J is strictly improved (up to tolerance ϵ) relative to the value of J when no open barriers are resolved; an infeasible patient has no prescribed barriers. In mathematical notation, this is equivalent to there existing some feasible feature vector $\mathbf{x}$ such that $J(\mathbf{x}; \mathbf{x}^c) > J(\mathbf{x}^c; \mathbf{x}^c) = J_1(\mathbf{x}^c)$.

Definition 2 (Actionable patient). Given a feasible patient with feature vector $\mathbf{x}^c$, we say that the patient is **actionable with likelihood** J^*, where $J^* := J(\mathbf{x}^*; \mathbf{x}^c)$ as computed via the BtD algorithm. For every $J' < J^*$, we say that the patient is **actionable above threshold** J'.

Algorithm 1 $\mathrm{BtD}(\mathbf{x}^c, \mathbf{l}, \epsilon)$

```
Initialize: z_0 ← J_1(x^c, x^c); J_max ← 1; J_best ← 0; and x_best ← x^c
while J_best < J_max do
    [feasible, J_1, J_2, x*] ← SP(z_0; x^c)
    if not feasible then
        break
    else  #  SP is feasible
        J ← J_1 · J_2 and J_max ← J_2
        if J > J_best then
            J_best ← J and x_best ← x*
        end if
    end if
    z_0 ← J_1 + ε
end while
return x_best, J_best
```

3. Empirical Results

In this section, we assess the performance of our proposed approach on real data from a large academic medical center. More specifically, we answer the following questions:

1. How does the composite objective function J perform in discriminating likelihood of discharge given a selected subset of barriers?
2. Does the BtD algorithm primarily identify Maybe patients as those who are actionable above certain thresholds as initially hypothesized? How does a patient's actionability relate to their initial discharge score, and how many barriers are typically prescribed?

3. What lower threshold $J_{\min}$ should be chosen for actionable patients to yield a practically meaningful number of patients on a daily basis? Which barriers tend to be prescribed as minimal barriers, and how should these be interpreted in this context?
4. What is the contribution of personalized resolution likelihoods?
5. How does the BtD approach compare to complete enumeration in terms of computational efficiency? How can our approach be modified in settings with other (e.g., tree-based) predictive models?

We conducted our analysis on the out-of-sample test sample, i.e., 27,601 observations (patient-days) from 6,123 unique patient hospitalizations. We implemented the BtD algorithm in `julia` version 1.10.4 and utilized `Gurobi` 11 as a mixed integer optimizer, with default parameters and a single thread, to solve the $\mathrm{SP}(z_0; \mathbf{x}^c)$ MIP instances within the BtD algorithm. All the experiments in this section were carried out on a laptop with processor 3.1 GHz Dual-Core Intel Core i5, memory 8 GB 2133 MHz LPDDR3. Experiments in Section 4 were run on a cluster computing environment to accommodate large-scale replications.

3.1. Efficacy of the Composite Objective Function

We begin by assessing how well the composite objective function J performs in capturing the idealized definition of MIN. Such idealized barriers should satisfy two conditions simultaneously: ($I1$) that they all must be resolved by the time of discharge and ($I2$) the patient must be discharged on the same calendar day that all of them are resolved. To test this, we performed the following analysis: for each test set observation, we randomly selected a subset of up to five of the patient's open barriers and calculated the objective function value J corresponding to the modified feature vector obtained by resolving the selected subset of barriers; then, retrospectively, we labeled each instance as follows:

$$\text{label} = \begin{cases} 1, & \textit{if } \text{both conditions } I1 \text{ and } I2 \text{ were satisfied} \\ 0, & \textit{otherwise.} \end{cases}$$

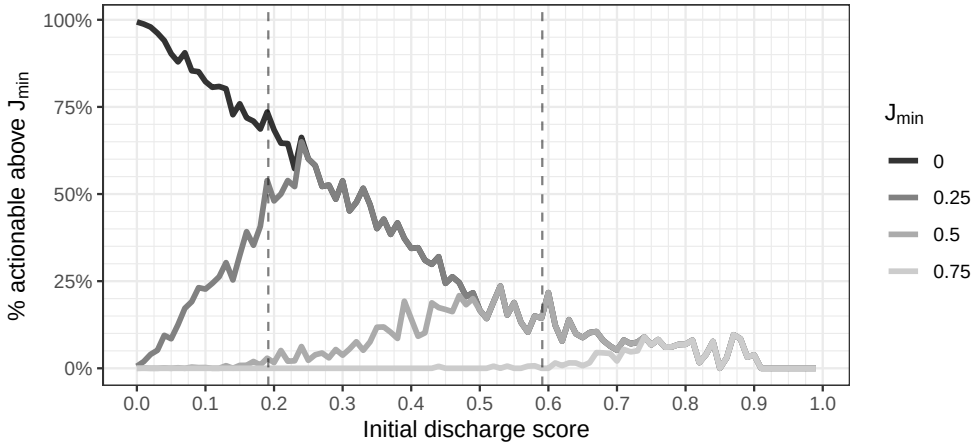
In this experiment 17.1% of observations met $I1$, 20.1% met $I2$ given that $I1$ was met, and 3.4% met both $I1$ and $I2$ (i.e., had a label of 1). We calculated the discriminative performance of J in terms of AUC for this labeled outcome. Overall, the AUC was 0.854 (95% CI [0.843,0.864]); within the Yes/Maybe/No subgroups, Yes patients had an AUC of 0.938 [0.924,0.952], Maybe patients 0.910 [0.899,0.922], and No patients 0.814 [0.799,0.829]. This suggests that J is particularly effective at discriminating between patients' labels within the Yes and Maybe subsets of patients. In the Supplemental Material, we include detailed analysis on the performance of J with respect to conditions $I1$ and $I2$ and as a function of the number of selected barriers.

3.2. Identifying actionable patients

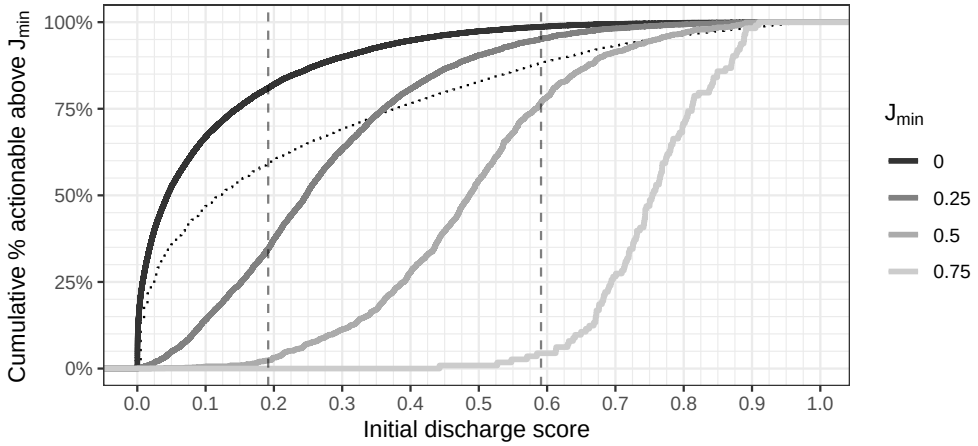
To better evaluate the BtD algorithm and assess our hypothesis that MAYBE patients account for the majority of actionable patients, we solved the BtD algorithm for all 27,601 observations. The corresponding results regarding the percent of patients who are actionable at different likelihood thresholds $J_{\min} \in \{0, 0.25, 0.5, 0.75\}$ are shown in Figure 2 (2(a) shows percent actionable for a given initial discharge score, while 2(b) shows the empirical cumulative density functions for actionable patients above thresholds).

Figure 2 **Percent of actionable patients at a given score and with given value of $J_{\min}$, where $J_{\min} \in \{0, 0.25, 0.50, 0.75\}$. Two vertical lines for Maybe patient score cutoffs are shown; note that $J_{\min} = 0$ corresponds to feasible patients.**

(a) Percent of actionable patients at a given initial discharge score (i.e., $J(\mathbf{x}^c; \mathbf{x}^c)$) and value of $J_{\min}$ (initial scores grouped to second decimal)



(b) Cumulative percent of actionable patients at a given discharge score and value of $J_{\min}$. The cumulative percent of *all* patients at a given score is shown as a dashed line for reference.



3.2.1. Feasibility: The percent of patients at or above a given discharge score who are feasible is shown with $J_{\min} = 0$ in Figure 2(a). We observe that the percent of feasible patients generally decreases linearly with increasing baseline discharge scores. As a given observation being feasible coincides with the existence of at least one barrier inMIN (per the BtD algorithm), this behavior is consistent with our expectations of how MIN would behave: for patients who are in their current state unlikely to be discharged, there likely exist some barriers to their discharge which could raise their discharge likelihood (while also being reasonably likely to be resolved, given that the patient is further from discharge); at the same time, patients with increasingly higher scores are more likely to be discharged in their current state, so it is less likely that there exists barriers which not only increase the patient's discharge likelihood, but importantly are also likely to be resolved. This is particularly important given the patient heterogeneity noted earlier—not every barrier is always resolved by discharge (and indeed, some barriers are resolved only in a small minority of patients), and some barriers only occur in specific patient subpopulations (which is particularly salient given that the data includes surgical patients across all services).

3.2.2. Actionability at different thresholds: Not surprisingly, increasingly larger minimum thresholds $J_{\min}$ affect the percent of patients who are actionable. We also observe that the percent actionable above threshold $J_{\min}$ peaks at approximately an initial discharge score of $J_{\min}$ across the different choices shown. Overall, the qualitative behavior seen here suggests that the optimization approach we have taken to identifying actionable barriers aligns with the clinical intuition we used to approach this problem, namely, that the composite objective would capture the trade-off between the likelihood of improving a patient's discharge and the likelihood of being able to resolve barriers to put the patient in that improved discharge state.

We also note that the changing thresholds $J_{\min}$ highlight the inherent tradeoff between the number of patients identified as actionable above such a threshold and how likely patients are to be discharged with such barriers resolved. We return to this point in more detail in Section 3.3.

3.2.3. Proportion of Maybe patients: Figure 2(b) confirms our initial hypothesis: while Maybe patients account for 28.8% of observations in the test set (Table 1), they represent a disproportionately large share of actionable patients across a range of thresholds. For example, at a threshold of $J_{\min} = 0.25$, Maybe patients account for 60.6% of actionable patients, while at a threshold of $J_{\min} = 0.5$, Maybe patients account for 74.5% of actionable patients. Not surprisingly, though, we do see that for $J_{\min}$ sufficiently large (in our case, $J_{\min} \geq 0.60$), Maybe patients are no longer the majority of those identified as actionable (the same is true of $J_{\min}$ sufficiently small, in our case $J_{\min} \leq 0.20$).

3.2.4. Number of prescribed barriers: Among all feasible patients, the number of open barriers is 11.5 on average (median of 10, range of 1-40), while the number prescribed for such patients is 3.0 on average (median of 2, range 1-27). This suggests that the BtD algorithm is effective at reducing the list of barriers for each patient to a more reasonable number (26.9% of the list of open barriers on average, median 21.4%, range [3.4%,100%]). Consistent with this behavior, the number of prescribed barriers differs noticeably across the patient groups: No patients are prescribed an average of 3.5 barriers (median 3, range [1,27]), Maybe patients 1.2 barriers (1, [1,4]), and Yes patients 1.0 barriers (1, [1,2]).

Finally, as the number of open barriers is negatively correlated with initial discharge score (correlation coefficient of -0.402 for feasible patients), we expected that the number of prescribed barriers would also be negatively correlated with initial discharge score; indeed, we found this to be the case (correlation coefficient of -0.462).

3.3. Model performance—Thresholds for actionability and Barriers selected

Given an overall understanding of how actionability relates to a patient's underlying discharge likelihood, we now focus on addressing some specific questions:

- What choice of $J_{\min}$ gives a reasonable number of actionable patients which is practically feasible to review on a daily basis?
- Which barriers are typically prescribed? What is a clinical interpretation of these barriers?
- As observed earlier, Maybe and Yes patients are often prescribed a single barrier. How often is this barrier the one with the highest personalized resolution likelihood?

3.3.1. Model performance across choices of $J_{\min}$: To assess the model performance, we consider several key metrics (aligning with those considered in Section 3.1):

($M1$) the percentage of actionable observations for whom selected barriers were all resolved by the time of discharge;

($M2$) the percentage of observations meeting criterion $M1$ who were discharged on the same day that all selected barriers were resolved; and

($M3$) the percentage of observations for which both $M1$ and $M2$ were satisfied.

We evaluated these across various thresholds $J_{\min}$; the corresponding results are shown in Table 3. We also include several additional variables, such as values of the composite objective function and the percent of actionable patients relative to all patients.

As expected, we see that the percent of actionable patients who have all prescribed barriers resolved **and** are discharged on the same day as their resolution (i.e., metric $M3$) increases as the threshold for actionability $J_{\min}$ is increased, highlighting the tradeoff between the number of actionable patients identified and the percent of those patients for whom the prescribed barriers

Table 3 Prescriptive model's test-set performance at different choices of $J_{\min}$. The percent shown next to the number actionable is relative to all test-set observations. Averages for J^*,J_1^*, J_2^*, and "# Barriers" are based on those values as computed using the BtD algorithm. Recall that J_1^* corresponds to a discharge likelihood and J_2^* to a barrier resolution likelihood.

	# Actionable		# Barriers		Average value			Metric (%)		
$J_{\min}$	above $J_{\min}$		Open	Prescribed	J_1^*	J_2^*	J^*	$M1$	$M2$	$M3$
0.0	18619	(67.5%)	11.50	3.02	0.25	0.67	0.18	68.7	33.2	22.8
0.1	10872	(39.4%)	9.42	2.00	0.36	0.75	0.28	74.1	35.9	26.6
0.2	6590	(23.9%)	8.85	1.67	0.45	0.80	0.37	76.8	40.5	31.1
0.3	3934	(14.3%)	8.36	1.45	0.53	0.84	0.45	78.9	45.6	35.9
0.4	2195	(8.0%)	7.85	1.33	0.61	0.87	0.53	79.6	49.3	39.3
0.5	1074	(3.9%)	7.33	1.22	0.68	0.90	0.61	79.7	54.3	43.3
0.6	483	(1.7%)	6.89	1.13	0.76	0.92	0.70	77.6	61.1	47.4
0.7	204	(0.7%)	6.98	1.08	0.82	0.94	0.77	78.9	72.7	57.4
0.8	64	(0.2%)	7.06	1.05	0.88	0.95	0.84	85.9	80.0	68.8

meet metric $M3$. In our setting, with a median daily surgical patient census of 190 patients in the test period, we believe that a reasonably sized list of actionable patients would account for 3-5% of patients (approximately 5-10 patients); therefore, $J_{\min} = 0.5$ would be a reasonable choice.

Overall, 67.5% of all patients are feasible. We believe that infeasible patients are likely those at an ambiguous stage in their care pathway, with some of the barriers they will encounter during their hospitalization not even been triggered yet (e.g., placing a request for a post-hospital facility). The reason we used the *trigger-resolution* framework to represent barriers was that introduction (trigger) of a barrier in itself will likely reduce the discharge prediction score, but it will also make possible a resolution whose contribution—if it occurs—to the prediction score will be higher than the score reduction caused by the triggering of the barrier. Indeed, if we compare feasible and infeasible patients, feasible patients tend to have more open barriers: overall an average of 11.5 barriers versus 7.0, difference of 4.5, 95% CI [4.4,4.6]. This difference is partly attributable to feasible patients having lower baseline discharge scores, therefore we also looked at Yes/Maybe/No patients separately as well:

- No patients: feasible patients have 3.0 more barriers on average relative to infeasible patients (12.3 versus 9.3, 95% CI [2.8,3.3]).
- Maybe patients: feasible patients have 1.1 more barriers on average (8.2 versus 7.1, 95% CI [1.0,1.3]).
- Yes patients: feasible patients have 1.4 more barriers on average (7.1 versus 5.7, 95% CI [1.1,1.6]).

To conclude the quantitative discussion of the model performance, it is important to compare the values of the objective functions and their respective metrics. Per Table 3, we have the following:

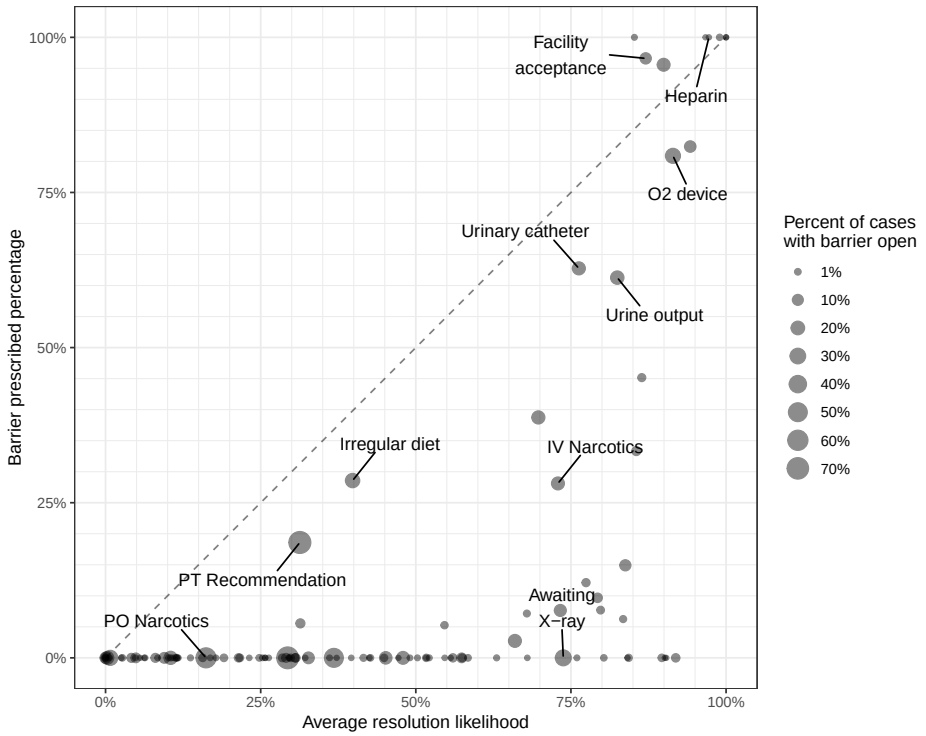
1. The value J_2^* corresponds to a barrier resolution likelihood and is therefore intended as a proxy of metric $M1$. We see that on average, J_2^* consistently overestimates $M1$ across most thresholds. The cause of this is likely multifactorial in nature. Firstly, note that J_2^* is defined as a product of individual likelihoods, while $M1$ captures a **joint** resolution percent; however, barrier resolution is likely **not** independent across barriers, and therefore the product of individual estimates performs poorly. Interestingly, if we restrict our attention to actionable patients with a single prescribed barrier, the average value of J_2^* still overestimates $M1$. This suggests that, even though the resolution likelihood models were calibrated on a barrier-by-barrier basis, the optimization model has induced a form of selection bias such that the selected barrier has an overly optimistic resolution estimate. Nevertheless, it is important to recall that Table 3 is only for *actionable* patients; therefore, patients with barriers with all low resolution likelihoods (and hence are infeasible) are not reflected here. Indeed, for lower values of $J_{\min}$, J_2^* is a well-calibrated proxy for $M1$.
2. The average score J_1^*, in contrast, underestimates the 24-hour discharge percentage $M2$ for small values of $J_{\min}$ and overestimates $M2$ for larger values of $J_{\min}$, and the latter discrepancy is to a larger degree than the one noted for $M1$ (hence $M3$ is also overestimated). In Section 4, we extend our model to better calibrate estimates of $M3$ using a robust optimization approach, particularly for higher values of $J_{\min}$ that are of practical interest.

3.3.2. Barrier prescription behavior: To facilitate a qualitative understanding of the BtD algorithm, we also considered which barriers were prescribed by the model, especially in relation to the prevalence of that barrier and the average resolution likelihood of that barrier. The results of this analysis are shown in Figure 3 for the subset of patients who are actionable at $J_{\min} = 0.5$ (the barriers in Table 2 are labeled for context).

First of all, we note that barriers with low average resolution likelihood (below approximately 35%) are rarely prescribed as MIN. (An example of such a barrier is "PO Narcotics", denoting that the patient is taking oral narcotic medication.) Further, with the exception of seven barriers, the model almost exclusively prescribes a lower percent of that barrier than the average resolution likelihood.

This figure also highlights the relationship between some of the barriers, and it also points to the observational nature of the data. For example, let us consider two barriers related to post-discharge accommodations: "Facility acceptance" (indicating that the patient is awaiting acceptance in a post-hospital facility, such as a skilled nursing facility, to which they will be transferred upon discharge) and "PT Recommendation" (indicating that the patient is still awaiting the physical therapy team to provide a recommendation on the appropriate post-discharge location for the

Figure 3 **Average barrier resolution likelihood and prescription percentage among patients who are actionable above $J_{\min} = 0.5$. Each point is a barrier and its size corresponds to the percent of patients with that barrier triggered. All percents are relative to this population only (with resolution/prescription percents relative to the subset of those with such a barrier). Barriers which are never open in this population are not shown. Abbreviations: PO="medication taken by mouth"; IV="intravenous"; PT="physical therapy"; O2="oxygen."**



patient). Among actionable patients above $J_{\min} = 0.5$, Facility Acceptance is the fifth most highly prescribed barrier among all 149 (on an absolute basis, prescribed to 10.6% of *all* actionable patients above $J_{\min} = 0.5$), highlighting the critical importance of post-discharge facilities for a nontrivial number of patients (for those with the barrier, it is prescribed 96.6% of the time).

In contrast, in this same group the PT Recommendation barrier is prescribed 18.6% of the time it is open for a patient. We believe this discrepancy is likely due to the fact that the PT Recommendation barrier is not being reliably documented as resolved by discharge (as, in reality, the patient likely did receive such a team's input before discharge), hence lower resolution likelihoods are estimated and, as a result, it is less likely to be prescribed. Overall, this highlights the inherent challenge with using EHR data, where documentation patterns and behavior evolve over time. Lastly, we believe that Figure 3 could also serve as the basis for further discussion with clinical teams on the ground to identify sources of the discrepancies between resolution versus prescription

percentages. For example, imaging studies (such as magnetic resonance imaging, or MRI) are often identified by care teams as a barrier to patient progression (Safavi et al. 2022). While our data confirms that imaging studies are often barriers, they are rarely prescribed (given low estimated resolution likelihoods), further highlighting the importance of incorporating resolution likelihood into the modeling framework.

We also evaluated the pairwise prescription behavior of the BtD algorithm by calculating the percent of patients prescribed barrier b' whenever that patient is prescribed b (i.e., the conditional prescription behavior given b). The corresponding results for barriers prescribed in at least 1% of test set observations (sorted by absolute prescription frequency) are shown in Figure SM9. We focus on interpreting two specific barriers which highlight the overall conditional prescription behavior: awaiting facility acceptance and low urine output. When the facility acceptance barrier is prescribed, it is often the only barrier. In contrast, when the low urine output barrier is prescribed, the urinary catheter barrier (indicating the patient has a urinary catheter in place) is often also prescribed. This model behavior is precisely what is desired, reflecting that the discharge prediction model realistically reflects interactions of different barriers, and these relationships are captured in the prescribed MIN set as well.

Finally, we conclude this qualitative assessment by noting that we made the deliberate choice to include both clinical and non-clinical barriers in our model. Examples of clinical barriers include things like "O2 Device" (indicating a patient needs oxygen support), while "Facility Acceptance" or "Awaiting MRI [Magnetic Resonance Imaging]" would be non-clinical, resource-focused barriers. This choice was made early in the model development stages, and the primary impetus for that was because, even if specific interventions might be targeted at non-clinical barriers (for example, by facilitating additional MRI resources on a given day), the context of what clinical barriers remain for a patient is critically important. For example, if a patient is prescribed both a clinical and a non-clinical barrier, and their non-clinical barrier is resolved but the clinical barrier remains, the patient is still not necessarily likely to be discharged. Therefore, we believe the approach we have taken to identify (possibly multiple) barriers is critical to highlight for potential users how their interventions might help a patient progress toward discharge.

3.3.3. Is the prescribed barrier the most likely to be resolved? To conclude our discussion, we also checked whether the prescribed barriers were the ones with the highest resolution likelihood for that patient. Overall, 82.3% of actionable patients (above threshold 0.5) have a single prescribed barrier; among these patients, 89.4% are prescribed the barrier with the highest resolution likelihood (9.4% the second most likely, and 1.2% the third or lower barrier by likelihood). This suggests that it is necessary to consider both barrier resolution and its effect on discharge (as

opposed to resolution alone) in any heuristic approach to solving BtD. (We return to this in our computational discussion in Section 3.5). Lastly, we note that if we instead consider all feasible patients (i.e., $J_{\min} = 0$), then the percent with a single barrier is 37.6% (and among these, the prescribed barrier is the most likely 73.9% of the time).

3.4. Contribution of Personalization

To better understand the value of personalization (based on individual resolution likelihood models as developed in Section 2.3), we evaluated its performance relative to a simpler approach. Namely, we replaced the personalized likelihoods with the baseline (aggregate) resolution likelihood for each barrier (based on the percent resolved by discharge in the training sample); as such, for a given barrier, we gave every observation with that barrier open the same estimated resolution likelihood. Using these, we reran the BtD algorithm; to allow a fair comparison, we selected that value of $J_{\min}^{\text{baseline}}$ so that the number of actionable patients at this threshold was the same as the choice of $J_{\min}^{\text{personal}} = 0.5$ in the original, personalized approach, i.e.,1073 patient-day observations. (The selected value was $J_{\min}^{\text{baseline}} = 0.41022$.)

Table 4 Comparison (%) of the performance of BtD using personalized versus baseline likelihoods.

Metric	Personalized	Baseline	p value
($M1$): All prescribed barriers resolved	79.7%	70.0%	< 0.001
($M2$): Discharged on the same day as prescribed barrier resolution, among those with criterion $M1$ met	54.3%	55.7%	0.581
($M3$): Both $M1$ and $M2$ met	43.3%	39.0%	0.048
($M2_{\text{next}}$): Discharged on the same day or next day as prescribed barrier resolution, among those with criterion $M1$ met	82.4%	82.3%	1.0
($M3_{\text{next}}$): Both $M1$ and $M2_{\text{next}}$ met	65.6%	57.6%	< 0.001

In Table 4, we compare the two models with respect to the retrospective performance metrics used earlier. In addition, we also considered how the choice of discharge time horizon impacts the criteria above. To do this, we created relaxed versions of the criteria, labeled $M2_{\text{next}}$ and $M3_{\text{next}}$, to be evaluated over a longer horizon, namely, whether a patient is discharged on the same day *or next day* following barrier resolution. Because some barriers (like facility acceptance) are more likely to have a day lag between resolution and when discharge is feasible (e.g., given need to arrange transport to the facility), $M3_{\text{next}}$ is likely more practically relevant.

We observe that personalized BtD significantly outperforms the baseline in terms of the metrics $M1$ and $M3$, meaning a higher patients have their prescribed barriers resolved, and a higher percent have both their barriers resolved and discharge shortly following that occurrence (statistical comparison are shown per a Fisher exact test, where we treat the two sets as independent samples).

Importantly, we also see there is no appreciable difference in the likelihood of patients to be discharged within 24 hours *given that the prescribed barriers are resolved*, suggesting that the incremental improvement we see in the percent of patients meeting $M3$ (the most important criterion in this setting) is attributable to the personalization of resolution likelihood (whereas $M2$ reflects the discharge likelihood model's performance, which one would expect to be similar).

The relative improvement in $M3$ for personalization relative to baseline is 9.2% (95% CI $[0.2\%, 18.7\%]$). In comparison, the relative value of personalization is larger for $M3_{\text{next}}$ with a relative percent increase of 18.9% (95% CI $[8.7\%, 30.4\%]$). This is particularly important as the same-or-next-day approach is likely more practically reasonable.

One of the reasons we believe that personalization is important in this context is the underlying patient heterogeneity noted earlier. What is a barrier that must be resolved in certain groups might not have a similar behavior in other groups, and therefore using additional covariate information is necessary to distinguish such variation in behavior observed in clinical practice.

Other actionability thresholds: We also evaluated the criteria $M3$ and $M3_{\text{next}}$ across all possible percentages of actionable patients (the preceding results equate to 3.9% actionable). The corresponding results are shown in Figure SM8. There are several notable observations:

1. The baseline model exhibits undesirable behavior for small percentages of actionability whereby the performance is *worse* when considering small subsets of patients. As having a small actionable percentage is precisely the regime of practical interest, this reinforces the importance of using the personalized approach.
2. For larger percentages of actionable patients, the personalized approach is indistinguishable from the baseline in terms of the criterion $M3$. However, there are actionable percentages where the baseline model is not feasible, in contrast to the personalized model, suggesting that the baseline approach is overly restrictive when using aggregated likelihood values.
3. For the same-or-next-day criterion $M3_{\text{next}}$, the personalized approach is uniformly superior to the baseline. Given the points raised about this metric above, the personalized approach is clearly preferable.

We close our discussion of personalization by noting that while we have focused on the comparison of a personalized approach relative to a baseline which uses aggregate likelihoods, many approaches to barrier identification do not incorporate *any* resolution likelihood information. For example, Safavi et al. (2019) display barriers which, if resolved, increase the discharge prediction above a certain threshold (i.e., if $J_2(\mathbf{x}') - J_2(\mathbf{x}^c) > \Delta$, where $\mathbf{x}'$ reflects a barrier change from $\mathbf{x}^c$); but they do not incorporate the likelihood of resolution at all. This is a common theme across discharge prediction models (Pahlevani et al. 2024).

3.5. Computational Efficiency

We conclude our assessment of the performance of the BtD algorithm by considering its computational efficiency. In the test set, the average time to solve the BtD algorithm per patient-day observation is 0.77 seconds, or an average total of roughly two minutes total time (139.6 seconds) per day (if the BtD algorithm is run sequentially; minimum by day of 30.8 seconds, maximum 193.3). This is well within the solution time required by the application. Importantly, the computational time required for BtD algorithm does not display any scaling behavior in the number of open barriers (see Table SM4), in contrast to an enumerative approach.

Resource-constrained settings: In applied settings where maintaining a commercial mixed-integer optimization solver license is cost-prohibitive, we suspect that it would be possible to improve an enumeration approach in several ways. For example, one could prune part of the search tree based on the barrier resolution likelihoods (especially if one has a threshold $J_{\min}$ already chosen based on an offline analysis). Alternatively, it is possible that one might restrict attention to only a **fixed** small number of possible barriers which, per the discussion in Section 3.2, in our setting could reasonably be around three or four. Finally, we expect that heuristics based on estimating the value of modifying a single barrier (in calculating J_1 for all possible single-barrier resolutions) would also work reasonably well, with the main caveat being that predictive models which capture feature interactions are likely not as well approximated with such an approach. In the Supplemental Material, we include a detailed assessment of how feasible enumeration is given pre-specified choices of $J_{\min}$.

Applicability to other model classes: We end our discussion of the computational results by discussing two important points regarding the results observed in our setting. Firstly, many of the barriers in our setting have a baseline resolution likelihood well below 100%. This suggests that, by the nature of the composite objective J and the choice of some $J_{\min} > 0$ to define actionability, the number of barriers chosen for actionable patients above this threshold will necessarily have a small number of barriers selected. Therefore, we believe that in settings where the barrier resolutions (in aggregate or on an individual-observation level) are higher overall, we expect the number of barriers prescribed per the BtD approach will also be higher, in which case the use of MIO modeling machinery is likely necessary. In our hospital-based application, we expect this to be the case as data quality improves (with technological refinements over time), implying that barrier resolution is more accurately recorded in patient records. (As noted in Section 3.3, this is likely contributing to low resolution likelihoods for certain barriers.)

Finally, and most importantly, our approach can also be applied in settings where the underlying predictive model J_1 is not a feed-forward neural network. In particular, our approach requires

being able to (in a practically acceptable amount of computational time) solve the problem of maximizing J_1. Similar MIO-formulation machinery as developed by Fischetti and Jo (2018) and Anderson et al. (2020) has also been used in the settings where J_1 is from different model classes. Examples of this include the following:

- Tree-based-models (classification trees, random forest, boosted trees): see Mišić (2020) and Biggs et al. (2022).
- Logistic regression: as logistic regression is a feed-forward neural network without any hidden layers, this can be seen as a special case of our approach.
- Black-box models: there has been a variety of work in recent years on approximating black-box models, especially in the area of explainable artificial intelligence (Guidotti et al. 2018). As part of this, a variety of techniques have been developed to estimate simpler models (either tree-based or neural-network-based) given a black-box model. Any such approximations could then be used directly as inputs to our approach with J_1 (in this case, the black box) replaced by its proxy $\widetilde{J}_1$.

Note that more sophisticated model classes for J_2 (than the regularized logistic regression used here) can be easily incorporated into our approach with no increase in complexity (as these likelihoods are inputs to the BtD algorithm). At the same time, it is possible to accommodate other choices of composite J instead of the product, such as a convex combination of J_1 and J_2.

4. Patient Prioritization and Model Robustness

In this section, we turn our attention to two additional aspects related to our modeling approach: an illustrated example use case for patient prioritization of post-hospital facility acceptance; and an analysis of model robustness (and a related extension). In the Supplemental Material, we also include two additional use cases: identifying bed-days attributable to specific barriers and visualizing patient progression.

4.1. Patient prioritization—The case of post-discharge facility acceptance

One important practical application of the proposed prescriptive framework is to utilize it for informing barrier-specific interventions that require coordination of scarce resources at the hospital level. Here we focus on how the tool might be used specifically for the barrier of "awaiting post-hospital facility placement" in which the care team has identified that a patient requires a facility (such as skilled nursing facility, rehabilitation hospital, or long-term acute care) and needs to be accepted by one. Even though many hospitals are investing in partnerships with post-acute facilities (or creating their own facilities), they often lack insight into the operational workflow needed to identify which patients (across an entire hospital) should be prioritized for placement. At the same

time, this process relies on manual escalations in cases where post-hospital facility bed need is one of the barriers significantly delaying a patient's discharge. We focus on how a hospital could use our approach to identify patients to prioritize considering all patients awaiting a facility.

One way of using the BtD approach for this problem is as follows (for a given day):

1. First, we filter the patients for whom BtD prescribed the facility acceptance barrier. Let J_i^{BtD} be the score obtained from the BtD algorithm for the ith patient.
2. Next, among these patients, we run a modified version of the BtD algorithm ("BtD-NoFacility") which finds the **next best prescription** that does not contain the facility acceptance barrier. Similarly, let $J_i^{\text{NoFacility}}$ be the score obtained from the BtD-NoFacility algorithm for the ith patient.
3. Finally, we mark as "prioritized" the patients for whom the BtD score is above a threshold $\overline{J}_{\text{priority}}$, but the BtD-NoFacility score is below that threshold, i.e., $J_i^{\text{BtD}} > \overline{J}_{\text{priority}} > J_i^{\text{NoFacility}}$. The threshold $\overline{J}_{\text{priority}}$ can be adjusted (calibrated) to increase or decrease the number of patients to be prioritized.

We performed this experiment for the test sample of 153 days (dates between 5/1/2019 and 9/30/2019). There were 4523 patient-day observations where facility barrier was among the prescribed barriers (16.3% of all 27,601 test observations, or 30 patients a day). To best reflect actual use, we did this in a sequential manner (day by day), ensuring that each patient appears in the prioritized list at most once (namely, the first time a patient satisfies the prioritization rule described above). Similarly, we included a patient in the non-prioritized list only once (again, only the first time the patient does not satisfy the prioritization rule). A patient can appear in both lists. We set $\overline{J}_{\text{priority}} = 0.45$ to yield an average of one patient per day to mark as prioritized.

Table 5 Using BtD to prioritize patients for facility barrier resolution ($\overline{J}_{\text{priority}} = 0.45$). One outlier (present in both groups) with a long length of stay was excluded from calculating averages in (5) and (6).

Metric	Prioritized	Non-Prioritized
(1) Number of patients	153	804
(2) Number (%) of patients where facility barrier is resolved prior to discharge	135 (88.3%)	665 (82.7%)
(3) Number (%) of patients where the patient is discharged on the same or next day that the facility barrier is resolved, given that facility barrier is resolved	121 (89.6%)	505 (75.9%)
(4) % with facility barrier resolved and discharged on the same or next day	79.1%	62.8%
(5) Mean number of days between resolution and discharge	0.44	1.36
(6) Mean number of days between prescription and resolution	2.60	5.95
(7) Number (%) of patients with facility as the only prescribed barrier	146 (95.4%)	303 (37.7%)

Table 5 compares various metrics between prioritized and non-prioritized observations. We see that the percentage of patients for whom the facility barrier is resolved by the time of discharge is

slightly higher in the prioritized group between prioritized and non-prioritized observations (88.3% versus 82.7%, though the difference is not statistically significant). Implicitly, prioritized patients have relatively high initial BtD score ($J^{\text{BtD}} > \overline{J}_{\text{priority}}$), which could only be possible if their facility barrier had a high resolution likelihood.

On the other hand, the percentage of patients who are discharged on the same day or next day after the facility barrier is resolved (given that it is indeed resolved) is significantly higher for prioritized patients compared to non-prioritized patients (89.6% versus 75.9%). Consequently, across all patients, the percent of patients who have their facility barrier resolved and are discharged within one calendar day is 79.1% of prioritized versus 62.8% of non-prioritized. This is the primary metric we use to evaluate the performance of the proposed prioritization policy (and is the facility-specific analogue of the metric $M3_{\text{next}}$ discussed in depth earlier).

We also observe that the proximity of facility barrier resolution to discharge, which is noticeably worse for non-prioritized patients than prioritized patients (1.36 days on average for resolution to discharge versus 0.44 days), lends support to the idea that the facility barrier was not the singular barrier preventing discharge for non-prioritized patients (otherwise one would expect discharge to follow shortly after resolution). This partially addresses a consistent concern raised in discussions with clinical stakeholders, namely, that prioritization of barriers (for the stated purpose of timely discharge) which does not actually result in a patient being discharge-ready can damage trust in any prioritization process. Importantly, prioritized patients very frequently have facility acceptance as their only prescribed barrier (95.4% versus 37.7% for non-prioritized), suggesting that the prioritization criteria is identifying patients with fewer barriers, though this was not explicitly included in the logic.

We close our discussion by noting the potential benefit of this prioritization scheme. The average time from being marked as prioritized to barrier resolution (among those with the barrier resolved) is 2.60 days, equating to a total of 354 bed-days in the test period (or an average daily census of 2.3 patients). This represents the potential benefit of such an intervention if this time interval could be shortened. While the effort required to shorten this interval is highly nontrivial, we believe that it is helpful to contextualize what the possible bed-days saved if such an effort was pursued (in this case, from prioritizing an average of one patient per day).

4.2. Model robustness

To assess the robustness of the proposed model, we performed a variety of sensitivity analyses. Here we focus on the following:

1. *Alternate 48-hour model:* how do prescriptions change if a longer planning horizon is considered (48 hours versus 24 hours)?

2. *Prediction uncertainty:* how do prescriptions change with errors in the discharge model? We also formulate, solve, and empirically evaluate a robust optimization variant of the model to directly incorporate model uncertainty.

To assess changes for a given patient prescribed a set S of barriers in a nominal model M with an alternative prescribed set S' per model M', we use three different metrics:

1. *Differences in barriers:* we evaluate $|S \cup S'| - |S \cap S'|$, or the number of barriers different between S and S' (we focus on the percent of differences of 0 [i.e., equality], 1, or 2).
2. *Feasibility concordance*: whether the models M and M' are in agreement about whether a patient is feasible, i.e., $\mathbb{1}(|S| > 0) = \mathbb{1}(|S'| > 0)$.
3. *Jaccard index*: we compute the Jaccard similarity index $|S \cap S'|/|S \cup S'|$.

4.2.1. Alternative 48-hour model: To better understand the implications of the modeling choice of predicting discharge within 24 hours, we also created a prediction model for discharge within 48 hours (using the same setup as the 24-hour model, again with appropriate post-training calibration; see Table SM2). Given this new model, we solved the BtD algorithm using the 48-hour analogue of J_1. To understand how prescriptions changed, we evaluated how concordant the two models are in terms of feasibility (as defined in Section 2.4.1), how the numbers of prescribed barriers compare, and the relationship between the prescriptions in the two models.

Feasibility concordance: In terms of feasibility, the two models are in agreement 89.3% of the time (57.7% of the time the patient is feasible in both models, and 31.5% in neither); 9.7% feasible in the 24-hour model but not the 48-hour model; and 1.0% feasible in the 48-hour model but not the 24-hour model. In other words, typically barriers are *more likely* to be prescribed when assessing discharge likelihood over a *shorter* time horizon (although this is not uniformly true). Intuitively, this occurs because discharge scores are higher overall for a longer prediction horizon; therefore, higher baseline prediction scores mean that the ability to increase a score (through prescribing barriers) is smaller, and any such score increases must be balanced with the barrier resolution likelihood. Put a different way, prediction over a longer time horizon reduces the discriminative power of the BtD approach. To best appreciate this, it is easier to think about an extreme modification of the model, where we predict discharge within 100 days. In this case, all baseline discharge prediction scores would already be close to 1 and the model would rarely prescribe any barriers (unless the barrier had nearly 100% resolution likelihood).

Number of barriers prescribed: The number of barriers prescribed in the 24-hour model tends to be the same (59.9%) or larger (33.9%), while the 48-hour model prescribes more barriers in 6.2% of instances. The average number prescribed in the 24-hour model is 2.0 barriers (standard deviation 2.6), while for the 48-hour model the average is 1.6 (2.3). Among the subset of feasible patients

from each approach, the average is 3.0 barriers (2.6) in the 24-hour model versus 2.8 (2.5) in the 48-hour model. Across all patients (feasible and infeasible), the difference in number prescribed (24-hour minus 48-hour) averages 0.4 (SD 0.9; median of 0, interquartile range $[0, 1]$).

Relation between prescriptions: Given the above observations, we evaluated whether the 48-hour prescribed barriers are a subset of the 24-hour barriers. Indeed, this is the case for 87.7% of observations: 41.2% of the time, the 48-hour model is infeasible (i.e., no barriers exist which can increase $J_1 \cdot J_2$ from its initial value, hence none is prescribed) and therefore the prescribed set is tautologically a subset; and 46.4% there are at least some prescribed barriers in the 48-hour model and they are a subset of the 24-hour barriers. We also compared the prescription percentage (relative to number of times as an open barrier) between the two models; see Figure SM10. On an individual barrier level, the concordance between the prescriptions in the two models ranges from 92.6% to 100.0% (median 100.0%, IQR $[99.8\%, 100.0\%]$).

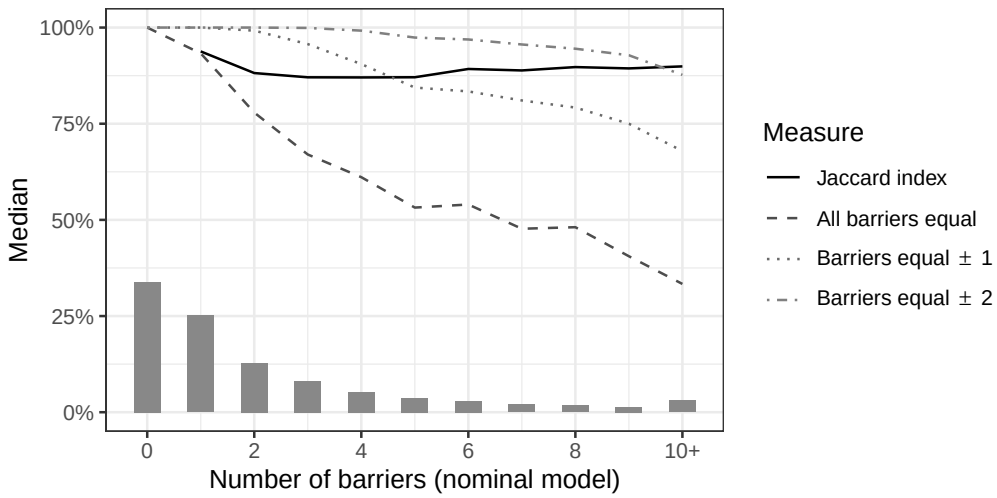
4.2.2. Changes in prescriptions given model estimation error: Our approach thus far has used the discharge prediction model (for J_1) to create point estimates of discharge likelihood changes given certain barriers are resolved. In reality, however, there is uncertainty in the estimated model, and it is possible that model uncertainty could translate into different prescriptions. Indeed, a disadvantage of most common neural networks is that, despite being efficient point predictors, they do not provide out-of-the-box output distributions that can be used to quantify prediction uncertainty. Moreover, traditional techniques for quantifying prediction uncertainty such as Bayesian optimization (Jospin et al. 2022) and variational inference (Blei et al. 2017) are not suitable to be used with neural networks because they quickly become intractable due to the large number of parameters. Gal and Ghahramani (2016) propose a tractable method, called Monte Carlo Dropout, for quantifying uncertainty in predictions of neural networks. The method relies on the use of dropout regularization (Srivastava et al. 2014).

To study how the MIN set changes with variation in the discharge prediction model, we apply this Monte Carlo Dropout technique. To proceed, first we train a neural network with a 20% dropout rate (the choice of 20% dropout rate was based on a common choice in the literature, see Gal and Ghahramani 2016, Gal et al. 2017, Verdoja and Kyrki 2020). We subsequently use this dropout-trained discharge prediction model (which we call the "nominal" model) to see how the prescriptions change when different versions of this dropout model are used. Specifically, for each observation in the test set, we perform the following 1,000 times:

1. Randomly silence 20% of the hidden nodes in the nominal model to create a new function $\widetilde{J}_1$ (in other words, we use a modified version of the nominal model to define $\widetilde{J}_1$).
2. Solve the BtD problem on $\widetilde{J} = \widetilde{J}_1 \cdot J_2$ to identify a MIN set $\widetilde{S}^*$ and the composite objective $\widetilde{J}^*$.

Then for each observation, we compute the Jaccard index relative to the nominal, non-silenced model (averaged across 1,000 runs); the percent of runs in which $\widetilde{S}^*$ equals the nominal MIN set S^*(or differs by at most 1 or 2 barriers); and the standard deviation of $\widetilde{J}^*$ as a function of the initial discharge probability.

Figure 4 **The median (across all test set observations) of Jaccard index and equality measures (computed across the 1,000 simulated dropout runs). Observations are grouped by the number of barriers prescribed in the nominal (non-silenced) dropout model. The histogram shows the number of observations for each specified number of barriers. Observations with more than 10 prescribed barriers are grouped with 10.**



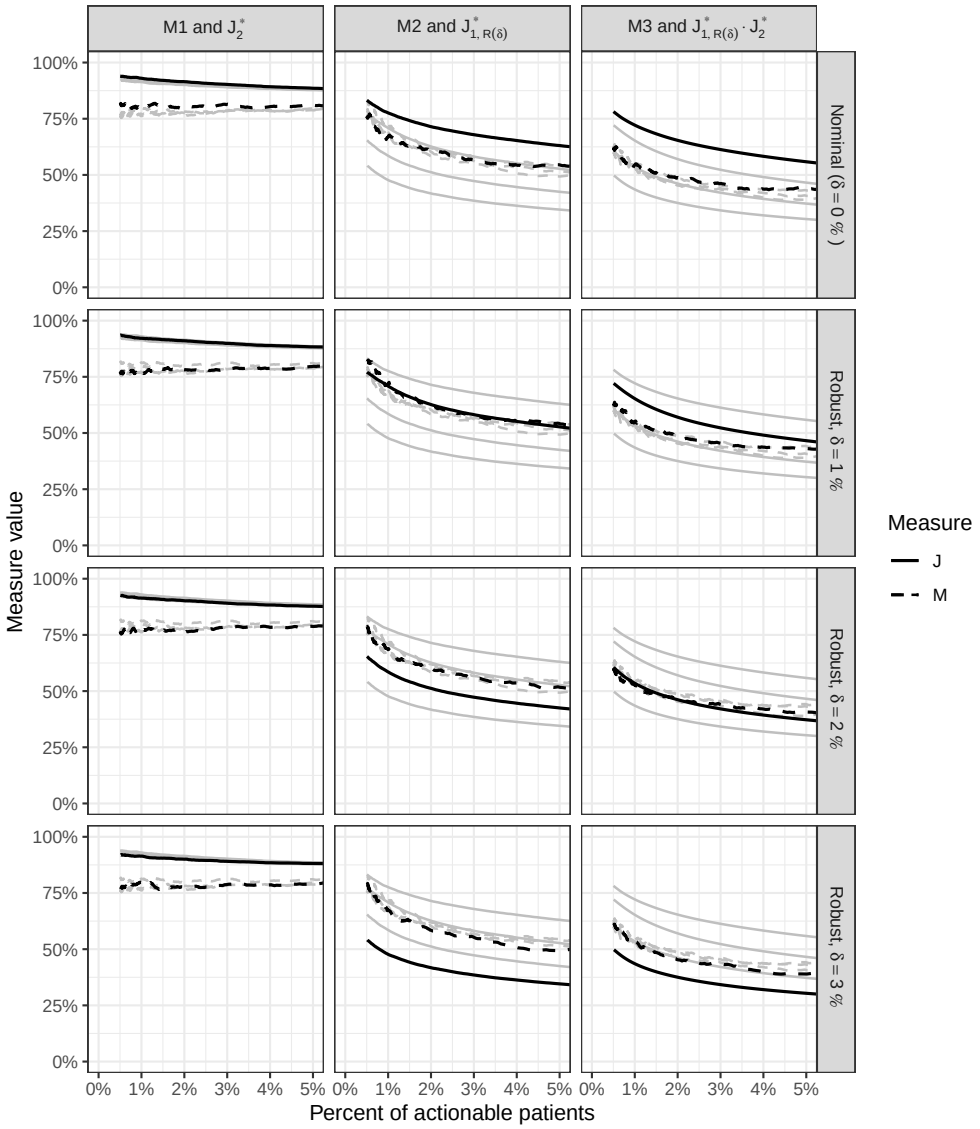
The medians (across all observations) of these measures are shown in Figure 4 (the distribution is shown in Figure SM11); the distribution of standard deviations of $\widetilde{J}^*$ is shown in Figure SM12 (the standard deviations of $\widetilde{J}_1^*$ are similar and are not shown). We observe that *for most patients* there is reasonably good agreement with the nominal model. For example, the median Jaccard index (across all patients feasible in the nominal model, i.e., prescribed at least one barrier) is 0.861 (IQR [0.801,0.960]), and it is consistent across the number of nominally prescribed barriers.

However, for some patients, the prescriptions clearly differ across the 1,000 samples, and even differences of one barrier could be practically meaningful whenever it is the one of specific interest (especially for patients with fewer barriers in the MIN set per the nominal model). This suggests it might be necessary to address model uncertainty directly.

Model extension—Worst-case dropout prediction: We proceed by doing precisely that, namely, incorporating model uncertainty for the discharge prediction into the BtD optimization framework itself. To do so, we consider a robust optimization perspective, replacing J_1 with a robust *worst-case* analogue $J_{1,R(\delta)}$ defined as the estimated discharge likelihood when a percent δ of hidden nodes are adversarially silenced ($\delta = 0\%$ corresponds to the nominal model). The corresponding

model reformulation is included in the Supplemental Material, Section SM5. As with other typical robust optimization models, its structure is comparable to the nominal model.

Figure 5 **The values of M1–M3 (percentages) and $J^*_{1,R(\delta)}$, J^*_2, and $J^*_{1,R(\delta)} \cdot J^*_2$ (averages) as a function of the percent of actionable patients for test set observations. δ denote the adversarial dropout percentage. Data is replicated across all rows in gray, with the given row shown in black to easily contextualize comparisons. Actionable percentages below 0.5% are not shown given the small sample sizes.**



To understand the performance of this alternative approach, we evaluate the same key metrics as in Table 3: above a given threshold $J_{\min}$, the percentage of patients meeting criteria M1–M3 and the corresponding values of $J^*_{1,R}$, J^*_2, and $J^*_{R(\delta)} = J^*_{1,R(\delta)} \cdot J^*_2$. Instead of comparing as a function of $J_{\min}$, we use the percent of actionable patients for a more direct comparison (as was done when

evaluating personalization in Section 3.4). Given that a percent actionable below 5% is the focus in Section 3, we restrict our attention to this range (for the full range, see Section SM5). The corresponding results are shown in Figure 5. The first observation is that the values of M1–M3, reflecting actual observed model performance, are generally consistent (though not identical) across different dropout percentages. However, the values of $J^*_{1,R(\delta)}$ are lower with increasing δ, as one would expect—the model incorporates pessimism in the estimated discharge likelihood (in contrast, J^*_2 is generally invariant). Even for a small choice of $\delta = 3\%$, the values of $J^*_{1,R(3\%)}$ are nearly half of the corresponding values of $J^*_{1,R(0\%)}$ for a given percent of actionable patients.

In terms of calibration, given a desired choice for percent of actionable patients, $\delta = 2\%$ gives excellent agreement between M3 and $J^*_{1,R(2\%)} \cdot J^*_2$, suggesting that the robust analogue of J^* can be used as a means of achieving calibration (in addition to addressing variance in the discharge prediction estimates). However, it is important to note that this calibration is not uniform across all possible choices of percent of actionable patients (see Figure SM6). Nevertheless, given that the focus in our setting is on identifying a small number of actionable patients (on the order of 3-5%), $\delta = 2\%$ appears to be a reasonable choice to achieve calibration if that is the desired objective.

Most importantly, we conclude our discussion of the worst-case dropout model by comparing how the actual *prescriptions* change as a function of δ, as opposed to the objective values alone. Summary statistics about feasibility concordance and differences in the prescribed barrier set are shown in Figure SM7. The robust models tend to have more infeasible observations primarily due to observations which were previously feasible (although, interestingly, some observations are feasible in the robust model but not in the nominal model). Importantly, the prescriptions are indeed different, even for $\delta = 1\%$. For example, 13.7% of observations have a different prescription in $J^*_{R(1\%)}$ than in J^* (due to both changes in feasibility and differences in which barriers are selected). **In other words, the robust model yields different MIN sets for some patients and does not only result in a modified objective function value.**

We also evaluated how specific barrier prescriptions changed as a function of δ by focusing on the five barriers with the largest absolute change in number of times prescribed (across all observations) for the nominal problem compared with $\delta = 10\%$ (see Figure SM7(c)). It is apparent that the robust model prescribes the "low urine output" barrier significantly fewer times (for $\delta = 10\%$, a *49.4% reduction*, from 10.4% of observations to 5.2%). Interestingly, the robust model prescribes awaiting facility acceptance to fewer patients as well. We believe that the robust dropout model is itself of independent interest and should be the focus of future study. In particular, it would be critical to study prospectively (in an actual implementation) how reliably prescriptions can be used for the robust and nominal approaches (to discern practical value of this modification). This could be done in a rapid-cycle improvement framework similar to what others have used (Horwitz et al. 2019) to guide nimble hospital operations.

5. Conclusion

As hospitals cope with increased congestion, timely discharges have become a pillar of efficient patient flow. However, central capacity planners have limited visibility into how to make targeted, small-scale interventions for patients who are most likely to get discharged within the next day. In this study, we overlay a combination of machine learning and discrete optimization models on a discharge prediction model to identify, for each patient, a subset of barriers (MIN) whose resolution are both necessary and sufficient for discharge, giving central planners clear guidance on how to best prioritize scarce resources. To the best of our knowledge, this is the first time that this modeling approach is applied in healthcare.

More formally, we formulated the problem of identifying minimal barriers as a prescriptive problem where the objective is to maximize the product of (i) the likelihood that the patient is discharged within 24 hours given that the selected barriers are resolved, and (ii) the likelihood that all the selected barriers are indeed resolved. We trained machine learning models to estimate the discharge likelihood and the likelihood of a barrier's resolution before discharge. We showed that the resulting prescriptive problem can be modeled as a mixed integer non-linear program which we then solved using the BtD algorithm.

Our computational experiments on real data corroborate (i) the utility of the proposed objective function in capturing the MIN, (ii) the effectiveness of the BtD algorithm in identifying the MIN, (iii) the significant impact of using personalized resolution likelihoods instead of baseline resolution likelihoods on the performance of the BtD algorithm, and finally (iv) the practical computational efficiency of the proposed BtD algorithm. Future work includes adding capacity constraints to our models. For example, some barriers in our model correspond to those with a "global" capacity constraint such as the number of available post-discharge facility beds or slots in an MRI machine in a given day. Adding such between-patient constraints raises a variety of important methodological questions such as scalability of various decomposition approaches. In addition, while our approach focused on using barrier resolution likelihood as the balancing measure for discharge likelihood, one could easily replace resolution likelihood with a more direct measure of "cost" required to resolve a barrier (such as relative effort required to resolve a barrier, or actual time required to resolve). We suspect that even in such cases, incorporating information about resolution likelihood (perhaps as a function of effort expended) will be crucial, and additional empirical work is needed to appropriately parametrize these more general models.

Acknowledgments

We acknowledge the constructive feedback from an anonymous associate editor and two reviewers. Copenhaver also thanks the Enterprise Research Infrastructure Services at Mass General Brigham for access to and use of the ERIS cluster computing environment without which this work would not have been possible.

References

Åhlin P, Almström P, Wänström C (2023) Solutions for improved hospital-wide patient flows–a qualitative interview study of leading healthcare providers. *BMC Health Services Research* 23(1):17.

Anderson R, Huchette J, Ma W, Tjandraatmadja C, Vielma JP (2020) Strong mixed-integer programming formulations for trained neural networks. *Mathematical Programming* 1–37.

Azari A, Janeja VP, Mohseni A (2012) Predicting hospital length of stay (phlos): A multi-tiered data mining approach. *2012 IEEE 12th International Conference on Data Mining Workshops*, 17–24 (IEEE).

Barnes S, Hamrock E, Toerper M, Siddiqui S, Levin S (2016) Real-time prediction of inpatient length of stay for discharge prioritization. *Journal of the American Medical Informatics Association* 23(e1):e2–e10.

Bertsimas D, O'Hair A, Relyea S, Silberholz J (2016) An analytics approach to designing combination chemotherapy regimens for cancer. *Management Science* 62(5):1511–1531.

Bertsimas D, Orfanoudaki A, Weiner RB (2020a) Personalized treatment for coronary artery disease patients: a machine learning approach. *Health Care Management Science* 23(4):482–506.

Bertsimas D, Pauphilet J, Stevens J, Tandon M (2020b) Predicting inpatient flow at a major hospital using interpretable analytics. *medRxiv* .

Bica I, Alaa AM, Jordon J, van der Schaar M (2020) Estimating counterfactual treatment outcomes over time through adversarially balanced representations. *arXiv preprint arXiv:2002.04083* .

Biggs M, Hariss R, Perakis G (2022) Constrained optimization of objective functions determined from random forests. *Productions and Operations Management* .

Blei DM, Kucukelbir A, McAuliffe JD (2017) Variational inference: A review for statisticians. *Journal of the American statistical Association* 112(518):859–877.

Chen L, Xu J, Lu Z (2018) Contextual combinatorial multi-armed bandits with volatile arms and submodular reward. *Advances in Neural Information Processing Systems* 31.

Chen W, Wang Y, Yuan Y, Wang Q (2016) Combinatorial multi-armed bandit and its extension to probabilistically triggered arms. *The Journal of Machine Learning Research* 17(1):1746–1778.

Conforti D, Guerriero F, Guido R, Cerinic MM, Conforti ML (2011) An optimal decision making model for supporting week hospital management. *Health Care Management Science* 14(1):74–88.

Copenhaver MS, Hu M, Levi R, Safavi K, Zenteno Langle AC (2019) Health system innovation: Analytics in action. *Operations Research & Management Science in the Age of Analytics*, 238–266 (INFORMS).

De Grood A, Blades K, Pendharkar SR (2016) A review of discharge-prediction processes in acute care hospitals. *Healthcare Policy* 12(2):105.

Fischetti M, Jo J (2018) Deep neural networks and mixed integer linear optimization. *Constraints* 23(3):296–309.

Franklin BJ, Mueller SK, Bates DW, Gandhi TK, Morris CA, Goralnick E (2022) Use of hospital capacity command centers to improve patient flow and safety: a scoping review. *Journal of Patient Safety* 18(6):e912–e921.

Franklin BJ, Yenduri R, Parekh VI, Fogerty RL, Scheulen JJ, High H, Handley K, Crow L, Goralnick E (2023) Hospital capacity command centers: a benchmarking survey on an emerging mechanism to manage patient flow. *The Joint Commission Journal on Quality and Patient Safety* 49(4):189–198.

Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* 33(1):1–22.

Gal Y, Ghahramani Z (2016) Dropout as a bayesian approximation: Representing model uncertainty in deep learning. *International Conference on Machine Learning*, 1050–1059 (PMLR).

Gal Y, Hron J, Kendall A (2017) Concrete dropout. *Advances in Neural Information Processing Systems* 30.

Gartner D, Kolisch R (2014) Scheduling the hospital-wide flow of elective patients. *European Journal of Operational Research* 233(3):689–699.

Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D (2018) A survey of methods for explaining black box models. *ACM computing surveys (CSUR)* 51(5):1–42.

Hastie T, Tibshirani R, Friedman JH (2009) *The elements of statistical learning: data mining, inference, and prediction* (Springer), 2nd edition.

Horwitz LI, Kuznetsova M, Jones SA, et al. (2019) Creating a learning health system through rapid-cycle, randomized testing. *New England Journal of Medicine* 381(12):1175–1179.

Huang Y, Li W, Macheret F, Gabriel RA, Ohno-Machado L (2020) A tutorial on calibration measurements and calibration models for clinical prediction models. *Journal of the American Medical Informatics Association* 27(4):621–633.

Jospin LV, Laga H, Boussaid F, Buntine W, Bennamoun M (2022) Hands-on bayesian neural networks—a tutorial for deep learning users. *IEEE Computational Intelligence Magazine* 17(2):29–48.

Karhade AV, Ogink P, Thio Q, Broekman M, Cha T, Gormley WB, Hershman S, Peul WC, Bono CM, Schwab JH (2018) Development of machine learning algorithms for prediction of discharge disposition after elective inpatient surgery for lumbar degenerative disc disorders. *Neurosurgical Focus* 45(5):E6.

Kim K, Lee C, O'Leary K, Rosenauer S, Mehrotra S (2014) Predicting patient volumes in hospital medicine: A comparative study of different time series forecasting methods. *Northwestern University, Illinois, USA, Scientific Report* .

Levin SR, Harley ET, Fackler JC, Lehmann CU, Custer JW, France D, Zeger SL (2012) Real-time forecasting of pediatric intensive care unit length of stay using computerized provider orders. *Critical Care Medicine* 40(11):3058–3064.

Liu V, Kipnis P, Gould MK, Escobar GJ (2010) Length of stay predictions: improvements through the use of automated laboratory and comorbidity variables. *Medical Care* 739–744.

McCoy TH, Pellegrini AM, Perlis RH (2018) Assessment of time-series machine learning methods for forecasting hospital discharge volume. *JAMA Network Open* 1(7):e184087–e184087.

Mišić VV (2020) Optimization of tree ensembles. *Operations Research* 68(5):1605–1624.

Morton A, Marzban E, Giannoulis G, Patel A, Aparasu R, Kakadiaris IA (2014) A comparison of supervised machine learning techniques for predicting short-term in-hospital length of stay among diabetic patients. *2014 13th International Conference on Machine Learning and Applications*, 428–431 (IEEE).

Na L, Carballo KV, Pauphilet J, Haddad-Sisakht A, Kombert D, Boisjoli-Langlois M, Castiglione A, Khalifa M, Hebbal P, Stein B, et al. (2023) Patient outcome predictions improve operations at a large hospital network. *arXiv preprint arXiv:2305.15629* .

Pahlevani M, Taghavi M, Vanberkel P (2024) A systematic literature review of predicting patient discharges using statistical methods and machine learning. *Health Care Management Science* 27(3):458–478.

Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, Liu PJ, Liu X, Marcus J, Sun M, et al. (2018) Scalable and accurate deep learning with electronic health records. *npj Digital Medicine* 1(1):18.

Ribeiro MT, Singh S, Guestrin C (2016) "Why should I trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.

Rubin DB (2005) Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association* 100(469):322–331.

Safavi KC, Bravard MA, Yun BJ, Levine WC, Dunn PF (2022) Design, implementation and impact of a new physician role to address capacity challenges at a large academic medical center. *Journal of Hospital Management and Health Policy* 6.

Safavi KC, Khaniyev T, Copenhaver M, Seelen M, Zenteno Langle AC, Zanger J, Daily B, Levi R, Dunn P (2019) Development and validation of a machine learning model to aid discharge processes for inpatient surgical care. *JAMA Network Open* 2(12):e1917221–e1917221.

Shalit U, Johansson FD, Sontag D (2017) Estimating individual treatment effect: generalization bounds and algorithms. *International Conference on Machine Learning*, 3076–3085 (PMLR).

Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* 15(1):1929–1958.

Steyerberg EW (2019) *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating* (Springer), 2nd edition.

Temple MW, Lehmann CU, Fabbri D (2015) Predicting discharge dates from the NICU using progress note data. *Pediatrics* 136(2):e395–e405.

Van Walraven C, Forster AJ (2017) The TEND (Tomorrow's Expected Number of Discharges) model accurately predicted the number of patients who were discharged from the hospital the next day. *Journal of Hospital Medicine* 13(3):158–163.

Verdoja F, Kyrki V (2020) Notes on the behavior of MC dropout. *arXiv preprint arXiv:2008.02627* .

Supplemental Material

This Supplemental Material contains the following additional tables, figures, and analysis for completeness:

1. features included in the data (see Table SM1);
2. the full list of 149 barriers;
3. information on performance of different classifier algorithms for discharge prediction (see Table SM2);
4. calibration curves for the discharge prediction score (see Figure SM1);
5. pairwise comparison of test set AUC, percent of observations with a given barrier, and percent resolved for that barrier;
6. details of the formulation of the problem of maximizing J_2 subject to lower bounds on J_1 as a MIP by using the approach of Anderson et al. (2020);
7. the performance of J as a function of the number of selected barriers;
8. additional data on the relative value of personalization in the BtD algorithm (Figure SM8);
9. discussion of two additional use cases for actionable barriers in identifying an estimate of the number of bed-days attributable to a specific barrier and in visualizing patient progression;
10. solve times for the BtD algorithm as a function of the number of open barriers;
11. discussion of improvements to a full enumeration approach based on various choices of $J_{\min}$;
12. the conditional prescription percentage for test set patients (Figure SM9);
13. the percent of time each barrier is prescribed (given that it is open) in the 24-hour and 48-hour models (*cf.* Section 4.2.1);
14. the distribution of metrics (Jaccard index, set agreement percentages, and standard deviation of $\widetilde{J}^*$) in the 1,000 simulated dropout models; and
15. the worst-case dropout estimation formulation and results (Section SM5).

SM1. Additional predictive modeling details

1.1. List of barriers

The list of 149 barriers (by thematic grouping) is as follows:

1. Bedside measurements: Cognition; High Pain; Level of consciousness; Orientation level; Speech
2. Bloodwork: Cryoprecipitate; Plasma; Platelets; RBC
3. Awaiting cardiac procedures: Cath; EP; Echocardiogram
4. Case management: Bed Hold; Discharge Services; High Risk
5. Consult services: Addiction Services; Allergy; Anesthesiology; Cardiology; Dentist; Dermatology; Endocrinology; Gastroenterology; General Surgery; Gerontology; Gynecology; Hematology; Infectious

Table SM1 Overview of the data sources used. "Categorical" indicates that only one of the values in a given list can be used for each patient, while several values can be used if the type is "Set."

Source	Description of data	Examples	Type
Demographics	Patient's age, sex, and marital status	60, Female, Married	Categorical
Encounter Info	Admission source for the current hospitalization	Physician referral, Outside facility	Categorical
	Patient class for the current hospitalization	Inpatient, Surgery admit	Categorical
	Surgery type(s) that the patient has had	Total knee replacement, Spinal fusion	Set
	Inpatient floor the patient is currently on	Ellison7, Lunder6	Categorical
	Surgical service currently taking care of the patient	Orthopedics, Neurosurgery	Categorical
	Primary treatment team	Team IDs	Categorical
	Attending physician	Employee IDs	Categorical
Flowsheets	Measurements of vital signs	Temperature, Blood pressure	Numeric
	Bedside nurse assessments	Pain score, fluid intake/output	Numeric
	PT/OT/SLP assessments	Assisstive device, recommendation	Set
	CM assessments	High risk assessment, Discharge barriers	Set
Labs	Lab measurements for a list of components	White blood cell, Hemoglobin	Numeric
LDAs	Placement/removal of lines/drains/airways	Urinary Catheter, Suction Drain	Set
Medications	Dose and route of the administration of medications	Narcotics, Antiemetics	Set
Orders	Placement and completion of physician consults	Physiotherapy, Cardiology, Psychiatry	Set
	Types and the current status of Imaging studies	MRI - Pending, XR - In progress	Set
	Dietary status of the patient	Regular, Sips of water, No oral intake	Set
	Placement and completion of blood work orders	Red Blood Cell, Plasma	Set
Facility referrals	Referral to and acceptance from outside facilities	Name/type of facility, status of referral	Set

Diseases; Internal Medicine; Interventional Radiology; Nephrology; Neurology; Neurosurgery; Nutrition Services; Oncology; Ophthalmology; Oral Maxillofacial Surgery; Orthopedic Surgery; Otolaryngology; Pain Management; Palliative Care; Physical Medicine Rehab; Plastic Surgery; Podiatry; Psychiatry; Pulmonology; Pyschology; Rheumatology; Smoking Cessation Program; Social Work; Surgery: Burn; Surgery: Neuro Spine; Surgery: Ortho Hand; Surgery: Ortho Spine; Surgery: Trauma and Acute Care; Thoracic Surgery; Urology; Vascular Surgery

6. Intake/output: Irregular diet; No oral intake; Emesis; Not passing flatus; Low Urine output
7. Post-acute placement: await facility acceptance; await home services acceptance
8. Imaging: CT; MRI; Other; US; XR
9. Labs: CRE - high; CRE - increased; GLU - high; GLU - low; HGB - decreased; HGB - low; INR - high; K - high; K - low; Lactate - high; NA - high; NA - low; Troponin - increased; WBC - high
10. Lines, drains, and airways: BiliaryDrain; ChestTube; EFMS; EpiduralCatheter; GiTube; NegPressureWoundTherapy; OpenSurgicalWound; PICC; PenroseDrain; SuctionDrain; SuprapubicCatheter; UrinaryCatheter
11. Medications: Antiemetic; Antipsychotic; Colonoscopy Prep; Diuretic; Heparin; IV Antibiotics; IV Fluid; IV Narcotic; Nebulized Meds; Nitroglycerin; PO Narcotic; Phenobarbital; Steroids; TPN; Tamsulosin; Tube feeds; Warfarin
12. Microbiology pending: Anaerobic; Blood; CSF; Fluid (not CSF); Fluid Culture Smear (not CSF); Fungal; Mycobacteria; Stool; Tissue; Urine; Wound
13. Occupational therapy: Final dispo recommendation

Table SM2 Performances of different classifier algorithms on the test sample

Prediction horizon	Classifier	AUC	Recall	Precision
24 hours	Classification Tree	0.811	0.545	0.521
	Logistic Regression	0.862	0.620	0.581
	Neural Network	**0.877**	**0.619**	**0.611**
48 hours	Neural network	0.871	0.740	0.727

14. Physical therapy: Activity; Final dispo recommendation; Sit to Stand Level of Assistance; Soft restraint L wrist; Soft restraint R wrist; Stand to Sit Device Used; Stand to Sit Level of Assistance; Supine to Sit Level of Assistance; Surface to Surface Device Used; Ambulation Assistance
15. Speech language pathology: Clinical Swallow Pharyngeal Phase; Clinical Swallow Response; Diet recommendation; Final dispo recommendation; Medication recommendation; Swallowing recommendation
16. Vital signs: BP; Cardiac rhythm; O2 Device; Pulse; Respiration; SpO2; Stool occurrence; Temperature

1.2. Out-of-sample prediction performance

For precision and recall calculations in Table SM2, we used the *unbiased* threshold score above which a patient is predicted to be discharged and below which is predicted not to be discharged. We defined the unbiased threshold as that which would equally balance model recall and precision in the training set. Moreover, when this threshold is used to convert prediction scores to binary classifications, the number of predicted discharges in a given subset is expected to be equal to the number of actual discharges in that subset, hence the name "unbiased threshold." For the 48-hour model used in the main text (Section 4), we only evaluated the neural network model.

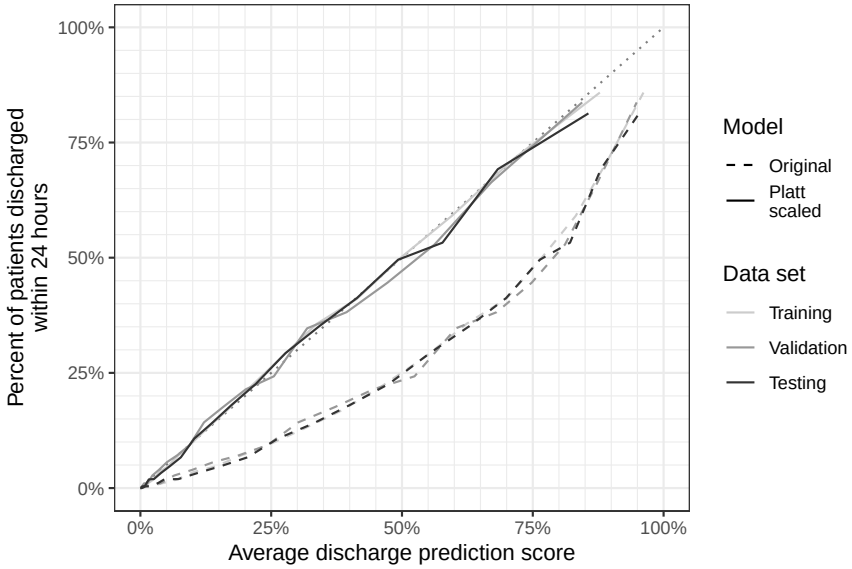
1.3. Model calibration

Calibration curves for the 24-hour discharge prediction model are shown in Figure SM1. The prediction model consistently overestimates discharge likelihood across all values of the scores. Therefore, we applied Platt scaling (estimated based on the training sample) to the final discharge predictions to obtain well-calibrated discharge probability estimates (p-value for Spiegelhalter test of 0.082, 0.004, and 0.228 on the training, validation, and test sets, respectively). The original Brier scores were 0.139, 0.143, and 0.145; and the final Brier scores were 0.109, 0.111, and 0.113.

1.4. Resolution likelihood model performance

In Figure SM2, we show the performance of the resolution likelihood models on the test set for the subset of 116 barriers for which personalized likelihoods are used (the remaining 33 use the baseline resolution percentage from the combined training and validation sets per the validation procedure described in the main text). The performance measure shown is AUC, and we consider the relationship between AUC and the underlying percent of observations with each barrier as well as the percent of observations with the barrier resolved (among those with that barrier). Each

Figure SM1 **Calibration curves for the discharge prediction scores on the three data sets. Predictions grouped into 20 equally sized groups for each of the data sets. Two models are shown: the original, and the version after Platt scaling based on the training set. A dotted line on the diagonal (perfect calibration) is included for reference.**



point corresponds to a barrier. Note that five barriers have test AUC strictly below 0.5 (although all training and validation AUCs were strictly larger than 0.5, by design; these three barriers correspond to those triggered in fewer than 2% of observations). In general, we see that there is a significant positive correlation between AUC and the percent of observations, as noted in the main text.

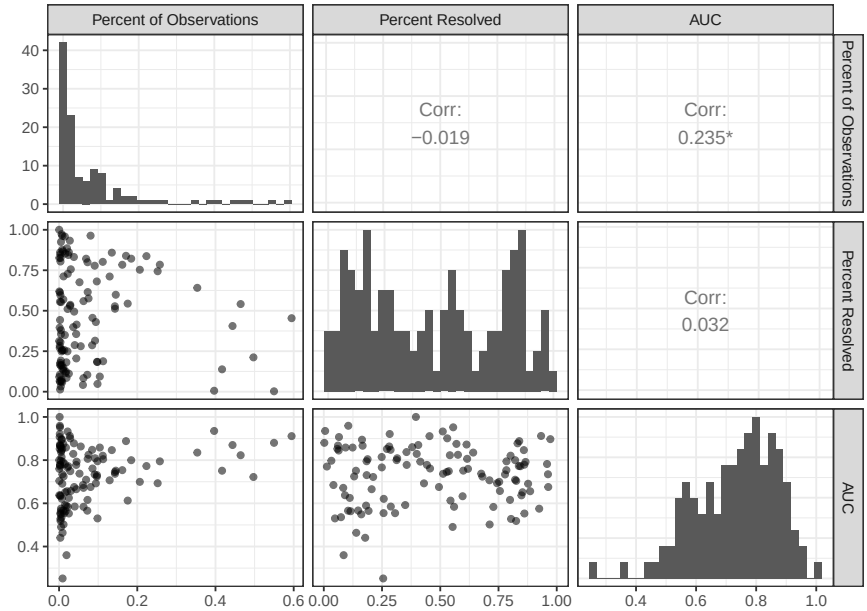
1.5. Performance of J as function of number of selected barriers

In Section 4, we assessed the performance of J for a randomly selected subset of up to five barriers (choosing from one through five with equal probability) to reflect a variety of scenarios including a single barrier being selected up to five barriers. (If there were fewer than five barriers, we sampled at most the number of open barriers.) To better understand how the performance of J varies as a function of the number of barriers, we also consider when k open barriers are selected, for $k \in \{1, 2, \ldots, 10\}$. In particular, for each k, we consider the subset of test set observations with at least k open barriers; for each observation, we randomly select a subset of k of those barriers, and then compute the corresponding value of J_1, J_2, and J, along with the label as defined in Section 3.

For each choice of k, we compare several metrics:

- Brier score (relative to the labeled outcome);
- AUC for J (relative to the labeled outcome);

Figure SM2 **Resolution likelihood models—Pairwise comparison of test set AUC, percent of observations with a given barrier, and percent resolved for that barrier. The diagonal shows histograms for each of those three measures; the upper diagonal displays correlation coefficients; and the lower diagonal shows scatter plots.**



- AUC for J_1 (relative to the discharged-same-day outcome, only on the subset of patients with all selected barriers resolved); and
- AUC for J_2 (relative to all-barriers-resolved outcome).

The corresponding results are shown in Figure SM3. Note that the Brier score for the labeled outcome (whether patient has all selected barriers resolved *and* is discharged within one day of all such barriers being resolved) decreases as the number of barriers increases. At the same time, the overall AUC for J generally increases, going from 0.799 for $k = 1$ to 0.946 for $k = 10$. Interestingly, the AUC for J_1 is generally decreasing, while the AUC for J_2 is increasing. One possible explanation for the decrease in AUC for J_1 is that with a larger number of selected barriers, more time is likely required to pass before barrier resolution (and hence the discharge model, which does not account for changes in non-barrier features, degrades in performance). Note that the increasing width of the J_1 confidence intervals is because this is calculated on an increasingly small subset of observations.

SM2. MIP formulation approach

Here we show the details of formulating the problem of maximizing J_2 subject to a lower bound on J_1 can be formulated as a linear mixed integer program (MIP), thereby making use of the formulation approach taken in Anderson et al. (2020). For expository purposes, we begin by examining the structure of a single-hidden-layer neural network to illustrate how the prediction score is calculated (the discharge prediction model as used in the main text has this structure).

Figure SM3 **Brier score and AUC metrics for composite objective as a function of the number of barriers selected (95% CIs are shown for each metric). The Brier score shown is for J (not J_1 nor J_2).**

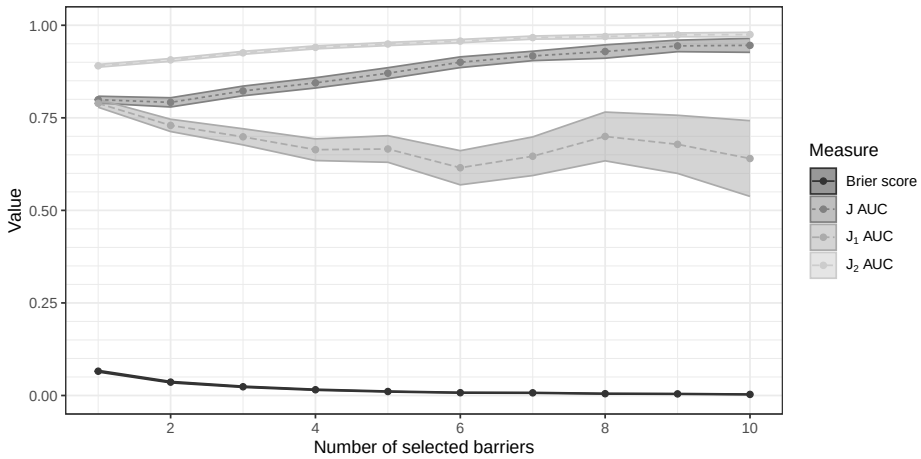
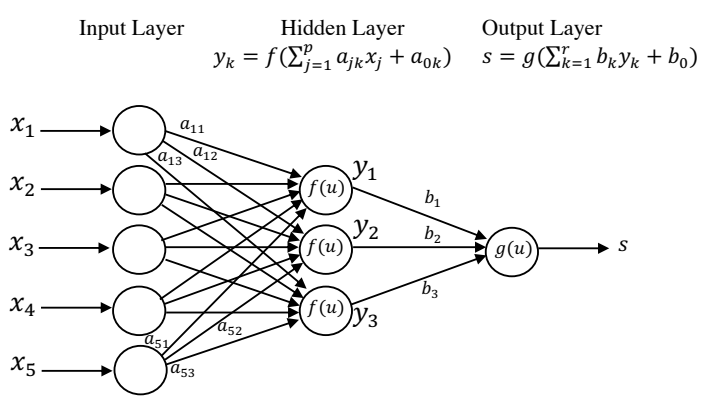


Figure SM4 shows an example of a (feed-forward) neural network with 5 input nodes, 3 hidden nodes and a single output node. In this mock example, $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5)$ is the input feature vector. The values a_{j1}, a_{j2}, a_{j3} are the coefficients of the jth feature in the first, second, and third hidden nodes, respectively. Moreover, a_{01}, a_{02}, a_{03} are the biases added to the the first, second, and third hidden nodes, respectively. The outputs of hidden nodes (y_1, y_2, y_3) are calculated as follows: $y_k = f\left(a_{0k} + \sum_{j=1}^{5} a_{jk}x_j\right) = \max\left\{0, a_{0k} + \sum_{j=1}^{5} a_{jk}x_j\right\}$. Similarly, b_1, b_2, b_3 are the coefficients of the first, second, and third hidden nodes in the output node, and b_0 is the bias added to the output node. The output prediction score z is calculated as follows: $z = g\left(b_0 + \sum_{k=1}^{3} b_k y_k\right) = \left(1 + \exp\left(-b_0 + \sum_{k=1}^{3} b_k y_k\right)\right)^{-1}$.

Figure SM4 **Network structure and parameters for an example feedforward neural network classifier**



To formalize the structure illustrated in Figure SM4 and adapt it to our problem setting, we partition the feature space into 4 components as described in Section 2 in the main text: $\mathbf{x} =$

$(x_1, x_2, ..., x_p) = (\mathbf{s}; d; \mathbf{t}; \mathbf{r}) = (s_1, s_2, ..., s_{p_s}; d; t_1, t_2, ..., t_{p_b}; r_1, r_2, ..., r_{p_b})$. In other words, we define 4 sets of indices $\mathcal{S} = \{1, 2, ..., p_s\}, \mathcal{D} = \{p_s + 1\}, \mathcal{T} = \{p_s + 2, ..., p_s + 1 + p_b\}$, and $\mathcal{R} = \{p_s + 1 + p_b + 1, ..., p_s + 1 + 2p_b\}$ that represent the set of indices of the static, days since admission, trigger and resolution features, respectively, in our feature space. In Section 3.1 in the main text, the trained neural network model had $p = p_s + 1 + 2p_b$ input nodes, r hidden nodes (on a single layer), and a single output node to predict the discharge likelihood within 24 hours for a given feature vector $\mathbf{x}^c = (x_1^c, x_2^c, \ldots, x_p^c)$ of size p. The following are parameters of the prediction model:

$$a_{0k},\ \forall k \in [r] : \text{Bias added to the } k^{th} \text{ hidden node}$$

$$a_{jk},\ \forall j \in [p] \text{ and } \forall k \in [r] : \text{Coefficient of the } j^{th} \text{ feature in the } k^{th} \text{ hidden node}$$

$$b_0 : \text{Bias added to the output node}$$

$$b_k,\ \forall k \in [r] : \text{Coefficient of the } k^{th} \text{ hidden node in the output node}$$

$$f(u) = \max\{0, u\} : \text{ReLU activation function for the hidden nodes}$$

$$g(u) = (1 + \exp(-u))^{-1} : \text{Logistic activation function for the output node}$$

Next, we introduce three sets of decision variables to express the problem as a MIP:

$$x_j,\ \forall j \in \mathcal{R} : \text{indicating the new value of the (resolution) feature } j$$

$$y_k,\ \forall k \in [r] : \text{indicating the output of the } k^{th} \text{ hidden node}$$

$$u_k,\ \forall k \in [r] : \text{indicating whether the } k^{th} \text{ hidden node is active, i.e., has positive output}$$

For a given feature vector and parameters, the output y_k of the kth hidden layer can be written as follows:

$$y_k = \max\{0, a_{0k} + a_{1k}x_1^c + a_{2k}x_2^c + ... + a_{pk}x_p^c\},\ \forall k \in [r] \tag{1}$$

which can be expressed linearly with the following sets of constraints and auxiliary binary variables u_k.

$$y_k \geq a_{0k} + a_{1k}x_1^c + a_{2k}x_2^c + ... + a_{pk}x_p^c,\ \forall k \in [r] \tag{2a}$$

$$y_k \leq a_{0k} + a_{1k}x_1^c + a_{2k}x_2^c + ... + a_{pk}x_p^c + M(1 - u_k),\ \forall k \in [r] \tag{2b}$$

$$y_k \leq Mu_k,\ \forall k \in [r] \tag{2c}$$

$$y_k \geq 0, u_k \in \{0, 1\},\ \forall k \in [r] \tag{2d}$$

(Here, M is an appropriately calibrated constant.) Similarly, the prediction score z, which is the output of the output layer, can be written as

$$z = \frac{1}{1 + \exp(-(b_0 + b_1y_1 + b_2y_2 + ... + b_ry_r))}$$

Note that if we would like this modified score to be greater than a pre-determined threshold z_0, then this can be written as a linear constraint:

$$b_1 y_1 + b_2 y_2 + ... + b_r y_r \geq \log\left(\frac{z_0}{1-z_0}\right) - b_0.$$

Finally, we incorporate the restrictions on the allowed changes to the feature vector. Specifically,

- $x_j \in \{0,1\}, \forall j \in \mathcal{R}$: we are only allowed to change the barrier resolution features from 0 to 1
- $x_{\mathcal{R}(j)} \leq x^c_{\mathcal{T}(j)}, \forall j \in [p_b]$: only a currently open barrier can be resolved,
- $x_{\mathcal{R}(j)} \geq x^c_{\mathcal{R}(j)}, \forall j \in [p_b]$: a barrier that is already resolved cannot be reverted back to being unresolved

where $\mathcal{R}(j)$ and $\mathcal{T}(j)$ indicate the jth element of the index set $\mathcal{R}$ and $\mathcal{T}$, respectively.

Note that the current features $\mathbf{x}^c$ are fixed parameters in our subproblem. Similarly, the personalized resolution log-likelihoods $\log(L_j)$ are parameters pre-calculated from the barrier resolution prediction models for each barrier. Finally, the coefficients of the neural network model $\mathbf{a} = (a_{jk})$, $\mathbf{b} = (b_k)$ are also the fixed parameters in the subproblem. Given a threshold discharge prediction score z_0, the following is the MIP formulation for the aforementioned subproblem, hereinafter referred to as $\mathrm{SP}(z_0; \mathbf{x}^c)$:

$$
\begin{aligned}
[\mathrm{SP}(z_0;\mathbf{x}^c)] \qquad & \max \sum_{j\in[p_b]} \log(L_j) x_{\mathcal{R}(j)} && && \text{(3a)}\\
& \text{s.t. } y_k - \sum_{j\in\mathcal{R}} a_{jk} x_j \geq \sum_{j\in\mathcal{S}\cup\mathcal{D}\cup\mathcal{T}} a_{jk} x^c_j + a_{0k} && \forall k \in [r] && \text{(3b)}\\
& y_k - \sum_{j\in\mathcal{R}} a_{jk} x_j + M u_k \leq \sum_{j\in\mathcal{S}\cup\mathcal{D}\cup\mathcal{T}} a_{jk} x^c_j + a_{0k} + M && \forall k \in [r] && \text{(3c)}\\
& y_k - M u_k \leq 0 && \forall k \in [r] && \text{(3d)}\\
& \sum_{k=1}^{r} b_k y_k \geq \log\left(\frac{z_0}{1-z_0}\right) - b_0 && && \text{(3e)}\\
& x_{\mathcal{R}(j)} \leq x^c_{\mathcal{T}(j)} && \forall j \in [p_b] && \text{(3f)}\\
& x_{\mathcal{R}(j)} \geq x^c_{\mathcal{R}(j)} && \forall j \in [p_b] && \text{(3g)}\\
& y_k \geq 0 && \forall k \in [r] && \text{(3h)}\\
& u_k \in \{0,1\} && \forall k \in [r] && \text{(3i)}\\
& x_j \in \{0,1\} && \forall j \in \mathcal{R} && \text{(3j)}
\end{aligned}
$$

In the above formulation, constraints (3b)-(3d) linearize the ReLU activation function, i.e., $y_k = \max\{0, a_{0k} + \sum_{j\in\mathcal{S}\cup\mathcal{D}\cup\mathcal{T}} a_{jk} x^c_j + \sum_{j\in\mathcal{R}} a_{jk} x_j\}$. Constraint (3e) imposes the restriction that the modified score has to be above the pre-determined threshold z_0. Constraints (3f) impose the condition

that a barrier cannot be resolved if it is not triggered in the first place and (3g) impose that a barrier cannot be reverted back to being "unresolved" if it is currently resolved. Finally (3h)-(3j) define the types and bounds of the decision variables.

Note that the formulation contains $2r + p_b$ variables and $4r + 2p_b + 1$ constraints. While the formulation explicitly includes variables x_j even for barriers j that are not triggered, a simple preprocessing step can eliminate these (as they will equal zero in any feasible solution). Because of the Platt scaling used, we replace z_0 by the inverse of the appropriate Platt scaling function applied to z_0. We set $\epsilon = 0.01$ for all analyses, equating to solving for $z_0 \in \{0, 0.01, 0.02, \ldots, 0.99\}$.

SM3. Additional use cases

3.1. Attributable bed-days

Identifying, in a scalable and transparent manner, which barriers prevent patients from being discharged is a challenging problem in hospital operations, particularly given the number of barriers present and the need to attribute such delays to patient-specific needs. Here we consider how our approach provides estimates of bed-days directly attributable to a barrier, thereby highlighting the opportunity available by addressing specific barriers.

In particular, we consider the subset S (among all test observations) of those which meet the following criteria: a single barrier is prescribed, and the patient is not discharged within 24 hours. Among this subset, we define an estimate of **delays attributable to barrier b**, measured in bed-days, as

$$\widehat{\text{Attrib}}_S(b) := \sum_{\substack{(\mathbf{x}^c, \mathbf{x}^*) \in S, \\ b \text{ is resolved} \\ \text{barrier}}} \left(J_1(\mathbf{x}^*; \mathbf{x}^c) - J_1(\mathbf{x}^c; \mathbf{x}^c) \right),$$

where the observation $(\mathbf{x}^c, \mathbf{x}^*) \in S$ corresponds to an observation with feature vector $\mathbf{x}^c$ and $\mathbf{x}^*$ is the corresponding optimal solution return by the BtD algorithm. The difference $J_1(\mathbf{x}^*; \mathbf{x}^c) - J_1(\mathbf{x}^c; \mathbf{x}^c)$ accounts for the improvement in the patient's 24-hour discharge score given the resolution of the prescribed barrier.

The corresponding estimated attributions are shown in Table SM3 for all barriers with estimates above 0.1% of bed-days in the test sample. Not surprisingly, the barrier of "awaiting post-hospital facility acceptance" (indicating there are active requests to place the patient into a post-hospital location such as rehab or a skilled nursing facility) accounts for the largest attributable number of bed-days (representing 0.9% of *all* bed-days in the test set). We suspect that this is a significant underestimate, highlighting the underlying challenges in attributing delays in care to specific barriers. For that reason, we believe this area remains ripe for further analytical investigation and modeling.

Table SM3 Attribution estimates (for the test set) across all barriers with estimate at least 20 bed-days. We restrict our attention to only observations where the patient is not discharged within 24 hours.

Barrier b	**# Observations where b is the only prescribed barrier**	$\widehat{\text{Attrib}}(b)$
Awaits post-hospital facility acceptance	1717	253.9
Irregular diet	871	118.9
Awaits Physical Therapy's post-hospital recommendation	555	52.6
Chest tube	167	40.8
Urinary catheter	262	36.1
Oxygen device need	252	28.3

3.2. Understanding patient progression

Here we highlight another way to use our modeling framework, in this case considering how a patient evolves in terms of their actionability and discharge likelihood throughout their hospitalization. More precisely, we consider two dimensions: a patient's Yes/Maybe/No category; and the number of prescribed barriers (for simplicity, grouped into four categories, $\{0, 1, 2, 3+\}$, where 0 denotes the patient is not prescribed any barriers).

For every patient, we observe how the patient's status changes from day to day. For example, a patient may be hospitalized with length-of-stay one day with the following sequence of states:

$$\text{Day 0: No (2 } \textit{prescribed}) \quad \longrightarrow \quad \text{Day 1: Yes (1 } \textit{prescribed}) \quad \longrightarrow \quad \texttt{Discharged}.$$

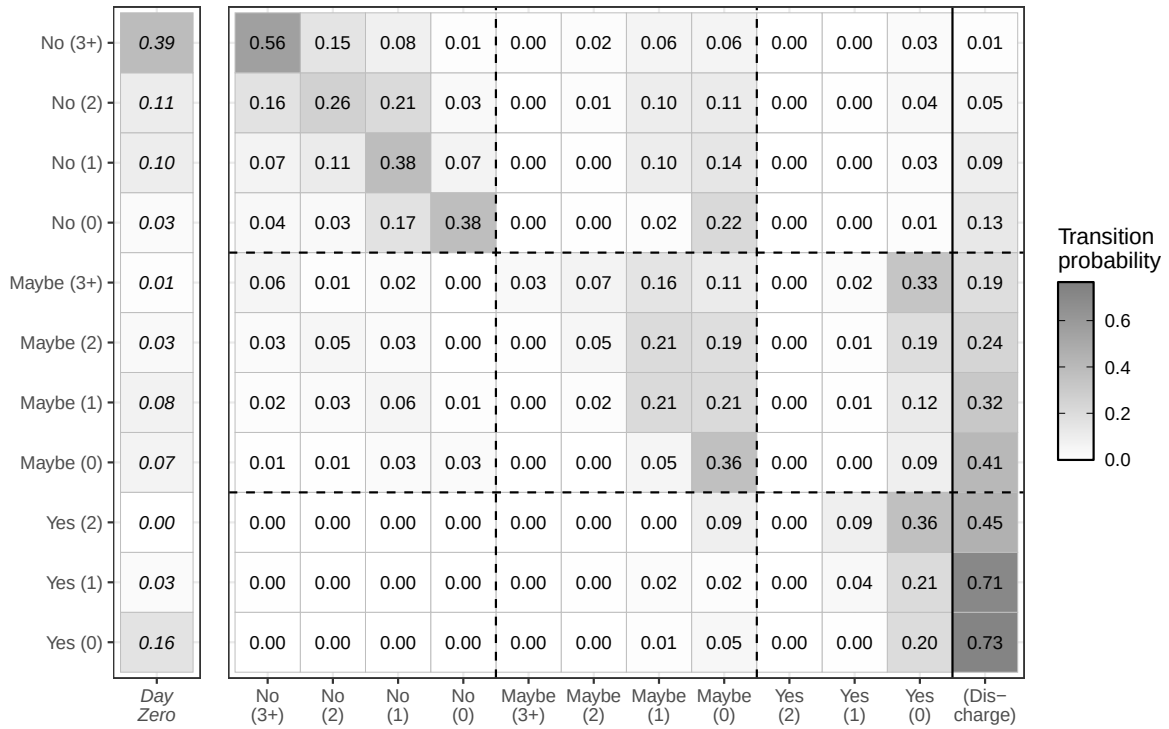
Based on the frequencies observed in the test set, we have a corresponding transition matrix with estimated probabilities of transitioning from one state to the next, shown in Figure SM5 with each row corresponding to a patient's state on day d and the columns their state on day $d+1$.

This graphical representation provides an easy way for non-technical users to understand how patients evolve with respect to prescribed barriers during their stay. For example:

- We see that Yes patients frequently discharge from their current state, with the second most frequent being a transition to "Yes (0 *prescribed*)."
- For No patients with three or more prescribed barriers, we also observe that they are most likely to stay in that state the following day (56% probability).
- For Maybe patients, the most likely subsequent state is "`Discharged`" whenever 0, 1, or 2 barriers are prescribed, whereas it is "Yes (0 *prescribed*)" when 3 or more are prescribed. Further, when Maybe patients stay in a Maybe state on the subsequent day it tends to be one with the **same number or fewer** prescribed barriers.

The initial starting frequencies also match intuition: the most frequent state on day zero is "No (3+ *prescribed*)" (at 39%), while the second most frequent is "Yes (0 *prescribed*)," highlighting the heterogeneous patient population under consideration.

Figure SM5 **Transition probabilities from various states determined by frequencies in the test set. Each row corresponds to an initial state, and the columns correspond to the patient's state on the following day. States are displayed in the format "Yes/Maybe/No group (# prescribed barriers)." Dashed lines are shown for visual reference in order to group states in Yes/Maybe/No groups. "(Discharge)" is treated as an absorbing state. The initial (day zero) state percentages are shown on the left for reference.**



SM4. Computations

4.1. Solve time for BtD algorithm

The average computation time for the BtD algorithm as applied to the test set is shown in Table SM4. We see that the average time is essentially independent of the number of open barriers (overall averaging less than 1 second), in contrast to an enumerative approach for which the computation time would scale exponentially (more on this in the following section).

4.2. Improvements to complete enumeration

Because of variation in the underlying resolution likelihood scores, it is typically not necessary to enumerate *all* subsets of a given patient's open barriers in order to compute the optimal value J^*, so long as one has chosen a value of $J_{\min}$ sufficiently large. To illustrate this point, let us consider an observation (patient-day) with b open barriers with resolution likelihoods $L_1, \ldots, L_b \in [0, 1]$. Instead of considering all 2^b possible subsets of open barriers to compute J^*, we can instead provide an upper bound on the size of subsets that need to be considered if one is only interested interested in identifying observations with $J^* > J_{\min}$, where $J_{\min}$ is fixed *a priori*. To do so, let

Table SM4 Computational performance of BtD algorithm as a function of the number of open barriers. Computation time is an average (in seconds). A row for the total is shown at the bottom.

# Open barriers	# Observations	Average time (s)
≤ 5	5004	0.85
6	2414	0.88
7	2660	0.88
8	2669	0.86
9	2432	0.83
10	2069	0.80
11	1867	0.79
12	1540	0.77
13	1223	0.74
14	1034	0.67
15	854	0.66
16	686	0.59
17	567	0.56
18	488	0.48
19	395	0.50
≥ 20	1699	0.38
(Any)	27601	0.77

$L_{(1)} \geq L_{(2)} \geq \cdots \geq L_{(b)}$ denote the sorted values of $L_1, \ldots, L_b$; then one must only consider subsets of the b barriers up to size $\bar{k}_{J_{\min}}$ defined as

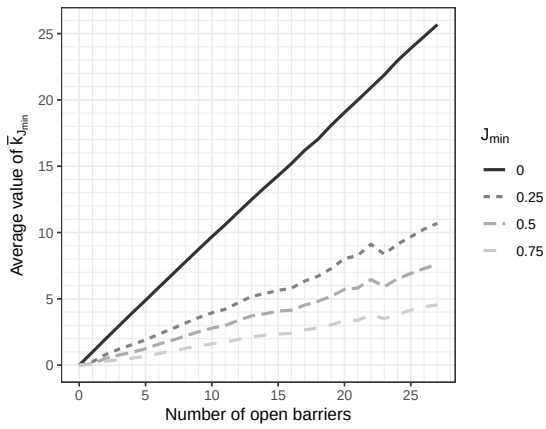
$$\bar{k}_{J_{\min}} := \max\left\{k : 0 \leq k \leq b, \prod_{i=1}^{k} L_{(i)} > J_{\min}\right\}.$$

(If $L_{(1)} \leq J_{\min}$, we define this quantity as 0.) This is a valid upper bound because any subset of $k > \bar{k}_{J_{\min}}$ barriers necessarily has $J_2 \leq J_{\min}$ and hence $J \leq J_{\min}$.

To illustrate how $\bar{k}_{J_{\min}}$ behaves as a function of the number of open barriers b, we show its average value (over the test set observations) in Figure SM4. Not surprisingly, if $J_{\min} = 0$, then $\bar{k}_0 \approx b$ whenever a patient has b open barriers (since the likelihoods are always strictly positive for all but 8 barriers; for the 8 barriers which are not triggered and resolved in at least one patient in each of the training and validation sets, we set the likelihoods to zero). For the choice of $J_{\min} = 0.5$ which is advocated for in the main text, we see that $\bar{k}_{0.5}$ is appreciably smaller than the number of open barriers on average; for example, for patients with 20 open barriers, the average value of $\bar{k}_{0.5}$ is around 4 barriers. Note that the behavior observed in this plot (and for $\bar{k}_{J_{\min}}$ more generally) is highly dependent upon the underlying resolution likelihoods. We noted earlier that the resolution likelihoods in our setting are commonly well below 1, suggesting that this simple adjustment to complete enumeration can be effective.

To further assess this claim, we also evaluated how $\bar{k}_{J_{\min}}$ evolves as a function of $J_{\min}$; the corresponding results are shown in Figure SM5. For the choice of $J_{\min} = 0.5$, we see that the average value of $\bar{k}_{0.5}$ across all test set observations is approximately 2.7 barriers, while the maximum is 16

Figure SM4 **Average value of $\bar{k}_{J_{\min}}$ for $J_{\min} \in \{0, 0.25, 0.5, .0.75\}$ as a function of the number of open barriers. Those with more than 27 barriers (0.9%) are not shown.**



barriers. If one uses this choice of $\bar{k}$ to enumerate all possible subsets of at most $\bar{k}$ barriers, then the number of subsets is on the order of $2^{20} \approx 10^6$ on average. Each subset requires computing the value of J_1 based on changes to that subset of barriers, and this is easy to do. **Therefore, an enumerative approach is feasible in our setting for the choice $J_{\min} = 0.5$.** (The important caveat, of course, is this relies on having performed a full analysis already which identified the choice of $J_{\min}$.)

SM5. Worst-case dropout prediction

We now turn our attention to a formulation of the BtD problem using the worst-case dropout prediction (main text Section 4.2.2). **We use the same notation as in the MIP formulation above in Section SM2.** The nominal value of J_1 is

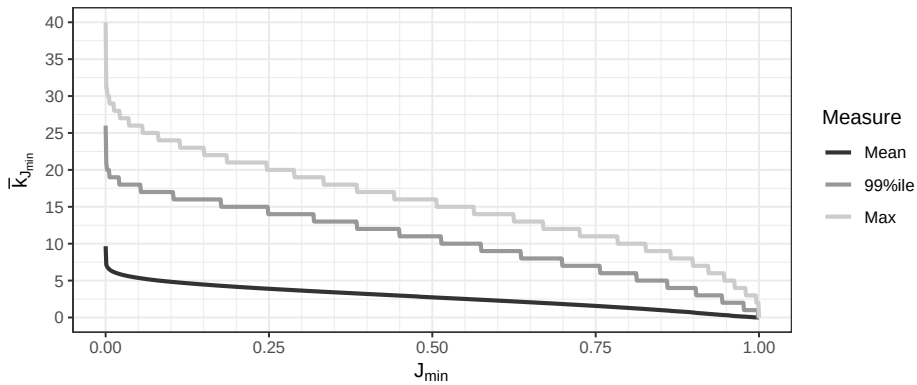
$$J_1(\mathbf{x}) = g\left(b_0 + \sum_k b_k y_k\right),$$

where g is the logistic function and $y_k = \max\left\{0, a_0 + \sum_j a_{jk} x_j\right\}$ (here we treat $\mathbf{a}$ and $\mathbf{b}$ as those from a dropout-trained model instead of the usual discharge prediction model). If a percentage $\delta \in \{0, 0.01, .0.02, \ldots, 0.99\}$ of hidden nodes are silenced (say, corresponding to indices $D \subseteq [r]$, with $|D| = \delta \cdot r$), this objective (with appropriate reweighting as is typical) becomes

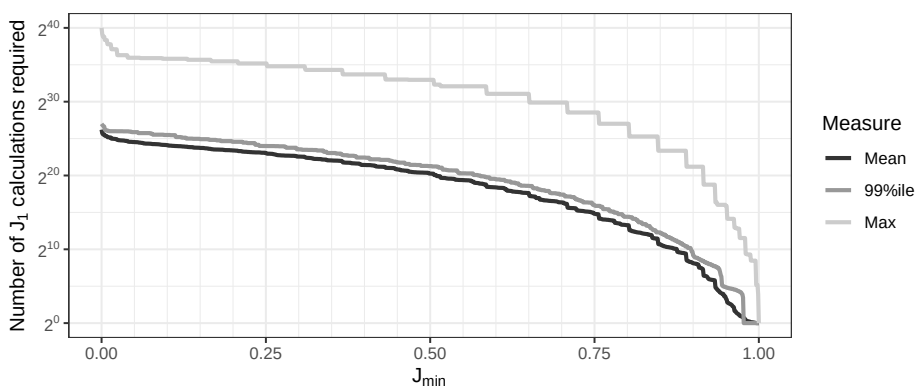
$$\widetilde{J}_1(\mathbf{x}) = g\left(b_0 + \frac{1}{1-\delta} \sum_{k \notin D} b_k y_k\right).$$

Figure SM5 **Behavior of $\bar{k}_{J_{\min}}$ as a function of $J_{\min}$ and corresponding number of computations required using this upper bound. Mean, 99th percentile (shown as "99%ile"), and max are across all test set observations.**

(a) $\bar{k}_{J_{\min}}$ as function of $J_{\min}$.



(b) Number of calculations of J_1 required (based on upper bound $\bar{k}_{J_{\min}}$); note the logarithmic scale.



Therefore, a worst-case (adversarial) silencing coincidences with finding the set of nodes that minimizes $\widetilde{J}_1$. This can be easily written as

$$J_{1,R(\delta)}(\mathbf{x}) = g\left(\min_{\substack{d:\\ \sum_k d_k=(1-\delta)r\\ d_k\in\{0,1\}}} b_0 + \frac{1}{1-\delta}\sum_k b_k y_k d_k\right).$$

Using the standard duality trick, the BtD model with J_1 replaced by its δ-worst-case analogue $J_{1,R(\delta)}$ yields the following variant of the subproblem $\mathrm{SP}(z_0;\mathbf{x}^c)$, with additions or modifications to

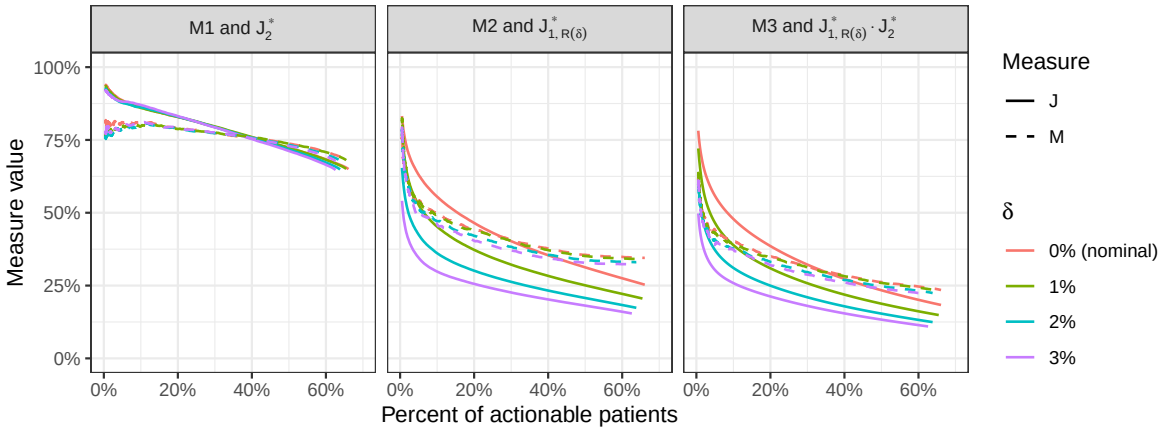
the nominal problem shown with a "$\star$" to the left of the constraint:

$$
\begin{aligned}
[\mathrm{SP}_{R(\delta)}(z_0;\mathbf{x}^c)] \quad & \max \sum_{j\in[p_b]} \log(L_j) x_{\mathcal{R}(j)} \\
& \text{s.t. } y_k - \sum_{j\in\mathcal{R}} a_{jk}x_j \geq \sum_{j\in\mathcal{S}\cup\mathcal{D}\cup\mathcal{T}} a_{jk}x_j^c + a_{0k} && \forall k\in[r] \\
& \quad y_k - \sum_{j\in\mathcal{R}} a_{jk}x_j + Mu_k \leq \sum_{j\in\mathcal{S}\cup\mathcal{D}\cup\mathcal{T}} a_{jk}x_j^c + a_{0k} + M && \forall k\in[r] \\
& \quad y_k - Mu_k \leq 0 && \forall k\in[r] \\
& \star \quad \frac{1}{1-\delta}\left[-r(1-\delta)\lambda - \sum_k \mu_k\right] \geq \log\left(\frac{z_0}{1-z_0}\right) - b_0 \\
& \star \quad b_k y_k + \lambda + \mu_k \geq 0 && \forall k\in[r] \\
& \quad x_{\mathcal{R}(j)} \leq x^c_{\mathcal{T}(j)} && \forall j\in[p_b] \\
& \quad x_{\mathcal{R}(j)} \geq x^c_{\mathcal{R}(j)} && \forall j\in[p_b] \\
& \quad y_k \geq 0 && \forall k\in[r] \\
& \quad u_k \in \{0,1\} && \forall k\in[r] \\
& \quad x_j \in \{0,1\} && \forall j\in\mathcal{R} \\
& \star \quad \lambda \in \mathbb{R} \\
& \star \quad \mu_k \geq 0 && \forall k\in[r]
\end{aligned}
$$

In other words, it has $r+1$ additional variables (λ and $\boldsymbol{\mu}$), $2r$ additional constraints, and one constraint modified. This is again a linear MIP which can be solved in the same way as the nominal problem. In the model used in the main text, $r=100$, so $\delta=0.05$ would correspond to five silenced nodes, for example.

Performance of $J^*_{R(\delta)}$ as a function of δ: In addition to the figure shown in the main text, we also compared the performance of the robust approach across different percentages of actionable patients. The corresponding results are shown in Figure SM6. We see that the agreement between M3—the percent of patients with all prescribed barriers resolved and discharged on the same day as resolution—and $J^*_{R(\delta)} = J^*_{1,R(\delta)} \cdot J^*_2$ is different across different actionable percentages, suggesting that if calibration is desired for a specific percentage, the appropriate choice of δ might differ (for example, the nominal model is already reasonably effective at 40% actionable, and $\delta = 2\%$ is effective for actionable thresholds below 5%). We also compared feasibility concordance, agreement in barrier set prescriptions, and changes in specific barrier prescriptions in the robust dropout model relative to the nominal model across thresholds $\delta \in \{0\%, 1\%, \ldots, 10\%\}$—see Figure SM7 (note that while most values plateau by $\delta = 10\%$, they do continue changing with larger choices of δ, which is not shown here for ease of interpretation). For changes in specific barriers, we restricted our

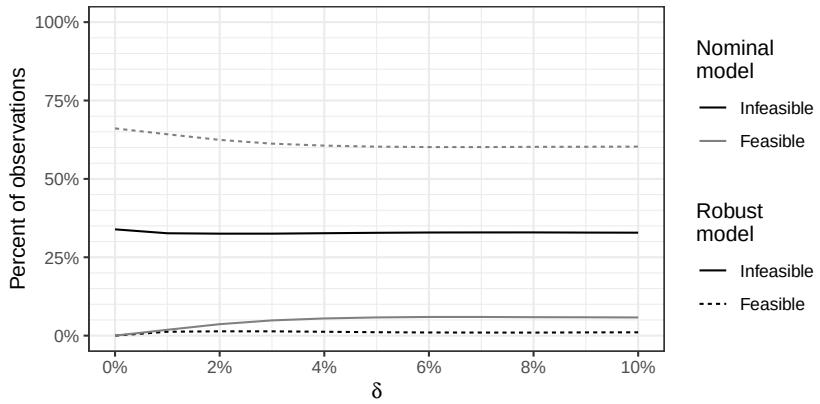
Figure SM6 **The values of M1–M3 (percentages) and $J^*_{1,R(\delta)}$, J^*_2, and $J^*_{1,R(\delta)} \cdot J^*_2$ (averages) as a function of the percent of actionable patients for test set observations. δ denote the adversarial dropout percentage. Actionable percentages below 0.5% are not shown given the small sample sizes.**



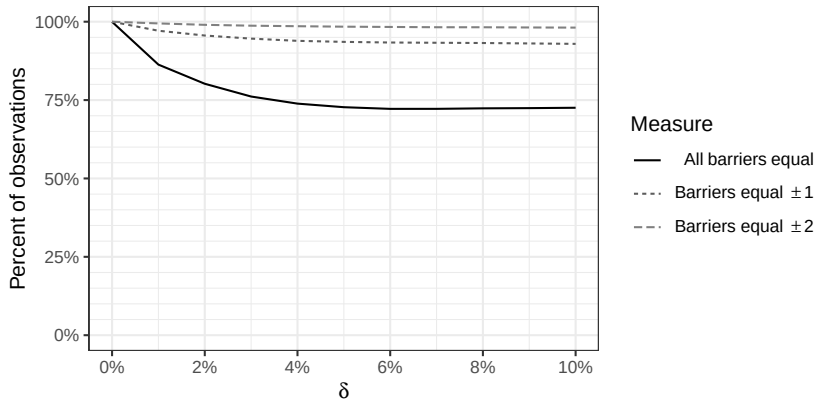
attention to the barriers with the five largest absolute changes in the number of times prescribed (and present percent relative to the number of times prescribed in the nominal).

Figure SM7 **Feasibility concordance, agreement in barrier set prescriptions, and changes in specific barrier prescriptions in the robust dropout model relative to the nominal model across thresholds $\delta \in \{0\%, 1\%, \ldots, 10\%\}$. Note that $\delta = 0\%$ is the nominal model. For changes in specific barriers, we restrict our attention to the barrier with the five largest absolute changes in the number of times prescribed (and present relative to number prescribed in the nominal).**

(a) Feasibility concordance between the models (by feasibility/infeasibility of the nominal and robust models)



(b) Percent of observations with equal prescription sets (and up to 1 or 2 differences)



(c) Percent prescribed (relative to nominal) in the five barriers with largest absolute changes in number of times prescribed

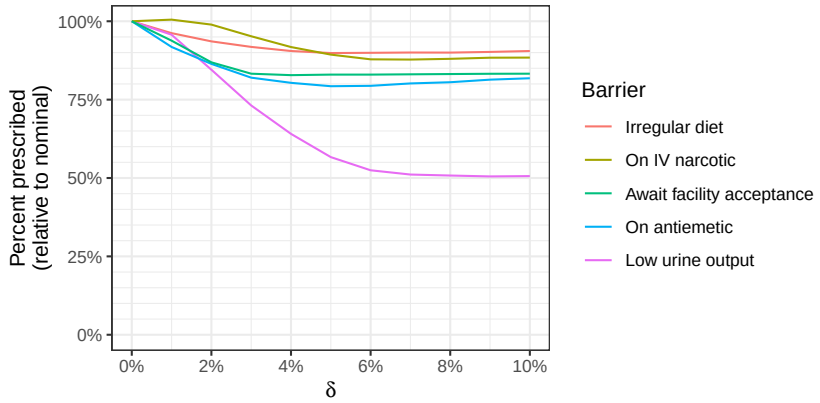


Figure SM8 **Comparison of the percent of patients meeting criteria $M3$ and $M3_{\text{next}}$ as a function of the percent of patients who are actionable from the personalized versus aggregated approaches (generalization of Table 4).**

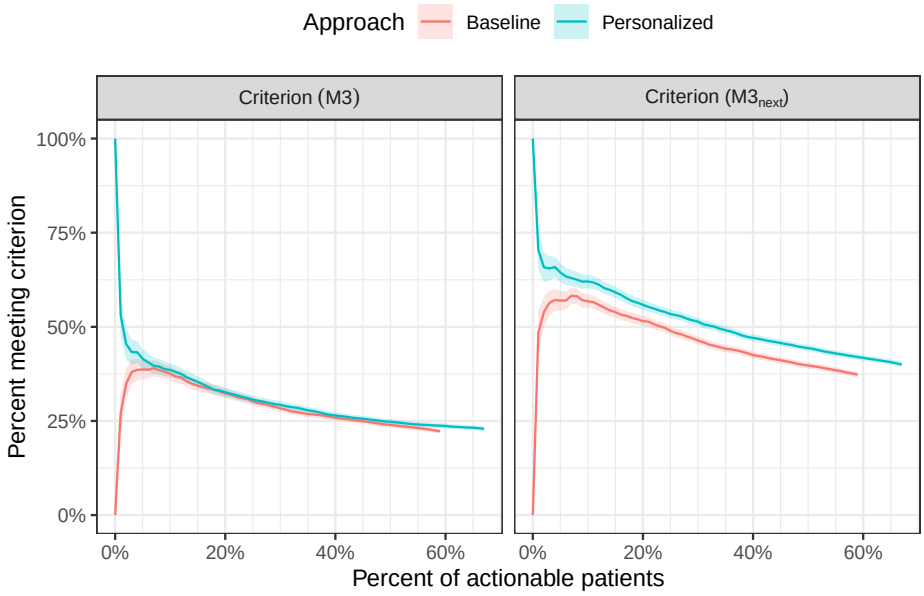


Figure SM9 **Conditional percentage of patients prescribed barrier b' (column) whenever that patient is prescribed b (row) (i.e., the conditional prescription behavior). Only barriers prescribed in at least 1% of test set observations are shown. Barriers are sorted by decreasing absolute prescription frequency (most frequent on bottom left). "LDA" = lines, drains, and airways.**

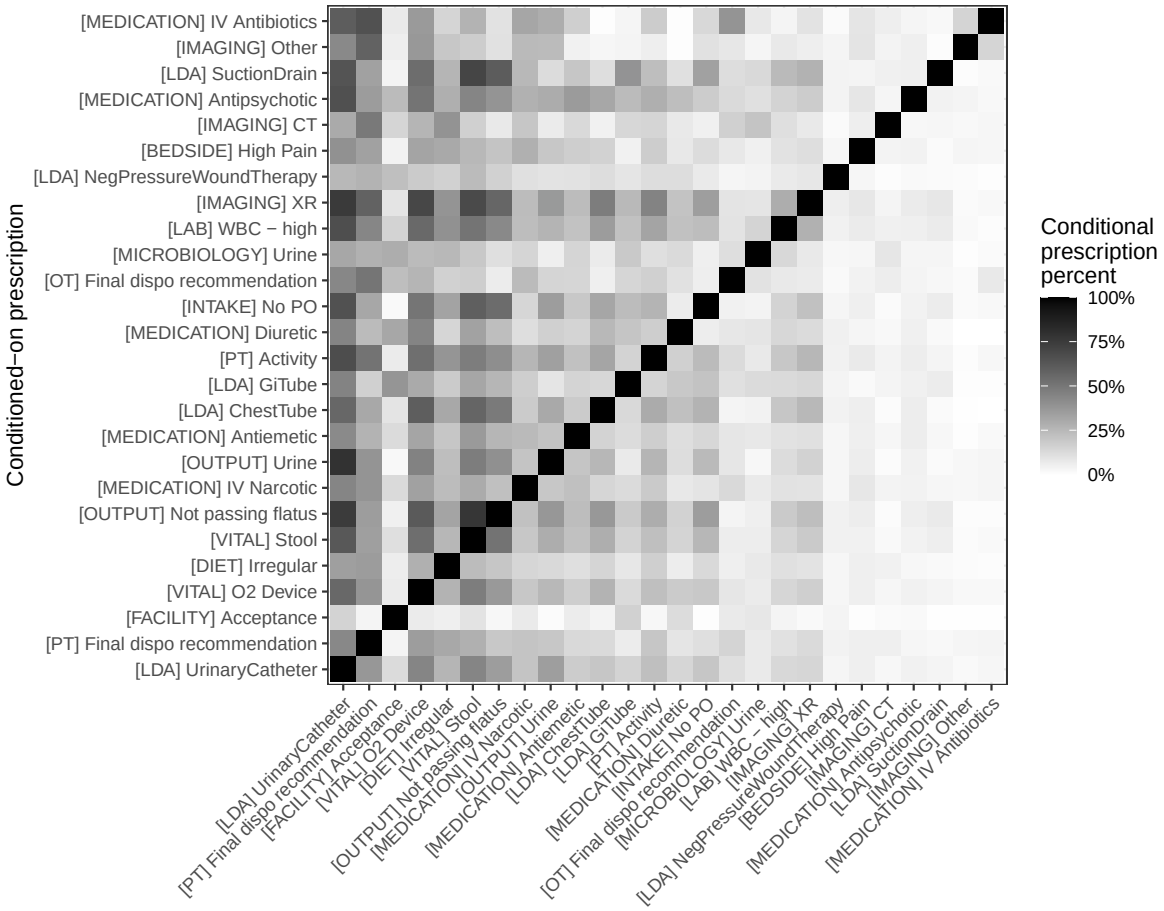


Figure SM10 **Comparison of the average barrier prescription percentage (among those with the barrier open) in the test set for the 24-hour and 48-hour models. Each point is a barrier and its size corresponds to the percent of patients with that barrier triggered. The same barriers as in the main text are labeled for reference (except "PO Narcotics", removed for legibility). Abbreviations: PO="medication taken by mouth"; IV="intravenous"; PT="physical therapy"; O2="oxygen."**

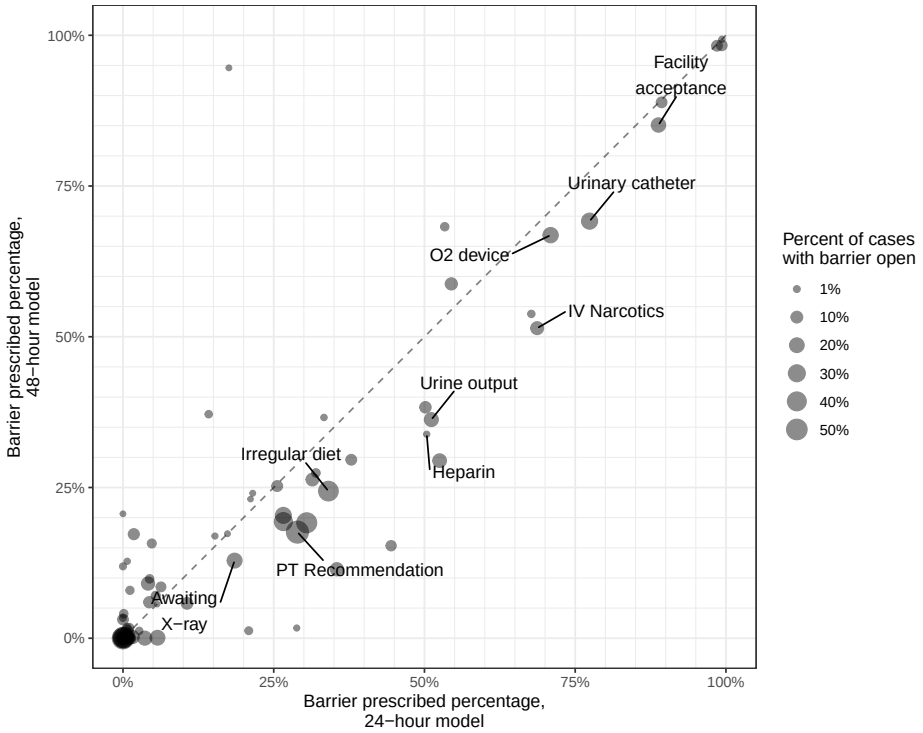


Figure SM11 **The distribution of measures (computed across the 1,000 simulated dropout runs) for all test set observations (medians reflect those reported in Figure 4). Observations are grouped by the number of barriers in the nominal (non-silenced) dropout model. The top plot shows a histogram reflecting the number of observations for each specified number of barriers. Observations with more than 10 prescribed barriers are grouped with 10. Whiskers extend to the 1st and 99th percentiles.**

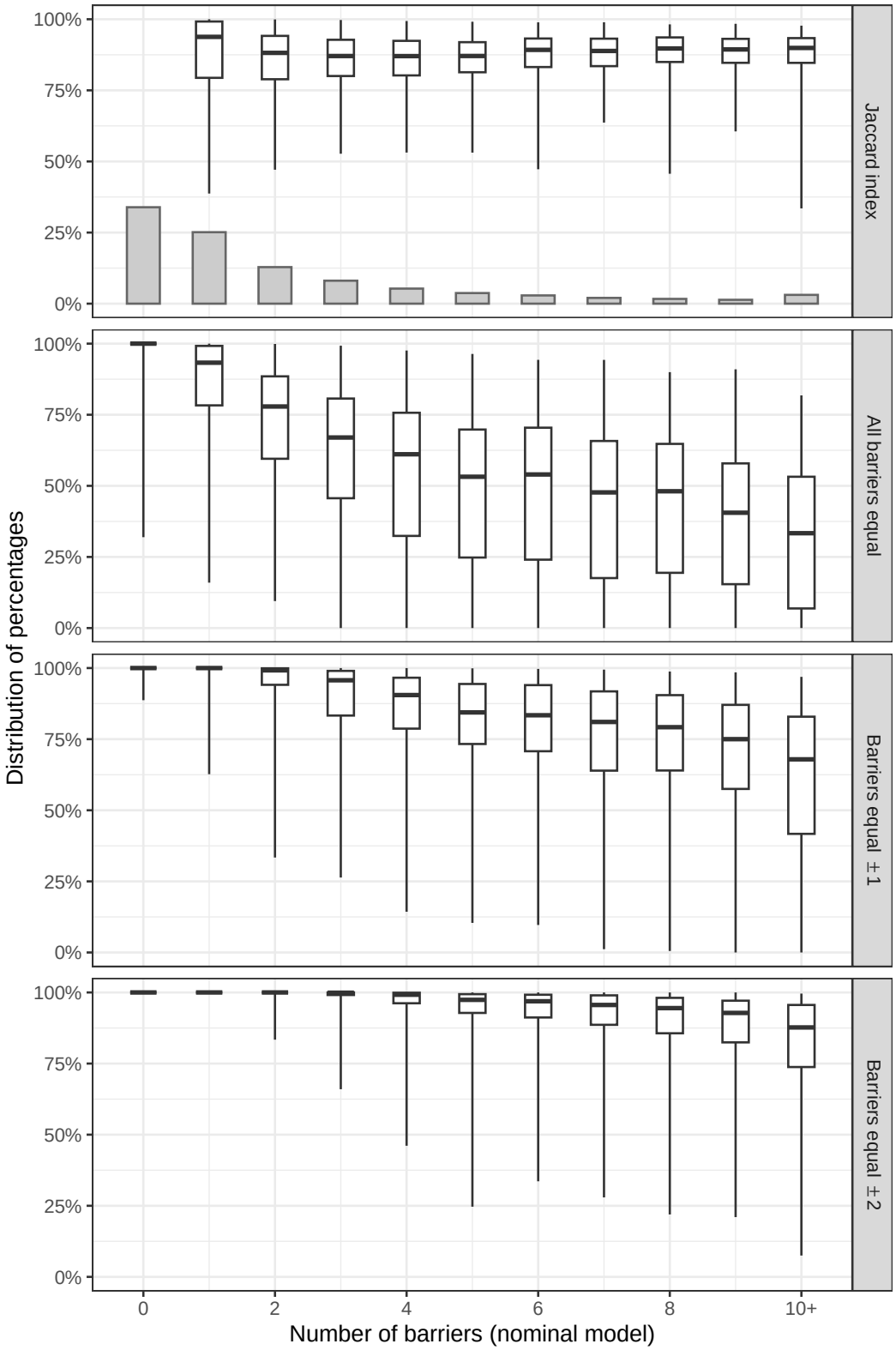
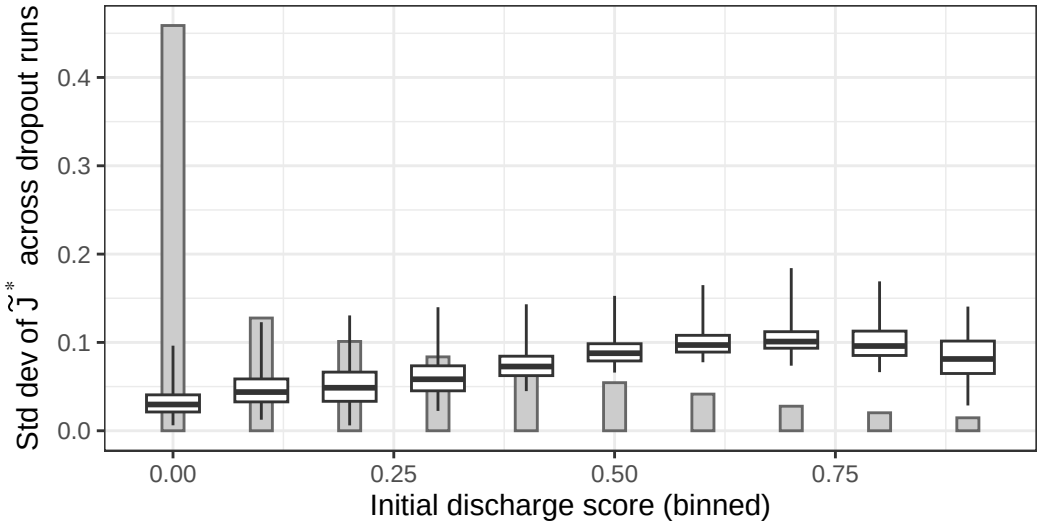



Figure SM12 **The distribution of the standard deviation of $\widetilde{J}^*$ (as measured across the 1,000 simulated dropout runs). Observations are grouped into 10 evenly-spaced groups by initial discharge score in the dropout model ($[0, 0.1)$, $[0.1, 0.2)$, and so on). The histogram reflects the number of observations for each specified number of barriers. Whiskers extend to the 1st and 99th percentiles.**