HHS Public Access

Author manuscript

IISE Trans Healthc Syst Eng. Author manuscript; available in PMC 2020 April 19.

Published in final edited form as:

IISE Trans Healthc Syst Eng. 2019 ; 9(2): 172–185. doi:10.1080/24725579.2019.1584133.

An Integrated Framework for Reducing Hospital Readmissions using Risk Trajectories Characterization and Discharge Timing Optimization

Adel Alaeddini[1], **Jonathan E Helm**[2], **Pengyi Shi**[3], **Syed Hasib Akhter Faruqui**[1]

[1]Department of Mechanical Engineering, University of Texas at San Antonio, San Antonio, TX-78249,

[2]Kelley School of Business, Indiana University, Bloomington, IN 47405

[3]Krannert School of Management, Purdue University, West Lafayette, IN 47907

Abstract

When patients leave the hospital for lower levels of care, they experience a risk of adverse events on a daily basis. The advent of value-based purchasing among other major initiatives has led to an increasing emphasis on reducing the occurrences of these post-discharge adverse events. This has spurred the development of new prediction technologies to identify which patients are at risk for an adverse event as well as actions to mitigate those risks. Those actions include pre-discharge and post-discharge interventions to reduce risk. However, traditional prediction models have been developed to support only post-discharge actions; predicting risk of adverse events at the time of discharge only. In this paper we develop an integrated framework of risk prediction and discharge optimization that supports both types of interventions: discharge timing and post-discharge monitoring. Our method combines a kernel approach for capturing the non-linear relationship between length of stay and risk of an adverse event, with a Principle Component Analysis method that makes the resulting estimation tractable. We then demonstrate how this prediction model could be used to support both types of interventions by developing a simple and easily implementable discharge timing optimization.

Keywords

Readmission Prediction; Cox Mixture Model; Kernel PCA; Discharge Decision Optimization; Expectation-Maximization Algorithm

1 Introduction

The current state of the art in readmission prediction lies in predicting readmission risk at the time of discharge. Most readmission risk prediction models predict only the cumulative 30 day readmission risk for a discharged patient (e.g., see the logistic regression model (Shulan et al. 2013), LACE and LACE+ index (van Walraven et al. 2010, 2012) in Canada,

Address(es) of author(s) should be given

PARR-30 (Billings et al. 2012) in the UK, and the HOSPITAL score (Donzé et al. 2013) in the United States. These models, however, make it difficult to accurately target interventions to prevent hospital readmissions given that the readmission could occur any time in the first 30 days after discharge. More recent risk prediction models have sought to predict the actual timing of the readmission within the 30-day window to integrate better with decision support frameworks designed for targeted, post-discharge interventions. For example, Helm et al. (2016) develops a Cox proportional hazard model to predict the time to readmission and integrates it with an optimization to target phone calls that can detect and mitigate readmission-causing conditions before the patient becomes so sick that they must be readmitted. Grzyb et al. (2017) suggest a multi-task Cox proportional hazard model to learn a shared representation across patient diagnoses to produce more accurate predictions of patients readmission times. However, the authors group patients by discharge codes, while our clustering algorithm developed in Section 2 takes a more general approach by grouping patients according to their similarity in the readmission timing. Vinzamuri and Reddy (2013) combine correlation based regularizers with Cox regression to handle correlated and grouped features which often appear in health care data. Shams et al. (2015) propose a tree-based classification method to estimate the probability of readmission that can directly incorporate a patient's history of readmission and risk factor changes over time. They validate their model using Veterans Health Administration (VA) data from inpatients hospitalized for heart failure, acute myocardial infarction, pneumonia, or chronic obstructive pulmonary disease. Hao et al. (2015) implements a random forest to predict a patient's hazard rate and time-to-readmission curves using the Maine Health Information Exchange (HIE) system via a real time provider portal.

While these models make important strides toward managing the readmission problem via post-discharge management, it has been shown that interventions during the hospital stay also impact readmission rates. For example, Anderson et al. (2012) found that patients discharged early (often evidenced by highly utilized inpatient units) exhibit increased risk of readmission, mortality, and other adverse outcomes. In contrast to early discharge, Anderson et al. (2011) hypothesizes that there may also be a phenomenon of keeping patients longer when occupancies are lower. This demonstrates the importance of discharge decision making in addition to post-discharge interventions. Bartel et al. (2014) shows that keeping a patient one extra day can reduce mortality risk by nearly 6%. Current techniques that predict readmission risk only at the time of discharge, however, are not adequate to support such a decision making framework. To target inpatient interventions, hospitals need a dynamic prediction model that updates readmission risk over the course of an inpatient stay. The medical literature also discovered similar relationship between LOS and patient outcomes; see, for example, Kuo and Goodwin (2011), Heggestad (2002). However, most of the prediction models focus on predicting the readmission probability; our method also predicts the readmission timing, which is important for hospital patient flow management.

In this paper, we first develop a readmission risk prediction model that combines both needs for a comprehensive readmission mitigation strategy: (1) the ability to predict post-discharge readmission timing, and (2) the ability to dynamically update this prediction as a patient progresses through inpatient treatment. Next, we develop a decision support tool by integrating the predictive model with the discharge optimization to balance the risks of early

discharge with hospital congestion. As a proof of concept, we demonstrate how the integrated framework could be used to determine when to discharge a patient based on hospital utilization and risk of discharge readmission. Specifically, we make the following contributions.

Prediction.

We develop a new model that extends the classical Cox proportional hazard model (COX 1972) to characterize and cluster trajectories of readmission risk as a function of discharge timing (length of stay). Our model dynamically predicts the risk of readmission *during* the patient's stay in the hospital. This prediction not only provides the readmission risk if the patient were to be discharged on the current day, but also the trajectory of risk over potential future days of the patient's hospital stay. This feature allows us to measure the benefit of keeping the patient in the hospital shorter or longer. Additionally, it clusters the patients based on a mixture model framework to better capture patient heterogeneity, allowing different shapes of the risk trajectory for different types of patients. To capture the non-linear relationship between patient's length of stay (LOS) and readmission risk, we use a kernel principle component analysis (KPCA) approach to extract non-linear features of key risk factors. Using inpatient data from a partner VA hospital, we demonstrate that the prediction power of our model outperforms existing methods. Farewell (1982) combines a logistic formulation for the probability of occurrence of an event with a parametric failure time distribution for the time of occurrence of the event. Kuk and Chen (1992) extend the model by using Cox's proportional hazards regression for the time of occurrence of the event. Rosen and Tanner (1999) present a mixture model which combines features of the Cox proportional hazards model with the mixtures-of-experts. Eng and Hanlon (2012) describe a Cox mixture model to cluster heterogeneity in time-to-event data and apply it to a cancer genomic study.

Decision Support.

We build a simple and easily implementable discharge timing optimization framework that demonstrates how our dynamic prediction model may be employed in practice to reduce inpatient readmissions while considering ward congestion. We use this framework to explore the effectiveness of discharge policies through various numerical experiments, and to generate structural insights regarding the discharge decision. Chan et al. (2012) consider the dynamic, state-dependent discharge decision in the ICU. They suggest developing a prediction model on patient outcome as an important area for future work, which is a major component in this paper. The discharge optimization we consider also broadly relates to the optimal control in service operations; see, for example, George and Harrison (2001) and Ata and Shneorson (2006). Chan et al. (2014) leverage a fluid model to determine whether a speedup service rate should be used, which is motivated from ICU management setting.

The rest of the paper is organized as follows. In Section 2, we develop the risk prediction model. In Section 3, we validate our prediction model using a data set from a partner hospital, and compare the performance of the prediction over several existing methods. In Section 4, we specify the discharge decision framework based on the prediction model and demonstrate numerical results under various parameter setting. Finally, we conclude our

paper in Section 5. The list of variables used in the proposed prediction and optimization framework is provided in Table 1.

2 Dynamic prediction of patient discharge risk

In this section, we develop a new model to predict time to readmission that extends the classical Cox proportional hazard model (COX 1972). Prior Cox-based readmission prediction models intend for LOS to be exogenous and determined at the time of discharge based on the physician's medical judgment. In contrast, our prediction model must be equipped with capabilities that can enable us to treat LOS as an endogenous variable, since LOS is controlled via the discharge decision in our optimization framework. For example, a patient who is newly admitted to the hospital should recover at a different rate than a patient that has already stayed for a number of days.

To capture the non-linear relationship between LOS and readmission risk, we extend the basic Cox model to include a non-linear feature extraction method based on KPCA (Schölkopf et al. 1997). That is, we transform a linear model into a nonlinear one by mapping patient risk factors into a higher dimensional space using a kernel function, and then use principle component analysis (PCA) to reduce the dimension of the transformed space by identifying only the most significant factors, i.e., the "representative" nonlinear features (Schölkopf and Smola 2002). Meanwhile, to capture the heterogeneity in the patient population, we personalize the Cox representation by formulating a mixture model that clusters patients into similar groups, which allows for different patient types to have different baseline hazard functions (Eng and Hanlon 2012). We develop an expectation maximization (EM) algorithm to simultaneously estimate the mixture probability parameters and the coefficients associated with the nonlinear features for each group.

2.1 Kernel principal component analysis (KPCA) for nonlinear feature extraction

KPCA has been frequently used for feature extraction and dimensionality reduction in healthcare analysis (Mikalsen et al. 2018, Motai et al. 2017, Yang et al. 2014). In the basic Cox model, the regression component is $X_i\beta = \sum_{k=1}^{D} \beta_k X_{i,k}$, where $X_i = \{X_{i,1}, \ldots, X_{i,D}\}$ are covariates directly from the data set. In our case, these covariates are patient age, gender, diagnosis, LOS, etc. The dimension D is usually large and creates computational challenges for parameter estimation. In addition, the linear form of $X_i\beta$, may not be rich enough to capture the nonlinear relationship between LOS and the cumulative readmission probability. To overcome these two issues, we apply the kernel PCA method.

To illustrate, how KPCA solves these issues consider the following. Suppose the original data contains n patients and D risk factors for each patient, so the original data space is $n \times D$. We map the original risk factor, X_i, from the D dimensional space onto a ñ dimensional space spanned by $\tilde{n}$ features $\Phi(X_i)$. Fortunately, in the KPCA method it is not necessary to directly compute Φ in the $\tilde{n}$-dimensional space. Instead, we apply the kernel trick, representing the inner product of the Φ's with a kernel function κ.

Still, this space is too large to compute the parameter estimation, especially for multiple iterations of the EM algorithm. Thus, we apply PCA to reduce the dimension of the data to

be K dimensional. PCA does this using the following linear transformation: $Z_i = A\Phi(X_i)$. In PCA, A is chosen to be the transformation that maximizes the trace of the covariance matrix of the factors in the smaller K dimensional space. This is the equivalent of finding the K largest eigenvectors, $(V_1, \ldots, V_K)$, of the covariance matrix of $\Phi(X_i)$. Since we are not working directly with Φ, it has been shown that it is suffcient to directly work on the projected feature Z_i, which can be represented as $Z_i = \phi(X_i)^T V = \sum_{j=1}^{m} a_{ji} \kappa(X_i, X_j)$ with the kernel function κ. The main steps of kernel PCA algorithm can be found at (Schölkopf and Smola 2002).

We should note that while KPCA is one of the most common feature extraction methods, one may consider other linear or nonlinear feature extraction methods suited for her/his specific dataset. In a numerical study not reported here, we have compared the performance of Gaussian Kernel PCA with that of polynomial kernel PCA and standard PCA, and found Gaussian KPCA is the most competitive feature extraction method for our dataset.

2.2 Mixture model framework to account for heterogeneity in patient population

Using $Z_i = A\Phi(X_i)$ to represent the KPCA transformation of the risk factors X_i, the time to readmission is modeled using a hazard rate function with a risk factor regression component:

$$h(t; Y_i) = h_0(t) e^{Z_i \beta}$$

where $h_0(t)$ denotes the baseline hazard function that captures the time component the readmission event, and $Z_i\beta = \sum_{k=1}^{K} \beta_k Z_{i,k}$ captures the dependence on patient-specific risk factors Z_i for any given patient i. The risk factors $Z_i = \{Z_{i,1}, \ldots, Z_{i,K}\}$ are K nonlinear features extracted from the kernel PCA method that will be specified in the next section. Note that $h_0(t)$ can follow any distribution or even be unspecified. In this study, we use a piecewise constant function to represent baseline hazard model. With the hazard rate function, we can calculate the probability density of a readmission event occurring at time t for patient i as

$$f(t; Z_i) = \left\{ h(t; Z_i) exp\left(- \int_0^t h(u; Z_i) du \right) \right\}^{\delta} \cdot \left\{ exp\left(- \int_0^t h(u; Z_i) du \right) \right\}^{1-\delta}. \quad (1)$$

Here, $\delta = 0$ if t_i, the observed readmission time for patient i, is censored or 1 if it is not. To explain (1), note that if the observed readmission time is not censored, then we use the readmission density function, $h(t; Z_i) exp\left(- \int_0^t h(u; Z_i) du \right)$; otherwise, we use the survival function (i.e. the probability that the patient has not yet been readmitted by time t), the second term of (1).

One drawback of the Cox model, including KPCA Cox, is that it assumes the dataset is homogeneous and therefore fits the same baseline hazard function to all patients, i.e. all patients have the same shape risk curve. To personalize the risk prediction, it is important to allow for different shaped curves to account for patient heterogeneity. To do so, we utilize a

mixture proportional hazard framework based on the nonlinear features extracted by the KPCA to identify clusters of patients with similar readmission risk trajectories.

Now assume that there are p different clusters of patients; we present a method for determining the appropriate number of clusters at the end of this section. For each cluster, we assume a separate Cox model with the hazard rate function defined as

$$h^j(t; Y_i) = h_0^j(t) e^{Z_i \beta^j} \tag{2}$$

for cluster j, where the baseline function $h_0^j(t)$ can be different for each cluster, and $\beta^j = \left\{\beta_1^j, \ldots, \beta_K^j\right\}$ are the regression coefficients for the j^{th} cluster. Let π_j for $j = \{1, \ldots, p\}$ be the probability that a randomly selected patient belongs to cluster j. The probability density of a readmission event occurring at time t for patient i can be represented as

$$f(t; Y_i) = \sum_{j=1}^{p} \pi_j f_j(t; Z_i), \quad \sum_{j=1}^{p} \pi_j = 1, \tag{3}$$

where $f_j(t; Z_i)$ is defined for each cluster individually as in Eq. 1, replacing $h(\cdot)$ with $h^j(\cdot)$.

We need to estimate $\Psi = (\pi_1, \ldots, \pi_{p-1}, \beta^1, \ldots, \beta^p)$, the set of mixture model parameters to be estimated. To do so, we use Expectation Maximization (EM) to maximize the complete data likelihood function (Eng and Hanlon 2012). There are two possible strategies for implementing the EM algorithm for estimating the parameters of the proposed mixture KPCA Cox model: (1) Use KPCA once to extract nonlinear features from the entire dataset, and then implement EM algorithm based on the extracted features. This strategy requires one only execution of KPCA algorithm and is computationally more efficient than the second strategy (which will be discussed next). However, this strategy may miss the features that could (locally) explain the variation in subpopulations, as it extracts the dominant features across the entire dataset; (2) Use separate KPCAs to extract nonlinear features from each of the clusters identified by the EM algorithm. This strategy requires several executions of the KPCA algorithm at each iteration and for each cluster of the mixture model characterized by the EM algorithm. Consequently, its computational complexity is significantly higher than the first strategy. Meanwhile, it can result it better predictive accuracy compared to the first strategy, as it extracts the nonlinear features locally for each cluster. In this study, we adopt the first strategy because it provided considerable computational advantage and marginal difference (with the second strategy) in terms of predictive accuracy. The major steps of the EM algorithm for estimating the parameters of the mixture KPCA Cox model is provided in Appendix 5.

Selecting the optimal number of clusters.

The appropriate number of clusters, p can be determined using cross-validation. For this purpose the dataset is partitioned into v parts, namely 5. In each iteration of the cross validation, $n - 2$ partitions are considered for training (estimating the parameters), one partition is used for validation (optimizing the (hyper) parameters), and one partition is

selected for testing (comparing with other methods). Next, the EM algorithm with a pre-specified number of clusters $p = \{1, 2, \ldots, 4\}$ is executed v times on the partitioned data to account for v possible combinations of partition assignment to train, validation, and test sets. These v prediction accuracy are calculated and averaged for each candidate choice of p, and the final p is selected to optimize the evaluation criterion. Here, we use the average prediction accuracy of the trained algorithm over the validation set at 5, 10, 20 and 30 days after discharge (we use the test set for comparison with other methods). Figure 1 shows the prediction accuracy with respect to different choice of $p = \{1, 2, \ldots, 4\}$ using a data set from a partner VA hospital, which indicates the optimal choice is $p = 3$.

3 Model validation and prediction results

Our analysis was performed based on data from a database of 2,443 patients with 3,093 admission/readmission records from a VA medical center in Michigan during 2011. The following ten factors are considered for building the predictive model: *(i) Length of stay (LOS)*, a continuous factor ranging from 0 to 30 days; *(ii) Gender*, a discrete factor with two possibilities of “male” and “female”; *(iii) Age*, a continuous factor ranging from 22 to 97; *(iv) Health insurance*, a discrete factor with two possibilities of “Insured” and “Not insured”; *(v) Eligibility*, a discrete factor with 16 levels; *(vi) Employment status*, a discrete factor with seven levels ranging from “Not employed” to “Active duty”; *(vii) Enrollment status*, a discrete factor with 9 levels ranging from “Unverified” to “Other”; *(viii) Source of admission*, a discrete factor with 3 levels including “Hospital”, “Nursing home care unit (NHCU)”, and “Domiciliary”; *(ix) Ward*, a discrete factor with 17 levels representing the 17 wards of the VA facility; *(x) Principal diagnoses*, s a discrete factor with 30 levels representing the 30 common diagnoses treated at the VA facility. Table 2 illustrate the distribution of data across diffrent risk factors. Considering numerous levels of the discrete risk factors, 80 variables are used for encoding the dataset, which signifies the need for an appropriate feature extraction method, i.e. KPCA.

3.1 Model validation

Based on the above dataset, we compare the performance of the proposed mixture KPCA cox model with a number of predictive models in the literature. We use five-fold cross validation for training, validation and testing of the comparing models. The methods used in our comparison along with their information are presented in Table 3. Among the comparing methods, the proposed mixture KPCA Cox and KPCA Cox are regression models and other methods are (binary) classification models. For the regression model we use a threshold (optimized based on cross validation) for converting the estimated risk into class (readmission vs no readmission) prediction. Meanwhile, For training the classification models, censored data are transformed into “not-readmitted” class if censoring time occurs after the target date, i.e. 30 day after discharge, or they will be eliminated if the censoring time occurs before the target day. Meanwhile, for training the regression models, censored data are directly incorporated into the models using the censoring variable. The accuracy of readmission prediction 5, 10, 20, and 30 days after the discharge date is used the evaluation criteria. The reason for selecting multiple (four) days (instead of one day) for performance comparison is that, using multiple days we can better characterize the trajectory of

readmission risk development at di erent days after discharge. For predicting readmission at different days after discharge, i.e. 5, 10, 20, and 30, the classification models need to be trained multiple times for each of the target days. Whereas, regression models only need to be trained once for the entire range of target days.

Figure 2 illustrates the accuracy of the comparing methods in predicting the probability of readmission at 5,10, 20, and 30 days after discharge based on five fold cross-validation. While, some the methods provides considerable performance at specific snap shots, i.e. Boosting outperforms other methods for 5 days readmission prediction, the proposed method illustrates the overall best and most robust performance across the prediction horizon. The considerable performance of the proposed mixture KPCA Cox model can be attributed to several factors: (1) Using days from discharge as a longitudinal variable in the model (through the proportional hazard function), (2) Utilizing mixture model to account for the heterogeneity in the readmission risk of subpopulations, (3) Employing KPCA to extract informative (nonlinear) features from the set of existing risk factors, and (4) Effectively using the information of the censored data for training the predictive model.

3.2 Prediction Result

In this study we use LOS as the discharge decision in our optimization framework to reduce the probability of readmission. Figure 3.a illustrates the Box plot of LOS for readmitted vs not-readmitted patients, 30 days after discharge, which shows not-readmitted patients in the dataset generally have a longer LOS compared to the readmitted patients. To better demonstrate the potential effect of increasing LOS on improving the probability of readmission, we use the proposed mixture KPCA Cox method to characterize the cumulative probability of readmission as a function of LOS based on a random sample of 1,032 patients from the study dataset. To implement the proposed Mixture KPCA Cox method, we use the same setting as discussed in Table 3. Figures 3.b–c illustrate the predicted trajectories for individual patients as well as the trajectories associated with the 3 clusters of the proposed mixture KPCA Cox model.

It may be worth mentioning that our approach can handle an arbitrary number of variables changing with time. However, during the inpatient stay, most of the data is static with time (e.g. admitting diagnosis, patient demographics, health history etc.). Literature has also found that LOS is an important factor in readmissions and is a variable we have access to in our dataset, hence we focus primarily on LOS. A further application of our model would be to consider data about interventions and events that occur during the patient’s stay, such as diagnostic tests ordered, additional surgeries performed, visits to ICU, etc. However, this data is not available in our current dataset and is particularly difficult to obtain in general.

4 A discharge optimization framework

In this section, we demonstrate how our prediction model can integrate seamlessly with decision support methods for reducing readmissions. Specifically, we consider discharge timing and its impact on readmissions as well as congestion in the inpatient unit. As discussed in the introduction, early discharge can reduce congestion in an overloaded hospital ward, yet it carries higher risk of readmission. In our discussions with hospital

managers, we found that the industry need involves two components: information on when readmissions will occur (prediction) and what to do with that information (optimization). With the prediction model alone, hospitals can begin taking better discharge actions. However, the question of how many patients and which patients to discharge on a given day requires complex integration of many different system components such as current and future predicted risk, current occupancy, future arrivals and future projected discharges. In this environment ad-hoc decisions may leave significant room for improvement. Thus, in this section we provide an example of such an optimization approach to showcase how our prediction model may be easily integrated with a decision support framework.

In our decision support framework, we balance the tradeoff between ward congestion (which also carries risks for patients) and readmissions by formulating an easy-to-use, static optimization problem. Solving this optimization gives us the optimal length of stay (LOS) for different types of patients, which then determines the discharge policy by patient type. A high-level map of the optimization algorithm is shown in Figure 4.

Note that this framework is just one example of how our novel prediction model could be integrated into decision models to improve hospital operations. The goal is to demonstrate the capabilities of our prediction model to support future decision support research in readmissions.

4.1 Optimization model for homogeneous patients

In this paper, we focus on a “threshold-type” discharge policy. That is, we discharge a patient when his or her 30-day readmission probability falls below a pre-set threshold s, e.g., 10%. Based on our personal communication with our partner hospital, this type of discharge policy is commonly used by physicians. Given a readmission risk trajectory, $f(l)$, that predicts patient readmission risk as a function of LOS l, the threshold s maps directly to the number of days each patient needs to spend in the hospital. For example, if the desired threshold $s = 10\%$ corresponds to $l_s = 5$ days, then this patient will be discharged after spending 5 days in the system. The goal is to choose the optimal s, or equivalently, the optimal l_s to strike a balance between system congestion and expected number of readmissions.

For the purpose of illustration, we begin with one class of patients in this section, i.e., every patient follows the same risk trajectory after being admitted. Mathematically, let X_k^j denote the number of patients who already spent j days in the hospital on day k. Then, the pre-discharge system state can be described as

$$\left(X_k^0, X_k^1, \ldots, X_k^J\right)',$$

where J is the maximum number of days a patient is allowed to stay in the hospital, e.g., $J = 30$. Since no patient spends more than l_s days in the system under our discharge policy, the number of patients in the system post discharge equals

$$X_k = \sum_{j=0}^{l_s - 1} X_k^j. \tag{4}$$

In steady state, each day the threshold discharge policy will only discharge patients who have reached the threshold LOS, l_s, on the current day. Hence, the pre-action number of patients in system in long-run equals

$$X_{k-} = \sum_{j=0}^{l_s} X_k^j. \tag{5}$$

To capture system congestion, we use the expected post-discharge queue length, defined as

$$\mathbb{E}[Q_k] = \mathbb{E}[X_k - N]^+, \tag{6}$$

where N is the number of inpatient beds available for use, and $x^+ = \max(x, 0)$ for any real number x. Note that our decision framework can be generalized to include other congestion measures; see the remark at the end of this section. Assume that the unit cost of congestion (i.e. having more patients than beds) is C, and the cost associated with each readmission event is R. We want to solve the following optimization problem:

$$l_s^* = \operatorname{argmin}_{l_s} R \cdot s\mathbb{E}[D_k] + C\mathbb{E}[Q_k], \tag{7}$$

where s is the predicted readmission probability corresponding to the decision variable l_s from the given risk trajectory $f(l)$, and $\mathbb{E}[D_k]$ is the expected number of discharges per day under the discharge policy with the readmission probability threshold being s.

To calculate the expected queue length in (6), we need to specify the distributions of X_k. To do so, note that for $0 \le j < ls$, X_k^j equals the number of patients who arrived to the system j days before, i.e.,

$$X_k^j = A_{k-j}^0, \tag{8}$$

where $A_\cdot^0$ denotes the new arrivals on a given day. Thus,

$$X_k = \sum_{j=0}^{l_s - 1} A_{k-j}^0. \tag{9}$$

If we assume patient arrivals follows a Poisson distribution with mean Λ each day, then $\sum_{j=0}^{l_s - 1} A_{k-j}^0$ follows a Poisson distribution with mean $l_s\Lambda$, and we can calculate the expected queue length as

$$\mathbb{E}\left[Q_k\right] = \mathbb{E}\left[Poiss\left(l_s\Lambda\right) - N\right]^+. \tag{10}$$

In addition, recall that in long run we discharge all patients whose length of stay is l_s on any given day. This means that these patients came l_s days ago following a Poisson distribution with rate Λ. Thus, $\mathbb{E}\left[D_k\right] = \Lambda$ is the daily discharge rate.As a result, we can further rewrite the optimization problem (7) as

$$l_s^* = argmin_{l_s} R \cdot s\Lambda + C \cdot \mathbb{E}\left[Poiss\left(l_s\Lambda\right) - N\right]^+. \tag{11}$$

Remark.

Our optimization framework is flexible to account for non-Poisson arrival process and other congestion measures besides the post-discharge queue length. For example, we could use the pre-action daily queue length

$$\mathbb{E}[Q(k-)] = \mathbb{E}\left[Poiss\left(\left(l_s + 1\right)\Lambda\right) - N\right]^+ \tag{12}$$

or even the time-dependent performance using the intra-day system dynamics, i.e.,

$$X(t) = X_k + A_{(k,t)} - D_{(k,t)},$$

with $A_{(k,t)}$ and $D_{(k,t)}$ denoting the number of arrivals and discharges that occurred between time k (midnight of day k) and t.

A numerical demonstration of solving optimization problem (11)—As a demonstration, we use the average predicted trajectory for group 2 and group 3 patients, respectively, as the readmission risk trajectory $f(l)$ (as a function of LOS l); see the red and green curves illustrated in Figure 3c. In the experiments, we set $N = 32$, $\Lambda = 6.25$, $R = 20$, and $C = 1$. Figure 5 shows the expected queue length $\mathbb{E}[Q_k]$ (equals the holding cost since $C = 1$), the expected readmission cost ($R \cdot s\Lambda$), and the total cost under discharge policies with different l_s.

For both Figures 5a and 5b, the minimum cost is achieved at $l = 4$ days. In other words, if every patient follows the same trajectory as the average curve for group 2 or group 3 and using objective in Equation (11), the optimal policy (among the threshold type policies we consider) is to discharge a patient when her LOS reaches $l_s = 4$ days.

4.2 Optimization model for multi-class model

Assume that there are M classes of patients, which we can obtain by grouping patients with similar predicted risk trajectories; see Figure 3c for an example. For each $m = 1, \ldots, M$ class, we set a threshold s_m, for the readmission probability. Equivalently, we set a LOS threshold l_m for each class. Then, on each day k, we discharge every m-class patient who has

spent l_m days in the system. Similar to the single-class model, it is not diffcult to show that the total number of patients in the system (post discharge) equals

$$X_k = \sum_{m=1}^{M} \sum_{j=0}^{l_m - 1} X_k^{m,j}, \tag{13}$$

where $X_k^{m,j}$ denotes the number of class m patients that have spent j days in the system. We can also show that X_k follows a Poisson distribution with mean

$$\sum_{m=1}^{M} l_m \Lambda_m,$$

where Λ_m is the daily arrival rate of class m patients arriving according to a Poisson distribution. Thus, the expected post-discharge queue length equals

$$\mathbb{E}[Q(k)] = \mathbb{E}\left[Poiss\left(\sum_{m=1}^{M} l_m \Lambda_m\right) - N\right]^+, \tag{14}$$

and we can formulate an optimization problem that is similar to (11), i.e.,

$$\left(l_1^*, \ldots, l_M^*\right) = \text{argmin}_{l_1, \ldots, l_M} R \sum_{m=1}^{M} s_m \Lambda_m + C\mathbb{E}[Q(k)], \tag{15}$$

where s_m is the readmission probability corresponding to l_m for class m patients, and $\mathbb{E}[Q(k)]$ is given in (14).

A numerical demonstration of solving optimization problem (15)—We consider three classes of patients as grouped in Figure 3.c. Figure 6 shows the expected queue length $\mathbb{E}[Q(k)]$, the expected number of re-admission events, and the total cost under threshold discharge policies with different l_m for each patient class. In the experiments, we set $N = 52$, the total daily arrival rate $\Lambda = 6.25$, $R = 20$, and $C = 1$. We estimate the daily arrival rate of each class from the data and get $\Lambda_m = 2.28, 2.81, 1.15$ for $m = 1, 2, 3$, respectively. We also impose an extra constraint that each patient needs to spend at least 3 days and at most 20 days in the system to ensure a reasonable readmission probability upon discharge and a reasonable cost (from the inpatient cost perspective). Solving the optimization problem gives us the optimal LOS $l_1^* = 20$, $l_2^* = 7$, and $l_3^* = 3$ days. In other words, if we have three types of patients, each of them following the trajectory as demonstrated in Figure 3c, under the objective function in Equation (15), we should keep group 1 patients as long as possible (20 days), while discharging low risk patients as earlier as possible (3 days). Note that this insight is not because group 1 patients have a high (absolute) readmission risk and group 3 patients have a low (absolute) risk; it is because the *marginal* benefit for keeping a group 1 patient is the highest. We explain this rationale in Section 4.3.

4.3 Insights for the optimized discharge threshold

In this section, we explain the insight behind the optimization problem for the single-class model; the multi-class model can be explained similarly (see the end of this section). Under the static discharge policy, each patient spends l_s days in the hospital. Thus, finding the optimal l_s is equivalent to finding the optimal *offered load*, defined as

$$B = l_s \Lambda .$$

Plugging B into Equation (11), we get the objective function $U(B)$ as

$$U(B) = R\Lambda \cdot f\left(\frac{B}{\Lambda}\right) + C \cdot \mathbb{E}[Poiss(B) - N]^+, \tag{16}$$

where $f(l)$ denotes the risk trajectory function with LOS $l = \frac{B}{\Lambda}$. When N and B are reasonably large, we can approximate $Poiss(B)$ with a Normal$(B, \sqrt{B})$ random variable (this Normal approximation has been shown to work well for other random variables as well). As a result, the mean queue length can be approximated by

$$\begin{aligned}
&\mathbb{E}[Poiss(B) - N]^+ \approx \mathbb{E}[Norm(B, \sqrt{B}) - N]^+ \\
&= \sqrt{B}[\phi(\alpha) - \alpha(1 - \Phi(\alpha))] \\
&= \sqrt{B}\phi(\alpha) - (N - B)(1 - \Phi(\alpha)),
\end{aligned}$$

where $\alpha = \frac{N - B}{\sqrt{B}}$, and $\phi(\cdot)$ and $\Phi(\cdot)$ denote the density and cumulative distribution function for a standard normal random variable. Thus, taking the derivative of $U(B)$ in 16, we get

$$\frac{dU(B)}{dB} \approx -\frac{R\Lambda\beta}{\Lambda} \cdot f'\left(\frac{B}{\Lambda}\right) + C\left(\frac{\phi(\alpha)}{2\sqrt{B}} + \sqrt{B}\phi'(\alpha)\alpha'(B) + (N - B)\phi(\alpha)\alpha'(B) + (1 - \Phi(\alpha))\right).$$

Note that

$$\alpha'(B) = -\frac{N + B}{2B\sqrt{B}} \approx -\frac{1}{\sqrt{B}}$$

when B is close to N (which is the case in most hospital units, since ward utilization is frequently above 90%), and

$$\phi'(\alpha) = -\alpha\phi(\alpha) .$$

Also, when B is large, $\phi(\alpha)/(2\sqrt{B}) \approx 0$. Thus, we can further simplify the derivative as

$$\begin{aligned}
\frac{dU(B)}{dB} &\approx -R\beta \cdot f'\left(\frac{B}{\Lambda}\right) + C\left(\alpha\phi(\alpha) - \frac{N - B}{\sqrt{B}}\phi(\alpha) + (1 - \Phi(\alpha))\right) \\
&= -R\beta \cdot f'\left(\frac{B}{\Lambda}\right) + C(1 - \Phi(\alpha)) .
\end{aligned}$$

Setting the derivative $\frac{dU(B)}{dB} = 0$, we get the (approximate) optimal B satisfying

$$R\beta \cdot f'\left(\frac{B}{\Lambda}\right) = C(1 - \Phi(\alpha)). \tag{17}$$

On the left-hand side, $\beta \cdot f'\left(\frac{B}{\Lambda}\right)$ is the marginal improvement in discharge risk for one patient when increasing l_s; on the right-hand side, $(1 - \Phi(\alpha))$ captures the marginal change in system congestion as a function of the the discharge threshold l_s. To see that latter, note that

$$1 - \Phi(\alpha) = \mathbb{P}(Z > \alpha) = \mathbb{P}\left(Z > \frac{N - B}{\sqrt{B}}\right) = \mathbb{P}(B + Z\sqrt{B} > N).$$

Examining the final term above, we see that $B + Z\sqrt{B}$ is a Normal random variable with mean B and standard deviation $\sqrt{B}$, i.e. the approximate offered load when threshold is l_s. Thus, $\mathbb{P}(B + Z\sqrt{B} > N)$ is the (approximate) probability of exceeding the hospital ward's capacity (e.g. probability of blocking incoming patients from the emergency department).

Quite intuitively, Equation (17) says that *the optimal threshold balances the risk reduction achieved by keeping a patient longer versus the subsequent increase in system congestion*. In the multiple patient class setting, the rationale is similar. That is, the optimal solution again seeks to balance the marginal risk reductions for different classes of patients with the subsequent increase in system congestion. As a simple heuristic, our model tells us that, when choosing a class of patients for early discharge, we would want to choose the one with the smallest *marginal* risk reduction, since all patients contribute to the system congestion equally, regardless of the class. This is precisely what we saw in the numerical experiments of the previous section. In the next section, we will obtain the optimal discharge policies under various system conditions, where the insights can be explained by Equation (17) and the rationale behind it.

4.4 A discharge decision support framework

To get clean structural insights, we focus on the *single-class* setting; experiments in the multi-class setting generate similar insights. First, we examine the impact of ward capacity on the optimal discharge threshold. Table 4 shows the optimal l_s^* for different values of N using the average risk trajectory for group 3 patients. As we can see from this table, the larger the ward capacity N, the larger the threshold l_s. This observation can be explained by Equation (17). As N increases, if l_s^* remains unchanged, the blocking probability on the right-hand side of (17) decreases. Thus, to reach a new balance in Equation (17), the optimal offered load, or equivalently, the optimal l_s^* needs to increase, which means we keep patients longer in the hospital. This insight indicates that larger wards may benefit by keeping patients longer, improving quality without significantly impacting overcrowding. Conversely, smaller units may wish to take a more aggressive discharge policy to avoid excessive congestion.

Second, we examine the impact of risk trajectory on the optimal discharge threshold. Table 5 shows the optimal l_s^* for different values of N, where we use either the average risk trajectory for group 2 patients or the average risk trajectory for group 3 patients. Surprisingly, although group 2 patients have higher readmission risks than group 3 patients, the optimal discharge threshold does not differ much under a given N. To explain this, we refer back to Equation (17), that is, the optimal discharge threshold level balances the *marginal reduction* in readmission risk, not the absolute value of the readmission risk, with the subsequent increase in system congestion. *Thus, our second conclusion is that, the marginal change in readmission risk matters the most in optimizing discharge thresholds*. If future prediction models are developed to increase the prediction accuracy, the focus should be on improving the prediction of the marginal change, not the absolute values.

Lastly, we examine the impact of cost parameters on the optimal discharge threshold. Keeping the unit holding cost $C = 1$, Table 6 shows the optimal l_s^* for different values of N under $R = 5$ and $R = 40$, where the average risk trajectory for group 3 patients is used (for the single patient class). Clearly, the optimal discharge threshold l_s^* under $R = 5$ is smaller than that under $R = 40$, given the same capacity N. Thus, when the readmissions are relatively cheap, the hospital should focus on reducing ward congestion. For example, readmissions for some conditions are penalized by Medicare whereas others are not. Hence, wards that serve a large population of penalized conditions would likely want to focus more on readmission risk than ward congestion than those that serve mostly non-penalized conditions.

Though the observation is intuitive, an important result is that our decision framework provides a tool to *quantify* the tradeoff between system congestion and readmission risk. Tables 4 and 6 provide a Pareto analysis for decision makers to select their desired performance in terms of ward congestion and readmission risk. *By using R and C as tuning parameters, our optimization framework provides decision makers with a simple tool to directly observe the tradeoff between system-level congestion and individual-level readmission risk and choose their desired operating regimes based on their own organizational goals*.

5 Conclusions

In this paper, we develop a new readmission prediction model that captures the endogenous impact of length of stay on readmission risk. Specifically, our new approach enables us to predict readmission risk on any day of hospital stay (and beyond the end of the stay as well), not just as a function of when the patient was actually discharged. Additional features include the ability to predict time to readmission (to enable targeted interventions), a clustering method that enables personalization of the shape of the time-to-readmission curve, and general non-linear feature extraction for risk factors. To do so, we integrate a kernel PCA method for feature extraction with a new implementation of the EM algorithm that simultaneously estimates clusters of similar patients as well as risk factor parameters. We demonstrate how this new model could be used in a discharge decision framework that balances risks of early discharge (conversely benefits of keeping a patient longer) with

hospital ward congestion. This type of decision support could not be supported with earlier attempts at readmission risk prediction, due to the need to control discharge timing. This work paves the way for future efforts in readmission reduction. Future research could include a completely personalized prediction model, where each patient has their own unique readmission risk trajectory. This would enable more targeted discharge decision making. Additionally, future prediction models could include dynamic updates based on new information obtained during the hospital stay, such as lab tests, visits to the ICU, blood transfusions, etc. Along these lines, major challenges include the amount and availability of data. Finally, more work on the decision support framework could include a dynamic discharge policy that accounts for the current state of the ward and the patients in it.

Appendix

Algorithm: Mixture Cox KPCA clustering

We define W_i, the membership random variable that indicates which cluster patient i belongs to. For patient i, we also have the KPCA transformation of the risk factors $Z_i = A\Phi(X_i)$, and the observed readmission or censoring time, t_i. Using this, we can define the complete likelihood function as:

$$L(\Psi; \{t_i\}, \{Z_i\}, \{W_i\})$$

$$= \prod_{i=1}^{n} \prod_{j=1}^{p} \left\{ \pi_j f_j(t_i; Z_i) \right\}^{\mathbb{1}\{W_i = j\}} \tag{18}$$

$$= \prod_{i=1}^{n} \prod_{j=1}^{p} \left[\pi_j \left\{ h_0^j(t_i) exp(Z_i \beta^j) exp\left(- \int_0^{t_i} h_0^j(u) exp(Z_i \beta^j) du \right) \right\}^{\delta_i} \left\{ exp\left(- \int_0^{t_i} h_0^j(u) exp(Z_i \beta^j) du \right) \right\}^{1-\delta_i} \right]^{\mathbb{1}\{W_i = j\}}$$

where, i is the patient index and j is the cluster index. The indicator denotes that if patient i belongs to cluster j then we should use the likelihood function based on the model for cluster j. Next we discuss our implementation of the EM algorithm to estimate the parameters for the model above.

Expectation Maximization Algorithm

E-step: For the l^{th} iteration of the EM algorithm, we take the expectation of the log of the complete likelihood function (18). To do so, we first need to calculate the posterior distribution of the cluster membership random variable, W_i, given the data and the model parameters from last iteration using Bayesian approach:

$$T_{j,i}^{(l)} = P\left(W_i = j \middle| t_i, Z_i; \Psi^{(l)}\right) = \frac{\pi_j^{(l)} f_j\left(t_i; Z_i, \beta^{p,(l)}\right)}{\sum_{j=1}^{p} \pi_j^{(l)} f_j\left(t_i; Z_i, \beta^{p,(l)}\right)}$$

Thus, $T_{j,i}^{(l)}$ is the posterior probability of patient i belonging to cluster j at the l^{th} iteration of the EM algorithm, which can be interpreted as the similarity of the patient i readmission risk trajectory to those of patients in cluster j. We now approximate the expected log-likelihood function with respect to the posterior distribution of W_i:

$$Q(\Psi | \Psi^{(l)}) = E\left[\log L\left(\Psi; \{t_i\}, \{Z_i\}, \{W_i\}\right)\right] = \sum_{i=1}^{n} \sum_{j=1}^{p} T_{j,i}^{(l)} \left\{ \log \pi_j + \delta_i \left(\log h_0^j(t_i) + Z_i \beta^j - \int_0^\infty h_0^j(t_i) exp\left(Z_i \beta^j\right) dt \right) + (1 - \delta_i) \left(- \int_0^\infty h_0^j(t_i) exp\left(Z_i \beta^j\right) dt \right) \right\}.$$

M-step: This step finds the parameters that maximize $Q(\Psi/\Psi^{(l)})$. Given that $\boldsymbol{\pi}_j$ and $\boldsymbol{\beta}_j$ all appear in separate linear terms in Equation 19, the maximization can be applied separately for $\boldsymbol{\pi}_j$ and $\boldsymbol{\beta}_j$.

- **Estimating $\boldsymbol{\pi}_j$:** The estimate of $\boldsymbol{\pi}_j$ has the same form as the maximum likelihood estimate (MLE) for the Dirichlet distribution, i.e.,

$$\pi_j^{(n+1)} = \text{argmax}_{\pi_j} Q(\Psi | \Psi^{(l)}) = \text{argmax} \left\{ \sum_{j=1}^{p} \sum_{i=1}^{n} T_{j,i}^{(l)} \log \pi_j \right\} = \frac{\sum_{i=1}^{n} T_{j,i}^{(l)}}{\sum_{i=1}^{n} \sum_{j=1}^{p} T_{j,i}^{(l)}}.$$

- **Estimating $\boldsymbol{\beta}^j$:** Directly estimating $\beta^{j,(n+1)} = \underset{\beta^j}{argmax} Q(\Psi | \Psi^{(l)})$ is difficult because we do not specify the form of the baseline hazard $h_0^j(t_i)$. Following the literature (Cox (1975), Breslow (1975)), we use the *partial* likelihood function and apply the Breslow approximation to deal with repeated readmission times. That is, we estimate $\boldsymbol{\beta}_j$ via

$$\beta^{j,(n+1)} = \operatorname{argmax}_{\beta^j} \log \prod_{i=1}^{n} \left\{ \frac{\prod_{k \in \mathfrak{I}(t_{(i)})} T_{j,i}^{(l)} \exp\left(Z_i \beta^j\right)}{\left(\sum_{k \in \mathfrak{R}(t_{(i)})} T_{j,k}^{(l)} \exp\left(Z_k i \beta^j\right)\right)^{|\mathfrak{I}(t_i)|}} \right\}^{\delta_i}$$

$$= \operatorname{argmax}_{\beta^j} \sum_{i=1}^{I} \delta_i \left\{ \sum_{k \in \mathfrak{I}(t_{(i)})} \log\left(T_{j,i}^{(l)}\right) + \sum_{k \in \mathfrak{I}(t_{(i)})} Z_i \beta^j \right. \tag{19}$$

$$\left. - \left|\mathfrak{I}(t_{(i)})\right| \log\left(\sum_{k \in \mathfrak{R}(t_{(i)})} T_{j,k}^{(l)} \exp\left(Z_k \beta^j\right) \right) \right\}.$$

Here, $t_{(i)}$ is the i^{th} ordered unique readmission time, I is the number of unique readmission times, $\mathfrak{I}(t_{(i)})$ is the set of individuals who were readmitted on t_i day after discharge, $\mathfrak{R}(t_{(i)})$ is the set of individuals not belonging to set $\mathfrak{I}(t_{(i)})$, and δ_i means that only contributions from uncensored readmission times are considered. Then, Equation (19) can be maximized using the Newton-Raphson algorithm.

References

Anderson David, Golden Bruce, Jank Wolfgang, Wasil Edward. 2012 The impact of hospital utilization on patient readmission rate. Health care management science 15(1) 29–36. [PubMed: 21882018]

Anderson David, Price Carter, Golden Bruce, Jank Wolfgang, Wasil Edward. 2011 Examining the discharge practices of surgeons at a large medical center. Health care management science 14(4) 338–347. [PubMed: 21674142]

Ata Barış, Shneorson Shiri. 2006 Dynamic control of an *M/M/*1 service system with adjustable arrival and service rates. Management Science 52(11) 1778–1791.

Bartel Ann P, Chan Carri W, Kim Song-Hee Hailey. 2014 Should hospitals keep their patients longer? the role of inpatient care in reducing post-discharge mortality. Tech. rep., National Bureau of Economic Research.

Billings John, Blunt Ian, Steventon Adam, Georghiou Theo, Lewis Geraint, Bardsley Martin. 2012 Development of a predictive model to identify inpatients at risk of re-admission within 30 days of discharge (parr-30). BMJ open 2(4) e001667.

Breslow Norman E. 1975 Analysis of survival data under the proportional hazards model. International Statistical Review/Revue Internationale de Statistique 45–57.

Chan Carri W, Farias Vivek F, Bambos Nicholas, Escobar Gabriel J. 2012 Optimizing intensive care unit discharge decisions with patient readmissions. Operations research 60(6) 1323–1341.

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

Cox David R. 1975 Partial likelihood. Biometrika 62(2) 269–276.

COX DH. 1972 Regression models and life-tables. Journal of .

Donzé Jacques, Aujesky Drahomir, Williams Deborah, Schnipper Je rey L. 2013 Potentially avoidable 30-day hospital readmissions in medical patients: derivation and validation of a prediction model. JAMA internal medicine 173(8) 632–638. [PubMed: 23529115]

Eng Kevin H, Hanlon Bret M. 2012 Discrete mixture regression models for heterogenous time-to-event data: Cox assisted clustering. Bioinformatics 30(12) 1690–1697.

Farewell Vern T. 1982 The use of mixture models for the analysis of survival data with long-term survivors. Biometrics 1041–1046. [PubMed: 7168793]

George Jennifer M, Harrison J Michael. 2001 Dynamic control of a queue with adjustable service rate. Operations research 49(5) 720–731.

Grzyb Megan, Zhang Amber, Good Cristina, Khalil Khaled, Guo Bochen, Tian Lu, Valdez Jose, Gu Quanquan. 2017 Multi-task cox proportional hazard model for predicting risk of unplanned hospital readmission. Systems and Information Engineering Design Symposium (SIEDS), 2017 . IEEE, 265–270.

Hao Shiying, Wang Yue, Jin Bo, Shin Andrew Young, Zhu Chunqing, Huang Min, Zheng Le, Luo Jin, Hu Zhongkai, Fu Changlin, et al. 2015 Development, validation and deployment of a real time 30 day hospital readmission risk assessment tool in the maine healthcare information exchange. PloS one 10(10) e0140271. [PubMed: 26448562]

Heggestad Torhild. 2002 Do hospital length of stay and staffing ratio affect elderly patients' risk of readmission? A nation-wide study of norwegian hospitals. Health services research 37(3) 647–665. [PubMed: 12132599]

Helm Jonathan E, Alaeddini Adel, Stauffer Jon M, Bretthauer Kurt M, Skolarus Ted A. 2016 Reducing hospital readmissions by integrating empirical prediction with resource optimization. Production and Operations Management 25(2) 233–257.

Kuk Anthony YC, Chen Chen-Hsin. 1992 A mixture model combining logistic regression with proportional hazards regression. Biometrika 79(3) 531–541.

Kuo Yong-Fang, Goodwin James S. 2011 Association of hospitalist care with medical utilization after discharge: evidence of cost shift from a cohort study. Annals of internal medicine 155(3) 152–159. [PubMed: 21810708]

Mikalsen Karl Øyvind, Soguero-Ruiz Cristina, Bianchi Filippo Maria, Revhaug Arthur, Jenssen Robert. 2018 An unsupervised multivariate time series kernel approach for identifying patients with surgical site infection from blood samples. arXiv preprint arXiv:1803.07879 .

Motai Yuichi, Siddique Nahian Alam, Yoshida Hiroyuki. 2017 Heterogeneous data analysis: online learning for medical-image-based diagnosis. Pattern Recognition 63 612–624.

Rosen Ori, Tanner Martin. 1999 Mixtures of proportional hazards regression models. Statistics in Medicine 18(9) 1119–1131. [PubMed: 10378260]

Schölkopf Bernhard, Smola Alexander, Müller Klaus-Robert. 1997 Kernel principal component analysis. International Conference on Artificial Neural Networks. Springer, 583–588.

Schölkopf Bernhard, Smola Alexander J. 2002 Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press.

Shams Issac, Ajorlou Saeede, Yang Kai. 2015 A predictive analytics approach to reducing 30-day avoidable readmissions among patients with heart failure, acute myocardial infarction, pneumonia, or copd. Health care management science 18(1) 19–34. [PubMed: 24792081]

Shulan Mollie, Gao Kelly, Moore Crystal Dea. 2013 Predicting 30-day all-cause hospital readmissions. Health care management science 16(2) 167–175. [PubMed: 23355120]

van Walraven Carl, Dhalla Irfan A, Bell Chaim, Etchells Edward, Stiell Ian G, Zarnke Kelly, Austin Peter C, Forster Alan J. 2010 Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. Canadian Medical Association Journal 182(6) 551–557. [PubMed: 20194559]

van Walraven Carl, Wong Jenna, Forster Alan J. 2012 Lace+ index: extension of a validated index to predict early death or urgent readmission after hospital discharge using administrative data. Open Medicine 6(3) e80. [PubMed: 23696773]

Vinzamuri Bhanukiran, Reddy Chandan K. 2013 Cox regression with correlation based regularization for electronic health records. Data Mining (ICDM), 2013 IEEE 13th International Conference on. IEEE, 757–766.

Yang Shanshan, Zheng Fang, Luo Xin, Cai Suxian, Wu Yunfeng, Liu Kaizhi, Wu Meihong, Chen Jian, Krishnan Sridhar. 2014 Effective dysphonia detection using feature dimension reduction and kernel density estimation for patients with parkinsons disease. PloS one 9(2) e88825. [PubMed: 24586406]

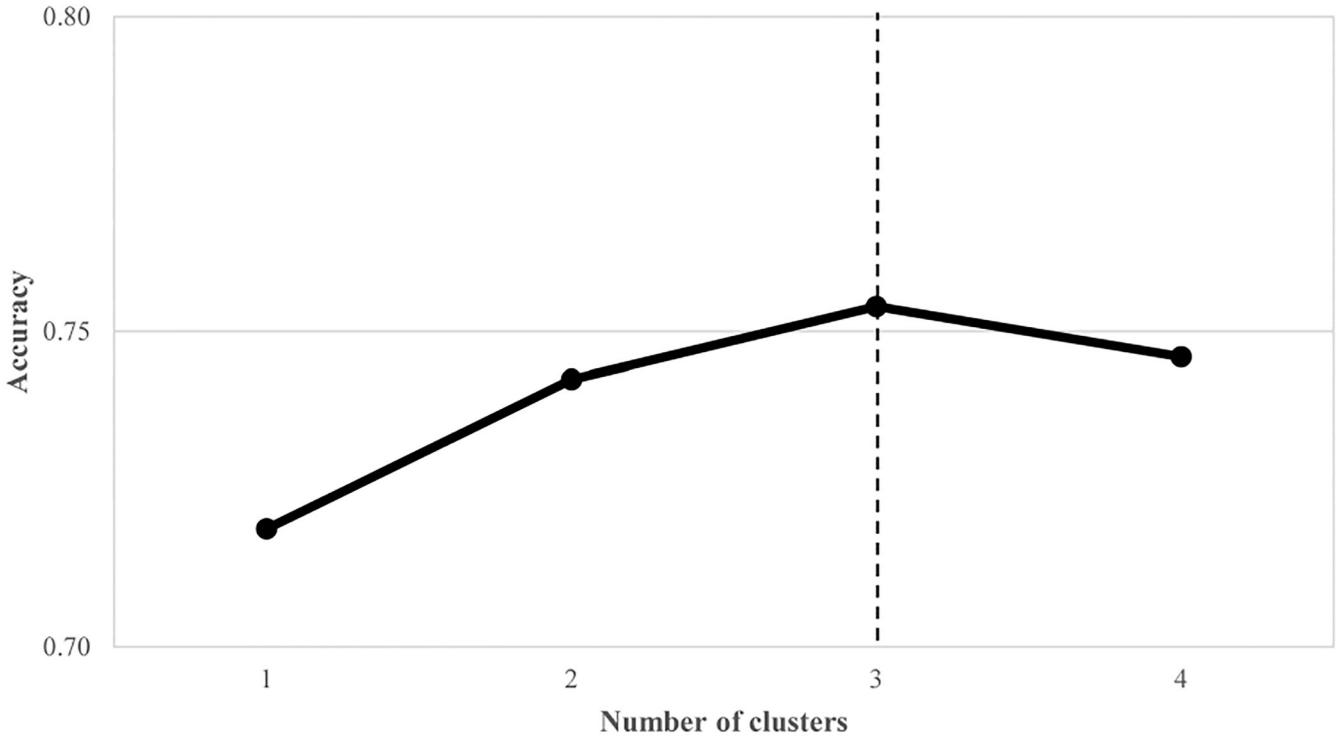
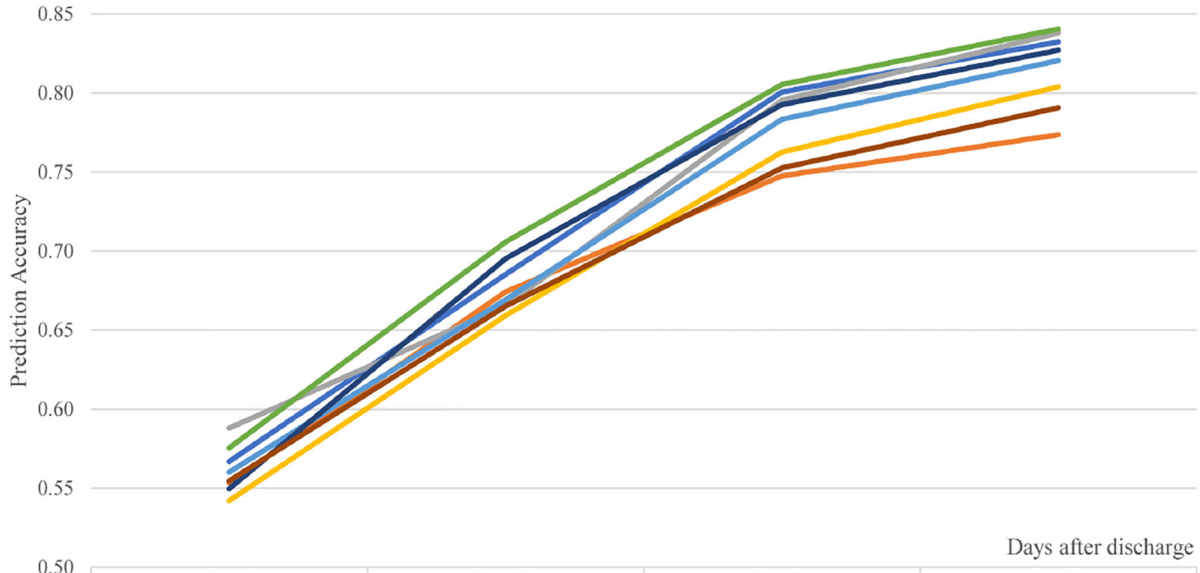
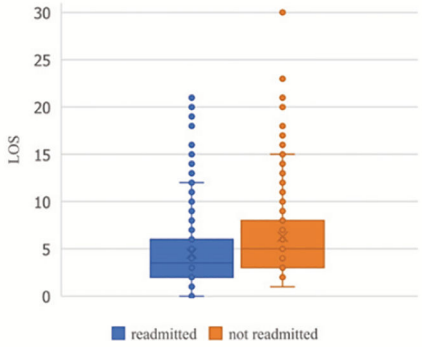


Fig. 1:
Prediction accuracy of Mixture KPCA COX model for various number of clusters.

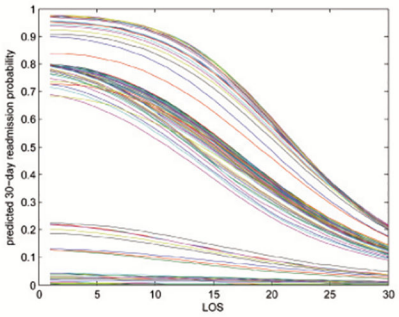


	Day 5	Day 10	Day 20	Day 30
SVM	0.567	0.685	0.801	0.832
Logistic Regression	0.553	0.674	0.748	0.774
Boosting	0.588	0.665	0.795	0.838
Classic Cox	0.542	0.659	0.763	0.804
KPCA_COX	0.560	0.669	0.783	0.821
M_KPCA_COX	0.575	0.706	0.805	0.841
NN	0.550	0.695	0.793	0.827
M Logistic	0.555	0.665	0.753	0.791

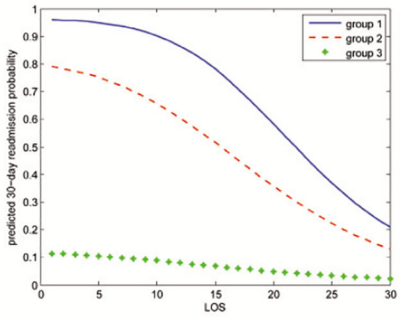
Fig. 2:
Prediction accuracy of the comparing methods for estimating readmission 5, 10, 20, and 30 days after discharge.



(a) Box plot of LOS for readmitted vs not-readmitted patients, 30 days after discharge



(b) Predicted trajectory for 1032 patients



(c) Average trajectory for the three groups based on the three clusters of the mixture KPCA cox model

Fig. 3: Predicted 30-day cumulative readmission probability against length of stay (LOS).

In Figure 3(a), we compare the LOS for different patients according to the readmission event 30 days after discharge. In Figure 3(b), each curve represents the predicted readmission probability trajectory for each of 1032 randomly selected patients. In Figure 3(c), we group the 1032 randomly selected patients into three groups based on the three clusters of the proposed mixture KPCA Cox model; the three curves in this plot correspond to the average readmission probability of the each group.

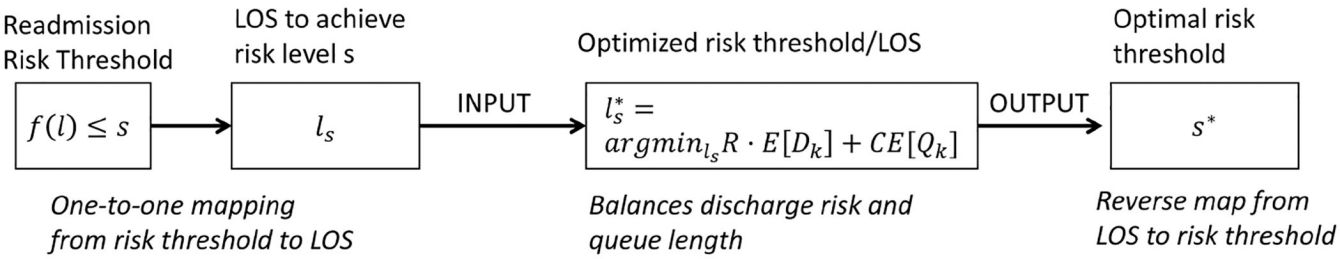


Fig. 4:
Flow chart of the optimization framework.

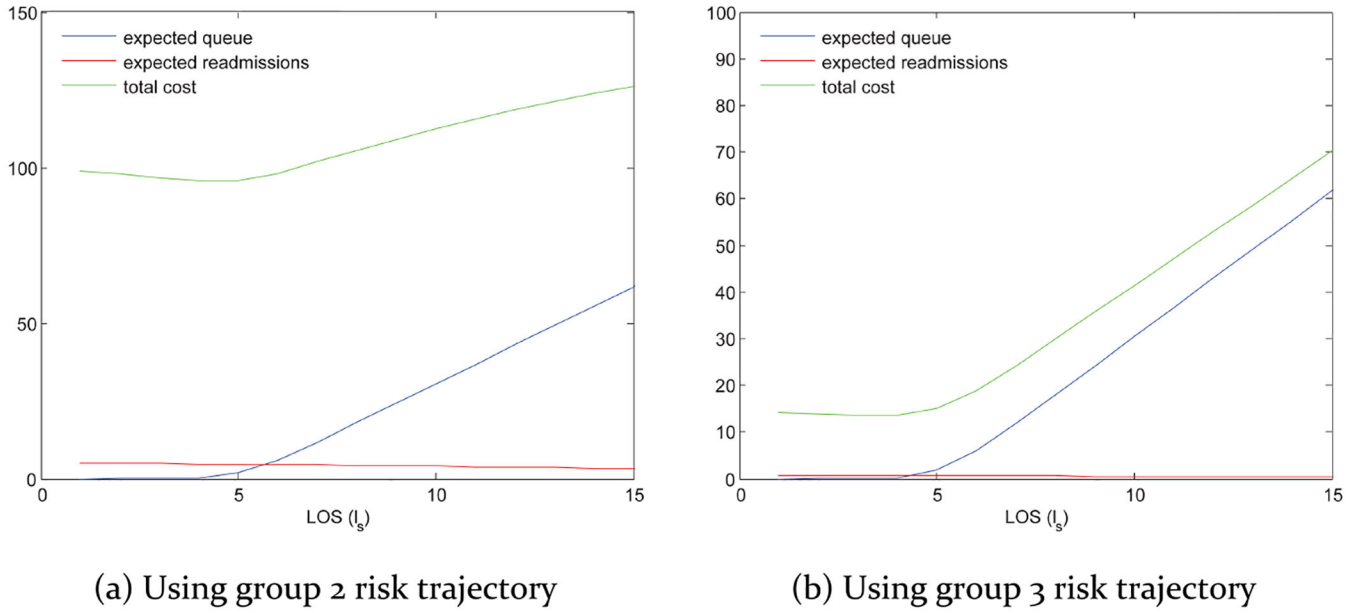


Fig. 5: Threshold discharge policy: expected queue length, the expected re-admissions, and the total cost.

We set $N = 32$, $\Lambda = 6.25$, $R = 20$, and $C = 1$. For the risk trajectory function $s = f(l)$, we use the average one from group 2 and group 3 patients, respectively, in the left and right plots. The x-axis denotes the LOS l, while the y-axis denotes the corresponding value of the expected queue length, the expected re-admissions, and the total cost for the blue, red, and green curve, respectively.

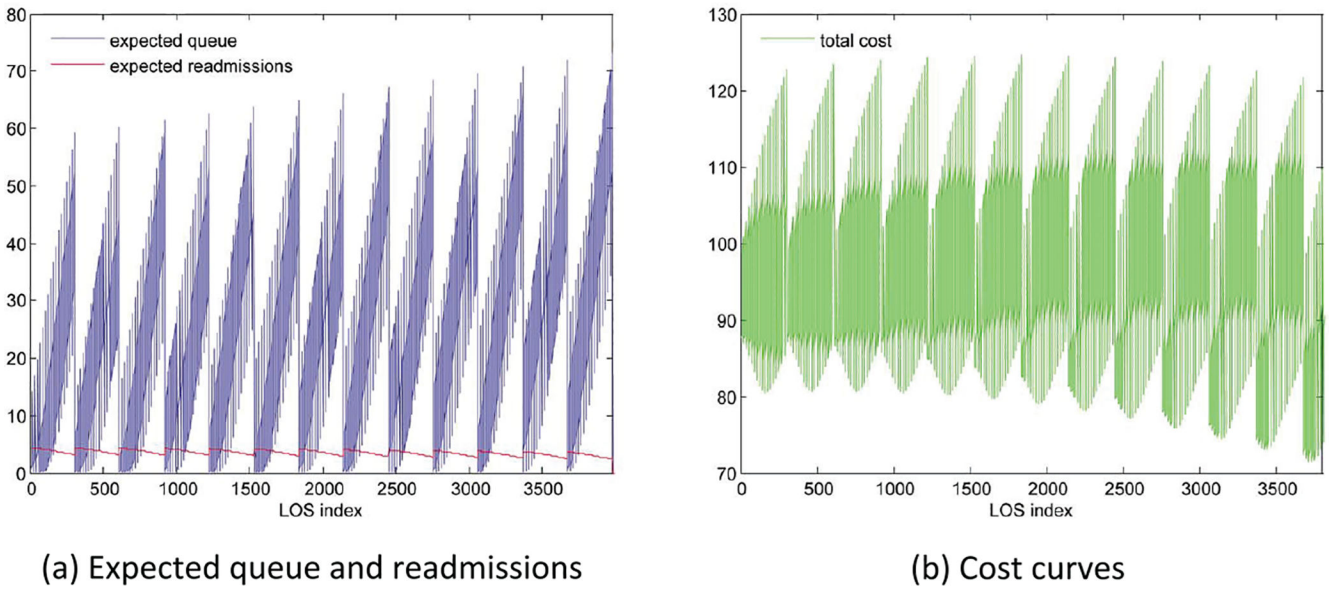


Fig. 6: Threshold discharge policy with multiple classes: expected queue length, expected number of re-admission events, and the total cost.

We set $N = 52$, $\Lambda = 6.25$, $R = 20$, and $C = 1$. The "LOS index" on the x-axis denotes an index of the combination of l_1, l_2, l_3, so that we can plot the cost against different choices of (l_1, l_2, l_3) on a two-dimensional figure. We impose a lower bound of 3 and an upper bound of 20 for each l_m, and the LOS index equals $17^2 \cdot l_1 + 17 \cdot l_2 + l_3$. The y-axis denotes the corresponding value of the expected queue length, the expected re-admissions, and the total cost for the blue, red, and green curve, respectively.

Table 1:

List of variables used in the proposed integrated framework.

Variable	Description
Readmission prediction model	
$X_{i,k}$	k^{th} risk factor $(K = 1,\ldots,D)$ of the i^{th} patient $(i = 1,\ldots,n)$.
Z_i	KPCA transformation of the risk factors Xi $(Z_i = \{Z_{i,1}, \ldots, Z_{i,K}\})$.
$h^j(t; Y_i)$	hazard function of patient i at time t when patient assigned to cluster j.
$\pi_{ij}, \sum_{j=1}^{p} \pi_{ij} = 1$	membership probability of patient i in cluster j.
$f(t; Y_i)$	readmission likelihood of patient i at time t.
Discharge optimization model	
X_k^j	number of patients who already spent j days in the hospital on day k.
$X_k = \sum_{j=0}^{l_s - 1} X_k^j$	total number of patients (post-discharge) in the hospital on day k.
X_{k-}	pre-discharge total number.
$\mathbb{E}[Q_k] = \mathbb{E}[X_k - N]^+$	expected queue length.
l_s^*	optimal discharge threshold.
$s = s(l)$	readmission probability (as a function of LOS l).
R and C	unit penalty cost for readmission and unit holding cost.

Table 2:

Distribution of data across different risk factors.

Risk Factor	Level	Overall	%
LOS	(Meam, Std)	(4.8,4.5)	
	(Q1,Q2,Q3)	(2,3,6)	
SEX	Female	2,929	94.70%
	Male	164	5.30%
Age	(Meam,Std)	(63.79,13.07)	
Health Insurance	Insured	2,774	89.69%
	Un-insured	319	1031%
Eligibility	10 = SC 50–100%	653	21.11%
	20 = Aid & Attendance	1,414	45.72%
	21 = Housebound	81	2.62%
	24 = POW	368	11.90%
	30 = SC 40–49%	81	2.62%
	31 = SC 30–39%	98	3.17%
	32 = SC 20–29%	69	2.23%
	33 = SC 10–19%	132	4.27%
	34 = SC less than 10%	160	5 17%
	40 = NSC - VA Pension	12	0.39%
	50 = NSC	6	0.19%
	101 = CHAMPVA	2	0.06%
	105 = Allied Veteran	10	0.32%
	106 = Humanitarian Emergency	1	0.03%
	107 = Sharing Agreement	4	0.13%
	109 = Tricare/ CHAMPUS	2	0.06%
Employment Status	Not Employed	1683	54.41%
	Retired	866	28.00%
	Employed Full Time	323	10.44%
	Unknown	56	1.81%
	Self Employed	60	1.94%
	Employed Part Time	102	3.30%
	Active Duty Military	3	0.10%
Enrollment Status	1 = Unverified	19	0.61%
	2 = Verified	3040	98.29%
	6 = Deceased	6	0.19%
	16 = Pending; Means Test required	3	0.10%
	17 = Pending, Eligibility is Unverified	1	0.03%
	19 = Not Eligible; Refused to pay co-pay	1	0.03%
	20 = Not Eligible; Ineligible Date	4	0.13%
	22 = Reject; Below Enrollment	2	0.06%
	23 = Other	17	0.55%

Risk Factor	Level	Overall	%
Source of Admission	Hospital	2910	94.08%
	NHCU	178	5.75%
	DOMICILIARY	5	0.16%
WARD	17 wards at the VA facility	Avg. Patients per ward = 189	
		(Min, Median, Max)=(2,87,818) patients in a ward	
Principle Diagnosis	30 Common diagnosis (ICD-9) at the VA facility	296, 292, 428, 295, 715, 303, 491, 291, 427, 486, 682, V57, 311, 786, 780, 599,162, 304, 250, 309, 414, 185, 996;276;V58, 560,38,285,584,410	

Table 3:

Comparing methods information.

No.	Method	Parameter Estimation Notes	
1	Support Vector Machine (SVM)	**a.**	Gaussian Kernal with = 1.6819
		b.	Cache limit: 5000
		c.	Method: Sequential Minimal Optimization (SMO)
2	Logistic Regression	**a.**	No interaction terms
3	Boosting	**a.**	Learners: Tree
		b.	Number of ensembles learning
		c.	cycles: 100
		d.	Method: AdaBoostM1
4	Classic Cox Proportional Hazard Regression	**a.**	Baseline hazard function: Nonparametric
		b.	Risk factor values at baseline hazard:0
		c.	Censoring considered
		d.	No interaction terms
5	Kernel Principle Component Analysis Cox (KPCA Cox)	**a.**	Gaussian Kernal with = 1.8147
		b.	13 principle components
		c.	Baseline hazard function: non parametric
		d.	Risk factor values at baseline hazard: 0
		e.	Censoring considered
6	Mixture Kernel Principle Component Analysis Cox (M KPCA Cox) *(Proposed)*	**a.**	Gaussian Kernal with = 1.8147
		b.	13 principle components
		c.	P=3 mixture components
		d.	EM algorithm for optimization
		e.	Mixture component selection criteria: cross validation
7	Neural Network (NN)	**a.**	Hidden layers: 1
		b.	Network training function: scaled conjugate gradient method
		c.	Perform function: cross entropy
8	Logistic Regression Mixture Model	**a.**	No interaction terms
		b.	P=3 mixture components
		c.	EM algorithm for optimization
		d.	Mixture component selection criteria: cross validation

Table 4:

Threshold discharge policy: impact of ward capacity *N*.

We set $\Lambda = 6.25$, $R = 20$, and $C = 1$. For the risk trajectory, we use the average one from group 3 patients. Column 1 denotes the capacity *N*, column 2 denotes the optimal discharge threshold l_s^*, column 3–5 denote the corresponding average queue length, probability of readmission, and optimal total cost under the optimal discharge threshold l_s^*, respectively.

N	l_s^*	$\mathbb{E}[Q_k]$	*s*	total cost
28	3	0.04	10.94%	13.71
32	4	0.22	10.72%	13.61
36	4	0.04	10.72%	13.44
40	5	0.17	10.46%	13.25
44	5	0.03	10.46%	13.11
48	6	0.13	10.17%	12.85
52	7	0.37	9.85%	12.69
56	7	0.10	9.85%	12.42

Table 5:

Threshold discharge policy: impact of risk trajectories.

We set $\Lambda = 6.25$, $C = 1$, and $R = 20$. For the risk trajectory, we use the average trajectory from group 2 or group 3 patients. The three columns under "group 2" or "group 3" show the optimal discharge threshold l_s^*, the average queue length $\mathbb{E}[Q_k]$, and the probability of readmission s, respectively.

	group 2			group 3		
N	l_s^*	$\mathbb{E}[Q_k]$	***s***	l_s^*	$\mathbb{E}[Q_k]$	***s***
28	4	0.87	76.42%	3	0.04	10.94%
32	4	0.22	76.42%	4	0.22	10.72%
36	5	0.65	75.17%	4	0.04	10.72%
40	6	1.41	73.68%	5	0.17	10.46%
44	7	2.51	72.00%	5	0.03	10.46%
48	7	1.07	72.00%	6	0.13	10.17%
52	8	1.95	70.08%	7	0.37	9.85%
56	9	3.11	67.95%	7	0.10	9.85%

Table 6:

Threshold discharge policy: impact of cost parameters.

We set $\Lambda = 6.25$, $C = 1$, and $R = 5$ or 40. For the risk trajectory, we use the average trajectory from group 3 patients. The three columns under "$R = 5$" or "$R = 40$" show the optimal discharge threshold l_s^*, the average queue length $\mathbb{E}[Q_k]$, and the probability of readmission s, respectively.

	$R = 5$			$R = 40$		
N	l_s^*	$\mathbb{E}[Q_k]$	s	l_s^*	$\mathbb{E}[Q_k]$	s
28	3	0.04	10.94%	3	0.04	10.94%
32	3	0.00	10.94%	4	0.22	10.72%
36	4	0.04	10.72%	5	0.65	10.46%
40	4	0.00	10.72%	5	0.17	10.46%
44	5	0.03	10.46%	6	0.49	10.17%
48	5	0.00	10.46%	6	0.13	10.17%
52	6	0.03	10.17%	7	0.37	9.85%
56	7	0.10	9.85%	8	0.82	9.51%