POMS PRODUCTION AND OPERATIONS MANAGEMENT

PRODUCTION AND OPERATIONS MANAGEMENT POMS

Vol. 25, No. 2, February 2016, pp. 233–257
ISSN 1059-1478 | EISSN 1937-5956 | 16 | 2502 | 0233
DOI 10.1111/poms.12377

Reducing Hospital Readmissions by Integrating Empirical Prediction with Resource Optimization

Jonathan E. Helm
Operations and Decision Technologies, Kelley School of Business, Indiana University, 1309 E. Tenth Street, Bloomington, Indiana 47405, USA, helmj@indiana.edu

Adel Alaeddini
Department of Mechanical Engineering, University of Texas, San Antonio One UTSA Circle, San Antonio, Texas 78249, USA
adel.alaeddini@utsa.edu

Jon M. Stauffer, Kurt M. Bretthauer
Operations and Decision Technologies, Kelley School of Business, Indiana University, 1309 E. Tenth Street, Bloomington, Indiana 47405, USA, stauffer@indiana.edu, kbrettha@indiana.edu

Ted A. Skolarus
Department of Urology, University of Michigan, VA, Health Services Research & Development (HSR&D) Center for Clinical Management Research, VA, Ann Arbor Healthcare System, 3875 Taubman Center, 1500 East Medical Center Drive, Ann Arbor, Michigan 48109, USA
tskolar@med.umich.edu

Hospital readmissions present an increasingly important challenge for health-care organizations. Readmissions are expensive and often unnecessary, putting patients at risk and costing $15 billion annually in the United States alone. Currently, 17% of Medicare patients are readmitted to a hospital within 30 days of initial discharge with readmissions typically being more expensive than the original visit to the hospital. Recent legislation penalizes organizations with a high readmission rate. The medical literature conjectures that many readmissions can be avoided or mitigated by post-discharge monitoring. To develop a good monitoring plan it is critical to anticipate the timing of a potential readmission and to effectively monitor the patient for readmission causing conditions based on that knowledge. This research develops new methods to empirically generate an individualized estimate of the time to readmission density function and then uses this density to optimize a post-discharge monitoring schedule and staffing plan to support monitoring needs. Our approach integrates classical prediction models with machine learning and transfer learning to develop an empirical density that is personalized to each patient. We then transform an intractable monitoring plan optimization with stochastic discharges and health state evolution based on delay-time models into a weakly coupled network flow model with tractable subproblems after applying a new pruning method that leverages the problem structure. Using this multi-methodologic approach on two large inpatient datasets, we show that optimal readmission prediction and monitoring plans can identify and mitigate 40–70% of readmissions before they generate an emergency readmission.

Key words: hospital readmissions; post-discharge patient monitoring; readmission risk profiling; Bayesian survival analysis; delay-time models of readmissions
History: Received: December 2013; Accepted: December 2014 by Tsan-Ming Choi, after 1 revision.

1. Introduction

Hospital and medical center readmissions is a serious health-care issue demanding increased attention as costs continue to rise and patient care suffers. Based on a report to Congress in 2008, over 17% of Medicare patients were readmitted in the first 30 days after discharge, accounting for more than $15 billion dollars per year (Foster and Harkness 2010). Not only are readmissions expensive, recent studies have also linked the rate of readmission to quality of care in medical centers (e.g., Halfon et al. 2006). Surprisingly, Foster and Harkness (2010) found that a significant percent of readmissions are avoidable through better post-discharge management; of the $15 billion spent, *$12 billion was associated with potentially preventable readmissions*. Current strategies to reduce readmissions focus on (i) identifying high-risk patients (e.g., Kansagara et al. 2011, Rosenberg et al. 2007, Wallmann et al. 2013), or (ii) developing an effective plan for post-discharge care (e.g., Jack et al. 2009). While these heuristic clinical approaches have proven effective in avoiding readmissions, there remains significant opportunity for an approach that combines

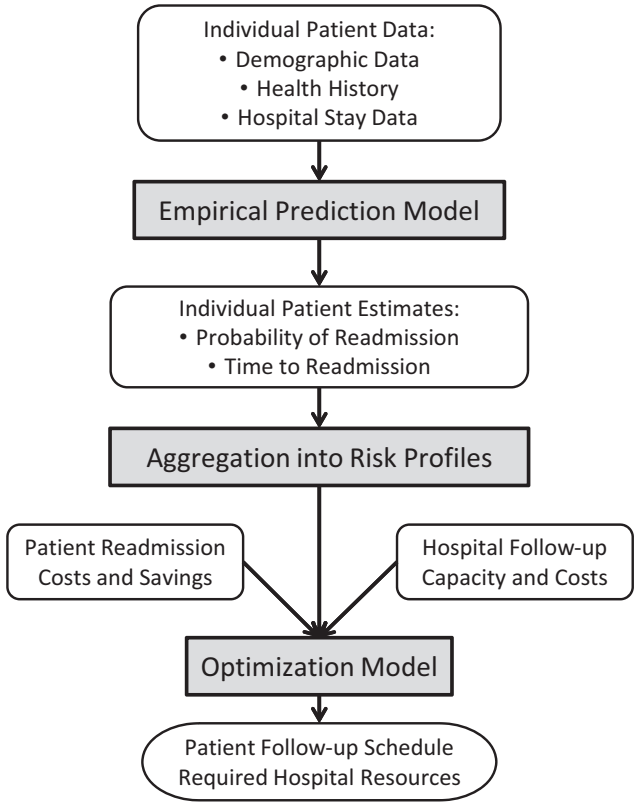
rigorous empirical modeling to predict time to readmission with optimization to design schedules and allocate staff for post-discharge monitoring. To have the largest possible impact on readmissions, it is necessary to know both when a patient is likely to be readmitted (empirical prediction model) and when to monitor that patient to identify the condition before it triggers a readmission (optimization model). This study represents a multi-methodology effort aimed at integrating clinical, statistical, and operations management techniques to (i) quantify post-discharge risk of readmission for each patient over time, (ii) to design optimal post-discharge treatment plans for early detection and avoidance of potential readmissions, and (iii) to allocate sufficient system capacity to be able to administer the optimal treatment plans for a cohort of patients.

Numerous efforts have focused on capturing the key dynamics of the readmission system (Desai et al. 2009, Kansagara et al. 2011). The study of readmission risk factors typically falls into three major categories: (i) patient attributes such as history of readmission, severity of illness, comorbidity, age, gender, life satisfaction, change in clinical variables, source of payment, etc. (e.g., Dunlay et al. 2009, Wallmann et al. 2013, Watson et al. 2011); (ii) factors targeting the predischarge process including length of stay, adequacy of discharge plan, nursing environment of the hospital, characteristics of the physician, etc., (e.g., McHugh and Ma 2013, Rosen et al. 2013); and finally (iii) factors targeting the post-discharge process including inadequacy of post-discharge planning and follow up, non-compliance with medication and diet, failed social support, impairment of self-care, etc. (e.g., Hernandez et al. 2010, Wallmann et al. 2013, Watson et al. 2011). Using the above risk factors a number of health-care systems have started implementing online readmission risk calculators. Some of these calculators may be found at http://riskcalc.sts.org/STSWebRiskCalc273/, by the Society of Thoracic Surgeons which predicts the risk of operative mortality and morbidity after adult cardiac surgery, and at http://www.readmissionscore.org, by the Center for Outcomes Research and Evaluation (CORE), which helps predict a patient's likelihood of readmission for heart failure within 30 days of discharge. Despite their benefits, these calculators have serious limitations. They (i) assume homogeneity of the population and hospital's performance; (ii) provide no estimate on time to readmission; and (iii) provide no guidance on how to use the estimates to make better care decisions. Our methods will address these deficiencies.

Recently, researchers have begun to investigate the impact of targeted discharge planning and post-discharge management on reducing readmissions, focusing on financial incentives/cost-effectiveness, pre-discharge patient education, and improved post-discharge management. In particular, several studies claim that post-discharge management can reduce readmissions by *12% to 30%* (see Gonseth et al. 2004) and *as high as 85%* (see Fonarow et al. 1997) by targeting high-risk populations (see Minott 2008, Wolinsky et al. 2009), telemonitoring (see Graham et al. 2012), and other monitoring strategies. By integrating patient risk calculations and empirical predictions of time to readmission with optimization methods to design monitoring plans, we capture both of the high-impact approaches (risk profiling and planned monitoring) from the medical literature in a quantitative framework for optimally designing these post-discharge monitoring plans that are currently designed using expert judgment or ad hoc approaches.

Figure 1 provides a high-level overview of our multi-methodology approach which uses both readmission prediction and follow-up schedule optimization to reduce readmissions. First the Empirical Prediction Model utilizes individual patient data to determine a probability of readmission and expected time to readmission for each patient. These patient and procedure specific readmission curves are then aggregated with *K*-mean clustering (or other methods) into several different risk profiles. The aggregated readmission curves for each risk profile, organization-specific resource capacity and

Figure 1 Multi-Methodology Model Overview



cost information, and medical procedure-specific readmission cost and savings information are all used as inputs into the Optimization Model. The Optimization Model uses these inputs to determine an optimal follow-up schedule for each patient risk profile and determine the number of resources the hospital or health system will need to execute all expected follow-up schedules.

While previous literature relies on a siloed approach, focusing either on predicting readmissions or on strategies to reduce readmissions, this research integrates the two using advanced mathematical, statistical, and operations management techniques combined with clinical expertise. We not only develop an integrative framework for investigating both aspects of readmission modeling simultaneously, we also contribute new methods to each of the areas. To the best of our knowledge, existing studies have not effectively considered heterogeneity among patient populations, and are not able to adapt population-based readmission estimates to individual patients. Further, previous readmission prediction models have only focused on small groups of patients with a single readmission triggering condition (e.g., elderly cardiovascular patients), and the results are often not generalizable to other cases (see Feudtner et al. 2009, Gonseth et al. 2004). In addition, most of the available studies have not effectively used the array of available machine-learning techniques to improve their results. This study addresses these deficiencies by enabling individualized readmission probability estimates and a generalizable method that can encompass diverse patient populations and multiple readmission causing conditions over an arbitrary time period. Further, existing models have been lacking comprehensive optimization approaches to design tailored post-discharge management plans. No literature to our knowledge captures, as we intend to do, the health-care organization's ability to support a large-scale implementation of a post-discharge management scheme that simultaneously solves for post-discharge monitoring timing and the organizational resource capacity needed to implement such schedules.

Finally, we demonstrate how this multi-methodology approach can be applied via an extensive case study and numerical analysis using two different datasets from (i) a partner hospital in Michigan including 2449 patients with 17 diagnoses, 3108 readmissions, and 15 demographic, socioeconomic, and clinical factors, etc. (ii) the State Inpatient Databases (SID) for 5000 patients diagnosed with bladder, kidney, and prostate cancer in 2009 along with other cancers (see http://www.hcup-us.ahrq.gov/db/state/siddist/SID_Introduction.jsp). The results for the two datasets were structurally similar, so for the purposes of cohesive exposition we focus on the results from the partner hospital in Michigan for this study.

Section 2 develops the empirical model to predict readmission occurrence and timing. Section 3 uses the predicted empirical readmission density from section 2 to develop a follow-up schedule for patients and staffing plan for a follow-up organization. Section 4 brings both components together, empirical prediction and resource optimization, in a case study using historical inpatient readmission data to design a practical post-discharge monitoring schedule and generate insights into tactical and operational management of post-discharge care. These results confirm the conjecture in the medical literature that between 12% and 85% of readmissions can be avoided or identified early through better post-discharge plans and show how to effectively design such plans. Section 5 concludes the study.

2. Stage 1: Empirical Modeling to Predict Time to Readmission

While a number of studies have focused on predicting whether or not a patient will be readmitted within 30 days (see van Walraven et al. 2010), there is only one article to our knowledge that focuses on predicting the time to readmission (see Yu et al. 2013). While Yu et al. (2013) shares similarities with our work, there are important differences in the two approaches. From a methodological perspective Yu et al. (2013), among other studies, does not consider which condition has caused the readmission, for example, infection, dehydration, kidney failure etc. We are able to capture this feature using a frailty approach to model these conditions as latent competing risks with stochastic dependence. In addition, Yu et al. (2013), among others, use a population-based approach based on equally weighted readmission records from a specific hospital to calculate the risk of readmission for that hospital's patients. However, we employ transfer learning to weight the readmission records in the dataset based on their similarity to the readmission record(s) for the patient of interest to: (i) further personalize the estimate and (ii) alleviate the problem of data scarcity. Finally, Yu et al. (2013), along with other readmission prediction models, gives the same importance to all readmission records regardless of how recently the readmission occurred. Our method assigns importance (weight) to the admission/readmission records based on record recency (more recent records get more weight) using an optimization process to choose the appropriate weights. This accounts for the phenomenon that each patient's health status and/or behaviors can change over time. From the specific modeling perspective, we use a Bayesian approach while Yu et al. (2013) uses a

classical approach. Further, we employ a parsimonious prior while Yu et al. (2013) employs a forward selection procedure for identifying the most important variables.

Understanding the time to readmission is critical to making clinically effective decisions to mitigate potential readmissions, such as when to follow up with a patient who has been discharged from the hospital. In this section, we develop empirical prediction models to accurately capture the probability distribution on time to readmission based on two different datasets (the State Inpatient Database (SID) as well as a dataset from a partner hospital in Michigan) to show that our methods can be used broadly (e.g., on SID) or tailored to a specific hospital.

Beyond exploring the new area of predicting time to readmission, we also address two other features that are prevalent in health care: (i) the need to personalize the prediction method to each individual patient, and (ii) scarcity of relevant data. The result of the empirical modeling in this section is a set of tailored probability distributions (one for each individual patient) that is personalized for each patient in our datasets. Our approach builds up the prediction model through three steps as shown in Figure 2. Step 1 (section 2.1) develops a general population estimate for time to readmission, which accounts for demographic, socioeconomic, health history, co-morbidity, the hospital the patient was treated at, and other relevant patient and system characteristics. In section 2.1, we also discuss how we are able to incorporate the cause of readmission into our prediction model. We do so by developing a Weibull regression model that incorporates observable and unobservable risk factors.

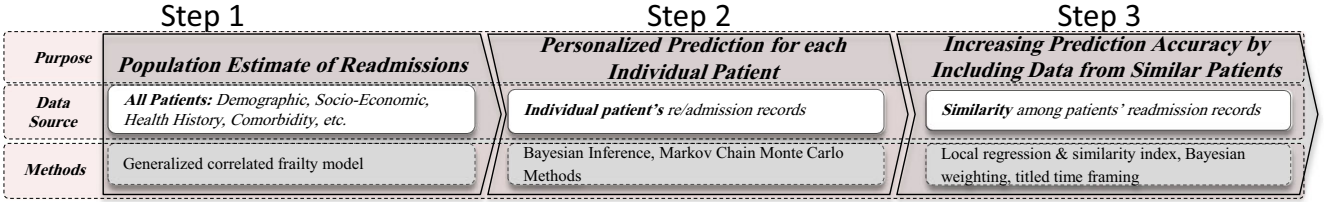
The model from Step 1 is then personalized in Step 2 (section 2.2) by parameterizing the Weibull model using each patient's personal history of hospital admissions and readmissions to date. Step 3 (also section 2.2) addresses the problem of data scarcity, which occurs when an individual has too few prior records to adequately parameterize the model with their individual data alone (Step 2) and when applying the method to a new hospital or group that has little relevant data. For example, when personalizing the readmission estimate, we are able to use data from all patients in our dataset (not just from the patient whose time to readmission curve is currently being estimated), by adding weights to the data records. Higher weights indicate a higher level of statistical similarity of any given patient in the dataset with the target patient. This approach, called transfer learning, and the specifics of calculating and incorporating weights are discussed in section 2.2.

In section 2.3 we discuss the methods and algorithms used to apply the approaches in sections 2.1 and 2.2 to our real-world datasets. While the data about a particular individual or specific hospital may be small, the overall dataset we intend the model to work with will be large. With large datasets, the common methods for prediction have significant drawbacks. For example, machine learning often suffers from results being difficult to interpret and sometimes yields patterns that are a product of random fluctuations, while more classical prediction models employ oversimplifying assumptions that lead to incorrect conclusions. To overcome these limitations, we develop an empirical prediction model that integrates both classical prediction methods with machine learning using a Bayesian framework. We conclude the section by comparing the accuracy of our prediction model against other commonly used prediction models in the literature.

2.1. Population-Based Model of Time to Readmission

We begin by building a population-based estimate of time to readmission in which we consider the impact of (i) time after initial discharge from the hospital, (ii) patient-specific risk factors impacting likelihood of readmission, and (iii) unobservable or random effects that capture patient heterogeneity. We capture time to readmission using a Weibull regression model. We begin with a hazard rate function, $h(t)$. In our optimization model presented in section 3, this hazard rate can be used to model the deterministic arrival rate function of a non-homogeneous Poisson process (NHPP) capturing the arrival of readmission-causing failures (as is common in delay-time analysis), however, the NHPP assumption is not necessary for estimation of the survival model. The probability that a patient has not yet been readmitted by time t is therefore given by $S(t) = exp\left(-\int_0^t h(u)du\right)$, which we call the survival function. The hazard function, however,

Figure 2 Framework for Predicting Patient Readmissions



depends not only on time but also on a set of K risk factors for readmission, $X = [x_1, \ldots, x_K]$. We consider the following factors available to us in the data: length of stay, gender, age, employment status, insurance coverage level, profession/military rank, ward(s) visited during inpatient stay, principal diagnosis, and source of admission, that is, VA hospital, nursing home, home, non-VA hospital. To tailor our hazard rate function to these patient characteristics, we employ a Weibull regression model which incorporates the important risk factors that affect probability of readmission as follows:

$$h(t|X) = h_0(t) \cdot exp(XB)\,, \tag{1}$$

where $h_0(t) = \rho t^{\rho-1}$ is a Weibull function. $B = [b_0, b_1, \ldots, b_K]'$ is a vector of K regression parameters (risk factor coefficients) to be estimated. However, not all of the risk factors affecting Equation 1 are easily known or even measurable. For example, patients can be readmitted for several different post-discharge complications—common ones include infection, dehydration, kidney failure, failure to thrive—where these conditions are all competing to cause a readmission, may exhibit stochastic dependence, and are not observable at the time a prediction is made. In the data, we are only able to observe the factor that *caused* the readmission, for example, infection, which is essentially the minimum failure time of all the latent risk factors that could cause readmission. To account for such latent competing risks and their stochastic dependence, we use a "frailty" approach to extend the Weibull regression model (see Clayton 1978, Hougaard and Hougaard 2000, Oakes 1989). If there are M frailty terms, $v_1, \ldots, v_M$, corresponding to the M latent risks, then the risk-specific hazard rate for the mth latent risk factor can be written as follows:

$$h_m(t|X, v_m) = h_{0,m}(t) \cdot exp(XB_m + v_m)\,. \tag{2}$$

Equation 2 is a generalization of Equation 1 to incorporate unmeasurable risk factors, v. Thus, we have a different hazard rate function for each of the competing risks that might cause a readmission—for example, $h_1(t|X, v_1)$ could be the hazard rate for infection, $h_2(t|X, v_2)$ could be the hazard rate for failure to thrive, etc. This allows the model to capture not only the time to readmission but also different time to readmission dynamics for different causes of readmission. This could potentially help clinicians better target diagnostic questioning and tests to look for specific readmission causing conditions at different times after discharge; an idea which is supported by the clinical literature (see Hu et al. 2014). Assuming that the vector of frailties v is drawn from a multivariate distribution with density $g(v_1, \ldots, v_M)$ and t_i for $i = 1, \ldots, M$ is the failure time for the ith readmission causing condition, then the unconditional (expected) survivor function can be calculated by integrating with respect to density g:

$$\begin{aligned} S(t_1,\ldots,t_M|X) &= \int \cdots \int S(t_1,\ldots,t_M|v_1,\ldots,v_M) \\ &\quad \times g(v_1,\ldots,v_M)dv_1\ldots dv_M \\ &= \int \cdots \int \Pi_{m=1}^{M} \{exp[-\int_0^{t_m} h_m(u|X,v_m)du]\} \\ &\quad g(v_1,\ldots,v_M)dv_1\ldots dv_M \\ &= \int \cdots \int \Pi_{m=1}^{M} \{exp[-t_m^{\rho_m} exp(XB_m+v_m)]\} \\ &\quad g(v_1,\ldots,v_M)dv_1\ldots dv_M. \end{aligned} \tag{3}$$

The first line takes the expectation of the joint survivor function over the frailty terms $v_1, \ldots, v_m$. The second line follows from the assumption made in the frailty literature that, conditional on the frailty, the risks for the different causes of readmission are independent (see Gordon 2002). Thus, the joint distribution of the times to readmission from each cause, $t_1, \ldots, t_m$, decomposes into the product of the marginal survival functions, $S_m(t_m) = exp[-\int_0^{t_m} h_m(u|X, v_m)du]$. The third line follows by integrating $h_m(t|X, v_m)$ from Equation 2, and the fact that, for our Weibull formulation $\int_0^t h_{0,m}(u)du = t_m^{\rho_m}$. Recalling that in a survival model, the density function is given by $f(t) = h(t)S(t)$, from Equations 2 and 3 the unconditional density function can be calculated as:

$$\begin{aligned} f(t_1,\ldots,t_M|X) &= \int \cdots \int f(t_1,\ldots,t_M|v_1,\ldots,v_M) \\ &\quad \times g(v_1,\ldots,v_M)dv_1\ldots dv_M. \\ &= \int \cdots \int h(t_1,\ldots,t_M|v_1,\ldots,v_M) \\ &\quad \times S(t_1,\ldots,t_M|v_1,\ldots,v_M) \\ &\quad \times g(v_1,\ldots,v_M)dv_1\ldots dv_M. \\ &= \int \cdots \int \Pi_{m=1}^{M} \{\rho_m t_m^{\rho_m-1} \cdot \\ &\quad \times exp(XB_m+v_m)\ exp[-exp(XB_m+v_m)t_m^{\rho_m}]\} \\ &\quad \times g(v_1,\ldots,v_M)dv_1\ldots dv_M. \end{aligned} \tag{4}$$

The first line follows by taking the expectation over the frailties as in Equation 3. The second line follows by applying the definition of f, that is, $f(t) = h(t) \times S(t)$.

Thus, the final result is just a product of the hazard functions for each latent factor, $h_m(t_m|X, v_m)$, with the conditional survivor function again by applying the assumption of independence of risks conditional on the frailty terms.

The final step is to calculate the marginal likelihood function for estimating the unknown parameters of the model. First, as is common in health-care data, we must account for the fact that some of the data we have obtained will be censored. In the case of readmissions, a logical choice of censoring limit is 30 days, given that the current policy only penalizes readmissions within 30 days. To incorporate censoring, let δ_{im} represent the censoring indicator for the individual $i = 1, \ldots, n$ and frailty $m = 1, \ldots, M$ that is one if the data are uncensored and zero if the data are censored. The following presents the standard form of the likelihood function with censoring, $L(B,\rho,v)$, that takes the form of the density, f, when data are uncensored, and the form of the survival function, S, when the data are censored:

$$\begin{aligned}
&L(B,\rho,v|data) = \\
&= \Pi_{i=1}^{n} \int \ldots \int \Pi_{m=1}^{M} (\rho_m t_{im}^{\rho_m - 1} \cdot exp(XB_m + v_m) \\
&\quad \times exp[-exp(XB_m + v_m) t_{im}^{\rho_m}])^{\delta_{im}} \cdot (exp[-t_{im}^{\rho_m} exp(XB_m + v_m)])^{(1-\delta_{im})} \\
&\quad \times g(v_1, \ldots, v_M) dv_1 \ldots dv_M \\
&= \Pi_{i=1}^{n} \int \ldots \int \Pi_{m=1}^{M} \left\{ (\rho_m t_{im}^{\rho_m - 1} exp(XB_m + v_m))^{\delta_{im}} \right. \\
&\quad \left. \times exp(-t_{im}^{\rho_m} exp(XB_m + v_m)) \right\} g(v_1, \ldots, v_M) dv_1 \ldots dv_M.
\end{aligned} \tag{5}$$

where t_{im} is the time of readmission due to latent factor m (or the censoring time, e.g., day 30) for data record (patient) i. The population model presented in Equations 2–5 has the nice properties of being able to account for the impact of time, patient observable risk factors, and unobservable competing risks (e.g., infection, failure to thrive, dehydration, etc.) on readmissions by estimating the unknown parameters from data.

2.1.1. Parameter Estimation for Time to Readmission. The next step is to estimate the unknown parameters, specifically the parameter of the marginal baseline hazard ρ, the coefficients of our patient risk factors, B, and frailties, v. Unfortunately, these parameters cannot be effectively estimated directly using conventional likelihood maximization methods because these methods cannot directly maximize the full likelihood, given data, and the small sample properties of these estimators have yet to be studied (Ibrahim et al. 2005). To avoid these pitfalls, we transform the hazard model presented in Equations 2–5 into a Bayesian model and use Markov Chain Monte Carlo (MCMC) methods to draw parameters from their posterior distributions (Gilks and Wild 1992). Not only does the Bayesian framework enable effective parameter estimation, it also facilitates our approach to addressing data scarcity issues when personalizing the prediction to individual patients/hospitals. This framework also permits right censored data as presented in Equation 5, where censoring occurs because we observe patients at a specific point in time and if they have not yet been readmitted or have left the database we do not observe their true time to readmission, but instead observe a survival time that is smaller than their time to readmission.

We used Winbugs software for Bayesian analysis and MCMC sampling. To set up the model in Winbugs, we used a rectangular data format with separate columns to represent regular readmission and censoring times. Individuals who are censored are given a missing value in the vector of readmission times, while individuals who actually get readmitted are given a zero in the censoring time vector. The truncated Weibull model based on Winbugs built in censoring function is used to include appropriate term(s) in the full conditional distribution (similar to the model explained in Equation 5). Detailed instructions and examples on Bayesian analysis of Weibull regression in censored survival analysis using Winbugs can be found at Spiegelhalter et al. (2003).

To transform the Weibull hazard model into a Bayesian framework, we first need to define a distribution that represents our current belief about the parameters to be estimated, which is called a Bayesian prior. Then, we define a likelihood function, $L(params|data)$, that calculates the probability that the chosen parameters are a good representation of the data. Finally, we multiply the prior by the likelihood function and then normalize to obtain a posterior distribution on our set of parameters that better fits the observed data. We generate the posterior distribution by sampling the prior distribution on parameters B, ρ, and v in a Monte Carlo fashion and calculating the posterior based on the data and the result of sampling the prior using Gibbs Sampling algorithm (see Gilks and Wild 1992).

To avoid over-fitting, it is important to choose an appropriate prior distribution. In particular, we want a prior distribution in which the mode of the parameters are likely to be near zero while the variance is monotonically decreasing with positive finite value at zero (see Gustafson 1997). This discourages the model from selecting too many variables in the data fitting process. From our numerical experiments setting, the prior distribution of the risk factors, B, to be independent double exponential

random variables, using multivariate normal distribution to represent the joint distribution of the frailties, v_m, and choosing a gamma distribution for the parameter of the marginal baseline hazard, ρ_m, performed well for our data. In the next section, we show how we can employ transfer learning to parameterize our Weibull regression model described in this section with enough data and yet still personalize the readmission prediction.

2.2. Personalizing Readmission Predictions and Data Scarcity

A key feature of health-care modeling is the need to personalize prediction and forecasting models to account for individual patient characteristics because aggregate population-based models do not often perform well on a patient-by-patient basis. Personalizing time to readmission predictions also enables us to develop risk profiles that allow us to tailor our monitoring decision framework in section 3 and significantly outperform a population-based monitoring plan. To obtain a population-based estimate, we simply use all the data available in our datasets in the likelihood function, L (Equation 5), when estimating the parameters. To tailor the estimate to one specific patient, we would only use that patient's data in calculating the likelihood function. While the population estimate is not discerning enough, a single patient would not have nearly enough data points (historical admission/readmission records) to adequately estimate the large number of parameters in Equation 5. To overcome this, we use an approach called transfer learning (see Pan and Yang 2010), which includes records from statistically similar patients in the likelihood function to increase the amount of data we can use to estimate the parameters for the particular individual.

In transfer learning, we calculate the posterior distribution on the parameters we wish to estimate, B,ρ,v, using all the data but giving more influence to data records based on how similar they are to the patient we are trying to estimate the parameters for. In estimating the time to readmission density function for patient i, we have a similarity/relevance weight for patient j that is given by $w_{i,j}$. For the data from patient j, we then take the likelihood to the power $w_{i,j}$. This way, data from a higher weight for patient j, indicating they are more similar to patient i, has more influence in the likelihood function.

The weighting scheme to calculate similarity and relevance of data records includes two factors: (i) Readmission record similarity: measuring how similar two patients in the data are; and (ii) Recency of readmission records: giving a higher weight to more recent admission/readmission records.

2.2.1. Record Similarity (W_1).

To calculate readmission record similarity, ($W1$), we first divide the index of risk factors 1, …, K into the set of indices that represent numeric factors, K_1, and categorical factors, K_2. Separating the factors into numeric and categorical groups and comparing only against other factors in the same group mitigates a potential bias if $|K_2| > > |K_1|$ (where $|\cdot|$ is the cardinality of the set) that could be introduced because the numeric factors will always be less than one, while the categorical factors will be either 0 or 1. Next, we use cosine similarity measure (Pang-Ning et al. 2006) for calculating the similarity among numeric factors, and simple matching (Sokal 1958) for categorical factors. Next, we use the weighted average of the numerical factors (with weight $|K_1|$) and categorical factors (with weight $|K_2|$) to provide the total similarity measure. We normalize the numerical risk factors for patient i to fall within the interval (0,1). Similarity for categorical risk factors for patients i and j, $x_{i,n}, x_{j,n}$, for $n \in K_2$, is represented by an indicator, $\mathbb{1}\{x_{i,n}, x_{j,n}\} = 1$ if the categorical factors are identical and 0 otherwise. The total similarity measure will then be given by:

$$W_{1_{ij}} = \frac{\omega_1 W_{11_{ij}} + \omega_2 W_{12_{ij}}}{\omega_1 + \omega_2} \tag{6}$$

where $\omega_1 = |K_1|$ and $\omega_2 = |K_2|$ are the number of factors in the numerical and categorical groups, respectively. $W_{11_{ij}} = \frac{\sum_{k\in K_1} x_{i,k}\cdot x_{j,k}}{\sqrt{\sum_{k\in K_1}(x_{i,k})^2}\cdot\sqrt{\sum_{k\in K_1}(x_{j,k})^2}}$, and $W_{12_{ij}} = \frac{\sum_{k\in K_2}\mathbb{1}\{x_{i,k},x_{j,k}\}}{|K_2|}$. In other words, to calculate the overall readmission record similarity, like factors are compared with like factors and then averaged according to how many of that type of factor appears in the data. In this way, the influence of the categorical variables on the continuous variables is reduced.

2.2.2. Record Recency (W_2).

For readmission record recency, we developed a tilted time framing mechanism that is closely related to exponentially weighted moving average (EWMA) smoothing. A weighting factor of W_2 is defined based on the generalized logistic function $W_2(t) = [1 + Q\cdot exp(-A(t-\Gamma)^{\frac{1}{\eta}}))]^{-1}$, where t is the date the readmission record occurred, Γ is the current date, and A,η,Q are the parameters of the logistic functions determining the shape and scale of the function (Richards 1959). These parameters can be determined by minimizing the mean squared error (MSE) of the estimated probabilities of readmissions (in the training dataset), and the respective empirical probabilities (in the validation dataset). The total weight applied to records from patient j being used to estimate time to readmission for

patient i, $w_{i,j}$, is determined by multiplying the similarity and recency measures.

This approach can also be used at the hospital level, employing information from health systems with available data on readmission rates to predict the rate of readmissions in new medical centers or those with insufficient data. We summarize the algorithm for empirically estimating our time to readmission density function in Table 1.

2.3 Analyzing Model Performance on Real Data

To demonstrate the effectiveness of our approach, we compare its predictive power to other effective prediction methods from the literature including: (i) classification and regression trees (CART), (ii) multilayer perceptron (MLP), (iii) logistic regression, (iv) a boosting algorithm (AdaBoost), and (v) Bayesian networks. All of the benchmark models are fixed-effect models as these are the most common in the readmission prediction literature. This permits a comparison of our novel method of employing random effects to readmissions with the literature standard.

For Classification and Regression Tree (CART), we used classification trees with pruning having a minimum 10 observations for the parent nodes, at least one observation per leaf, and the same observation weight. For the Multilayer Perceptron Neural Net (MLP) we used one hidden layer with learning rate of 0.3 and momentum of 0.2. For multinomial logistic regression, we used 0.25 as the minimum significance levels for the variables to remain in the model with backward-forward model selection strategy. For Boosting, we employed ADABOOST M1 PART with Decision Stump classifier and weight threshold = 100. For Bayesian Net, we consider a simple Bayes estimator with $\alpha = 0.5$ and a hill climbing search algorithm.

Table 1 High-Level Procedure of the Readmission Prediction Framework

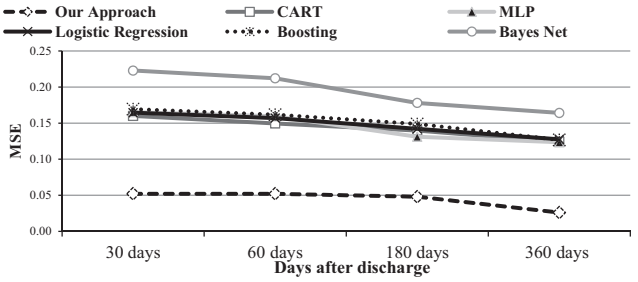
	Readmission prediction procedure
Input	Readmission data, Prior dist. of hazard function parameters and risk factor covariates
Output	Posterior dist. of hazard function parameters and risk factor covariates Empirical time to readmission density function
Procedure	1. Set prior dist. for risk factors, frailties, and marginal baseline hazard (e.g., double exponential, multivariate normal, gamma) 2. Set W2 function parameters (record recency weight) 3. **FOR** patient i in dataset 4. Calculate the total weight of all records in the database with respect to the last readmission event of patient i 5. Replicate each record according to the least common multiplier of the calculated weights 6. Find the posterior distribution of the model parameters based on the weighted dataset using Gibbs sampling 7. Calculate the posterior distribution of readmission/no readmission based on optimal parameters **NEXT PATIENT** Return Posteriors

The analysis was performed based on data from a database of 2449 patients with 3108 admission/readmission records, and 15 demographic, socioeconomic and patient health-related factors from a medical center in Michigan over the years 2006–2011; and from the SID database, though the analysis is presented only for the Michigan hospital since the results and insights were similar for the SID database. After initial analysis, the following nine variables were incorporated into the model (several factors were eliminated): length of stay, gender, age, employment status, insurance coverage level, profession/military rank, ward(s) visited during inpatient stay, principal diagnosis, and source of admission, that is, VA hospital, nursing home, home, non-VA hospital. We used threefold cross validation for evaluating model performance. That is, we divided the data into three separate datasets, one for training, one for validation, and one for testing; repeating this procedure three times. For each of the three repetitions of the threefold cross validation, the data were randomly divided with 60% for training, 20% for validation, and 20% for testing.

We used an iterative optimization process (in our case simulated annealing) to determine the optimal value of the parameters for the weight function, $W_2(t)$, based on the validation dataset. The objective function of the simulated annealing algorithm was taken as the MSE of the estimated probabilities of readmission for patients in the validation set using the model parameterized in the training dataset vs. the respective empirical probabilities in the validation dataset. However, weighting parameters complicate the likelihood function and consequently affect the efficiency of the Bayesian updating procedure. To deal with this, we employ a simple, yet effective technique of replicating admission/readmission records based on their weights. For example, if there are two records where one of them has weight two, while the other is a regular record with weight one, we may replicate the first one twice and leave the last one without replication. The appropriate number of replications for each record in more complex scenarios can be achieved by using the least common multiple of the weights.

Figure 3 presents the MSE in predicting individual patient's probability of readmission based on application to the testing data in the threefold cross-validation using approximately 620 records (20% of the total 3108 readmission records). Four time snapshots (30, 60, 180, and 360 days after discharge) were used to

Figure 3 Mean Squared Error (MSE) of the Comparing Methods



compare our model against its counterparts. It is clear that our method significantly outperforms other leading prediction methods for all time horizon predictions.

3. Stage 2: Tactical and Operational Planning Model for Design and Optimization of Post-Discharge Follow-Up Planning

Recent studies have showed that targeted post-discharge telephonic outreach, telemedicine follow-ups, and dedicated discharge management teams can reduce the number of readmissions (e.g., Cardozo and Steinberg 2010, Melton et al. 2012, de Toledo et al. 2006). Our clinical collaborator and medical colleagues have also expressed an interest in more effective monitoring approaches employing a variety of methods including phone calls and office visits. In this section, we design a tactical planning model that integrates foreknowledge of patient readmission rates from section 2 with an optimization model to determine (i) what days a follow-up organization (clinic, telemedicine, etc.) should be open, (ii) how many resources to allocate to each specialty/type of follow-up, and (iii) when to schedule discharged patients for a follow-up.

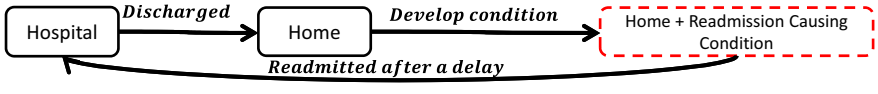
3.1. A Delay-Time Model of the Post-Discharge Patient Health Condition

The accurate empirical forecast of patient readmission timing from section 2 paves the way for improved planning and scheduling of at risk patients to avoid or mitigate readmissions before they occur. In particular, the forecasting models form the basis of a delay-time model of patient readmission dynamics. Delay-time models have been employed extensively in machine maintenance literature (e.g., Keller 1974, Luss 1976) but are not as common in health-care applications. Let $\mathcal{J}$ be the set of patient types (e.g., bladder cancer, hip replacement, etc.), and $p_j(t)$ be the empirical probability distribution developed in section 2 representing the time from when a patient of type $j \in \mathcal{J}$ is discharged until they are readmitted. In addition to $p_j(t)$, we also consider how long the patient had a detectable condition before getting readmitted to the hospital, and thus presented an opportunity to detect the condition before it triggered an emergency readmission. To do so, we consider that after a type j patient is discharged they may develop a condition at a later time that will eventually result in readmission. Once they develop a condition, there is a delay, which we call $D_j \sim F_j(t)$, from the time the patient develops the condition to the time they get readmitted to the hospital. If a patient has an "inspection"—for example, follow-up office visit or phone call—after they develop the condition but before they are readmitted then the doctor can take steps to avoid or mitigate the readmission. Figure 4 depicts the readmission dynamics at a high level.

Traditional delay-time models start with a distribution on the time that a condition develops and combine it with the delay until that condition causes a readmission. This study takes the reverse approach, since time to readmission is directly observable from the data, whereas the time when the patient develops the condition is not. Thus, the methods in section 2 are capable of directly calculating the distribution on the timing of a readmission and not the time when the patient developed the condition. Thus, we directly employ the time to readmission curve from the estimation model and use the delay-time formulation to calculate the time when the condition develops.

Let T be the length of the planning horizon (e.g., 30 days after discharge) and let $\mathcal{T} = \{0, 1, 2, \ldots, T\}$ be the set of days in the planning horizon, where the patient's discharge day is 0 (i.e., zero days after discharge). The probability that a patient has developed a condition but not yet been readmitted by t_1 days after discharge (i.e., the condition is detectable) is given by $\int_{t_1}^{T} p_j(s)[1 - F_j(s - t_1)]ds$. The other important dynamic in the readmission system is that the number of discharges and health-care resource capacities vary by day of week, leading to a weekly repeating cycle of $d_0 = 0, \ldots, 6$ for days of the week in the discharge cycle and $d_1 = 0, \ldots, 6$ for days of the week in the resource capacity cycle. For the rest of the

Figure 4 Post-Discharge Patient Flow with Readmissions



study, we will use t_i to refer to the patient's 30 day planning horizon after discharge, and d_i to refer to the cycle of the health-care system (discharges or capacity). Letting $Y^j_{d_0}$ be the random variable for the number of type j patients who are discharged on day d_0 of the planning horizon, we can show that the number of patients that have developed a condition, but have not yet been readmitted (I) by day t_1, $N^j_{I,d_0}(t_1)$, has a Binomial mixture distribution with mean

$$\mathbb{E}[N^j_{I,d_0}(t_1)] = \mathbb{E}[Y^j_{d_0}] \int_{t_1}^{T} p_j(s)(1 - F_j(s - t_1))ds. \quad (7)$$

It is easy to show that, for discharges that follow a non-homogeneous Poisson process, this distribution follows a Poisson distribution.

Now we can calculate the number of readmissions of type j patients who were discharged on day d_0 that are averted or caught early (A) by scheduling a follow-up on day t_1, $N^j_{A,d_0}(t_1)$. This depends on the type of inspection and how well it can detect a readmission causing condition. The set of inspection types is given by $\mathcal{R}$. In our case, we consider one type of perfect and imperfect inspection corresponding to an office visit or a phone call, respectively (i.e., $\mathcal{R} = \{per, imp\}$). These are two options that are currently in use in an ad hoc manner and are being considered for implementation by our clinical co-author. Other options include telemedicine and home visits.

Let r_k be the probability of detecting a condition using follow-up method $k \in \mathcal{R}$ (for perfect inspection $r_{per} = 1$ and imperfect $r_{imp} < 1$). To calculate the probability of detecting a condition at the n + 1st follow-up, first note that it is only necessary to know the history of follow-ups following the last perfect inspection. This is because a perfect inspection "wipes the slate clean" by detecting all conditions that have occurred and not yet caused a readmission prior to that inspection. Thus, let t_0 be the time of the most recent perfect inspection before time t with $t_0 = 0$ if there has not yet been a perfect inspection. If we let $\mathcal{P}(\mathcal{T})$ be the power set of $\mathcal{T}$ minus the null set (since the null set lacks meaning in our context), then we can represent the history of imperfect inspections for a patient that had their last perfect inspection at time t_0 as $\tau \in \mathcal{P}(\mathcal{T}) = \{t_0, t_1, t_2, \ldots, t_n : t_i < t_{i+1},\ i = 0, \ldots, n-1\}$. We define the set of feasible actions that can follow a particular history, τ, as

$$\begin{aligned}\mathcal{A}(\tau) = &\{(per, t) : T \geq t > \max_{s\in\tau} s\} \\ &\cup \{(imp, t) : T \geq t > \max_{s\in\tau} s\} \cup \{End\}\end{aligned} \quad (8)$$

which represents all possible timing/type combinations for the next follow-up after the sequence of follow-ups, τ, where (imp,t) and (per,t) represent an imperfect and perfect inspection, respectively, at time t and "End" represents the action of doing no more follow-ups for the remainder of the planning horizon. For purposes of exposition, we allow for a slight abuse of notation by defining $\ddot{t}(a)$ and $\ddot{k}(a)$ to be the time of the inspection associated with action a (e.g., t_{n+1} from above) and the type of inspection associated with action a (e.g., perfect or imperfect), respectively. Further, let $\ddot{n}(\tau)$ be the number of imperfect inspections in history τ and $\tilde{t}_i(\tau)$ be the time of the ith imperfect follow-up in history τ for $i = 1, \ldots, \ddot{n}(\tau)$ and $\tilde{t}_0(\tau)$ be the time of the most recent perfect inspection follow-up. We use the convention (with abuse of notation) that $\tilde{t}_0(\tau) = -\infty$ if there has not yet been a perfect inspection. It is important to note that $\tilde{t}_{\ddot{n}(\tau)}(\tau)$ represents the time of the most recent follow-up in the follow-up history τ, as this particular follow-up will be referred to frequently. The probability of detecting a condition in a type j patient at the n + 1st follow-up, given history τ and action $a \in \mathcal{A}(\tau)$ is given by

$$\begin{aligned}\rho_j(a,\tau) = r_{\ddot{k}(a)} \Bigg(&\int_{s=\ddot{t}(a)}^{T} p_j(s)[F_j(s - \tilde{t}_{\ddot{n}(\tau)}(\tau)) \\ &- F_j(s - \ddot{t}(a))]ds \\ &+ \sum_{i=0}^{\ddot{n}(\tau)-1} \int_{s=\ddot{t}(a)}^{T} (1 - r_{imp})^{i+1} p_j(s) \\ &[F_j(s - \tilde{t}_{\ddot{n}(\tau)-i-1}(\tau)) - F_j(s - \tilde{t}_{\ddot{n}(\tau)-i}(\tau))]ds \Bigg)\end{aligned} \quad (9)$$

The multiplier outside the parenthesis accounts for the detection rate of the type of follow-up indicated by action a. Inside the parenthesis, the first term accounts for all the conditions that have developed since the last inspection (that occurred at time $\tilde{t}_{\ddot{n}(\tau)}(\tau)$) and hence could not be detected at prior inspections. The second term accounts for the probability of detecting all conditions that were (i) first detectable during the imperfect inspection at time $\tilde{t}_{\ddot{n}(\tau)-i}(\tau)$, but (ii) were not detected at that follow-up or any of the subsequent i imperfect inspection follow-ups (hence $(1 - r_{imp})^{i+1}$) because of the failure of the imperfect inspection, and (iii) were not readmitted before time $\ddot{t}(a)$. Thus, they would have a chance to be detected at the inspection at time $\ddot{t}(a)$ indicated by action a. The sum over i adds one term for each imperfect inspection that has occurred since the most recent perfect inspection. The term where $i = \ddot{n}(\tau) - 1$ is slightly different because $\tilde{t}_0(\tau)$ represents the beginning of the string of imperfect inspections, being either the most recent perfect inspection ($\tilde{t}_0(\tau) > 0$) or when the patient was discharged from the hospital if there has not yet been a perfect inspection ($\tilde{t}_0(\tau) = -\infty$). In the

case where $\tilde{t}_0(\tau) = -\infty$ (no perfect inspection yet), then $F_j(s - \tilde{t}_0(\tau)) = 1$, which corresponds to the fact that failures only begin to arrive after the patient is discharged (since the discharge process is considered a perfect inspection). Thus, $1 - F_j(s - \tilde{t}_1(\tau))$ represents conditions that were first detectable at the first imperfect inspection and have not yet caused a readmission.

Equation 9 gives the probability of detecting a readmission causing condition for an individual patient. We now link this detection probability to the decision of how many appointment slots to reserve for following up with a cohort of patients that are discharged from the hospital. With $a \in \mathcal{A}(\tau)$ being the timing and type of the next follow-up we define this decision variable as $\Theta^{j,a}_{d_0,\tau}$, which is the number of phone call slots (if $\ddot{k}(a) = imp$) or office visit appointments (if $\ddot{k}(a) = per$) to reserve on day $\ddot{t}(a)$ for type $j \in \mathcal{J}$ patients who were discharged on day $d_0 = 0, \ldots, 6$ and have a follow-up history of $\tau \in \mathcal{P}(\mathcal{T})$. Each follow-up has the opportunity of detecting a condition before it causes an emergency readmission. The result may still be a readmission, but a less costly one. Our assumption is that detecting the condition before it becomes an emergency readmission averts some or all of the cost of an emergency readmission for that patient. This is what we mean by *averting a readmission*. Letting $\mathbb{1}$ be the indicator function, and recalling that $Y^j_{d_0}$ is the number of patients of type j discharged on day d_0, the number of additional readmissions averted by appointment allocation $\Theta^{j,a}_{d_0,\tau}$ is given by

$$N^{\Theta^{j,a}_{d_0,\tau}}_{A,j} = \sum_{\ell=1}^{Y^j_{d_0} \wedge \Theta^{j,a}_{d_0,\tau}} \mathbb{1}\{\text{Condition Detected at time } t_{n+1} \text{ for Patient } \ell\}, \tag{10}$$

where $\wedge$ represents the minimum operator. As before, this also follows a Binomial mixture, which can be seen by conditioning on $Y^j_{d_0}$ and applying the law of total probability. Further, it can be shown that

$$\mathbb{E}[N^{\Theta^{j,a}_{d_0,\tau}}_{A,j}] = \mathbb{E}\left[Y^j_{d_0} \wedge \Theta^{j,a}_{d_0,\tau}\right] \rho_j(a,\tau) \tag{11}$$

Because the Θ's capture both the follow-up day, $\ddot{t}(a)$, and the discharge day, d_0, the optimization will reveal not only the capacity to allocate to follow-ups on each day $d_1 = 0, \ldots, 6$ (by summing the decision variable over $\{\ddot{t}(a) : (\ddot{t}(a) + d_0) \bmod 7 = d_1\}$) but also the optimal timing of each follow-up for a patient of type j determined by the history τ for all non-zero Θ's. The following sections will analyze the mix and volume of follow-up types to determine how these different methods (e.g., phone, office visit, etc.) can be used most effectively in practice. The goal is (i) to develop an optimization model that identifies optimal placement of both types of follow-up, (ii) understand the structure of the placement of these follow-ups to provide heuristics or rules of thumb that could be employed in designing post-discharge follow-up plans even in the absence of optimization.

3.2. Optimal Design of a Post-Discharge Monitoring Organization

While there is a rich literature on delay-time models in machine maintenance, the optimization model we develop adds new system-level decision-making capability not previously considered by allowing the optimization to consider not only the timing of inspections but also how much capacity to reserve for those inspections over a planning horizon. Adding to the complexity, each day has a stochastic number of discharges (potential patients to follow up with) with a different distribution for each day of the week as well as a different cost for scheduling follow-ups by day of week. Finally, we develop a complex cyclostationary equilibrium model with parameters varying over a seven-day horizon, but allowing patients to be scheduled for inspection up to 30 days after their discharge. Next, we present the notation followed by the model formulation.

Let $c^{j,k}_{d_1}$ be the cost per appointment of type $k \in \mathcal{R}$ for patient type j on day $d_1 = 0, \ldots, 6$ (i.e., Sun-Sat). $u_{j,a}$ is the usage requirement of follow-up resource indicated by action $a \in \tau$ for a type j patient (e.g., some patients need 15-minute slots, whereas others require 20- or 30-minute slots). β_j is the average benefit of averting a readmission by early detection for a type j patient. Let $\tilde{C}_{d_1,k}$ be the maximum capacity of a type k follow-up resource that can be reserved on day d_1. To capture the cyclically time-varying cost of reserving capacity for follow-ups, we define a function that maps the follow-up day, t (on the scale of $\mathcal{T}$), after a discharge on day $d_0 \in \{0, \ldots, 6\}$, to the day of week the appointment would occur and corresponding cost:

$$\hat{c}_j(a, d_0) = c^{j,\ddot{k}(a)}_{(d_0 + \ddot{t}(a)) \bmod 7}. \tag{12}$$

Finally, we need notation to capture the day of the week that capacity is being reserved on to account for the costs that differ by day. Let $\mathcal{M}(d_0, d_1) = \{t \in \mathcal{T} : (t + d_0) \bmod 7 = d_1\}$ be the set of possible follow-up days of a patient discharged on day d_0 that map to the same day of the week, d_1. In this set, d_0 is that day of the week of discharge (e.g., $d_0 = 0$ represents a Sunday discharge, …, $d_0 = 6$ represents a Saturday discharge) and $t + d_0$ represents the (relative) date t days after

discharge. Therefore, $(d_0 + \ddot{t}(a)) \bmod 7$ returns the day of the week for that date that is t days after the patient's discharge. $d_1 \in \{0, \ldots, 6\}$ represents a day of the week (e.g., $d_1 = 1$ is Monday), so if $(t + d_0) \bmod 7 = d_1$, then this implies that t days after discharge day d_0 (e.g., Friday) is the d_1 (e.g., Monday) day of the week. Thus, $\mathcal{M}(d_0, d_1)$ is all the days within the planning horizon (e.g., $t = 1, \ldots, 30$) that correspond to the day of week d_1. For example, if $d_0 = 5$ (i.e., Friday) and $d_1 = 1$ (i.e., Monday), $\mathcal{M}(d_0, d_1) = \{3, 10, 17, 24\}$ as 3 days after Friday is Monday, 10 days after Friday is also Monday, etc. PROGRAM 1 provides an optimal design for staffing and scheduling of a follow-up organization.

PROGRAM 1:

$$\min_{\Theta} \mathbb{E}\left[\sum_{j\in\mathcal{J}}\sum_{d_0=0}^{6}\sum_{\tau\in\mathcal{P}(\mathcal{T})}\sum_{a\in\mathcal{A}(\tau)} \hat{c}_j(a,d_0)\Theta^{j,a}_{d_0,\tau}\cdot u_{j,a} - \beta_j N_{A,j}^{\Theta^{j,a}_{d_0,\tau}}\right] \tag{13}$$

s.t.

$$\tilde{C}_{d_1,k} \geq \sum_{j\in\mathcal{J}}\sum_{d_0=0}^{6}\sum_{\tau\in\mathcal{P}(\mathcal{T})}\sum_{a\in\mathcal{A}(\tau):\ddot{k}(a)=k,\ddot{t}(a)\in\mathcal{M}(d_0,d_1)} \Theta^{j,a}_{d_0,\tau}\cdot u_{j,a} \quad \forall d_1 \in \{0,\ldots,6\}, k \in \mathcal{R} \tag{14}$$

$$\sum_{a\in\mathcal{A}(\tau)} \Theta^{j,a}_{d_0,\tau} \leq \Theta^{j,(imp,\tilde{t}_{\ddot{n}(\tau)}(\tau))}_{d_0,\tau\setminus\{(imp,\tilde{t}_{\ddot{n}(\tau)}(\tau))\}} \quad \forall j \in \mathcal{J}, d_0 = 0,\ldots,6, \tau \in \mathcal{P}(\tau) : |\tau| > 1 \tag{15}$$

$$\sum_{\tau\in\mathcal{P}(\tau):\tilde{t}_{\ddot{n}(\tau)}<t_1} \Theta^{j,(per,t_1)}_{d_0,\tau} \geq \sum_{k\in\mathcal{R}}\sum_{t_2=t_1+1}^{T} \Theta^{j,(k,t_2)}_{d_0,\{t_1\}} \quad \forall j \in \mathcal{J}, d_0 = 0,\ldots,6, t_1 = 2,\ldots,T-1 \tag{16}$$

The objective function minimizes the cost of staffing the follow-up clinic minus the cost of readmissions averted. Equation 14 ensures that capacity constraints for each resource are respected. This is complicated by the fact that we are considering a system that functions on a weekly repeating cycle. Thus, we have different costs and discharge distributions for each day of the week, but the pattern repeats each week. However, patient health status does not evolve on a weekly repeating cycle but instead plays out over a longer (non-repeating) horizon, which we denote by $\mathcal{T}$ and can be thought of as the 30-day readmission window for example. In order to ensure that enough capacity is allocated on a given day of the week, take Monday, for example, we must account for patients that are scheduled the first Monday after their discharge, as well as those that will be scheduled the second Monday after their discharge, and the third Monday after their discharge and so forth. This is the reason for choosing the follow-up day such that $(\ddot{t}(a) + d_0) \bmod 7$ is equal to the day whose capacity is the focus of the constraint. We also consider all patients who were discharged on different days of the week (the sum over d_0) since the discharges vary by day of week in the cyclostationary health-care system.

Finally, Equations 15 and 16 are critical constraints linking sequential follow-ups together, ensuring that the follow-up time sequence follows the correct pattern. In particular, it should only be possible to schedule as many follow-ups at time t_1 in the sequence $\tau \cup \{t_1\}$ as the most follow-ups that were scheduled in any previous visit in τ. This is because the probability of detection for each patient is dependent on all the previous inspection times, and so to obtain the correct detection probability it is necessary that each patient being scheduled in a particular time sequence has taken every visit in the sequence. Equation 15 ensures that the number of inspections scheduled following a history τ (LHS of the equation) is no more than the number of inspections scheduled in the most recent imperfect inspection in τ (RHS of the equation). Equation 16 ensures the same for perfect inspections. The reason we need two separate constraints to capture this criterion is because a perfect inspection "resets" the inspection history, and thus must be treated differently.

The stochastic optimization model in PROGRAM 1 suffers from an extremely large state space, action space, and number of constraints. Both the number of decision variables and number of constraints in Equation 16 are exponential in the length of the planning horizon because the probability of detection at each visit depends on the history of all imperfect inspections since the last perfect inspection. Further, the fact that the number of discharges on a given day can be stochastic makes $\mathbb{E}[N_{A,j}^{\Theta^{j,a}_{d_0,\tau}}]$ a non-linear function of the decision variable, $\Theta^{j,a}_{d_0,\tau}$.

3.2.1. Computational Challenges. It is easy to see that the number of constraints represented by Equation 15 is exponential because of $\mathcal{P}(\mathcal{T})$. Proposition 1 shows that the number of decision variables in PROGRAM 1 is also exponential in the length of the planning horizon.

PROPOSITION 1. *Let T be the length of the planning horizon. The number of decision variables in PROGRAM 1 is $|\mathcal{J}||\mathcal{R}| \cdot 7(2^{T+1} - (T+2))$.*

PROOF. At each time point, t, in the planning horizon there are two choices: schedule a perfect inspection or an imperfect inspection. The outcome of the

choice depends on the history of imperfect inspections and the time of the most recent perfect inspection prior to t. If the last perfect inspection were at time $t-1$, there is only one possible history (i.e., 2^0) that has an inspection at time t. If the last perfect was at time $t-2$, there are two possible (i.e., 2^1) histories *per*, *imp* or *per*, *None*. If the last perfect inspection was at time s then there are 2^{t-s-1} possible histories. If there was no previous perfect inspection, there are 2^{t-1} histories. Thus, at time t, there are a total of $\sum_{s=0}^{t-1} 2^s = 2^t - 1$ possible histories and there are $|\mathcal{R}|$ possible actions at time t (in our case $|\mathcal{R}| = 2$: perfect or imperfect inspection). Thus, there are $|\mathcal{R}|(2^t - 1)$ decision variables at time t. Summing this over all time points in the planning horizon yields $|\mathcal{R}| \sum_{t=0}^{T}(2^t - 1) = |\mathcal{R}|(2^{T+1} - (T+2)$. Finally, this pattern repeats for each patient type (of which there are $|\mathcal{J}|$) and days in the cost/discharge cycle ($d = 0, \ldots, 6$). □

For a 30-day planning horizon, the number of decision variables significantly exceeds commercial solver limits. To handle the problem size, we introduce a new, network flow-based method for transforming the stochastic optimization problem into a tractable linear deterministic one in the next section. We show that this transformed problem is actually a weakly coupled network flow problem (defined in section 3.3), which can be decomposed into several independent networks with a few linking constraints. Not only does the network flow formulation allow for much faster solution approaches, the network flow structure inherently captures constraints of Equations 15 and 16, leaving only a small number of linking constraints (Equation 14) that weakly couple otherwise independent networks for each patient type and day of week.

Finally, by applying pruning methods to our decomposed network models, we are able to solve even large problems very effectively. This allows us to design complete patient follow-up monitoring schedules as well as staffing plans for a realistically sized follow-up organization. The development of such techniques has the potential to impact many areas by providing a methodology for solving multi-dimensional, large state space stochastic optimization problems that are common to the health-care domain and elsewhere.

3.3. Network Flow-Based Transformation of the Stochastic Optimization Model

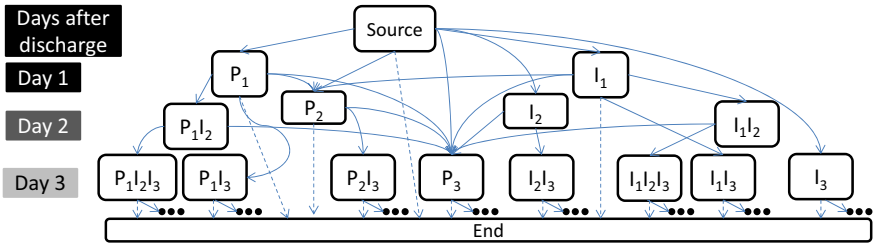
Methods in the literature have been developed for transforming these stochastic queueing network problems into deterministic ones that admit tractable optimization methods. Unfortunately, prior transformations in this vein of literature (such as Helm and Van Oyen 2014, Helm et al. 2013, and Deglise-Hawkinson et al. 2013) fail when applied to solve reasonably sized post-discharge monitoring problems. After applying previous transformation methods, we could solve problems in a reasonable time for a 15-day planning horizon, as shown in Figure 6, whereas our goal is to develop an optimal schedule to reduce 30-day readmissions.

In this section, we show that our problem has a special structure that allows us to decompose our problem into a set of independent network flow models with a small number of linking side constraints. Similar to the naming convention of Gocgun and Ghate (2012), we call this a weakly coupled network flow problem. We develop methods to solve large problem sizes that are intractable for general linear programming representations of the readmission problem that do not exploit the weakly coupled network structure. Our particular model of post-discharge monitoring exploits the fact that we can decompose the monitoring problem along days of the repeating discharge and follow-up cost cycle (e.g., days of the week), along patient types, and along probabilistic sample paths in the case of stochastic discharges. If there are n patient types and a maximum of m possible discharges on any given day then we would solve $7 \cdot m \cdot n$ smaller networks (each of which typically solves in seconds) instead of one large network (that is too large to input into commercial solvers). Using Lagrangian relaxation on the linking constraints, we can decouple these $7 \cdot m \cdot n$ networks and solve them in parallel, allowing for much larger problems to be solved. In fact, we are able to solve the 30-day horizon problem relatively quickly, even though the number of potential decisions is extremely large.

We first present the network model with deterministic discharges and then extend the model to incorporate stochastic discharges. To do so, we show that the non-linear stochastic objective, the expected number of readmissions averted for any given capacity limit (i.e., $\mathbb{E}\left[Y^j_{d_0} \wedge \Theta^{j,a}_{d_0,\tau}\right]$ from Equation 11), can be calculated exactly using a new stochastic branching method that we develop. This method maintains the weakly coupled network structure and allows us to decompose the problem and tractably solve a set of smaller network flows with side constraints. A further convenient feature of our stochastic branching method is that the number of "stochastic" branches taken determines the optimal amount of capacity to reserve on each day.

3.3.1. Deterministic Discharges. We begin by describing the individual decoupled networks, one for each discharge day and patient type, which are the building blocks of the full, weakly coupled

Figure 5 Decoupled Network Formulation Representing PROGRAM 1 for One Discharge Day and Patient Type and Three Follow-Up Days. Several copies of this network will be solved in parallel with linking constraints moved to the objective by taking the Lagrangian



network formulation. A truncated example of a decoupled network is depicted in Figure 5.

In this figure, P_t and I_t represent a perfect or imperfect inspection, respectively, on day t after discharge. Each node in the network represents reserving capacity for a particular inspection on a particular day, given a specified history of inspections up to that point. For example, node $P_2I_4I_7$ represents scheduling an imperfect inspection on the 7th day after discharge, given that the most recent perfect inspection was scheduled on day 2 and there was also an imperfect inspection scheduled on day 4. The amount of flow that passes through a node represents the amount of capacity to reserve on that day for that type of inspection.

Each arc adds a new inspection (time and type) to a previous sequence of inspections (the node the arc originates from), by connecting the originating node with a node that adds the new inspection to the previous sequence (if the inspection is imperfect) or starts a new sequence (if the inspection is perfect). We denote this concatenation/revision by the operator $\oplus$. Node τ represents inspection history $\tau \in \mathcal{P}(\mathcal{T})$ up to time $\tilde{t}_{i(\tau)}(\tau)$. Each arc from node τ represents an action, $a \in \mathcal{A}(\tau)$. Hence, from node τ, arcs connect to all future times, $t > \tilde{t}_{i(\tau)}(\tau)$, and all future inspection sequences that can occur starting from the sequence τ. For patient type j, the cost of each arc is the cost of having an inspection on day t (if the arc enters a node associated with inspection type k on day t) minus the expected reward of averting a readmission. The expected reward that makes up part of the arc cost, $\mathbb{E}[\beta_j N_{A,j}^{\Theta^{j,a}_{d_0,\tau}}]$ calculated from Equation 11, depends on the history of inspections since the most recent perfect inspection (the node the arc originates from), and the timing and type of the new inspection (the node the arc terminates at). Thus, arc $(\tau,\tau\oplus a)$ for a type j patient who was discharged on day d_0 would have cost

$$\delta^{j,d_0}_{\tau,\tau\oplus a} = \hat{c}_j(a,d_0) - \beta_j \rho_j(a,\tau), \tag{17}$$

where $\hat{c}_j(a,d_0)$ is given by Equation 12 and $\rho_j(a,\tau)$ is given by Equation 9. We now present the weakly coupled network flow formulation, where a small number of side constraints couple the otherwise independent networks described above. Let x^{j,d_0}_{iz} be the decision variable for the number of type j patients on day d_0 moving from inspection history i to inspection history $z = i\oplus a$ for some action a. Because the arc cost depends on all three parameters—patient type, inspection history, day of discharge—this actually represents an arc from node (j,d_0,i) to node (j,d_0,z). We differentiate arcs for each patient type and discharge day in our notation to highlight how the problem is weakly coupled and how to decompose the full problem into smaller subproblems. b^{j,d_0}_τ is the net traffic demand for node (j,d_0,τ), where b^{j,d_0}_n is equal to the number of discharges at the source and sink nodes for patient type j and discharge day d_0 and is zero for all other nodes. With $|S|$ being the cardinality of set S, for $|\mathcal{J}| = 2$ and a discharge day cycle of length 7, there are $7{\cdot}2 = 14$ source nodes and sink nodes. Finally, we define the set of inspection history transitions as $A = \{(\tau, \tau\oplus a) :\ \tau \in \mathcal{P}(\mathcal{T}), a \in \mathcal{A}(\tau)\}$.

Network Flow Formulation of PROGRAM 1

$$\min_x \sum_{j\in\mathcal{J}} \sum_{d_0=0}^{6} \sum_{(i,z)\in A} x^{j,d_0}_{iz}\,\delta^{j,d_0}_{iz} \tag{18}$$

s.t.

$$\sum_{(i,n)\in A} x^{j,d_0}_{in} - \sum_{(n,z)\in A} x^{j,d_0}_{nz} = b^{j,d_0}_n \qquad \forall n \in \mathcal{P}(\mathcal{T}), j \in \mathcal{J}, d_0 = 0,\ldots,6 \tag{19}$$

$$\sum_{j\in\mathcal{J}} \sum_{d_0=0}^{6} \sum_{t\in\mathcal{M}(d_0,d_1)} \sum_{\{\tau\in\mathcal{P}(\mathcal{T}):\tilde{t}_{i(\tau)}(\tau)<t\}} x^{j,d_0}_{\tau,\{t\}} \le \tilde{C}_{d_1,per} \qquad \forall d_1 = 0,\ldots,6 \tag{20}$$

$$\sum_{j\in\mathcal{J}} \sum_{d_0=0}^{6} \sum_{t\in\mathcal{M}(d_0,d_1)} \sum_{\{\tau\in\mathcal{P}(\mathcal{T}):\tilde{t}_{i(\tau)}(\tau)<t\}} x^{j,d_0}_{\tau,\tau\cup\{t\}} \le \tilde{C}_{d_1,imp} \qquad \forall d_1 = 0,\ldots,6 \tag{21}$$

$$0 \le x^{j,d_0}_{iz} \le UB^j_{iz} \quad \forall (i,z) \in A, j \in \mathcal{J}, d_0 - 1,\ldots,6$$

Equation 18 is the objective, which is a min-cost flow. Equation 19 is the flow conservation constraint. Equation 20 is the capacity constraint for perfect inspections. Every time there is a perfect inspection at time t, it wipes the information set clean except for the perfect inspection at time t. Hence, the arc $x^{j,d_0}_{\tau,\{t\}}$ represents a perfect inspection at time t given a history of τ. t is chosen so that it falls on day of the week d_1 (see the sum over $t \in \mathcal{M}(d_0, d_1)$. To capture every possible perfect inspection at time t, we also sum over all possible imperfect inspection histories in which all the inspections occur before t, given by $\{\tau \in \mathcal{P}(\mathcal{T}) : \tilde{t}_{\ddot{n}(\tau)}(\tau) < t\}$, and recalling that $\tilde{t}_{\ddot{n}(\tau)}(\tau)$ is the most recent inspection (largest inspection time) in τ. Finally, there is a sum over all possible discharge days, d_0, since patients discharged on any of the discharge days in the cycle may be scheduled for day d_1 and use up some of the capacity. Equation 21 accomplishes the same purpose except for imperfect inspections. All components are the same as Equation 20, except in the index of the decision variable, the information set is changed by appending the new inspection at time t to the former history τ; hence the arc $x^{j,d_0}_{\tau,\tau\cup\{t\}}$ goes from inspection history τ to inspection history $\tau \mathrm{U} \{t\}$ instead of erasing the old history as with the perfect inspection.

3.3.1.1. Side Constraint Elimination and Network Decomposition: Next, we discuss an efficient solution method that is capable of solving large-scale instances of the weakly coupled Network Flow Formulation of PROGRAM 1. Inspection of the problem structure reveals that the constraint set is block diagonal, with the blocks linked together by a relatively small number of the system capacity constraints defined in Equation 20 and 21. Further, each block is an independent set of network flow constraints. Fortunately, there are only a few of these side constraints relative to the number of constraints that fit the network flow structure. By taking the Lagrangian relaxation of these constraints it is possible to iteratively solve pure network problems and use the subgradient method until the algorithm converges to an acceptable tolerance (see Fisher 1985, Geoffrion 1974). Further, Equations 20 and 21 are the only constraints that link patient types and discharge days to one another because of the sum over $j \in \mathcal{J}$ and $d_0 = 0, \ldots, 6$. Once these constraints are taken into the objective function, the full network problem can now be decomposed into subproblems with one pure network subproblem for each patient type and discharge day. The subproblem looks the same as the original except with a fixed j and d_0 in Equation 19 and in the objective as well as the Lagrangian of the side constraints.

By taking advantage of this problem structure, we are able to solve large-scale problems to optimality in most cases, and near optimality in the remaining few cases. Specifically, we relax the capacity constraints Equations 20 and 21 into the objective function with a corresponding vector υ of Lagrange multipliers. The vector υ has $|\mathcal{R}| \cdot 7$ elements (where in our case $|\mathcal{R}| = 2$, corresponding to perfect and imperfect inspection types), one multiplier per capacity constraint in Equations 20 and 21. The resulting relaxed problem then further decomposes into $7 \cdot |\mathcal{J}|$ independent network flow subproblems, one for each patient type and discharge day of week, that can be solved very quickly using subgradient optimization (see Held et al. 1974) to search for the optimal multipliers.

3.3.2. Stochastic Discharges Network Model. In this section, we extend the deterministic model of section 3.3.1 to consider the case of a stochastic number of patients being discharged from the hospital. Unfortunately, the presence of stochastic discharges not only destroys the network structure, but it also eliminates the linearity of the model. From the network standpoint, with deterministic discharges the number of patients scheduled was the same as the number discharged. With stochastic discharges, you must pay for the capacity reserved but you only get benefit for the number of patients that actually get scheduled. Thus, the concept of arcs from the deterministic model has no meaning in the stochastic case because flow means something different for cost (amount of slots reserved) and for benefit (number of patients seen times benefit). The capacity reservation remains linear in amount of appointment slots reserved but from Equation 11 the benefit from capacity reserved is

$$\mathbb{E}\left[Y^j_{d_0} \wedge x^{j,d_0}_{\tau,\tau\oplus a}\right]\rho_j(a,\tau) = \rho_j(a,\tau)\sum_{n=0}^{x^{j,d_0}_{\tau,\tau\oplus a}} \mathbb{P}(Y^j_{d_0} \geq n), \quad (22)$$

which is no longer linear, nor even convex. To overcome this challenge and restore both linearity and the network structure, we employ a stochastic branching method we call *sample path decomposition*.

3.3.2.1. Sample Path Decomposition: To incorporate stochastic discharges, we further decompose the network along sample path realizations and then send exactly one unit of flow through each subproblem. Let ω_n be the realization of the discharge random variable $Y^j_{d_0}$ that corresponds to the number of discharges being larger than or equal to n. Thus, $\mathbb{P}(\omega_n) = \mathbb{P}(Y^j_{d_0} \geq n)$. If the maximum possible number of discharges is M, then we decompose each deterministic subproblem (with a source supply of $b^{j,d_0}_0 \geq 1$) into M subproblems with a source supply of

$b_0^{j,d_0,\omega_n} = 1$ for $n = 1, \ldots, M$. Then, the cost of each arc (similar to Equation 17) is given by

$$\delta_{\tau,\tau\oplus a}^{j,d_0,\omega_n} = \hat{c}_j(a,d_0) - \beta_j \mathbb{P}(\omega_n)\rho_j(a,\tau), \qquad (23)$$

Now x_{iz}^{j,d_0,ω_n} is the decision variable indicating whether the nth patient of type j discharged on day d_0 moves from inspection history i to inspection history $z = i \oplus a$ for some action a. The objective becomes $\min_x \sum_{j\in\mathcal{J}} \sum_{d_0=0}^{6} \sum_{n=1}^{M} \sum_{(i,z)\in A} x_{iz}^{j,d_0,\omega_n} \delta_{iz}^{j,d_0,\omega_n}$. Note this objective calculates exactly the desired objective of Equation 22. The flow balance constraints remain the same except for an addition of the ω_n index to Equation 19. Recall that the sample path decomposition now defines flows as actual patients arriving to be seen rather than slots reserved, but it is necessary to pay for all slots reserved, whether or not patients are seen. However, in each subproblem the full price of capacity is charged for each unit of flow sent across the arc, but the benefit is discounted by the probability of the patient showing up, $\mathbb{P}(\omega_n)$.

The final challenge is how to deal with capacity constraints. Again the flows represent patients seen, but we must reconcile this with the fact that capacity constraints refer to appointments reserved. This is handled as follows. For the nth subproblem (ω_n), the solution of the network flow subproblem will send zero flow through a resource on a particular day (even if positive flow was sent at a ω_m for $m < n$), indicating that the probability that the nth patient will show up is not worth the cost of reserving an appointment for that patient. Thus, for each day of the planning horizon ($d_1 = 0, \ldots, 6$), let m be the last network with positive flow on that day. Then, m is the amount of capacity that should be reserved. Summing the benefit from each network matches the definition of the objective in Equation 22, in which the LHS is the the expected benefit from reserving $x_{\tau,\tau\oplus a}^{j,d_0}$ capacity and the RHS is exactly the calculation that we get by summing the non-zero flows from the ω_m subproblems. The costs also match, since each flow is charged the full cost of the capacity reserved. Thus, the capacity calculation in the Lagrangian relaxation will be correct as well.

After applying the sample path decomposition, it is now possible to regain linearity and the pure network form while only increasing solution times linearly in M (the maximum possible discharges), although using parallel computing solution times may not increase at all. This is true because it is only necessary to solve M additional subproblems (ω_i for $i = 1, \ldots, M$) for each original subproblem from the deterministic optimization and these subproblems can be solved in parallel similar to the previously described decompositions. Using the pruning methods developed in the next section, however, it will become clear that in most cases it will not be necessary to solve anywhere near M subproblems to incorporate stochasticity as $\mathbb{P}(\omega_i)$ will quickly become so small that all branches can be pruned, thereby eliminating the need to solve any ω_j for $j > i$.

3.3.3. Pruning Arcs/Nodes

Even with the weakly coupled network formulation, the size of the problem still hinders us from solving a full 30-day planning horizon as the number of arcs and nodes grows too large for commercial solvers. Computational results supporting this are presented in section 3.3.4. Fortunately, the special structure of our problem enables us to prune a large number of arcs and nodes from the network without impacting the optimal solution. The following theorem formalizes this notion by guaranteeing that certain arcs will never be taken in the optimal solution.

THEOREM 1. *In the Network Flow Formulation of PROGRAM* 1 (*stochastic and deterministic*), *an arc with positive cost will have zero flow in the optimal solution to the network flow.*

PROOF. We prove the result by showing that any solution with a positive cost arc can be improved by redirecting the flow away from the arc. Thus, in any optimal solution no flow will be placed along positive cost arcs. Consider such a solution where arcs (τ_1, τ_2) and (τ_2, τ_3) both have non-zero flow and the costs of arc (τ_1, τ_2) is $c_{12} > 0$ and the cost of arc (τ_2, τ_3) is $c_{23} < 0$. If we send flow on arc (τ_1, τ_2) and (τ_2, τ_3) along the arc (τ_1, τ_3) instead we gain $c_{1,2}$ because we are no longer sending flow along the positive arc. Further, the cost of arc (τ_1, τ_3) is better than the cost of (τ_2, τ_3) (i.e., $c_{13} < c_{23}$) because removing the inspection in between the inspection at τ_1 and the one at τ_3 only increases the value of the inspection at τ_3, since it is less likely that a previous inspection has caught the potential readmission. This can be seen directly from Equation 9. Also from Equation 9, the costs of all subsequent inspections after node τ_3 are also either improved or not impacted because having one fewer inspection in the history of inspections does not decrease the value of all subsequent inspections. Thus, redirecting the flow on the positive valued arc results in a lower overall cost flow. □

Not only does Theorem 1 enable the pruning of a large portion of the arcs and nodes for any given network, it also leads to the result that the optimal schedule of a longer planning horizon can often be obtained by solving a network for a much shorter planning horizon. The following corollary states this formally.

COROLLARY 1. *Let* $V^*_{j,d_0}(t)$ (*deterministic*) *or* $V^*_{j,d_0,\omega}(t)$ (*stochastic*) *be the optimal value of subproblem of the network flow formulation of Program* 1 *for type* j *patients discharged on day* d_0 (*with sample path* ω *for stochastic*) *under a planning horizon of length* t. S *be the time s.t.* $\max_{t>S}\{\beta_j\rho_j((per,t),\{0\}) - \hat{c}_j((per,t),d_0)\} < 0$ (*deterministic*) *or* $\max_{t>S}\{\beta_j\mathbb{P}(\omega)\rho_j((per,t),\{0\}) - \hat{c}_j((per,t),d_0)\} < 0$ (*stochastic*). *Then* $V^*_{j,d_0}(t) = V^*_{j,d_0}(S)$ (*similarly* $V^*_{j,d_0,\omega}(t) = V^*_{j,d_0,\omega}(S)$) $\forall t > S$. *Further, the optimal schedules will be the same for all* $t > S$.

PROOF. We show the result for the deterministic case because the stochastic case is identical except for a different reward term. It can be verified from Equation 9 that the largest possible marginal benefit that can be gained from scheduling an inspection on day t is if the inspection scheduled is a perfect inspection and is the first inspection of a patient's monitoring regime. To see this let $\tilde{t}_{\ddot{n}(\tau)+1} = t$ and note that for all $\tau \neq \{0\}$ (i.e., schedules in which t is not the first inspection of the monitoring regime)

$$\begin{aligned}\rho_j((per,t),\{0\}) &= \int_{s=t}^{\infty} p_j(s)[1-F_j(s-t)]ds \\ &\quad - \hat{c}_j((per,t),d_0) \\ &\geq \int_{s=\tilde{t}(a)}^{T} p_j(s)\sum_{i=0}^{\ddot{n}(\tau)}[F_j(s-\tilde{t}_{\ddot{n}(\tau)-i}(\tau)) \\ &\quad - F_j(s-\tilde{t}_{\ddot{n}(\tau)-i+1}(\tau))]ds \geq \rho_j(a,\tau)\end{aligned}$$

The first inequality follows because $\sum_{i=1}^{\ddot{n}(\tau)}[F_j(s-\tilde{t}_{\ddot{n}(\tau)-i}(\tau)) - F_j(s-\tilde{t}_{\ddot{n}(\tau)-i+1}(\tau))] + F_j(s-\tilde{t}_{\ddot{n}(\tau)}(\tau)) \leq 1$. The second inequality follows because $r^{imp} \leq 1$. Thus $\rho_j((per,t),\{0\})$ is the largest possible benefit of having an inspection on day t. If this benefit is less than the cost of scheduling an inspection on day t for all $t > S$, then all the arcs for days $t > S$ will have positive cost and thus will be pruned by Theorem 1. Thus, the network for horizon length S will be identical to the pruned network for horizon length $t > S$ and hence will have the same optimal solution. Corollary 1 can be used to easily compute off-line the finite horizon length needed to achieve an infinite horizon optimal and thus greatly reduce computation times. In section 3.3.4, we show that Theorem 1 and Corollary 1 have a profound impact on solution times, enabling us to solve problems that were intractable in the original formulation and even in the network formulation of PROGRAM 1. □

3.3.4. Computational Results. This section discusses solution times for the various approaches to solve the optimization problem described above. For the original PROGRAM 1, the optimization fails to solve except for instances with small planning horizons (small T). Figure 6 demonstrates the benefit of the network transformation as well as pruning. These computation times are for one iteration of the subgradient optimization for the Lagrangian relaxation of the network flow formulation of PROGRAM 1 solved on a computer with an Intel i5-3230M @ 2.6GHz processor with 8GBs of RAM. The total times are linear in the number of iterations of the subgradient optimization, which are still typically <30 seconds for a 30-day planning horizon.

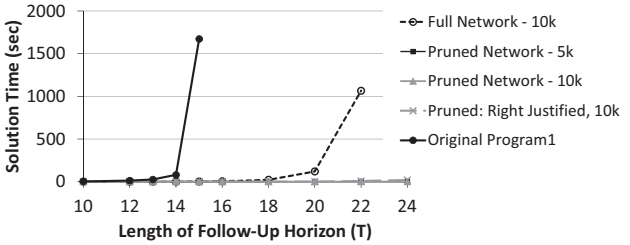
Figure 6a demonstrates the impact of pruning on solution times. Network subproblem decomposition and pruning enabled by Theorem 1 and Corollary 1 can solve large problem instances that quickly become intractable for the standard formulation and even the full network formulation. It also shows that pruning is effective regardless of the skewness of the time to readmission density ($p_j(t)$), demonstrated by the "Right Justified" curve, in which the peak of $p_j(t)$ was shifted to the right. Figure 6b demonstrates the insight provided by Corollary 1, as it is clear that once the model exceeds a fixed planning horizon length the number of solution arcs stops growing. This is because all future days (e.g., beyond 10 days for Pruned Network 5K Benefit) do not have enough benefit and therefore get pruned. The "Right Justified" curve also shows that pruning can happen at the beginning of the planning horizon as well as the end, as all arcs are pruned before day 10.

4. Case Study: Numerical Analysis and Insights

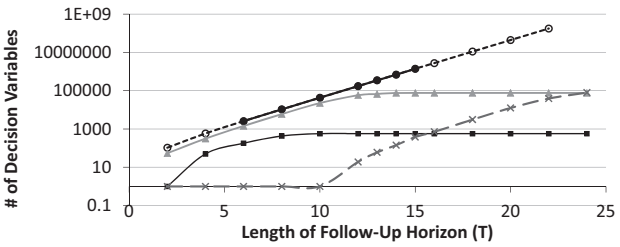
Given the increasing attention on hospital readmissions by both policymakers and health-care administrators, a primary goal of this study is to provide hospitals with an effective data-driven method for reducing readmissions. Here, we present results illustrating the effectiveness of our approach at achieving this goal. We focus our discussion on the size of the reductions possible, the value of being able to predict patient readmission risk profiles, and insights from the patient follow-up schedules.

In this section, we employ the empirical prediction model developed in section 2 to generate inputs to the optimization model. For the patient readmission curves, we randomly selected 10% of patients from the full dataset described in that section. We then run a series of experiments to generate insights into the impact of optimal post-discharge monitoring, risk profiling, stochasticity of discharges, time-varying

Figure 6 Comparing Pruned vs. Unpruned Networks at Different Benefit Levels for Averting Readmissions (β_j) and Different Planning Horizon Lengths (*T*). The same legend is used for both (a) and (b)



(a) CPLEX solution times



(b) Number of arcs

capacity limits and costs, and the benefit of averting a readmission on resulting monitoring schedules, staffing plans, and effectiveness of the post-discharge monitoring system. We find that risk profiling (e.g., categorizing patients as high, medium, and low risk for readmission) has a major positive impact on the effectiveness of post-discharge monitoring.

The prediction model provides both the time to readmission density function, $p_j(t)$, as well as a risk profiling approach for separating patients into different risk categories for readmission. This empirical density function and classification scheme are then used to build an optimization model for post-discharge monitoring. The discharge pattern, X^j_{d0}, is based on 1.5 months of discharge data from our partner hospital. We based the delay-time function on a survey of five surgeons in which we asked them to estimate how long a patient might stay at home with different conditions (e.g., infection, dehydration, etc.) before being readmitted to the hospital. The averages for major causes of readmission are as follows: (Infectious, 48 hours), (Metabolic, 48 hours), (Failure to thrive, 72 hours), (Urinary, 48 hours), (Hematologic, 48 hours), (Cardiac, 24 hours), (Pulmonary, 48 hours), (Gastrointestinal, 48 hours), (Neuro/psych/MSK/Oto/Optho, 48 hours), (Vascular, 24 hours), (Wound related/hematoma, 72 hours). We used these data to fit a discrete delay-time function with a mean of 36 hours. The detection probability of a phone call was placed at 40% based on discussions with our clinical co-author, but was also varied in the sensitivity analysis.

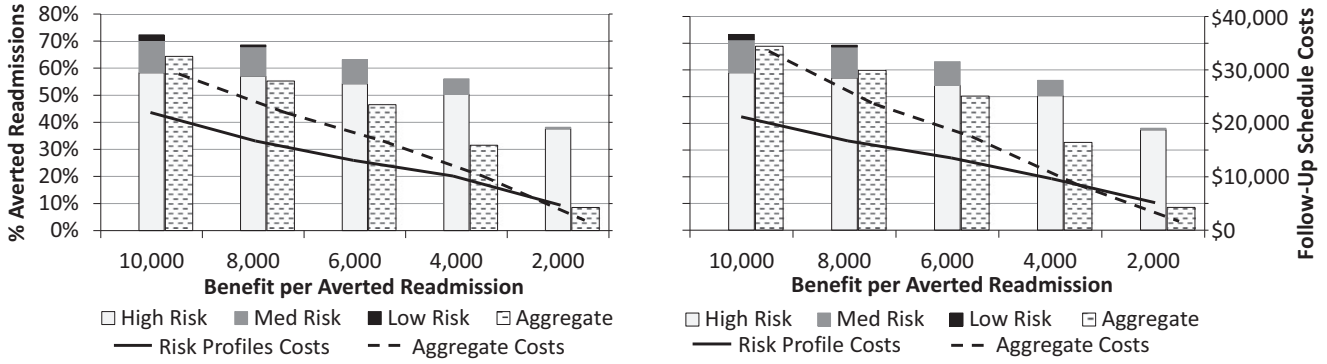
The costs for phone calls and office visits were determined through discussions with our clinical co-author and verified with Medicare reimbursement structures. For example, a phone call from a nurse practitioner is reimbursed at the rate of $25 for a call between 11 and 20 minutes. The office visit is more complex because diagnostic tests may need to be ordered in addition to compensating the doctor for their time. Our clinical co-author estimates a follow-up office visit would cost anywhere between $100 and $500 per visit depending on how many diagnostics were needed, but most likely closer to $100–$200. We then varied the costs of these tests by day of week to capture the fact that some days, such as weekends, are undesirable for performing follow-ups from the medical professional and patient standpoint. The benefit of averting a readmission (β_j) is varied throughout the case study as a means of providing sensitivity analysis and insight. Recall that the parameter β_j represents a weighted average of those readmissions that are completely eliminated and those readmissions whose length and/or severity are reduced by early detection of the readmission triggering condition.

4.1. Numerical Analysis

4.1.1. Risk Profiling. To develop a risk profile, we used the empirical prediction model from section 2 to generate personalized time to readmission curves for each patient in our dataset. Using K-means clustering based on total probability of readmission within 30 days we separated the patients into high, medium, and low risk patients. The groups had an average 30-day readmission probability of 72% for high risk, 18% for medium, and 4% for low. In our dataset, 20% of patients were high risk, 24% were medium risk, and 56% were low risk.

Figure 7 reports the percent of readmission triggering conditions that were detected before causing an emergency readmission by optimally scheduling patient follow-ups. Figure 7 compares a schedule based on optimizing three-patient risk profiles with a schedule based on optimizing a single population aggregate readmission curve. The bars in Figure 7 correspond to the percent averted readmissions on the left y-axis and the solid and dotted lines correspond to the follow-up schedule costs on the right y-axis. Our results show that we are able to reduce the number of readmissions by roughly 40–70%, depending on the cost benefit per averted readmission. Notice that the vast majority of averted readmissions are for patients in the high risk category. While this behavior

Figure 7 Comparing Effectiveness of Monitoring Plans with and without Risk Profiling. (a) Capacity Constrained (b) No Capacity Constraint



(a) Capacity constrained optimization

(b) No capacity constraints in optimization

is as one would expect, it is important to note that it is not a simple matter to accurately predict readmission risk profiles. The techniques and empirical results from section 2 allow hospitals to predict risk profiles for their discharged patients. The optimization model in section 3 then takes advantage of these results to reduce readmissions by targeting the right patients at the right time for follow-up phone calls and office visits.

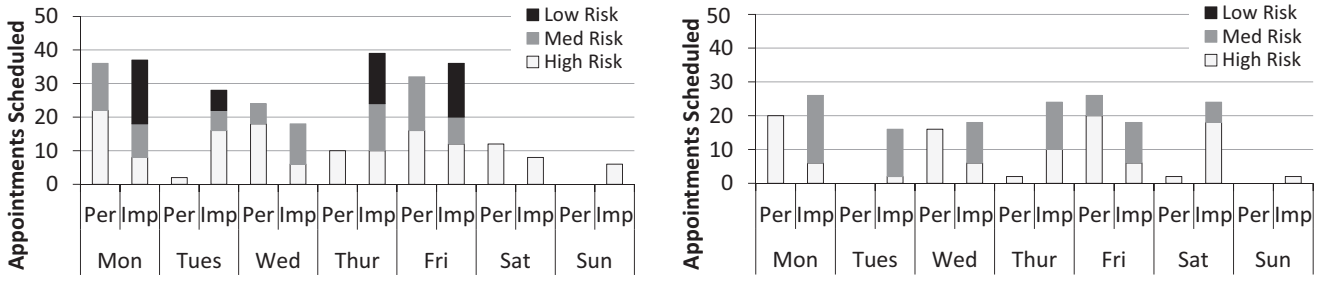
From Figure 7 it is clear that risk profiling has a significant impact on the effectiveness of a post-discharge scheduling and staffing plan. The risk profiled approach (the solid bars) averts a much higher percent of readmissions—increasing from 9% to 39% averted, that is, more than *four times as many readmissions averted* at $2K benefit—when the benefit of an averted readmission is lower by focusing on the high-risk patients that provide the most benefit per inspection. When the benefit of an averted readmission grows, the difference in readmissions averted shrinks but the risk profiled plan (solid line) does so at significantly lower cost—2/3 *the cost*—than the aggregate plan (dashed line), again by using targeted follow-ups and not wasting too much effort on low-risk patients. Risk profiling is also more effective when capacity is more tightly constrained (Figure 7a) than when it is not (Figure 7b). This is noticeably illustrated when the benefit per averted readmission is $10k or $8k. At both levels, the percentage of averted readmissions gap between the three risk profiled bars and the aggregate bar is larger with capacity constraints in place.

Also note how few readmissions are averted from the low risk profile even at a benefit of $10k per readmission averted. It requires planning for 56 phone calls a week to avert 0.2 readmissions. From this, we could draw the conclusion that low risk patients need not be planned for but instead should be contacted only when extra time is available, which can smooth the workload and fill in gaps left by no-shows or inability to fill all slots due to stochastic discharges. Our empirical prediction model could provide a printout each day of the optimal low risk patients to target and the order in which to call them if the nurse/doctor has time in their schedule.

Figure 8 reports the number of perfect and imperfect follow-ups staffed by day of the week and by patient risk profile at a benefit of $10K and $6K per averted readmission. In both cases, no perfect follow-ups are scheduled for low risk patients, only phone calls. The majority of perfect follow-ups are scheduled for Monday, Wednesday, and Friday. Monday and Friday are higher because, in order to avoid the weekend, the patients whose optimal follow-up time would have been on a Saturday get pushed to Friday and Sunday visits get pushed to Monday. Because the time to readmission distribution is mostly concave, the "next best" follow-up time is adjacent to the best one. Wednesday is used because it is a cheap day and spaced out from Monday and Friday, as scheduling back to back perfect inspections is not very beneficial. At $10K benefit, appointments are reserved for high risk patients on every day of the week but Sunday (the most expensive day), because scheduling an inspection at the "right time" is very beneficial due to the high likelihood that these patients will eventually be readmitted. For the other groups, however, the schedule sticks to the Monday, Wednesday, Friday staffing schedule mentioned previously. At the lower $6K benefit, even the high risk patients are mostly staffed for on Monday, Wednesday, Friday and low risk patients are not staffed for at all. Even with this comparatively light staffing plan, we are still averting around 65% of all possible readmissions because of the focused effort enabled by our empirical risk profiling scheme.

Figure 9 shows the uncapacitated follow-up schedule for a high risk patient in contrast to medium risk

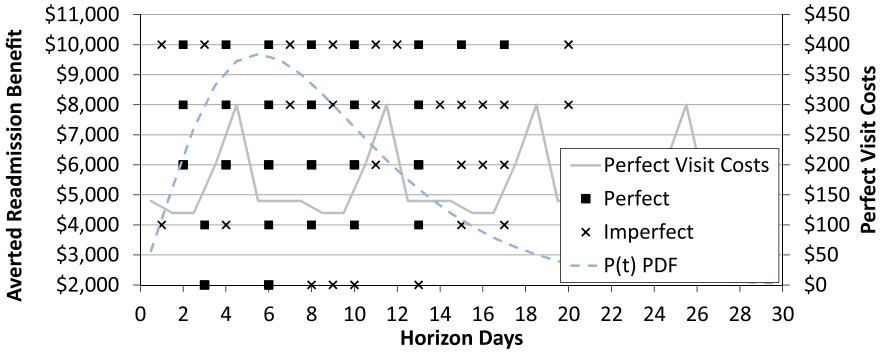
Figure 8 Weekly Staffing Requirements for Follow-Ups Under Different Benefit Levels of Averting a Readmission



(a) \$10K per readmission averted

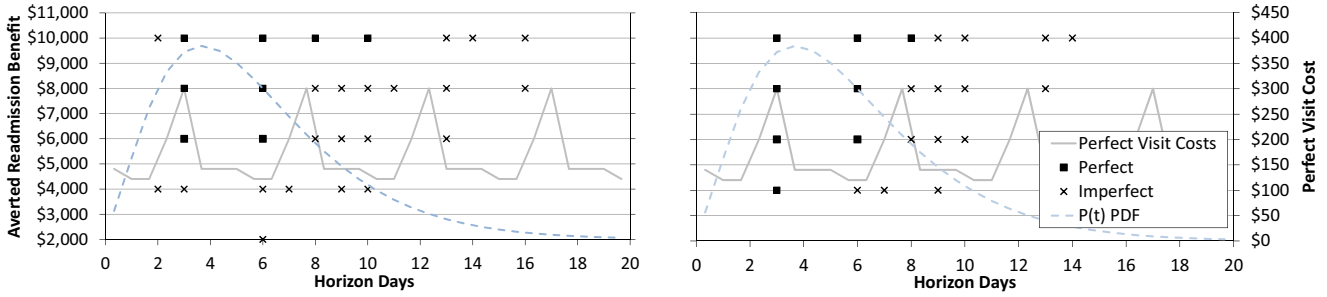
(b) \$6K per readmission averted

Figure 9 Post-Discharge Patient Follow-Up Schedule for High Risk Patients for Different Levels of Benefit of Averting a Readmission. "×" is a phone call and "square" is an office visit. The solid line is the cost profile for office visits and the dashed line provides a structural representation of the time to readmission density, $p_j(t)$.



patients in Figure 10 for (a) uncapacitated optimal and (b) capacitated optimal solutions. In all cases, almost all follow-ups are scheduled within two weeks of discharge. This is a significant finding because the current standard after high-readmission rate surgeries is to follow-up 2–3 weeks after discharge. However, by this point most of the patients who would need to be readmitted have already been readmitted. Follow-ups in the optimal schedules typically begin within 2 or 3 days of discharge, no matter the averted readmission benefit (except \$2000), always centered around the peak of the readmission curve. In general, it is best to place office visits (perfect inspections) close to the peak of the readmission curve, with phone calls before or after or both. As the benefit of averting a readmission increases, the more the optimal schedule will suggest calling patients further out after discharge. For medium risk patients in Figure 10a only one follow-up phone call and no office visits are scheduled at the lowest benefit per averted readmis-

Figure 10 Patient Follow-Up Schedule over a 30-Day Horizon (Tuesday Discharge) for a Medium Risk Patient Considering: (a) Uncapacitated (b) Capacitated Optimizations. "×" is a phone call and "square" is an office visit. Dashed line gives the shape of the time to readmission pdf $p_j(t)$ and solid line is office visit cost by day of week.



(a) No capacity constraints in optimization

(b) Capacity constrained optimization

sions, with more office visits gradually being added as the benefit of averting a readmission increases. The high risk patient is monitored closely over the first 17–20 days, while the medium risk patient generally receives a few office visits during their peak-risk period shortly after discharge followed by a series of phone calls.

To generate Figure 10b, we created capacity limits based on discussions with our clinical co-author to identify reasonable capacities for office visits and phone calls. The effect of capacity limits by day of week does two things to the schedule. First, it spreads out the workload across the days of the week, with less clustering around Monday, Wednesday, and Friday. Secondly, as can be seen in Figure 10, capacity limits can cause one type of follow-up to be replaced with another. At the \$4000 benefit level, for example, the first two phone calls are replaced by an office visit since phone calls were capacity constrained on that particular day of the week. At \$10,000, the fourth office visit is replaced by two phone calls. This indicates that it is very important to have a follow-up at the right time and that, even though the phone call has at most a 40% detection rate, adding several phone calls at key times can be a good surrogate for an office visit or *vice versa*.

4.1.2. Stochastic Discharges. The final component we investigate is the impact of stochastic discharges. We present the case where there are no capacity constraints to focus on the impact of stochasticity. Interestingly, uncertainty regarding how many patients will be discharged, and thus need follow-ups, does not significantly change the timing of the follow-ups. Adding uncertainty does, however, change the staffing level across the week, encouraging the model to move some capacity from office visits (more expensive) to phone calls (less expensive) and from more expensive days to cheaper days (Figure 11). The overall staffing level for expensive office visits is also lower in the stochastic case because expensive capacity is eliminated unless there is a high expectation the slot will be filled.

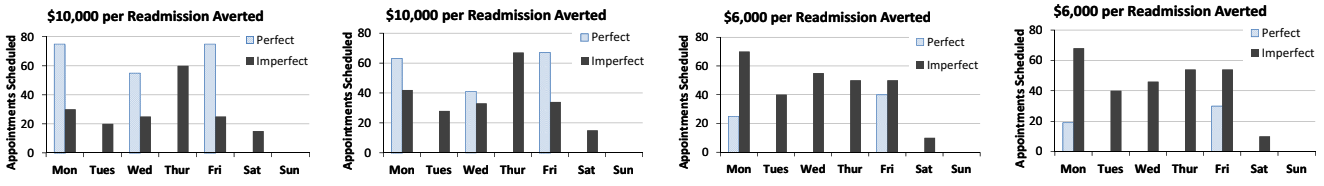
4.2. Sensitivity Analysis

The optimal monitoring schedules and the associated number of readmissions averted are, of course, subject to the model inputs. Most figures in section 4.1 illustrate a sensitivity analysis on the key input, benefit of an averted readmission. We did this because each hospital and medical procedure will have a different value for the benefit of an averted readmission. In this section, we perform a sensitivity analysis and examine the impact of other important model inputs; specifically the number of risk profiles, the detection rate of an imperfect inspection, the structure of the delay-time distribution, and how errors in the empirical prediction model propagate through the optimization model.

First, we identify the impact of the number of risk profile groups on the model solution by dividing the patients in our dataset into two, three, four, and five different risk profiles and comparing it to the aggregated or one profile for the entire group. The results are shown in Figure 12. The percentage of averted readmissions is broken out in the bar chart for each risk profile. This illustrates that almost all averted readmissions are from the high risk profile. Please note that the number of patients in each risk profile changes as the number of profiles are increased. For example, the number of patients in the Med-High risk profile with only two profiles is much larger than the number of patients in the Med-High risk profile when there are five risk profiles.

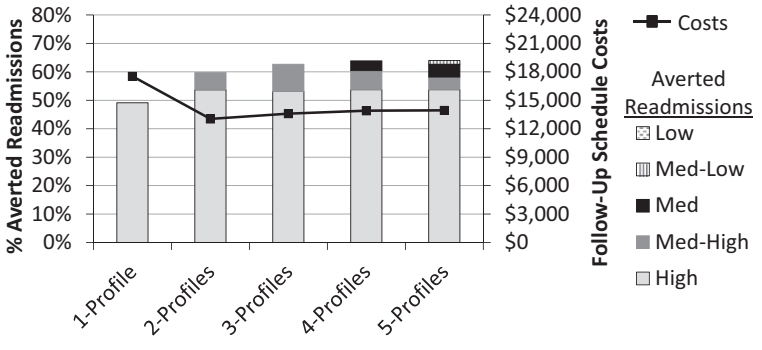
Using two risk profiles, instead of leaving all patients in one profile, improves averted readmissions by 21% and reduces follow-up costs by 25.6%. Moving from two to three risk profiles averts 6.1% more readmissions. However, the cost of these follow-ups increased 4.3%. Adding a fourth and fifth risk profile provides minimal benefit in averted readmissions (1–2%) and follow-up costs actually increase slightly. Figure 12b illustrates why we choose three risk profiles for our numerical analysis in section 4.1 The modeled solution value (Benefits − Total Cost) increases from one to two to three risk profiles. However, after three risk profiles the solution value levels

Figure 11 Comparing a Staffing Plan Having Deterministic Discharges with a Plan Having Stochastic Discharges with the Same Mean for \$10K (a, b) and \$6K (c, d) Benefit of an Averted Readmission

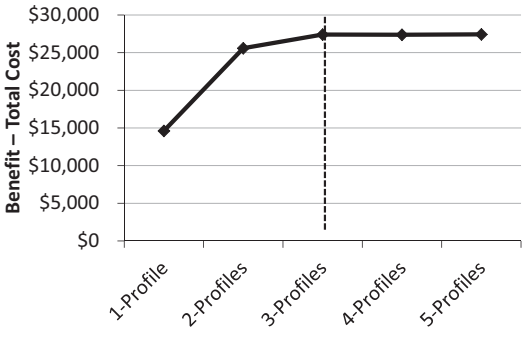


(a) Det. disch. 10K benefit **(b)** Stoch. disch. 10K benefit **(c)** Det. disch. 6K benefit **(d)** Stoch. disch. 6K benefit

Figure 12 Comparing Cost and Effectiveness of Averted Readmissions Across Differing Numbers of Risk Profile Groups

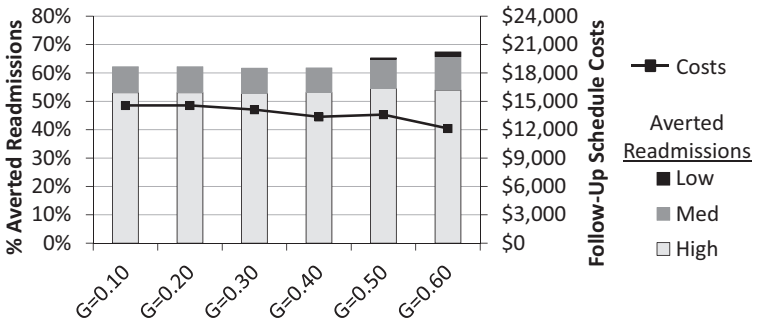


(a) Detailed Analysis

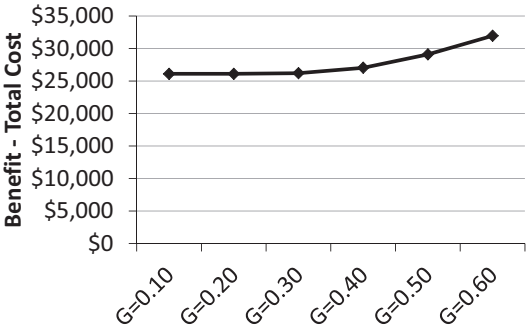


(b) Modeled Solution Analysis

Figure 13 Comparing Cost and Effectiveness of Averted Readmissions Across Differing Imperfect Detection Levels



(a) Detailed Analysis



(b) Modeled Solution Analysis

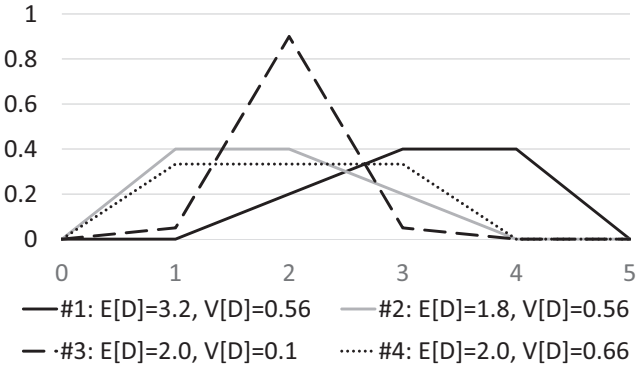
off, varying <1% thereafter. Thus, three risk profiles provide the best balance of simplicity and efficacy.

All of the results in the sections above use a phone call detection probability of 40%, however in Figure 13 we investigate the impact efficacy of a phone call by varying the detection probability from 10% to 60%. As expected, Figure 13b shows the optimization model solution value (Benefits − Total Cost) increasing as the detection probability improves. The results in Figure 13a illustrate the general trend that costs decrease and averted readmissions increase as the detection probability of a phone call increases, though this is not universally the case. Note that there is a slight dip from 62.2% averted readmissions to 61.7% when detection probability moves from 20% to 30%. This occurs because of the discrete nature of timing and type of follow-up. At <30% detection probability, no phone calls are performed at all due to lack of effectiveness, only office visits. At 30% detection probability, the optimal solution starts to replace some of the office visits with phone calls, leading to an overall better solution (through lowered follow-up staffing costs) but slightly lower detection rate. Another key takeaway is that the objective value is fairly robust to the detection probability, which means that we can develop good solutions with a wide variety of imperfect detection probabilities. Note the averted readmission percentage only varies from 62% to 67% and the cost from \$12K to \$14.5K as the detection percentage varies up to 60%.

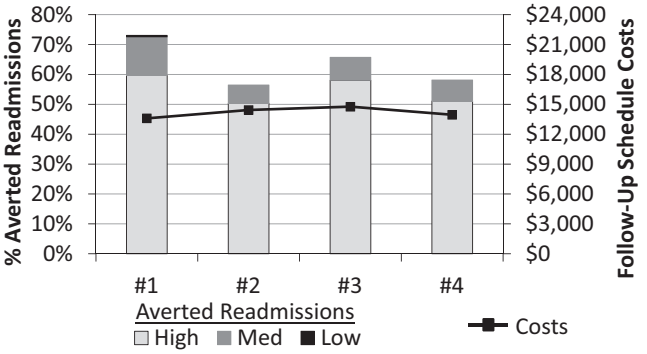
We also analyzed the impact of changes in the delay-time distribution on the optimal solution results. Figure 14a shows the different delay-time probability mass functions that were tested. These delay-time variations are based on our survey of surgeons that indicated that delay times for typical readmission-causing conditions tend to range from 1 to 4 days, with the average delay time between 2 and 3 days. In this analysis, we examine the (i) sensitivity to mean/skewness with curves #1 and #2, as these two have the same variance but different means, and (ii) sensitivity to variance with curves #3 and #4, which have the same mean but different variances.

The results in Figure 14b behave as expected. When the variance is held constant, the distribution with the higher mean (Curve #1) averts 17% more readmissions and has a lower cost than the distribution with the lower mean (Curve #2). When the mean is held constant, the distribution with the lower variance (Curve #3) averts 7.6% more readmissions than the

Figure 14 Curves #1 and #2 are Used to Explore the Impact of the Mean and Skewness of D. Curves #3 and#4 are used to explore the impact of variance of D, having the same mean but different variance. The mean (E[D]) and variance (V[D]) of each distribution are given in legend (a)



(a) Delay-time probability mass functions



(b) Cost and effectiveness of averted readmissions

distribution with the larger variance (Curve #4). The follow-up costs are slightly higher in Curve #3 with a lower variance because more office visits were scheduled. This occurs since office visits on the optimal follow-up day are more likely to find readmission conditions with a low delay distribution variance.

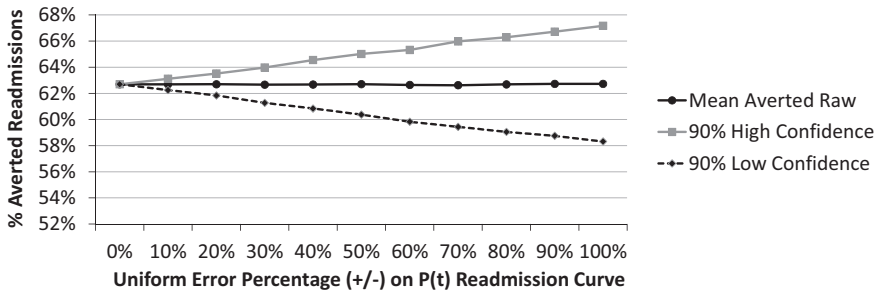
Finally, we analyzed how errors in the empirical prediction of the time to readmission pdf, $p_j(t)$, would impact the optimization model results. To test this, we created a simulation to act in the same way a hospital system would use our method in practice. Using past data, a hospital system would determine the optimal follow-up schedule for high, medium, and low risk patients. There may be error in estimating the readmission curve, $p_j(t)$. Therefore, our simulation used the optimal follow-up schedule and costs for the high, medium, and low risk profiles with an averted readmission benefit of \$6000 and a 40% imperfect detection rate. The simulation then randomly perturbed the readmission pdf, $p_j(t)$, by an error factor that was generated as a percentage in a positive or negative direction where the percentage was drawn from a uniform random variable. We then calculated the readmissions averted based on the fixed follow-up schedule and these randomly perturbed pdf curves. The result of 1000 of these simulations can be seen in Figure 15. This illustrates the robustness of the optimization model to variation in the readmission pdf, $p_j(t)$, since the averted readmissions stay between 58% and 68% even as the allowable random error approaches 100%.

4.3. Clinical Insights and Opportunities for Post-Discharge Inspection Approaches to Minimize the Burden of Readmissions

Using a combination of perfect (i.e., in-person office visits) and imperfect inspections (i.e., automated screening and live phone calls) to identify patients who are susceptible to readmission is clinically feasible and made possible through the integration of empirical prediction models (section 2), optimization (section 3), and clinical knowledge. The statistical and optimization tools developed in this study can offer hospitals the ability to plan for special care at the time of discharge with the confidence that they will have the resources to support each individual's follow-up needs.

Implementing post-discharge inspection approaches based on the models developed herein will reduce the burden of readmissions in at least two important ways leading to measurable decreases in readmissions, subsequent length of stay and avoidable spending.

Figure 15 Impact of Errors in the Empirical Prediction of Time to Readmission pdf, $p_j(t)$, Propagating into the Optimization Model Results



First, by predicting resource needs for patient follow-ups based on readmission risk, clinical deterioration can be efficiently detected early through targeted office visits and telephone calls to avoid readmission or at least decrease its intensity. Many times, patients are unaware of their clinical deterioration or simply take a 'wait-and-see' approach while their condition continues to worsen requiring more intense and expensive care later on. For example, failing to identify an infection early so oral antibiotics can be given in the outpatient setting will eventually result in an emergency department visit, intravenous antibiotics, and hospital readmission. Second, through identifying patients at highest risk of readmission, hospitals would be able to plan for readmissions and develop efficient systems to deal with patients who return for inpatient care rather than funneling them through expensive and resource constrained emergency departments. This is especially important since most current approaches to post-discharge care are often 'one-size fits all' with customary outpatient follow-up in two to three weeks regardless of an individual patient's readmission risk. As shown above, our analyses suggest that tailored post-discharge care optimized to each patient's clinical scenario and their health-care system has the potential to significantly reduce readmission risk over time.

By taking into consideration the daily risk of clinical deterioration and its detection, our models are able to align appropriate inspection efforts to get patients the supportive care they need in a timely fashion. Although readily measured, the often cited 30-day readmission metric is much too blunt to direct large scale clinical efforts to curtail readmissions. For these reasons, the optimization tools developed herein have the potential to direct guidelines for scheduling a range of follow-up intensities based on risk factors, as well as patient characteristics, and hospital characteristics so as to reach the broadest group of health-care organizations.

Finally, by using a State Inpatient Database (SID) and a partner hospital for the validation portion of this project, we were able to develop tools based on readily available hospital discharge data thereby increasing the generalizability of this work and its subsequent implementation.

5. Conclusion

While reducing hospital readmissions is one of the most pressing challenges for the US healthcare system, few efforts harness the intersection of medicine, statistics, and operations management to develop innovative approaches to help solve the dilemma. The comprehensive multi-methodology approach combining empirical models to predict readmission timing with optimization models to detect readmittable conditions before they cause a readmission has the potential to transform the way organizations approach reducing the burden of readmissions on patients and health-care organizations. Our approach integrates machine learning and classical prediction models along with transfer learning to generate accurate and personalized predictions of time to readmission and enables classification of patients into risk profiles. The results of the empirical model are then integrated with methods to transform a large-scale optimization based on stochastic delay-time models into a weakly coupled network flow model with tractable subproblems that can be solved as independent network flow models. In a case study based on data from a partner hospital, we show that the empirical model outperforms other leading techniques in terms of predicting time to readmission and our optimization model demonstrates the ability to *avert 40–70% of potential readmissions* using simple and implementable post-discharge follow-up schemes that simultaneously provide risk profiling, readmission timing prediction, patient monitoring schedule design, and staff planning. While this work focuses solely on post-discharge patient management, future research in readmissions may include integrating our work with the impact of pre-discharge actions, such has hospital length of stay, on readmissions.

Acknowledgments

The authors thank the senior editor and two anonymous reviewers for their valuable comments and suggestions which greatly improved the paper. Ted Skolarus gratefully acknowledges grant support he received from VA HSR&D Career Development Award—2(CDA 12-171).

References

Cardozo, L., J. Steinberg. 2010. Telemedicine for recently discharged older patients. *Telemed. J. E-Health* **16**(1): 49–55.

Clayton, D. G. 1978. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* **65**(1): 141–151.

Deglise-Hawkinson, J., M. P. Van Oyen, B. Roessler. 2013. Stochastic modeling and optimization in a decision support system for phase 1 clinical trial management. Working paper.

Desai, M. M, B. D. Stauffer, H. Feringa, G. C. Schreiner. 2009. Statistical models and patient predictors of readmission for acute myocardial infarction a systematic review. *Circ. Cardiovasc. Qual. Outcomes* **2**(5): 500–507.

Dunlay, S. M., M. M. Redfield, S. A. Weston, T. M. Therneau, K. H. Long, N. D. Shah, V. L. Roger. 2009. Hospitalizations after heart failure diagnosis: A community perspective. *J. Am. Coll. Cardiol.* **54**(18): 1695–1702.

Feudtner, C., J. E. Levin, R. Srivastava, D. M. Goodman, A. D. Slonim, V. Sharma, S. S. Shah, S. Pati, C. Fargason, M. Hall. 2009. How well can hospital readmission be predicted in a cohort of hospitalized children? a retrospective, multicenter study. *Pediatrics* **123**(1): 286–293.

Fisher, M. L. 1985. An applications oriented guide to lagrangian relaxation. *Interfaces* **15**(2): 10–21.

Fonarow, G. C., L. W. Stevenson, J. A. Walden, N. A. Livingston, A. E. Steimle, M. A. Hamilton, J. Moriguchi, J. H. Tillisch, M.A. Woo. 1997. Impact of a comprehensive heart failure management program on hospital readmission and functional status of patients with advanced heart failure. *J. Am. Coll. Cardiol.* **30**(3): 725–732.

Foster, D., G. Harkness. 2010. *Healthcare Reform: Pending Changes to Reimbursement for 30-day Readmissions*. Thomson Reuters, Ann Arbor, MI.

Geoffrion, A. M. 1974. *Lagrangean Relaxation for Integer Programming*. Springer, Berlin.

Gilks, W. R., P. Wild. 1992. Adaptive rejection sampling for gibbs sampling. *Appl. Stat.* **41**(2): 337–348.

Gocgun, Y., A. Ghate. 2012. Lagrangian relaxation and constraint generation for allocation and advanced scheduling. *Comput. Oper. Res.* **39**(10): 2323–2336.

Gonseth, J., P. Guallar-Castillón, J. R. Banegas, F. Rodríguez-Artalejo. 2004. The effectiveness of disease management programmes in reducing hospital re-admission in older patients with heart failure: A systematic review and meta-analysis of published reports. *Eur. Heart J.* **25**(18): 1570–1595.

Gordon, S. C. 2002. Stochastic dependence in competing risks. *Am. J. Polit. Sci.* **46**: 200–217.

Graham, J., J. Tomcavage, D. Salek, J. Sciandra, D. E. Davis, W. F. Stewart. 2012. Postdischarge monitoring using interactive voice response system reduces 30-day readmission rates in a case-managed medicare population. *Med. Care* **50**(1): 50.

Gustafson, P. 1997. Large hierarchical bayesian analysis of multivariate survival data. *Biometrics* **53**: 230–242.

Halfon, P., Y. Eggli, I. Prêtre-Rohrbach, D. Meylan, A. Marazzi, B. Burnand. 2006. Validation of the potentially avoidable hospital readmission rate as a routine indicator of the quality of hospital care. *Med. Care* **44**(11): 972–981.

Held, M., P. Wolfe, H.P. Crowder. 1974. Validation of subgradient optimization. *Math. Program.* **62**(6): 1265–1282.

Helm, J. E., M. P. Van Oyen. 2014. Design and optimization methods for elective hospital admissions. *Oper. Res.* **6**(1): 62–88

Helm, J. E., M. P. Van Oyen, T. R. Rohleder. 2013. Priority scheduling in a queueing network with an application to itinerary completion at destination medical centers. Working paper, Mayo Clinic.

Hernandez, A. F., M. A. Greiner, G. C. Fonarow, B. G. Hammill, P. A. Heidenreich, C. W. Yancy, E. D. Peterson, L. H. Curtis. 2010. Relationship between early physician follow-up and 30-day readmission among medicare beneficiaries hospitalized for heart failure. *JAMA* **303**(17): 1716–1722.

Hougaard, P., P. Hougaard. 2000. *Analysis of Multivariate Survival Data*, Vol. 564. Springer New York.

Hu, M., B. L. Jacobs, J. S. Montgomery, C. He, J. Ye, Y. Zhang, J. Brathwaite, T. M. Morgan, K. S. Hafez, A. Z. Weizer, M. Hu, B. L. Jacobs, J. S. Montgomery, C. He, J. Ye, Y. Zhang, J. Brathwaite, T. M. Morgan, K. S. Hafez, A. Z. Weizer, S. M. Gilbert, C. T Lee, M. S. Lavieri, J. E. Helm, B. K. Hollenbeck, T. A. Skolarus. 2014. Sharpening the focus on causes and timing of readmission after radical cystectomy for bladder cancer. *Cancer* **120**(9): 1409–1416.

Ibrahim, J. G., M. Chen, D. Sinha. 2005. *Bayesian Survival Analysis*. Wiley Online Library, Hoboken, NJ.

Jack, B. W., V. K. Chetty, D. Anthony, J. L. Greenwald, G. M. Sanchez, A. E. Johnson, S. R. Forsythe, J. K. O'Donnell, M. K. Paasche-Orlow, C. Manasseh, et al. 2009. A reengineered hospital discharge program to decrease rehospitalization: A randomized trial. *Ann. Intern. Med.* **150**(3): 178.

Kansagara, D., H. Englander, A. Salanitro, D. Kagen, C. Theobald, M. Freeman, S. Kripalani. 2011. Risk prediction models for hospital readmission: A systematic review. *JAMA* **306**(15): 1688–1698.

Keller, J. B. 1974. Optimum checking schedules for systems subject to random failure. *Manage. Sci.* **21**(3): 256–260.

Luss, H. 1976. Maintenance policies when deterioration can be observed by inspections. *Oper. Res.* **24**(2): 359–366.

McHugh, M. D., C. Ma. 2013. Hospital nursing and 30-day readmissions among medicare patients with heart failure, acute myocardial infarction, and pneumonia. *Med. Care* **51**(1): 52–59.

Melton, L. D., C. Foreman, E. Scott, M. McGinnis, M. Cousins. 2012. Prioritized post-discharge telephonic outreach reduces hospital readmissions for select high-risk patients. *Am. J. Manag. Care* **18**(12): 838.

Minott, J. 2008. Reducing hospital readmissions. *Acad. Health* **23**(2): 1–10.

Oakes, D. 1989. Bivariate survival models induced by frailties. *J. Am. Statist. Assoc.* **84**(406): 487–493.

Pan, S. J., Q. Yang. 2010. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **22**(10): 1345–1359.

Pang-Ning, T., M. Steinbach, V. Kumar, et al. 2006. Introduction to data mining. *Library of Congress*.

Richards, F. J. 1959. A exible growth function for empirical use. *J. Exp. Bot.* **10**(2): 290–301.

Rosen, A. K., S. Loveland, M. Shin, M. Shwartz, A. Hanchate, Q. Chen, H. M. A. Kaafarani, A. Borzecki. 2013. Examining the impact of the AHRQ patient safety indicators (psis) on the veterans health administration: The case of readmissions. *Med. Care* **51**(1): 37–44.

Rosenberg, M. A., E. W. Frees, J. Sun, P. H. Johnson, J. M. Robinson. 2007. Predictive modeling with longitudinal data: A case study of Wisconsin nursing homes. *N Am. Actuar. J.* **11**(3): 54.

Sokal, R. R. 1958. A statistical method for evaluating systematic relationships. *Univ. Kans. Sci. Bull.* **38**: 1409–1438.

Spiegelhalter, D., A. Thomas, N. Best, D. Lunn. 2003. WinBUGS user manual version 1.4 (January 2003). Available at http://www.politicalbubbles.org/bayes_beach/manual14.pdf

de Toledo, P., S. Jiménez, F. del Pozo, J. Roca, A. Alonso, C. Hernandez. 2006. Telemedicine experience for chronic care in copd. *IEEE Trans. Inf. Technol. Biomed.* **10**(3): 567–573.

Wallmann, R., J. Llorca, I. Gómez-Acebo, Á. C. Ortega, F. R. Roldan, T. Dierssen-Sotos. 2013. Prediction of 30-day cardiac-related-emergency-readmissions using simple administrative hospital data. *Int. J. Cardiol.* **164**(2): 193–200.

van Walraven, C., I. A. Dhalla, C. Bell, E. Etchells, I. G. Stiell, K. Zarnke, P. C. Austin, A. J. Forster. 2010. Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. *Can. Med. Assoc. J.* **182**(6): 551–557.

Watson, A. J., J. O'Rourke, K. Jethwani, A. Cami, T. A. Stern, J. C. Kvedar, H. C. Chueh, A. H. Zai. 2011. Linking electronic health record-extracted psychosocial data in real-time to risk of readmission for heart failure. *Psychosomatics* **52**(4): 319–327.

Wolinsky, F., S. Bentler, E. Cook, E. Chrischilles, L. Liu, K. Wright, J. Geweke, M. Obrizan, C. Pavlik, R. Ohsfeldt, et al. 2009. A 12-year prospective study of stroke risk in older medicare beneficiaries. *BMC Geriatr.* **9**(1): 17.

Yu, S., A. van Esbroeck, F. Farooq, G. Fung, V. Anand, B. Krishnapuram. 2013. Predicting readmission risk with institution specific prediction models. 2013 IEEE International Conference on Healthcare Informatics (ICHI), IEEE, Philadelphia, PA, pp. 415–420.