Operations Research Models for Reducing Hospital Readmissions

by

Xiang Liu

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Industrial and Operations Engineering)
in The University of Michigan
2019

Doctoral Committee:

Assistant Professor, Jonathan E. Helm, Co-Chair
Associate Professor, Mariel S. Lavieri, Co-Chair
Professor, David C. Musch
Assistant Professor, Cong Shi
Associate Professor, Ted A. Skolarus
Professor, Mark P. Van Oyen

Xiang Liu

liuxiang@umich.edu

ORCID iD: 0000-0003-2224-1254

ACKNOWLEDGEMENTS

First, I would like to thank my advisors, Dr. Mariel Lavieri and Dr. Jonathan Helm. They never hesitated to give me support and advice that helped me tremendously. They are the best advisors I could ever have.

Secondly, I would like to thank my dissertation committee members and collaborators in IOE, Urology, and Ophthalmology: Dr. Cong Shi, Dr. Mark Van Oyen, Dr. Ted Skolarus, Dr. David Musch, Dr. Joshua Stein, and many other medical fellows who provided great help towards the completion of my degree.

Lastly, I would like to thank my family and friends. Without their love and support, I could not have gone this far. Special thanks to my wife Manqi Li and my daughter Ansheng Liu: you are wonderful! Thank you to my parents Dr. Yayun Li and Chunsheng Liu: you are the best mom and dad in the world. I would also like to thank my dogs Playdoh and Hunio: thank you for putting smiles on my face every day.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

ABSTRACT

Hospital readmissions are burdensome and costly to both healthcare providers and patients. In the U.S., one in five Medicare patients is readmitted within 30 days of discharge. We study how to use operations research models to reduce hospital readmissions. Our approach focuses on both the hospital operations level and the policymaker system level. We develop a delay-time optimization framework to maximize the detection of post-operative complications via post-discharge checkups. Then we study how to design a bundled payment policy to balance and incentivize pre- and post-discharge readmission reduction efforts. We build a readmission prediction model using laboratory values observed during the index hospitalization. Ultimately, we provide novel methods for reducing readmissions in the continuum of care spanning between the pre- and post-discharge stages, at the hospital and policymaker levels.

CHAPTER I

Introduction

1.1 Background

In the United States, hospital readmissions are heavily scrutinized as a driver of healthcare costs. According to Weinberger et al. (1996), up to half of all hospitalizations are readmissions. Furthermore, it is estimated that up to 75% of all readmissions are preventable by patient education, pre-discharge assessment, and domiciliary aftercare (Benbassat and Taragin, 2000). In effect, preventable hospital readmissions represent approximately $25 billion in annual healthcare costs (PwC Health Research Institute, 2010). One in eight Medicare patients are readmitted within 30 days of discharge after surgery (PerryUndem Research & Communications, 2013), and 56.5% of readmitted Medicare patients are readmitted through the Emergency Department (ED) (Kocher et al., 2013), contributing to high costs. These readmissions represent not only preventable healthcare costs, but also a tremendous burden on patients and their families.

In order to address this problem, policies such as the Affordable Care Act (ACA) have been implemented (Koh and Sebelius, 2010). Following the ACA, the Centers for Medicare and Medicaid Services (CMS) now penalize hospitals with worse than expected 30-day readmission rates (Joynt and Jha, 2012). For example, Section 3025 of the Affordable Care Act added Section 1886(q) to the Social Security Act establishing the Hospital Readmissions Reduction Program (HRRP). This program requires CMS to reduce payments to the Inpatient Prospective Payment System (IPPS) hospitals with excessive readmission rates beginning in October 2012 (James, 2013). To provide stronger readmission reduction incentives, the CMS is gradually shifting its reimbursement schemes from Fee-For-Service to Pay-For-Performance and Bundled Payment. Among these reimbursement schemes, the Bundled Payment is believed to be most effective at providing incentives to reduce readmissions (Andritsos and

Tang, 2018; Guo et al., 2016). The CMS has established the Bundled Payment for Care Improvement (BPCI) Initiative in 2013. Under BPCI, a hospital receives a bundled payment for all costs incurred during an episode of care. Specifically, Model 2 of BPCI defines an episode of care to be "the inpatient stay in an acute care hospital plus the post-acute care and all related services up to 90 days after hospital discharge" (CMS, 2018a). These circumstances encourage healthcare professionals to more actively search for and implement solutions to minimize hospital readmissions (Wong et al., 2013).

Various methods have been proven effective at reducing readmissions. Nonetheless, such methods are costly(Jack et al., 2009), as they require hospitals to exert readmission reduction effort in two stages of care, namely the inpatient stay stage (before discharge) and the post-discharge follow-up stage (after discharge). For instance, a hospital can extend the length of stay to further stabilize a patient's condition in the inpatient stay stage. It can also perform follow-up checkups (e.g., office visits and telephone calls) and treatments in the post-discharge stage. There is a lack of a systematic approach for hospitals and policymakers to manage readmissions. Moreover, the proliferation of Electronic Health Records and rich data therein provides an opportunity to leverage analytics to predict the readmission risk and better target interventions.

This dissertation aims to create new models to study how hospitals and public health policymakers can reduce hospital readmissions to mitigate the readmission crisis. Figure 1.1 gives an overview of the thesis.In Chapter II, we present our model of post-discharge monitoring as well as analytical and numerical results. In Chapter III, we study how hospitals balance readmission reduction efforts and how a policymaker designs an effective reimbursement and penalty program to incentivize readmission reduction. Chapter IV presents a model using pre-discharge laboratory data to forecast readmissions. Finally, we conclude the thesis and discuss future research work in Chapter V.

1.2 Post-discharge Monitoring

Post-discharge checkup policies can reduce readmissions through early detection of health conditions, however, the methods behind designing effective checkup policies are poorly understood.

In practice, checkup policies implemented by hospitals are designed and based on unsupported heuristics. For example, current practice recommends that doctors first follow-up with cystectomy (a major surgery for bladder cancer) patients with

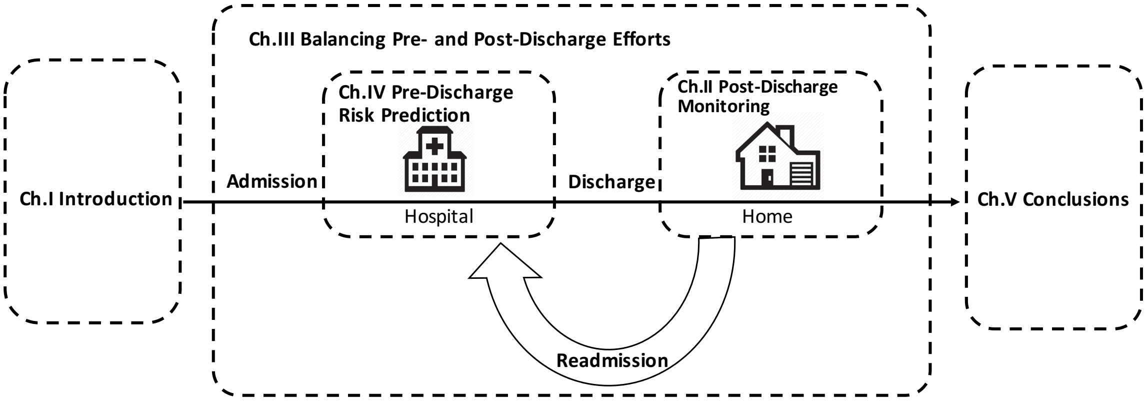

Ch.III Balancing Pre- and Post-Discharge Efforts
Ch.IV Pre-Discharge Risk Prediction
Ch.II Post-Discharge Monitoring
Ch.I Introduction
Admission
Discharge
Hospital
Home
Ch.V Conclusions
Readmission

Figure 1.1: Overview of the Dissertation

an office visit approximately two weeks after their hospital discharge; however, 40% of readmitted cystectomy patients are readmitted within one week of discharge, and as many as 67% of readmitted cystectomy patients are already readmitted before the first scheduled office visit (Hu et al., 2014; Skolarus et al., 2015). If post-surgical complications can be detected and treated promptly, many readmissions can be avoided. This chapter seeks to reclaim this missed opportunity by identifying the optimal timing as well as the type of checkups to perform after discharge. It also provides guidance for how many visits would be most effective. This will give healthcare professionals an increased chance of detecting a patient's health condition before it causes a readmission.

We develop and analyze a new delay-time analysis model to identify the optimal type and timing of checkups to implement post-discharge monitoring plans. By analyzing the structure of the optimal policies, we develop checkup schedules that can detect up to 43.7% more readmission-causing conditions before they result in a readmission. Further, we uncover simple rules of thumb that can help doctors design and improve monitoring plans even in the absence of advanced computer software or complex computations.

1.2.1 Contributions

This work is published in Liu et al. (2018a) and Krishnan et al. (2016). The main contributions of this chapter are summarized as follows: we develop new extensions of the traditional delay-time framework, providing new insights into the structure of delay-time machine maintenance problems and broadening the scope of problems in which delay-time analysis can be applied. In particular, we analyze the optimal structure of the checkup policies without assuming a specific parametric family. We show that imperfect checkups (such as phone calls) can affect the timing and detection probability significantly by considering the detection rate of checkups. Moreover, we incorporated various sources of data to estimate the hidden time-to-develop the condition distributions using the numerical inverse Laplace transform.

In addition to theoretical implications, this study contributes beneficial insights for physicians and other healthcare decision makers to help them improve post-discharge monitoring for patients. The application of our model and findings has the potential for broad impact including reduced hospital readmissions, improved quality of patient care, improved patient satisfaction, and reduced healthcare costs, all without overburdening clinicians (as clinician burden is often a major barrier to implementation of new healthcare practices). This is achievable by aligning checkup policy design with a number of key insights, namely: timing of checkups is the most important factor, checkup timing should be adjusted according to checkup detection rates, and checkup quantity is more important than checkup quality.

1.3 Chapter III: Balancing Pre- and Post-discharge Efforts

To incentivize hospitals to reduce readmissions, the Centers for Medicare and Medicaid Services (CMS) have established the Hospital Readmissions Reduction Program (HRRP) to penalized hospitals with excessive readmission rates. Moreover, the CMS has been experimenting with different reimbursement schemes, such as Pay-For-Performance (P4P) and Bundled Payments (BP), to provide stronger financial incentives for hospitals to reduce readmissions.

The battle against readmissions requires hospitals to exert readmission reduction effort in two phases of care: pre-discharge (during the inpatient stay) and post-discharge follow-up (after the patient has left the hospital). For instance, before discharge, a hospital can extend the length of stay to further stabilize a patient's condition. After discharge, the hospital can perform follow-up checkups (e.g., office visits, telephone calls, and e-visits) and treatments to prevent readmissions. Though proven effective, these readmission reduction efforts can be overly costly. For example, the Reengineered Hospital Discharge Program (Project RED) conducted a randomized clinical trial and found that readmission reduction could substantially impact on health care financing (Jack et al., 2009). Due to the excessive financial burden of the required efforts and investments, the momentum of readmission reduction has stalled since the implementation of the HRRP, as reported in a JAMA study (Desai et al., 2016).

As a policymaker, the CMS faces challenging decisions when designing the BP policy and readmission penalty programs to properly incentivize readmission reduction. If the cost (or penalty) of a readmission is small, hospitals may be unmotivated to take action. If the costs of readmission reduction measures (e.g., extended length of stay and intensive post-discharge follow-up care) are too expensive, hospitals may give up. In addition to the cost and penalty structures, the length of the readmission penalty window and length of an episode of care also play an important role. A New England Journal of Medicine article (Joynt and Jha, 2012) argued that hospitals have little control over readmissions that occur more than seven days after discharge, therefore policymakers should consider limiting the readmission penalty window. Further complicating the matter are factors such as the baseline readmission risk of the patient cohort, the type of the complications that cause readmissions, and the effectiveness of the post-discharge treatments in preventing readmissions.

This chapter studies operational factors that are critical to effective bundled payment policy design. We study key policy-level decisions such as designing readmission penalty programs, subsidizing post-discharge follow-up treatments, and shortening/extending the readmission penalty window length. Specifically, we study how a hospital may allocate readmission reduction efforts between the pre- and post-discharge phases of care. In the pre-discharge phase, the hospital exerts effort to reduce the readmission risk of a patient cohort. In the post-discharge phase as a parameter, the hospital provides post-discharge follow-up care to prevent readmission. This integrated two-stage framework enables us to analytically study how the CMS should design a bundled payment policy to align incentives.

1.3.1 Contributions

This work is to be submitted for publication in Liu et al. (2018b). This chapter makes theoretical contributions to the machine maintenance literature. We develop a novel Strengthen Then Maintain (STM) framework and study how a decision maker balances efforts between the strengthening stage and the maintaining stage. The maintaining stage is modeled as a discrete time finite horizon Markov Decision Process (MDP) machine maintenance problem. We provide a closed-form expression for the cost-to-go function for the machine maintenance MDP under an engaged policy. We prove a theoretical bound on the optimality gap for the closed-form solution when an engaged policy is not optimal. By studying the closed-form expression analytically, we demonstrate how the cost-to-go is affected by the failure rate of the machine (which is analogous to the patient in a healthcare setting). By integrating the two stages, we study how an entity in charge of the maintenance of a machine should allocate efforts between the two stages – in the strengthening stage, the failure rate can be reduced; and in the maintaining stage, the machine is maintained accordingly. The analytical results of the STM framework can be generalized to many machine maintenance problems.

We also uncover novel insights for designing an effective Bundled Payment policy to reduce readmissions. Our analytical results suggest that hospitals have more financial incentives if the readmission penalty window is shortened, the cost of post-discharge follow-up treatment is reduced, the cost/penalty of readmission is increased, and the post-discharge treatment efficacy is improved. For patients that are likely to experience acute events leading to a swift readmission, a more effective mechanism is to shorten the penalty window. For other cohorts, subsidizing inpatient and outpatient efforts may be effective. We believe that our model is the first study that analytically addresses how the penalty window length impacts the incentives for readmission reduction. Unlike many game-theoretical studies which use stylized functional forms, our model is less restrictive with minimal assumptions imposed on the functional form. At the core of our model is a Markov Decision Process model, which directly captures the patient's deterioration and the cost structures.

1.4 Chapter IV: Pre-discharge Risk Prediction

Radical cystectomy has one of the highest rates of complications and readmissions of any surgical procedure, with 25% of patients experiencing unplanned readmission within 30 days (Borza et al., 2017; Stimson et al., 2010; Hu et al., 2014; Skolarus et al., 2015). These high readmission rates, coupled with increasing policy focus on reducing readmissions, have motivated investigations into identification and optimization of patients at highest readmission risk. However, the ability to predict readmission using traditional administrative data is limited. This limitation makes it unclear where and when to focus resources, leaving readmission rates largely unchanged (Minnillo et al., 2015; James et al., 2016).

In this chapter, we used data from electronic health record to examine whether incorporating dynamic laboratory data into readmission prediction models improved risk stratification after radical cystectomy. Specifically, we assessed daily post-operative values for commonly obtained laboratory tests, and used machine learning techniques to compare values between readmitted and non-readmitted patients. We characterized the trajectory of laboratory values obtained in complete blood counts, basic metabolic panels, and coagulation studies during the index hospital stay. The framework showcases that common postoperative laboratory values may have discriminatory power to help identify patients at high risk of readmission after radical cystectomy.

1.4.1 Contributions

This work is to be submitted for publication in Kirk et al. (2018). In this work, we combined a logistic regression model and a support vector machine (SVM) model to incorporate longitudinal clinical data. By combining a regression model and a machine learning model, the framework is able to predict readmissions without loss of interpretability. Moreover, the SVM model can handle missing values. This is especially important, as missing values are very common in real-world clinical datasets. This study demonstrates the unique promise of readily available, dynamic data to inform risk stratification of patients most likely to be readmitted after cystectomy. Incorporating available, dynamic sources of physiological data (such as laboratory values) into prediction algorithms could enable more accurate identification and targeting of patients at greatest readmission risk.

CHAPTER II

Post-Discharge Monitoring

ABSTRACT: Hospital readmissions affect hundreds of thousands of patients every year, negatively impacting patients and placing a tremendous burden on the national healthcare system. Post-discharge checkup policies can reduce readmissions through early detection of health conditions, however, the methods behind designing effective checkup policies are poorly understood. Under current practice, up to 67% of readmitted patients return to the hospital before their first scheduled office visit. This work aims to develop effective checkup plans to monitor patients following hospital discharge using a variety of checkup methods including phone calls and office visits. We develop and analyze a new delay-time analysis model to identify the optimal type and timing of checkups to implement post-discharge monitoring plans. By analyzing the structure of optimal policies, we develop checkup schedules that can detect up to 43.7% more readmission-causing conditions experienced by readmission-bound patients. Further, we uncover simple rules of thumb that can help doctors design and improve monitoring plans even in the absence of advanced computer software or complex computations.

2.1 Introduction

In the United States, hospital readmissions are heavily scrutinized as a driver of healthcare costs. According to Weinberger et al. (1996), up to half of all hospitalizations are readmissions. Furthermore, it is estimated that up to 75% of all readmissions are preventable by patient education, pre-discharge assessment, and domiciliary aftercare (Benbassat and Taragin, 2000). In effect, preventable hospital readmissions represent approximately $25 billion in annual healthcare costs (PwC Health Research Institute, 2010). One in eight Medicare patients are readmitted within 30 days of discharge after surgery (PerryUndem Research & Communications, 2013), and 56.5% of readmitted Medicare patients are readmitted through the Emergency Department (ED) (Kocher et al., 2013), contributing to high costs. These readmissions represent not only preventable healthcare costs, but also a tremendous burden on patients and their families.

In order to address this problem, policies such as the Affordable Care Act (ACA) have been implemented (Koh and Sebelius, 2010). Following the ACA, the Centers for Medicare and Medicaid Services (CMS) now penalize hospitals with worse than expected 30-day readmission rates (Joynt and Jha, 2012). For example, Section 3025 of the Affordable Care Act added Section 1886(q) to the Social Security Act establishing the Hospital Readmissions Reduction Program. This program requires CMS to reduce payments to the Inpatient Prospective Payment System (IPPS) hospitals beginning in October 2012 (James, 2013). These circumstances encourage healthcare professionals to more actively search for and implement solutions to minimize hospital readmissions (Wong et al., 2013).

Numerous interventions have been proposed to prevent readmissions (including better pre-discharge care and improved discharge instructions). Post-discharge checkups such as phone calls, home visits, and office visits have been independently shown in the clinical literature to significantly reduce hospital readmissions (Dudas et al., 2001; Wong et al., 2013) and offset increases in demand for physician services (Green et al., 2013). The purpose of these checkups is to detect developing conditions before they worsen and cause either an unnecessary trip to the ED and/or an inpatient readmission.

Although checkups can mitigate the readmissions crisis, the methods behind designing effective checkup policies are poorly understood. Specifically, healthcare providers remain uncertain about how many checkups to schedule, what types of checkups to schedule, and when to schedule those checkups. In practice, checkup policies currently implemented by hospitals are designed and based on unsupported heuristics. For example, current practice recommends that doctors first follow-up with cystectomy (a major surgery for bladder cancer) patients with an office visit approximately two weeks after their hospital discharge; however, 40% of readmitted cystectomy patients are readmitted within one week of discharge, and as many as 67% of readmitted cystectomy patients are already readmitted before the first scheduled office visit (Hu et al., 2014; Skolarus et al., 2015). Our research seeks to reclaim this missed opportunity by identifying the optimal timing as well as the type of checkups to perform after discharge. It also provides guidance for how many visits would be most effective. This will give healthcare professionals (both clinicians and non-physicians) an increased chance of detecting a patient's health condition before it causes a readmission.

Because most hospitals do not yet have a systematized mechanism for managing follow-ups for their cohort of patients, much of the follow-up decision making is left to the treating surgeon, and it is typically determined on a case-by-case basis. This work seeks to improve the efficacy of these personalized follow-up plans. This approach has been confirmed as having low barriers to implementation relative to a larger scale, system-wide approach that considers costs and savings relative to total hospital resources. This is because medical professionals currently make decisions on a per-patient basis (hence no major culture change required) by weighing the expected benefit (e.g. early detection, readmission reduction, improved quality, etc.) versus the amount of time the practice is able/willing to commit to follow-ups. Cost-based calculations are not frequently used in these individual patient decisions, in part because it is difficult to assign a monetary value to early detection of a condition. This chapter provides analytical, data-based methods and decision guidelines (medical professionals are comfortable with both) to better personalize these decisions that doctors already make on a daily basis.

To provide contextual grounding for our practice-focused readmission detection approach, we develop our models in close collaboration with a urological practice, with a focus on cystectomy, which is one of the highest readmission rate surgeries in the U.S. Other papers have shown similarities in the readmission characteristics of cystectomy patients and other types of surgical patients (Jacobs et al., 2017). This approach could hence be generalizable to other types of surgery and other patient conditions by changing the model parameterization based on historical data, as long as the processes for follow-ups and underlying disease dynamics remain similar. More information about the key assumptions that must be verified before applying our models to other diseases is provided in subsequent sections.

The post-discharge monitoring process after cystectomy proceeds as follows. At the time of discharge, a monitoring schedule is determined by the discharge team and the patient is made aware of when they will be receiving phone calls and when they are scheduled to return for an office visit to check on their recovery. During a phone call or office visit, the patient will be tested to see if they have developed a condition that is likely to lead to readmission. For cystectomy, the two most common conditions are infection and failure to thrive (unable to eat enough food), which account for the majority of readmissions (Hu et al., 2014; Skolarus et al., 2015). These conditions exhibit important characteristics that are suited to early detection and mitigation: (1) these types of conditions are readily detectable via phone call, telemedicine, or office visit, (2) the window for detection is long enough to make a follow-up potentially effective (e.g. patients stay at home with an infection for several days before becoming sick enough for readmission), and (3) early detection can be effective in mitigating the condition on an outpatient basis or at the very least result in a reduced cost ED visit or readmission (e.g. providing antibiotics to treat infection, or early detection means the condition is less serious when treatment begins leading to reduced cost and better patient outcomes).

If a condition is detected early by a follow-up, steps to mitigate the condition can be immediately undertaken. These steps can include starting antibiotics to eliminate infection, or IV treatment for patients suffering from failure to thrive. Hence, early detection may avoid the readmission entirely, prevent an expensive ED visit, or at the very least lessen the time and cost of overcoming the condition while improving the quality of the outcome by catching the condition before it becomes too severe. At the suggestion of our clinical collaborator, we do not attempt to directly quantify the monetary value of such outcomes in our model, but instead leave the decision to the clinician/practice as to the amount of follow-up effort that is reasonable relative to the increased likelihood of early detection.

To capture this personalized follow-up process, we develop a delay-time modeling approach adapted from the machine maintenance literature to analyze and optimize post-discharge checkup policies. Several unique features of readmission dynamics require new extensions of the traditional framework, providing new insights into the structure of delay-time machine maintenance problems and broadening the scope of problems in which delay-time analysis can be applied. In addition to theoretical implications, this chapter contributes beneficial insights for physicians and other healthcare decision-makers to help them improve post-discharge monitoring for patients.

As a proof of concept, we calibrate, test, and validate our models on nationwide data for cystectomy patients. Cystectomy, often performed on bladder cancer patients, is a type of surgery that involves removal of all or part of the urinary bladder. Cystectomy patients experience one of the highest readmission rates of any surgery, as approximately 25% of cystectomy patients are readmitted within 30 days of discharge from the hospital (Hu et al., 2014; Jacobs et al., 2013).

The structure of this chapter is as follows. In Sections 2.3 and 2.4, we develop and analyze our model to understand key properties of the optimal checkup policies. We identify the importance of checkup timing, and how checkup timing is affected by the stochasticity of how long patients are ill prior to readmission (delay-time), as well as the detection rate of checkups. In Section 2.5, we verify our findings through numerical analyses by applying our model to the national State Inpatient Database patient cohorts. The numerical analyses also demonstrate that our model is robust to the system parameters and consistently outperforms current checkup policies. Using the same number of checkups, current practice (which is expected to detect only 16% of the conditions experienced by readmitted patients) can be improved by up to 43.7%. In Section 2.6, we summarize the theoretical and practical implications of our study. In particular, we highlight how our model provides valuable extensions to the traditional delay-time analysis framework and how our findings can effectively detect readmission-causing conditions and improve the quality of patient care, thereby mitigating the national readmissions crisis.

2.2 Literature Review

Readmissions play a critical role in recent clinical literature. It is estimated that up to 75% of readmissions are preventable by patient education, pre-discharge assessment, and domiciliary aftercare (Benbassat and Taragin, 2000), and post-discharge checkups such as phone calls, home visits, pharmacists' visits, and doctors' office visits can significantly reduce hospital readmissions (Dudas et al., 2001; Wong et al., 2013; D'Amore et al., 2011; Bellone et al., 2012; Costantino et al., 2013). Within the healthcare operations research literature, models have been created to improve post-discharge health outcomes, including reducing readmissions and mortality rates: Bartel et al. (2014) analyzes how the initial hospitalization length of stay impacts post-discharge mortality rate; Chan et al. (2012) studies the impact of ICU discharge strategies on readmissions; Kim et al. (2014) analyzes how ICU admission control strategies impact readmission rate. Bayati et al. (2014) builds a classification model to predict readmissions and analyzed intervention decisions. However, this work does not address the timing of interventions. Leeds et al. (2015) conducts a statistical analysis to study how surgeons make discharge decisions and the effect of decision-support tools for discharge. None of those models directly address how patients should be monitored after hospital discharge. To address that question, two areas in the operations research literature are especially relevant to our study: (1) machine maintenance and inspection, and (2) disease screening.

Machine maintenance and inspection: The literature of machine maintenance and inspection is very well-established. Literature surveys (Barlow and Proschan, 1996; Wang, 2002) categorize maintenance policies into two groups: preventive maintenance (PM) and corrective maintenance (CM). Our problem aligns more closely with PM frameworks since PMs proactively prevent failure, whereas CMs are only performed after failures occur. PMs can be scheduled in the following fashion: (1) age-dependent policies perform PM at a fixed time T; (2) periodic and sequential policies schedule multiple PMs in fixed or variable intervals; and (3) failure limit policies perform PMs when the failure rate of a machine exceeds a predetermined threshold. The dynamics of machine deterioration are typically modeled by (1) Markovian processes (Sim and Endrenyi, 1993), (2) semi-Markovian processes (Milioni and Pliska, 1988; Yeh, 1997), (3) hidden Markov models (White, 1977), and (4) delay-time models (Wang, 2012). More specifically, Sim and Endrenyi (1993) models the deterioration as a continuous time Markov chain and considers multiple failure types and repair/maintenance actions. They minimize the long-run average down-time and cost, which is not suitable for our problem. Yeh (1997) uses phase-type distributions to approximate general distributions of a semi-Markovian model. They develop algorithms for optimal state-age-dependent policies that also minimize long-run average cost. White (1977) develops a POMDP model for the machine inspection/maintenance problem which minimizes the long-run average cost. These models are not suitable for our problem because they assumed Markovian deterioration and optimized long-run average cost and down-time.

Wang (2012) gives a thorough survey on delay-time models, which are a special case of semi-Markovian models with three states. Traditional delay-time analysis is based on renewal theory and reliability which assumes the unit lifetime has an increasing failure rate. The goal of those models is typically to determine an inspection schedule that minimizes long-run costs (Christer and Jack, 1991; Jardine and Tsang, 2005) or minimizes expected down-times (Dagpunar, 1994) given identical units that can be replaced. Our problem necessitates several extensions: (1) unlike interchangeable machine components, patients cannot be "replaced"; (2) our objective is to maximize the probability of a checkup (inspection) detecting a patient's condition; (3) readmission rates depend on time since discharge, so we have a time-varying failure rate; and (4) existing models do not allow for policies composed of different types of inspections with varying inspection detection rates (Christer, 1999). Monitoring policies composed of inhomogeneous checkups (e.g. phone calls, office visits, etc.) are particularly valuable because empirical evidence indicates that policies consisting of mixed checkup methods are more effective than policies consisting of only one checkup method (Holland et al., 2005; Wong et al., 2013).

Close to our work is Milioni and Pliska (1988), where a semi-Markovian model with three states was used to model machine deterioration and catastrophic failure (i.e. no repair/replacement after failed). They considered two objectives: minimize the cost of inspections, false positives, and treatment; and minimize the probability of failure. Existence of optimal solutions and algorithms for solving the problems were established. However, the authors did not provide insights into the structure of the optimal policies. Moreover, they assumed perfect inspections in the sick state. Although this model is somewhat similar to our model, the key difference is that this model is still a long-run steady state planning model in both objective functions.

Disease screening: Within the healthcare operations research field, models have been developed to establish medical inspection schedules that detect the onset and progressions of diseases such as chlamydia infection (Teng et al., 2011), diabetes (Brandeau et al., 2004), AIDS (Sanders et al., 2005; Jónasson et al., 2017), hepatitis (Fu et al., 2012), breast cancer (Ayer et al., 2012; Brailsford et al., 2012; Ayer et al., 2015; Maillart et al., 2008), colorectal cancer (Harper and Jones, 2005; Güneş et al., 2015; Erenay et al., 2014), cervical cancer (Myers et al., 2000), prostate cancer (Pinsky, 2004; Tsodikov et al., 2006; Zhang et al., 2012a), bladder cancer (Kent et al., 1989), and glaucoma (Helm et al., 2015). Delay-time models are used to model hepatitis progression (Fu et al., 2012) and vascular patency loss (Zhang et al., 2012b). Most of the models are based on discrete time Markovian assumptions (Ayer et al., 2012; Myers et al., 2000; Kent et al., 1989; Ayer et al., 2015; Maillart et al., 2008; Erenay et al., 2014; Zhang et al., 2012a), which do not fit into our problem since the deterioration dynamics of the readmitted patients are not necessarily Markovian.

Bavafa et al. (2017) studies a three-state continuous time Markov model in the context of primary care routine visits. The authors examine the effectiveness of office visits as well as e-visits as a cost-effective preventative action. However, the model assumes Markovian deterioration and focuses on steady-state planning from the perspective of the primary care providers. Fu et al. (2012) applies delay-time models on hepatitis screening. However, they focus on optimal statistical estimation rather than the optimal monitoring schedule planning. Closest to our work is Zhang et al. (2012b), where follow-up checkups are scheduled to minimize the probability that the time between patency loss and its detection exceeds some length of time. The results on the timing of checkups under the assumptions of deterministic delay-time and Weibull-distributed failure rate are generally consistent with our findings. However, the authors consider perfect checkups only and do not consider general distributions. Their work focuses on the timing of checkups only and does not study how quantity, quality, or mix of different checkups impact monitoring schedules.

Moreover, they estimate the distributions using maximum likelihood methods assuming Erlang and exponential distributions, whereas we use best-fit distributions obtained directly from the data. The novelties of our work leverage the composition of different checkup methods (e.g. office visits and phone calls) and address the tradeoffs in scheduling checkups with both perfect and imperfect inspections under inhomogeneous failure rates. Our work differentiates from Zhang et al. (2012a) in the following aspects. 1) In contrast to their model, we analyze the optimal structure of the checkup policies (consisting of perfect checkups) without assuming a specific parametric family. 2) For imperfect checkups, we show that imperfect checkups (such as phone calls) can affect the timing and detection probability significantly by considering the detection rate of checkups. Moreover, 3) we incorporated various sources of data to estimate the hidden time-to-develop the condition distributions using numerical Laplace inverse transform. Helm et al. (2016) developed a mixed integer programming (MIP) approach to solving a planning problem for how many healthcare professionals to staff to implement a follow-up program. This model, however, assumed a homogeneous population(s) of patients and was designed as a static planning model for a cohort of patients taking the hospital's perspective. Our model, on the other hand, is patient centered and can be tailored based on each individual's projected readmission density curve – focusing on the operational level rather than a steady-state planning model. Our delay-time modeling approach also enables us to identify structural properties, which is not possible using their MIP formulation.

2.3 Model for Optimizing Post-discharge Checkup Policies

In this section, we develop and analyze a general model for designing monitoring plans for patients after they are discharged from the hospital. First, we introduce our model notation and parameters. A summary of the notation is presented in Table 2.1).

2.3.1 Delay-Time Model for Readmissions

Based on our field research, the dynamics of an inpatient readmission occur as follows. After a patient is discharged, he/she may develop a readmission-causing condition. When this condition first develops, it does not necessarily cause an immediate readmission (e.g. an infection). Instead, the patient's condition will degrade over time, eventually becoming so severe that he/she must return to the hospital and be readmitted. These dynamics are identical to those found in machine failure models,

ρ	The random variable representing time-to-readmission (time between discharge and readmission) given no checkups
$g_\rho(\cdot)$	The probability density function of τ
D	Delay-time, i.e., the length of time prior to τ that the illness was detectable by a checkup; this is equivalent to the amount of time that a patient is in the ill state
$f(\cdot)$	The probability density function of D; accordingly, $F(\cdot)$ is the cumulative distribution function of D
δ	The time when the condition developed, i.e., when the illness is first detectable by a checkup
$g_\delta(\cdot)$	The probability density function of δ
t_i	The time when checkup i is performed
T	The latest time following discharge that readmissions are tracked until; thus, this also represents the latest time during which a checkup can be placed
m	The number of different checkup methods available
$y_{ij} \in \{0, 1\}$	An indicator variable that denotes whether checkup method $j \in \{1, ..., m\}$ is used at t_i
$r_j \in [0, 1]$	The detection rate of checkup method j, i.e. if checkup method j is performed when a patient is ill, then the checkup will detect the illness with probability r_j
$r_{(i)} \in [0, 1]$	The detection rate of the checkup employed at t_i $\forall\, i \in \{1, ..., n\}$
$w_j \in \mathbb{N}$	The maximum number of times checkup method j can be used
$\Pi = (t_1, ..., t_n$ $, y_{11}, ..., y_{nm})$	A checkup policy
$N_i^\Pi \in \{0, 1\}$	An indicator variable that denotes whether or not an illness is detected at t_i given policy Π

Table 2.1: Model Notation and Parameters

which have been shown in the machine maintenance literature to be well modeled by a delay-time model. Unlike Markovian models, our model handles general distributions under mild conditions (see Section 2.3.4). Moreover, since our problem has a short planning horizon (30 days) and a transient nature (patient-centric not steady-state planning), continuous delay-time models allow us to keep track of how long a patient has been in each state and we can tailor the objective function as we shall see later. As seen in Figure 2.1, we consider individual patients stochastically progressing through three sequential states upon discharge: healthy, ill, and readmitted. Thus, within the framework of traditional delay-time analysis models used in preventative maintenance, the patient represents the system, illnesses represent defects, and readmissions represent failures.

Remark II.1 ("Ill State"). It is important to note here that the ill state is defined as identifying a patient in a state that causes them to be at risk for a future readmission. This includes conditions such as infection and failure to thrive, but also includes conditions such as when the patient has failed to fill a prescription, is taking their medicine incorrectly, or has not understood or followed post-discharge treatment plans such as exercise or nutritional guidelines. Both medical and compliance issues can be checked for during a phone call or office visit and incorporated into our modeling framework.

At time 0, we assume a patient is discharged in a healthy state. After a stochastic amount of time, δ, the patient develops a detectable condition and is considered to be in the ill state (the first black dot in Figure 2.1). We denote this time δ as the time-to-develop the condition. Following a period of time (delay-time), D (between the first and second black dots in Figure 2.1), the patient's condition worsens to the point where he/she is readmitted to the hospital. We denote this time-to-readmission as $\rho = \delta + D$ (the second black dot in Figure 2.1). Lastly, we let T denote the length of our model's planning horizon (e.g. $T = 30$ days). Clinical literature and policy both support a finite horizon model as the Centers for Medicare and Medicaid Services specify that hospital admissions only qualify as readmissions if they occur within 30 days of discharge.

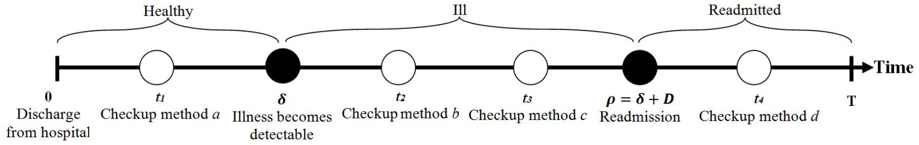


Figure 2.1: Patient State Progression and Checkup Policy

At the point of a patient's discharge, the case manager needs to determine the post-discharge checkup plan for the patient for the next 30 days. Given n checkup opportunities, our goal is to place a checkup at each time t_i, $i \in \{1, ..., n\}$ (white circles in Figure 2.1), to maximize the probability of detecting the patient in the ill state. While there is a possibility of a competing risk of patient mortality, 30-day mortality rates post-discharge are very small relative to readmission rates

In addition to choosing checkup times, decisions must be made regarding what type of checkup method (e.g. phone calls, home visits, doctors' office visits) to use at each checkup time, t_i. Given m different checkup methods, the indicator variable $y_{ij} \in \{0, 1\}$ denotes whether checkup method $j \in \{1, ..., m\}$ is used at time t_i. In Figure 2.1, $y_{1a} = y_{2b} = y_{3c} = y_{4d} = 1$. To model checkup method resource limitations, let w_j denote the maximum number of times checkup method $j \in \{1, ..., m\}$ can be used.

As mentioned in the contextual grounding of Section 2.1, we are developing this research to help personalize monitoring plans for each patient at the provider/practice level. Thus, we allow these constraints to be tailored to what the clinician believes is an appropriate level of checkup intensity (i.e. how many office visits and phone calls they are able/willing to make). For example, our clinical collaborator indicates that most surgeons would typically be willing to do one office visit, two in cases where they are more concerned about the patient, and a maximum of three where the patient's condition indicates very high risk. These determinations, however, are typically made by the clinician based on a medical and historical knowledge of the patient and their condition and are difficult to quantify in a cost-based or constraint-based structure. Further, budgets for checkups are not typically considered when making individual checkup decisions for specific patients, hence the inclusion of costs does not fit the current practice and would provide barriers given that many clinicians are averse to such an approach in individual patient decision making. Hence, we allow the provider/practice to determine how many office visits and phone calls (i.e. w_j's) they believe to be appropriate on a patient-by-patient basis and enter this number as a model parameter. The model also allows for clinicians to perform sensitivity analysis to determine, for example, the marginal benefit of an extra phone call or office visit compared to their base resource allocation.

To account for the differences in checkup methods, e.g. an office visit is more effective than a phone call, we let the detection rate $r_j \in [0, 1]$ denote the probability that method $j \in \{1, ..., m\}$ will detect a condition when the patient is in the ill state (i.e. true-positive). If $r = 1$, then we say that the checkup is a perfect checkup. If $r < 1$, we say that the checkup is an imperfect checkup. The detection rate accounts for the chance that a condition is present and yet is not detected. This could be due to an inability to detect illness based on the questions asked, poor patient responsiveness, or other reasons. Patients not answering the phone can also be considered, but based on discussions with a company that provides automated phone calls to detect readmittable conditions (www.cloud9hcs.com), they achieve full patient responses to their readmission detection scripts (questions) in greater than 85% of their phone calls. We do not consider false-positives in this model.

Each checkup policy is now defined as, $\Pi = (t_1, ..., t_n, y_{11}, ..., y_{nm})$. Further, let $N_i^{\Pi} \in \{0, 1\}$ be the indicator variable denoting whether or not the patient is detected in an ill state at time t_i, given policy Π. Our objective is to select the checkup policy that maximizes the probability of detecting the patient in an ill state (detection probability in shorthand):

$$\max_{\Pi} \sum_{i=1}^{n} \mathbb{E}[N_i^{\Pi}]. \tag{2.1}$$

2.3.2 Optimization Formulation

The time-to-develop the condition, δ, is described by a differentiable probability density function $g_\delta(\cdot)$, which is assumed to be independent of delay-time, D. This assumption is necessary for the mathematical formulation and is present in all related machine maintenance literature. We also confirm statistical independence of these two random variables in Section 2.5.1 using historical data. D has PDF $f(\cdot)$, CDF $F(\cdot)$, and complementary cumulative distribution function (CCDF) $\bar{F}(\cdot)$. Furthermore, the time-to-readmission, ρ, has probability density function $g_\rho(\cdot)$, which is the convolution of δ and D. The checkup optimization can be expressed as follows:

$$\max_{\substack{t_1,\ldots,t_n \\ y_{11},\ldots,y_{nm}}} \quad \sum_{i=1}^{n}\sum_{\beta=1}^{m} y_{i\beta} r_\beta \sum_{s=1}^{i} \int_{t_{s-1}}^{t_s} g_\delta(k)\bar{F}(t_i - k)\, dk \prod_{q=s}^{i-1}\sum_{\alpha=1}^{m} y_{q\alpha}(1 - r_\alpha) \tag{2.2}$$

$$\text{s.t.} \quad \sum_{l=1}^{m} y_{il} = 1, \quad \forall\, i \in \{1, ..., n\} \tag{2.3}$$

$$\sum_{i=1}^{n} y_{il} \leq w_l, \quad \forall\, l \in \{1, ..., m\} \tag{2.4}$$

$$0 \leq t_i < t_{i+1} \leq T, \quad \forall\, i \in \{1, ..., n-1\} \tag{2.5}$$

where $t_0 = 0$ and the empty product, Π, equals 1.

The first term in the objective, $y_{i\beta} r_\beta$, accounts for the detection rate of the method used for checkup i. The second term represents the probability that the patient developed the condition between checkups $(s-1)$ and s and is still not readmitted by checkup i. The last term (the product) represents the probability that checkups $s, ..., (i-1)$ all failed to properly detect the patient's existing condition. The constraint of Eq. (2.3) ensures that only one checkup method is utilized at each checkup time, Eq. (2.4) ensures that checkup method resource capacities are not violated, and Eq. (2.5) ensures proper ordering of the checkups.

Remark II.2. Note that our objective function only considers the probability of detection and does not account for how early the condition was detected. We chose this objective for several reasons. First, it is intuitive for the clinical audience and captures the essence of the post-discharge monitoring goal - to detect conditions and prevent readmissions. Second, there is no data, to our knowledge, that captures quantifies the benefits of capturing a condition earlier versus later. To capture this relationship, it would require the estimation of the detection probability as a function of the earliness of detection, which is difficult given the lack of delay-time related data. Nevertheless, capturing conditions early would likely be beneficial. It is possible to modify our objective function to achieve this, given proper data on the benefits of early detection.

2.3.3 Solution Approach

We solve this program numerically by dividing it into subproblems and enumerating all feasible y vectors. For each subproblem, we implemented an algorithm that combines a genetic algorithm (GA) with an ascent algorithm in the following fashion. The GA is used to generate solutions through random initialization, mutation, and crossover.

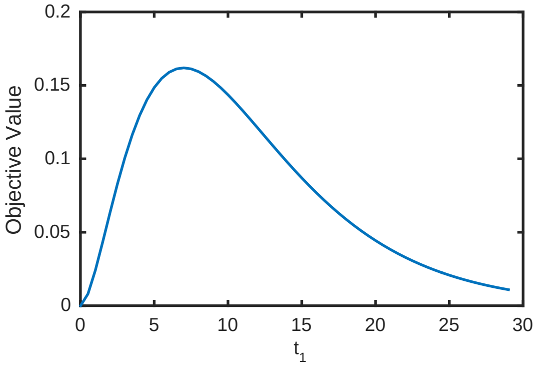
```
initialization
  Generate 200 solutions$^1$ according to the recursive construction described in Proposition
  II.4. Each of the 200 initial seeds is generated assuming a deterministic delay-time
  randomly sampled from the true delay-time distribution; $t \leftarrow 1$
while Not converged $^2$ or $t \leq 200$ or gradient norm $\leq 10^{-5}$ do
    1. Keep the top 25 fittest solutions and eliminate the rest 175 solutions
    2. Randomly mate solutions from the 25 solutions to generate 175 offspring solutions
    3. Mutate 20 randomly selected solutions by randomly permuting the timing
    4. Apply 5 iterations of gradient ascent to each solution
    5. $t \leftarrow t + 1$
end
Note:
$^1$ : A solution, for a $n-$checkup problem, is a $n$ dimensional vector. For example, $n = 2$,
  $(t_1, t_2)$ is a valid solution where $t$ is the timing of checkups. The fitness of a solution is its
  detection probability (i.e. the objective value).
$^2$ : We say the algorithm converged if the change in population average fitness is less than
  $10^{-5}$ from $t$ to $t + 1$.
```

Algorithm 1: Solution Procedure

In each generation, after the genetic operations, an ascent algorithm is applied to each of the solutions in the solution pool for no more than five iterations with decreasing step size. The master algorithm stops if the gradient is sufficiently small or the maximum number of iterations is reached. Note that the ascent algorithm alone is sufficient to find local optima if the distributions are differentiable with support on $(0, +\infty)$. The GA component is added to encourage escaping from local optima in the search for a global optimum and to handle distributions that are not differentiable and/or have finite support.

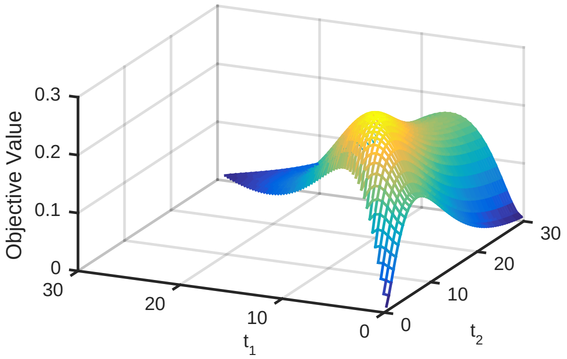
Remark II.3. Note that the objective function is not necessarily concave. For example, when the delay-time is deterministic and we are optimizing for only one perfect checkup, the concavity of the objective function is equivalent to the concavity of the probability density function of the time-to-develop the condition. However, under reasonable parameterizations in our numerical analysis, we found that our problem tends to have a unique optimum near the mode of the time-to-readmission curve (see

Figures 2.2 and 2.3). Hence the first order necessary conditions we analyze below provide strong intuition regarding the region of interest for scheduling checkups.



Note. time-to-develop the condition $g_\delta \sim \text{gamma}(1.81, 5.08)$, delay-time $D \sim \text{exponential}(2.35)$, one checkup with perfect detection rate

Figure 2.2: Objective Value for One Checkup



Note. time-to-develop the condition $g_\delta \sim \text{gamma}(1.81, 5.08)$, delay-time $D \sim \text{exponential}(2.35)$, two checkups with perfect detection rate

Figure 2.3: Objective Value for Two Checkups

2.3.4 Moving Parts and Assumptions

Our model consists of three moving parts that require estimation. The estimation of these moving parts is crucial and challenging due to data scarcity and censoring. In this section, we discuss each of the moving parts and modeling assumptions surrounding them. Later in Section 2.5, we discuss the estimation in detail and conduct sensitivity analysis

- Detection rate of imperfect checkups (r)

  The detection rate of an imperfect checkup is defined as the probability of detecting an existing condition. In our numerical analyses, we consider $r =$

0.6 for phone calls as a baseline and conduct sensitivity analyses by varying r between 0.2 and 1. In Section 2.4.3, we analyze the impact of detection rate (r) by studying gamma g_δ distributions.

- Time-to-develop the condition distribution (pdf: g_δ)
  The time-to-develop the condition distribution is the probability density of developing a readmission-causing condition after discharge. In order to establish the First Order Necessary Condition, we require g_δ to be continuously differentiable with support on $[0, T]$. In Section 2.4.1, we analyze the structure of the checkup timing assuming g_δ is unimodal. However, in Section 2.4.2, the unimodality assumption is relaxed.

- Delay-time distribution (pdf: f)
  The delay-time distribution is the probability density of the time between condition onset and readmission. We assume that the delay-time is independent of the time-to-develop the condition. In order to establish the First Order Necessary Condition, we assume f to be continuously differentiable with support on $[0, T]$. In Section 2.4.3, we analyzed the impact of detection rate (r) by studying exponentially distributed delay-time. Table 2.3 shows the results of the sensitivity analyses using different delay-time distributions.

Our model also assumes that 1) the 30-day post-discharge mortality rates are small relative to 30-day hospital readmission rates and therefore can be neglected; 2) the post-discharge checkup plan is not dynamically modified or updated; and 3) the planning horizon is finite (i.e., 30 days).

2.4 Structural Properties

In this section, we analyze special cases to develop structural insights, which are extended to more general cases through numerical analyses in Section 2.5. We first focus on the timing of checkups. Then we examine how different features such as stochastic delay-time, D, and different detection rates imply small modifications to the general timing structure. The analysis in Sections 2.4.1-2.4.3 serves to develop intuition into rules of thumb that are combined to design a practical, implementable policy for providers/practices described in Section 2.4.4, with each section providing a key building block. The overarching goal is to provide guidance toward a practical policy that is effective based only on historical data without relying on the optimization itself.

2.4.1 General Checkup Timing in Optimal Policies

We later show through numerical analyses (Section 2.5) that checkup timing has the highest impact on detecting an ill patient, so we begin our analysis with this feature. To understand the general structure of checkup timing, we analyze the case of current practice where standard protocol dictates a single doctor's office visit ($n = 1$). We begin by assuming a deterministic delay-time, $D = z \geq 0$, and a perfect detection rate. We later generalize these analytical results through numerical analyses in Section 2.5. The objective function for this special case can be rewritten as follows:

$$\max_{t_1} \ \mathbb{E}[N_1^{\Pi}] = \max_{t_1} \int_0^{t_1} g_\delta(k)(1 - F(t_1 - k))\, dk = \max_{t_1} \int_{t_1 - z}^{t_1} g_\delta(k)\, dk \tag{2.6}$$

The second equality follows from the fact that the deterministic delay-time, $D = z \geq 0$, implies $F(t_1 - k) = 1$, if $t_1 - k \geq z$, and $F(t_1 - k) = 0$ otherwise.

Differentiating the objective function with respect to t_1 yields the following First Order Necessary Condition (FONC) for optimality

$$0 = \frac{\partial}{\partial t_1} \int_{t_1 - z}^{t_1} g_\delta(k)\, dk \implies g_\delta(t_1 - z) = g_\delta(t_1) \tag{2.7}$$

Based on results from our data on readmitted cystectomy patients, we also leverage the fact that the time-to-develop the condition of readmitted patients, $g_\delta(k)$, is unimodal. By unimodality of $g_\delta(k)$, the condition $g_\delta(t_1 - z) = g_\delta(t_1)$ implies that $(t_1 - z)$ is before the mode of $g_\delta(k)$ and t_1 is after the mode of $g_\delta(k)$. Thus, the probability density of developing a condition at $t_1 - z$ must equal the probability density of a condition developing at t_1. In practical terms, this informs decision-makers that, given only one checkup opportunity, they should schedule the checkup a little bit ($< z$) after the time when conditions are most likely to develop.

Next, consider a more aggressive approach with n checkups. The following proposition shows that the general multivariate optimization can be transformed into a univariate optimization, focused only on the time of the first checkup. The proposition indicates the best way to achieve maximum coverage of high risk times in a patient's post-discharge recovery. Specifically, we want our checkups to cover as much of the period of time when the patient is at highest risk of having a readmission-causing condition as possible. This results in the following two insights. First, if checkups are too close (i.e. spaced closer than z time units), there is unnecessary overlap in the coverage (i.e. two checkups covering the same time period). Better coverage can be achieved by spacing them further apart without any loss in detection (since delay-time is deterministic). Second, we want the checkups to cover the high-risk period (i.e. the time window containing the highest time-to-develop the condition density), hence it is best to center all of the checkups around the mode of the time-to-develop the condition distribution, since the density is decreasing monotonically on either side of the mode.

Proposition II.4. *If the delay-time is deterministic ($D = z$ with probability 1) and the time-to-develop the condition g_δ is unimodal, then (1) it is sufficient to optimize t_1 only; (2) the checkups are spaced z days apart equidistantly ; and (3) the densities of developing the condition are equal at $t_1 - z$ and t_n*

$$\max_{t_1,\ldots,t_n} \sum_{i=1}^{n} \mathbb{E}[N_i^{\Pi}] = \max_{t_1} \int_{t_1-z}^{t_1+(n-1)z} g_\delta(k)\, dk \tag{2.8}$$

$$s.t.\ g_\delta(t_1 - z) = g_\delta(t_1 + (n-1)z) \tag{2.9}$$

$$t_{i+1} = t_i + z, \quad \forall\, i \in \{1, ..., n-1\} \tag{2.10}$$

Proof. We first show that $(t_n - t_1) = (n-1)z$. In other words, the time between the first and last checkups is exactly $(n-1)z$. □

The structure of our objective function appropriately avoids double counting the detection of conditions. To see how, notice that under the assumptions of deterministic delay-time ($D = z$) and perfect detection rates, the objective function in Eq. (2.2) becomes

$$\sum_{i=1}^{n} \int_{t_{i-1}}^{t_i} g_\delta(k)\bar{F}(t_i - k)\, dk = \sum_{i=1}^{n} \int_{\max(t_{i-1}, t_i - z)}^{t_i} g_\delta(k)\, dk \tag{2.11}$$

Thus, only the earliest successful checkup contributes a positive amount to the objective function. For example, if a condition was present during a time interval $(\delta,\ \delta+D)$ and three checkups were scheduled at some arbitrary times $t_i,\ t_j,\ t_k \in (\delta,\ \delta + D)$, then only the checkup at $\min\{t_i, t_j, t_k\}$ contributes a positive amount to the objective function.

This implies that an optimal solution must be such that the intervals $(t_i - z, t_i)$ are disjoint for all i. To see why, consider an arbitrary checkup schedule that has nondisjoint intervals. Suppose the smallest index corresponding to nondisjoint intervals is $j < n$ such that $(t_j - z, t_j)$ and $(t_{j+1} - z, t_{j+1})$ are nondisjoint. Then, $t_j = t_{j+1} - z + \gamma$ with $\gamma \in (0, z)$. We can construct another solution that is strictly better, by increasing t_{j+1} by $(z - \gamma)$. This increases the objective value by a non-negative amount:

$$\begin{cases} \int_{t_{j+1}}^{\min(t_{j+2}-z, t_{j+1}+z-\gamma)} g_\delta(k)\, dk, & \text{if } j \leq n-2 \\ \int_{t_{j+1}}^{t_{j+1}+z-\gamma} g_\delta(k)\, dk, & \text{if } j = n-1 \end{cases} \tag{2.12}$$

If $j = n - 1$, then the change in the objective value is strictly positive. Similarly, if $j \leq n - 2$ and the upper limit of the integral in Eq. (2.12) is $t_{j+1} + z - \gamma$, the change in the objective value is strictly positive and the adjustment of t_{j+1} leaves the intervals $(t_{j+1} - z, t_{j+1})$ and $(t_{j+2} - z, t_{j+2})$ disjoint. The last case we need to consider is if $j \leq n - 2$ and the upper limit of the integral in Eq. (2.12) is $t_{j+2} - z$. In this case, there is a non-negative change in the objective value and the intervals $(t_{j+1} - z + (z - \gamma), t_{j+1} + (z - \gamma))$ and $(t_{j+2} - z, t_{j+2})$ become nondisjoint, so we can repeat the steps above. This process terminates in finite iterations and results in a strictly positive change in the objective value. Thus, we can conclude that an optimal solution must satisfy $(t_n - t_1) \geq (n - 1)z$. Figure 2.4 illustrates how Eq. 2.12 is derived in the case of $j \leq n - 2$ and $t_{j+2} - z \geq t_{j+1} + z - \gamma$.

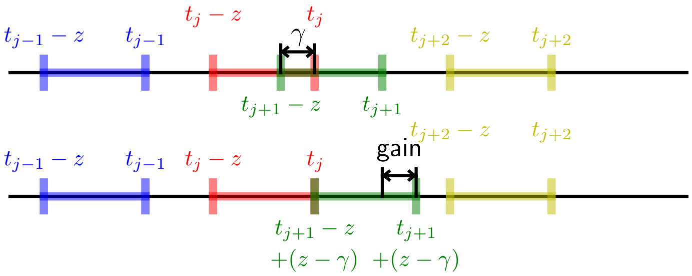


Figure 2.4: Schematic Sketch for Eq. 2.12

We will now argue that an optimal solution cannot have $(t_n - t_1) > (n - 1)z$. Combining this with our previous finding yields our desired result that an optimal solution must satisfy $(t_n - t_1) = (n - 1)z$. If $(t_n - t_1) > (n - 1)z$, then $\exists\, i \in \{1, ..., n - 1\}$ such that $t_{i+1} - t_i = z + \gamma$, with $\gamma > 0$. In other words, there is at least one pair of consecutive checkups that are spaced farther than z apart. A checkup schedule with this property is necessarily suboptimal because the objective value can be improved by adjusting either t_i or t_{i+1} (without changing any other checkup times), depending on their relative positions to the mode of $g_\delta(\cdot)$.

In particular, if $t_i < t_{i+1} \leq$ mode of $g_\delta(\cdot)$, we can increase the objective value by shifting the checkup i from t_i to $t_i + \epsilon$, where $\epsilon \in (0, \gamma]$. This increases the objective value by $\int_{t_i}^{t_i+\epsilon} g_\delta(k)dk - \int_{\max(t_i-z,t_{i-1})}^{\max(t_i+\epsilon-z,t_{i-1})} g_\delta(k)dk$. Observe that the second term is integrated over the interval $[\max(t_i - z, t_{i-1}), \max(t_i + \epsilon - z, t_{i-1})]$, which has length $\leq \epsilon$. Since the second integral interval is to the left of the first integral interval, and these two intervals are to the left of the mode, it follows that $\int_{t_i}^{t_i+\epsilon} g_\delta(k)dk - \int_{\max(t_i-z,t_{i-1})}^{\max(t_i+\epsilon-z,t_{i-1})} g_\delta(k)dk > 0$. By symmetry, if $t_i > t_{i+1} \geq$ mode of $g_\delta(\cdot)$, we can shift checkup $i+1$ from t_{i+1} to $t_{i+1} - \epsilon$, where $\epsilon \in (0, \gamma]$, to achieve a nonnegative improvement . If $t_i <$ mode of $g_\delta(\cdot) < t_{i+1}$, we can achieve a nonnegative improvement by moving t_i to the right (if $g_\delta(t_i) \geq g_\delta(t_{i+1})$) or moving t_{i+1} to the left (if $g_\delta(t_i) < g_\delta(t_{i+1})$). The improvement is strictly positive if g_δ is strictly unimodal, i.e., has a unique mode.

We can now conclude that an optimal solution must satisfy $(t_n - t_1) = (n-1)z$. Given our previous result that an optimal solution must have checkup times such that the intervals $(t_i,\ t_i + z)\ \forall\ i$ are disjoint, this implies that an optimal solution must be of the form $t_i = t_{i-1} + z,\ \forall\ i \in \{2, ..., n\}$. This is equivalent to letting $t_i = t_1 + (i-1)z,\ \forall\ i\ \in \{2, ..., n\}$. Note that this only holds assuming the delay-time is deterministic. This proves that $\max_{t_1} \int_{t_1-z}^{t_1+(n-1)z} g_\delta(k)\, dk$ is in fact optimal.

Remark II.5. If the distribution of the time-to-develop the condition is right/left skewed (yet still unimodal), this does not affect our optimality results at all, since our results assume nothing about the skewness of the curve. The checkups would still be centered around the mode, even though the mode will be later/sooner in the 30-day readmission window. If the distribution is not unimodal, then alternative optima might exist. Nonetheless, some of the properties from Proposition II.4 still hold. For example, under the assumptions of bimodal distribution and deterministic delay-time, we know the following: (1) if there was only one checkup to place, Proposition II.4 still holds; (2) if there were multiple checkups, checkups are placed no closer than z days apart (might be farther than z days apart depending on the shape of the bimodal curve). For the general case with multiple modes, the First Order Necessary Conditions still hold and the problem can still be solved numerically.

From Proposition II.4, we see that the problem effectively becomes the single checkup problem while letting $D = nz$. Thus, an optimal solution in the case of perfect inspection checkups and deterministic delay-times must satisfy the following conditions

$$g_\delta(t_1 - z) = g_\delta(t_1 + (n-1)z) \tag{2.13}$$

$$t_{i+1} = t_i + z, \quad \forall\, i \in \{1, ..., n-1\} \tag{2.14}$$

Reducing the n-dimensional optimization problem to a univariate optimization problem makes these conditions especially valuable because these univariate optimizations are easy to solve using ascent search or binary search even without specialized computer software. This can be achieved by solving the univariate FONC equation (which is in the form of $\psi(t_1) = 0$) using binary search since $g_\delta(t_1 - s) - g_\delta(t_1 + (n-1)z)$ is monotone increasing for a unimodal function), ascent search, or Newton's method. Furthermore, the conditions imply that an optimal policy schedules one contiguous block of checkups with the checkups collectively covering a time of length nz. Practically speaking, this informs decision-makers that if they have n perfect checkups (e.g. doctors' office visits), then the checkups should be scheduled surrounding the time when conditions develop most frequently such that there are z (delay-time) time units between each checkup.

2.4.2 Effect of Stochastic Delay-time on Optimal Checkup Timing

Proposition II.4 gives us the block structure of an optimal checkup policy with deterministic delay-time, D. In this section, we investigate how stochastic D affects the spacing of checkups within the block of checkups. First, relaxing the assumption that $D = z$, the objective function becomes

$$\max_{t_1,\ldots,t_n} \sum_{i=1}^{n} \int_{t_{i-1}}^{t_i} g_\delta(k)[1 - F(t_i - k)]\, dk \tag{2.15}$$

which for $n = 1$ equals $\int_0^{t_1} g_\delta(k)[1 - F(t_1 - k)]\, dk$, resulting in the following FONC:

$$0 = \frac{\partial}{\partial t_1} \int_0^{t_1} g_\delta(k)[1 - F(t_1 - k)]\, dk \implies g_\delta(t_1) = \int_0^{t_1} g_\delta(k) f(t_1 - k)\, dk = g_\rho(t_1) \tag{2.16}$$

Notice that the RHS of Eq. (2.16) is the formula for the probability density associated with a readmission occurring at t_1. This implies that at an optimal t_1, the marginal rate of developing a condition (i.e. the marginal increase in patients who could be detected if t_1 was increased) is equal to the marginal rate of a readmission occurring (i.e. the marginal lost patients that would be readmitted if t_1 was increased). Both results extend our intuition from Section 2.4.1 to the case of stochastic delay-time.

Generalizing the FONC to an arbitrary number of checkups yields

$$\int_{t_{i-1}}^{t_i} g_\delta(k) f(t_i - k)\, dk = g_\delta(t_i) F(t_{i+1} - t_i), \quad \forall\, i \in \{1, ..., n-1\} \tag{2.17}$$

$$\int_{t_{n-1}}^{t_n} g_\delta(k) f(t_n - k)\, dk = g_\delta(t_n) \tag{2.18}$$

The intuition behind these equations is similar to when $n = 1$ in that the optimal solution balances the marginal rate of catching a condition with the i^{th} checkup with the marginal rate of missing a later condition. The LHS of Eq. (2.17) is the probability of checkup i detecting a condition developed between t_{i-1} and t_i. Since the perfect checkup at t_{i-1} ensures t_i will only detect conditions between t_{i-1} and t_i, the LHS of Eq. (2.17) can be thought of as the marginal benefit of moving inspection i slightly to the right from t_i to $t_i + \epsilon$ (as $\epsilon \to 0^+$), and therefore capturing more conditions that could have developed between t_i and $t_i + \epsilon$. This is essentially the marginal opportunity cost. The RHS of Eq. (2.17) is the probability of t_{i+1} missing the condition developed after t_i. This is analogous to lost sales, in that it represents the marginal rate of patients developing a condition at t_i and being readmitted before the next inspection at t_{i+1}.

Rearranging the terms of Eq. (2.17) implies the timing between inspections follows a newsvendor-type solution:

$$t_{i+1} - t_i = F^{-1}\left(\frac{\int_{t_{i-1}}^{t_i} g_\delta(k) f(t_i - k)\, dk}{g_\delta(t_i)}\right) \tag{2.19}$$

The structure of Eq. (2.19) closely resembles the equation for the optimal stocking quantity in traditional newsvendor problems. This highlights the inherent tradeoff between (1) scheduling checkups closer together to increase the likelihood of detecting illnesses that develop between the checkups and (2) scheduling checkups farther apart to have the opportunity to detect more illnesses by covering a wider span of time. Both of these tradeoffs are inherently linked to the density of the delay-time function, F. Thus, the distance between any two checkups is determined by a solution where the delay-time density functions as the demand function.

It is worth noting that one can construct a recursive algorithm to solve the optimization in light of Eq. (2.19). For instance, given $t_0 = 0$ and t_1, one can determine $t_2 = t_1 + F^{-1}\left(\frac{\int_{t_0}^{t_1} g_\delta(k) f(t_1 - k)\, dk}{g_\delta(t_1)}\right)$. Recursively, one can determine $t_3, ..., t_n$. This reduces the problem to a univariate optimization where t_1 is the only decision variable. Moreover, an optimal solution must exist since we are maximizing a continuous function over a compact set. For the general case with stochastic delay-time and imperfect checkups, our solution procedure utilizes this recursive construction to generate the initial solution seeds.

If the solution to the FONCs is not unique, then one can solve the following univariate maximization to generate the optimal checkup policy.

$$\max_{t_1} \quad \sum_{i=1}^{n} \mathbb{E}[N_i^{\Pi(t_1)}] \tag{2.20}$$

$$\text{s.t. } t_1 \in [0,T] \tag{2.21}$$

In this optimization problem, the checkup policy $\Pi(t_1)$ is drawn from a set of potential candidates based on the FONCs:

$$\Pi(t^1) = [t_1, t_2, ..., t_n]^T \tag{2.22}$$

$$= \begin{bmatrix} t_1 \\ t_1 + F^{-1}\left(\dfrac{\int_{t_0}^{t_1} g_\delta(k) f(t_1 - k)\, dk}{g_\delta(t_1)}\right) \\ t_2 + F^{-1}\left(\dfrac{\int_{t_1}^{t_2} g_\delta(k) f(t_2 - k)\, dk}{g_\delta(t_2)}\right) \\ \vdots \\ t_{n-1} + F^{-1}\left(\dfrac{\int_{t_{n-2}}^{t_{n-1}} g_\delta(k) f(t_{n-1} - k)\, dk}{g_\delta(t_{n-1})}\right) \end{bmatrix} \tag{2.23}$$

Remark II.6. The analyses in this section are based on the KKT conditions, which assume (1) g_δ has support on $[0,T]$; (2) f has support on $[0,\infty)$; and (3) g_δ and f are continuously differentiable. It is worth highlighting that these results do not require unimodality.

2.4.3 Effect of Imperfect Inspection Checkups on Optimal Checkup Timing

As previously mentioned, hospitals have various checkup methods available with differing detection rates. Hence, it is valuable from both a practical and a theoretical perspective to understand how the optimal timing of checkups is affected by the detection rates of the checkups. For the purpose of exposition, we let $r_{(i)}\ \forall\ i \in \{1,...,n\}$ denote the detection rate of the checkup method employed at time t_i. We begin by considering the case where $n = 2$ and $r_{(1)} = r_{(2)} = r$. This yields the following objective value

$$r\int_0^{t_1} g_\delta(k)[1-F(t_1-k)]\, dk + (1-r)r\int_0^{t_1} g_\delta(k)[1-F(t_2-k)]dk + r\int_{t_1}^{t_2} g_\delta(k)[1-F(t_2-k)]\, dk \tag{2.24}$$

We can then derive FONCs as follows

$$\int_0^{t_1} g_\delta(k) f(t_1 - k)\, dk = g_\delta(t_1)\left(F(t_2 - t_1) + (1-r)(1 - F(t_2 - t_1))\right) \tag{2.25}$$

$$(1-r)\int_0^{t_1} g_\delta(k) f(t_2 - k)\, dk + \int_{t_1}^{t_2} g_\delta(k) f(t_2 - k)\, dk = g_\delta(t_2) \tag{2.26}$$

The intuition behind these equations are similar to the perfect checkup case in Eq. (2.17) and (2.18). The LHS of Eq. (2.25) is the probability density of detecting a condition that developed between 0 and t_1, i.e. marginal rate of gain in terms of detection. The RHS of Eq. (2.25) is the marginal density of missing a condition developed after t_1, i.e. loss sales. To see this, note the term $g_\delta(t_1)F(t_2 - t_1)$ appears and has the same intuition as in Eq. (2.17), i.e., the condition developed after t_1 but the patient was readmitted before t_2. However, the inspection at t_2 could also miss an extant condition due to the imperfect detection. This event is captured by the term $g_\delta(t_1)(1-r)(1 - F(t_2 - t_1))$, which implies the condition was detectable at time t_2 but failed to be detected. Eq. (2.26) represents the tradeoff between lost sales (RHS) and marginal change in detection (LHS). The RHS of Eq. (2.26) is the marginal density of a condition developing at time t_2, i.e. lost sales as before since any conditions developing after t_2 will not be detected. The first term on the LHS is the density of a condition being detectable at time t_2 that developed on 0 to t_1 and was missed by the inspection at time t_1, i.e. the marginal change in detection for conditions missed by the first inspection. The second term on the LHS is the probability density of detecting a condition that developed between t_1 and t_2.

Using the FONCs, we next show that as r increases, the two checkups move farther apart. Hence, by improving the detection rate of a particular method, the doctors should place inspections farther apart and can cover a larger time period in which to catch potentially developing conditions. The intuition behind this is that with a poor detection rate, a subsequent inspection can catch a condition that was previously missed if placed closer to the previous inspection. This comes at the expense of covering less overall timespan, as placing this inspection earlier will miss the opportunity to catch later developing conditions. As the detection rate increases, however, there is a smaller benefit of catching conditions missed by a previous inspection, since fewer patients are missed the first time.

For the analysis, let t_1^* and t_2^* be the optimal values of t_1 and t_2, respectively. To show this property analytically, we first introduce an inequality that relates the probability densities of developing the condition and readmission.

Definition II.7. Assuming g_ρ and g_δ are differentiable, the *delayed readmission log-*

likelihood inequality at time t is defined as $\frac{\mathrm{d}}{\mathrm{d}t}\log g_\delta(t) = \frac{g'_\delta(t)}{g_\delta(t)} \leq \frac{\mathrm{d}}{\mathrm{d}t}\log g_\rho(t) = \frac{g'_\rho(t)}{g_\rho(t)}$.

This inequality states that, at time t, the derivative of the log-likelihood of developing the condition is less than or equal to the derivative of the log-likelihood of readmission. This is similar to previous results we have seen relating the density functions of time-to-develop the condition, δ, and time-to-readmission, ρ. The following remark shows that this condition holds for Erlang and exponential distributions. The condition has been verified numerically for other distributions we use in our numerical studies (see Table 2.3 in Section 2.5). As we shall see in our numerical analyses, the shape of Erlang distributions resembles the observed time-to-develop the condition, and an exponential distribution is actually the best fit distribution for the delay-time.

Remark II.8. If the time-to-develop the condition follows an Erlang distribution with scale μ and shape parameter k_δ (Erlang(k_δ, μ)), and the delay-time follows Erlang(k_D, μ), then the time-to-readmission follows an Erlang(k_ρ, μ) where $k_\rho = k_\delta + k_D$. The delayed readmission log-likelihood inequality becomes $(k_\delta - 1)t^{-1} \leq (k_\rho - 1)t^{-1}$, which holds $\forall t > 0$.

The following lemma shows that, as the detection rate increases, the first inspection will be placed closer to the patient's time of discharge (i.e. moved earlier).

Lemma II.9. *If the delayed readmission log-likelihood inequality holds, then t_1^* decreases in r.*

Using this lemma, we now prove An increase in r causes the LHS of Eq. (2.25) to become smaller than the RHS, so t_1^* is no longer optimal. After simple algebraic manipulation of Eq. (2.25), r can be expressed in terms of t_1^* as:

$$\begin{aligned} r(t_1^*) &= \frac{g_\delta(t_1^*) - \int_0^{t_1^*} g_\delta(k) f(t_1^* - k)\, dk}{g_\delta(t_1^*)\left[1 - F(t_2^* - t_1^*)\right]} \\ &= \frac{1}{1 - F(t_2^* - t_1^*)} - \frac{\int_0^{t_1^*} g_\delta(k) f(t_1^* - k)\, dk}{g_\delta(t_1^*)[1 - F(t_2^* - t_1^*)]} \end{aligned} \tag{2.27}$$

Differentiating this function with respect to t_1^* yields

$$\begin{aligned} \frac{\partial r}{\partial t_1^*} &= -\frac{f(t_2^* - t_1^*)}{[1 - F(t_2^* - t_1^*)]^2} - \frac{\int_0^{t_1^*} g_\delta(k) f'(t_1^* - k)\, dk + g_\delta(t_1^*) f(0)}{g_\delta(t_1^*)[1 - F(t_2^* - t_1^*)]} \\ &+ \frac{f(t_2^* - t_1^*) \int_0^{t_1^*} g_\delta(k) f(t_1^* - k)\, dk}{g_\delta(t_1^*)[1 - F(t_2^* - t_1^*)]^2} + \frac{g'_\delta(t_1^*) \int_0^{t_1^*} g_\delta(k) f(t_1^* - k)\, dk}{g_\delta(t_1^*)^2[1 - F(t_2^* - t_1^*)]} \end{aligned} \tag{2.28}$$

Then, notice the following:

$$g_\rho(t_1^*) = \int_0^{t_1^*} g_\delta(k) f(t_1^* - k)\, dk \tag{2.29}$$

$$g_\rho'(t_1^*) = \int_0^{t_1^*} g_\delta(k) f'(t_1^* - k)\, dk + g_\delta(t_1^*) f(0) \tag{2.30}$$

We are interested in the situation where the derivative in Eq. (2.28) is non-positive. Plugging Eq. (2.29) and (2.30) into Eq. (2.28), this is equivalent to saying:

$$0 \geq -\frac{f(t_2^* - t_1^*)}{[1 - F(t_2^* - t_1^*)]^2} - \frac{g_\rho'(t_1^*)}{g_\delta(t_1^*)[1 - F(t_2^* - t_1^*)]} + \frac{f(t_2^* - t_1^*) g_\rho(t_1^*)}{g_\delta(t_1^*)[1 - F(t_2^* - t_1^*)]^2} + \frac{g_\delta'(t_1^*) g_\rho(t_1^*)}{g_\delta(t_1^*)^2[1 - F(t_2^* - t_1^*)]} \tag{2.31}$$

Combining like terms, we have

$$0 \geq \frac{f(t_2^* - t_1^*)}{[1 - F(t_2^* - t_1^*)]^2}\Big(\frac{g_\rho(t_1^*)}{g_\delta(t_1^*)} - 1\Big) + \frac{1}{g_\delta(t_1^*)[1 - F(t_2^* - t_1^*)]}\Big(\frac{g_\delta'(t_1^*) g_\rho(t_1^*)}{g_\delta(t_1^*)} - g_\rho'(t_1^*)\Big) \tag{2.32}$$

$$\implies \frac{f(t_2^* - t_1^*)}{[1 - F(t_2^* - t_1^*)]}\Big(1 - \frac{g_\rho(t_1^*)}{g_\delta(t_1^*)}\Big) \geq \frac{1}{g_\delta(t_1^*)}\Big(\frac{g_\delta'(t_1^*) g_\rho(t_1^*)}{g_\delta(t_1^*)} - g_\rho'(t_1^*)\Big) \tag{2.33}$$

Multiplying both sides by $g_\delta(t_1^*)$ yields

$$\Big(g_\delta(t_1^*) - g_\rho(t_1^*)\Big)\frac{f(t_2^* - t_1^*)}{[1 - F(t_2^* - t_1^*)]} \geq \frac{g_\delta'(t_1^*) g_\rho(t_1^*)}{g_\delta(t_1^*)} - g_\rho'(t_1^*) \tag{2.34}$$

From Eq. (2.25), it follows that $g_\rho(t_1^*) \leq g_\delta(t_1^*)$. Then, the LHS of Eq. (2.34) is positive. Hence, it is sufficient to show that the RHS of Eq. (2.34) is negative. That is, it is sufficient that

$$\frac{g_\delta'(t_1^*) g_\rho(t_1^*)}{g_\delta(t_1^*)} \leq g_\rho'(t_1^*) \iff \frac{g_\delta'(t_1^*)}{g_\delta(t_1^*)} \leq \frac{g_\rho'(t_1^*)}{g_\rho(t_1^*)} \tag{2.35}$$

The above inequality holds as a result of the delayed readmission log-likelihood inequality, which completes our proof.

Leveraging Lemma II.9, we next show that the gap, $t_1^* - t_2^*$, widens as r increases. Notice that the optimal timing t_1^* and t_2^* is the solution to the FONCs, i.e. Eq. (2.25) and (2.26). For general delay-time and time-to-develop the condition distributions, the FONCs are essentially a set of integral equations without a closed form solution. In the following theorem, we consider the case where the delay-time is exponential and the time-to-develop the condition is Erlang so that the time-to-readmission is in closed form since the convolution of exponential and Erlang distributions is an Erlang distribution. The structure and shape of the Erlang and exponential distributions are close to what is observed in practice through our numerical analyses (see Figure 2.7 in Section 2.5.1). With exponential-Erlang distributions, Eq. (2.26) effectively becomes a polynomial where t_1^* can be directly expressed in closed form.

Theorem II.10 now shows, for the case of Erlang and exponential densities for δ and D, that the two tests move farther apart as the detection rate increases. This result is later generalized in our numerical study.

Theorem II.10. *If the time-to-develop the condition follows Erlang(k, μ) and the delay-time follows exponential(μ), then $t_2^* - t_1^*$ strictly increases in r.*

We begin with the following technical lemma.

Lemma II.11. *If the delayed readmission log-likelihood inequality holds, then $\dfrac{g_\rho(t)}{g_\delta(t)}$ increases in t.*

Proof.

$$\frac{\partial}{\partial t}\left(\frac{g_\rho(t)}{g_\delta(t)}\right) = \frac{g_\delta(t)g_\rho'(t) - g_\rho(t)g_\delta'(t)}{g_\delta^2(t)} \tag{2.36}$$

$$= \frac{\dfrac{g_\rho'(t)}{g_\rho(t)} - \dfrac{g_\delta'(t)}{g_\delta(t)}}{\dfrac{g_\delta^2(t)}{g_\rho(t)g_\delta(t)}} \geq 0 \tag{2.37}$$

The last inequality follows from the delayed readmission log-likelihood inequality. □

Using this lemma, we now prove Theorem II.10.

Proof. Without loss of generality, assume $\mu = 1$. For $\mu \neq 1$, the problem can be scaled. We then rewrite Eq. (2.26) as follows:

$$g_\rho(t_2^*) - g_\delta(t_2^*) = r\int_0^{t_1^*} g_\delta(s)f(t_2^* - s)ds \;\Leftrightarrow\; \frac{e^{-t_2^*}t_2^{*k}}{k!} - \frac{e^{-t_2^*}t_2^{*k-1}}{(k-1)!} = r\int_0^{t_1^*}\frac{e^{-t_2^*}s^{k-1}}{(k-1)!}ds \tag{2.38}$$

$$\frac{e^{-t_2^*}t_2^{*k}}{k!} - \frac{e^{-t_2^*}t_2^{*k-1}}{(k-1)!} = r\frac{e^{-t_2^*}t_1^{*k}}{k!} \;\Leftrightarrow\; t_2^{*k} - kt_2^{*k-1} = rt_1^{*k} \tag{2.39}$$

$$\Leftrightarrow\; t_1^* = \left(\frac{t_2^{*k} - kt_2^{*k-1}}{r}\right)^{\frac{1}{k}} \tag{2.40}$$

The first and second derivatives of t_1^* with respect to t_2^* are

$$\frac{\partial t_1^*(t_2^*)}{\partial t_2^*} = \frac{(t_2^* - k + 1)\left(\frac{t_2^{*k-1}(t_2^*-k)}{r}\right)^{\frac{1}{k}}}{t_2^*(t_2^* - k)} \text{ and } \frac{\partial^2 t_1^*(t_2^*)}{\partial t_2^{*2}} = -\frac{(k-1)\left(\frac{t_2^{*k-1}(t_2^*-k)}{r}\right)^{\frac{1}{k}}}{t_2^{*2}(t_2^* - k)^2} \tag{2.41}$$

Based on the first and second derivatives, we show that $t_1^*(t_2^*)$ has the following properties: (1) t_1^* strictly increases in t_2^*; (2) $t_1^*(t_2^*)$ is concave; (3) $\lim_{t_2^*\to+\infty}\frac{\partial t_1^*(t_2^*)}{\partial t_2^*} = (1/r)^{\frac{1}{k}} > 1$; and (4) $\frac{\partial t_1^*(t_2^*)}{\partial t_2^*} > 1, \forall t_1^*, t_2^*$.

For (1), notice that $(t_2^* - k)$ has to be strictly positive for $t_1^* \in \mathbb{R}^+$. Hence, $\frac{\partial t_1^*(t_2^*)}{\partial t_2^*} > 0$, which implies t_1^* strictly increases in t_2^*. For (2), since $(t_2^* - k) > 0$, it is clear that $\frac{\partial^2 t_1^*(t_2^*)}{\partial {t_2^*}^2} < 0$ for $k > 1$, integer. Hence $t_1^*(t_2^*)$ is concave. To see (3), for $k > 1$ and $r \in (0,1)$, we have

$$\begin{aligned}
\lim_{t_2^*\to+\infty}\frac{\partial t_1^*(t_2^*)}{\partial t_2^*} &= \lim_{t_2^*\to+\infty}\frac{(t_2^*-k+1)\left(\frac{{t_2^*}^{k-1}(t_2^*-k)}{r}\right)^{\frac{1}{k}}}{t_2^*(t_2^*-k)} && (2.42)\\
&> \lim_{t_2^*\to+\infty}\frac{(t_2^*-k)({t_2^*}^{k-1}(t_2^*-k))^{1/k}}{t_2^*(t_2^*-k)}\left(\frac{1}{r}\right)^{1/k} && (2.43)\\
&= \lim_{t_2^*\to+\infty}\left(\frac{{t_2^*}^k - k{t_2^*}^{k-1}}{{t_2^*}^k}\right)^{1/k}\left(\frac{1}{r}\right)^{1/k} && (2.44)\\
&= \lim_{t_2^*\to+\infty}\left(1-\frac{k}{t_2^*}\right)^{1/k}\left(\frac{1}{r}\right)^{1/k} = \left(\frac{1}{r}\right)^{1/k} > 1 && (2.45)
\end{aligned}$$

Finally, (4) follows from properties 2 and 3. Given the four properties above, Figure 2.5 sketches $t_1^*(t_2^*)$ schematically.

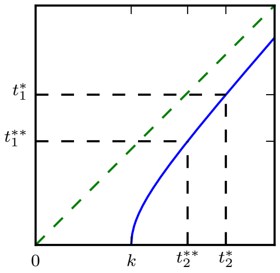


Figure 2.5: Schematic Sketch of t_1^* as a Function of t_2^*

Consider optimal t_1^* and t_2^* with detection rate r. As r increases, t_1^* decreases (Lemma II.9). By property 1, t_2^* also decreases. Denote the new optimal solution as t_1^{**} and t_2^{**}. As shown in Figure 2.5, since the slope of $t_1^*(t_2^*)$ is always strictly greater than one, it follows that $t_2^* - t_2^{**} < t_1^* - t_1^{**}$. Therefore $t_2^{**} - t_1^{**} > t_2^* - t_1^*$ as desired, which completes our proof. □

Proposition II.12. *Under the assumptions of Theorem II.10, if the detection rate changes from r to $r+\epsilon, (\epsilon > 0)$, then the increase in the gap between the two checkups is bounded above by $1-r$ (if $k = 1$) or $2(r+\epsilon)k$ (if $k \geq 2$).*

Proof. For Erlang-exponential distributions, Eq. (2.25) and (2.26) (FONCs) become:

$$e^{t_2}(k-t_1) - e^{-t_1}kr = 0 \tag{2.46}$$

$$r{t_1}^k = {t_2}^k - k{t_2}^{k-1} \tag{2.47}$$

Suppose r increases to $r+\epsilon$, from Theorem II.10, we know that t_1 moves to $t_1 - x, x > 0$.

Suppose t_2 moves to $t_2 - y, y > 0$, at the new optimum, Eq. (2.26) becomes:

$$r{t_1}^k = {t_2}^k - k{t_2}^{k-1} \tag{2.48}$$

$$(r+\epsilon)(t_1-x)^k = (t_2-y)^k - k(t_2-y)^{k-1} \tag{2.49}$$

For $k = 1$, we have

$$(r+\epsilon)x = y - \epsilon t_1 \tag{2.50}$$

We would like to express $x - y$ as a function of r and ϵ then put lower and upper bounds on it.

Lower bound: One trivial lower bound is $x - y \geq 0$ (result of Theorem II.10)

$$(r+\epsilon)x - y = \epsilon t_1 \tag{2.51}$$

$$\Leftrightarrow (r+\epsilon)x - (r+\epsilon)y \geq \epsilon t_1 > 0 \tag{2.52}$$

Upper bound: From Eq. (2.46) we know $t_1 < k = 1$. Also, we know that $x \leq t_1$. So

$$(r+\epsilon)x - y = \epsilon t_1 \tag{2.53}$$

$$-y = \epsilon t_1 - (r+\epsilon)x \tag{2.54}$$

$$x - y = x - (r+\epsilon)x + \epsilon t_1 \tag{2.55}$$

$$x - y \leq (1-r-\epsilon)t_1 + \epsilon t_k \leq (1-r) \tag{2.56}$$

For $k \geq 2$:

$$(t_2-k)t_2^{k-1} = rt_1t_1^{k-1} \tag{2.57}$$

Since $t_2 > t_1$ and $k \geq 2$, we have $t_2^{k-1} \geq t_1^{k-1}$. Then

$$t_2 - k \leq rt_1 \leq rk \quad (\text{since } t_1 < k) \tag{2.58}$$

$$t_2 \leq (r+1)k \tag{2.59}$$

Now we bound t_1.

$$e^{t_2}(k - t_1) - e^{-t_1}kr = 0 \tag{2.60}$$

$$(k - t_1) - e^{-t_1}kr \leq 0 \tag{2.61}$$

$$(k - t_1) - kr \leq 0 \tag{2.62}$$

$$(1 - r)k - t_1 \leq 0 \tag{2.63}$$

$$t_1 \geq (1 - r)k \tag{2.64}$$

The bounds for t_1 and t_2 at the new equilibrium (i.e. $t_1 - x$ and $t_2 - y$ are optimal for $t + \epsilon$):

$$t_2 - y \leq (1 + r + \epsilon)k \tag{2.65}$$

$$x - t_1 \leq -(1 - r - \epsilon)k \tag{2.66}$$

Combine the two inequalities, we have

$$x - y - t_1 + t_2 \leq 2(r + \epsilon)k \tag{2.67}$$

Therefore, the desired upper and lower bounds are

$$0 \leq x - y \leq 2(r + \epsilon)k \tag{2.68}$$

□

In practical terms, checkups should be placed farther apart as the detection rates improve. This is because when the detection rate is relatively low, there is a benefit to scheduling checkups that "overlap" each other in case a checkup fails to detect an existing illness. However, this benefit diminishes as the detection rate improves, so the checkups spread farther apart from one another. This allows the checkup schedule to cover a wider range of potential readmissions without losing detection quality.

2.4.4 From Theory to Practice: Implementable Policies from Modeling Insights

Through the prior analysis, we have captured the key factors affecting the efficacy of post-discharge checkup policies. To summarize the analytical insights of the previous section into practical rules of thumb, we now illustrate how to design a simple checkup policy for doctors and discharge planners. Suppose a patient is to be discharged and a post-discharge follow-up plan needs to be determined by the case manager. The case manager first decides the aggressiveness of the follow-up plan, i.e., how many office visits and phone calls to use. This can be done by evaluating the patient's readmission risk using existing risk calculators (Hu et al. (2014)). Given the estimates of the time-to-develop the condition density curve and the delay-time D (later in Section 2.5.1 we estimate the densities using historical data), the next step is to determine the timing of checkups.

From the analyses in Sections 2.4.1 and 2.4.2 and Proposition II.4, the checkups should be placed approximately z days apart (z being the average delay-time) such that the first and the last checkups are at the same height on the time-to-develop the condition curve (one on either side of the mode). Finally, from Theorem II.10, the case manager adjusts the spacing of checkups according to the detection rate of the checkups: higher detection rate spreads the checkups farther apart. For instance, the case manager should make less frequent contact with the patient if he/she believes that the patient was well educated for the diagnosis and understands what post-operative complications might happen (this translates to a higher detection rate); or the case manager may want to make frequent contact if he/she believes that the patient is less responsive to phone calls or is less adherent to the follow-up appointments (this translates to a lower detection rate). In the next section, we generalize the analytical insights using numerical studies to deepen the understanding of how to empirically estimate model parameters, of the impact of office visit and phone call sequencing, and of quantity versus quality of checkups.

2.5 Numerical Analyses

In this section, we conduct extensive numerical analyses on cystectomy readmissions from a regional hospital as well as the national State Inpatient Database (SID) to address the key questions that arise in post-discharge checkup policies: when to schedule checkups, how many checkups to schedule, and what types of checkups to schedule. First, we study two-checkup policies with one phone call and one office visit, which are consistent with current practice at our partner hospitals. We show that our approach improves the detection probability upon current practice by up to 43.7% when applied to readmitted patients. We test the robustness of our model with different exponential and gamma delay-time distributions. We also verify the delayed readmission log-likelihood inequality defined in Section 2.4.3. Next, we examine more aggressive checkup plans with more checkups to develop insights into: (1) optimal checkup timing and sequencing, (2) effects of varying the detection rate, and (3) checkup quantity vs. quality. We then validate our work by applying the optimal policies found to a different subset of patients and show that our results continue to hold. We conclude this section by summarizing rules of thumb that can be easily implemented by healthcare professionals to develop post-discharge checkup policies that have the potential to improve detection of readmission causing conditions.

2.5.1 Data and Model Parametrization

The numerical analyses in this section are based on two datasets. The first dataset contains delay-time information of 327 cystectomy patients discharged from our partner hospital between 2007 and 2012. The information in the dataset includes the following: date of discharge from the hospital, date of first contact with the healthcare provider after discharge, who initiated the contact, what the chief complaint was, date of readmission, what condition caused the readmission, and when the condition was first experienced. By computing the difference between the date of readmission and date of condition onset, we obtain the delay-time for each patient in this cohort. The data was manually collected by a medical student and a medical fellow at our partner hospital by going over medical charts and reviewing each patient's triage notes upon readmission. This patient cohort consisted of 79 female and 248 male patients between 37 and 91 years old (mean = 65.9, standard deviation = 11.2). Among the 327 patients, 63 patients (19%) were readmitted within 30 days of discharge. We used this database to obtain data on the delay-time random variable and the time-to-develop the condition random variable. Note that we focus on the readmitted patients only and exclude the patients who were not readmitted from our analysis. We also ignore the intervention and prevention effect of the checkups a patient received, which, at our partner hospital, typically included a phone call and a follow-up office visit on the 2nd and 12th day after discharge respectively.

We acknowledge there are many empirical challenges with this type of data and we do not address them all in this chapter. One of the key challenges is the estimation of the distributions. Since we only used readmitted patients in our estimation, it is likely that the estimated distributions differ from the ones parameterized using all patients, including readmitted and non-readmitted patients. In addition, since we ignored the intervention and prevention effect of existing checkups, our estimated distributions could be biased. Next, we provide an initial approach addressing how incorporating both readmitted and non-readmitted patients might affect our model's performance. Notice that results presented in that appendix are obtained from a limited case study on a very specific dataset. Nevertheless, empirical estimation is not the primary focus of our study and the remaining empirical challenges are left to future work.

At our partner hospital, current practice is to place a phone call on day 2 and an office visit on day 12 after discharge. These checkups could bias the data and results as there could be endogeneity induced by current checkup practice. We considered four types of patients in our chart review cohort: (1) patients who were not going to be readmitted regardless of checkups and intervention (non-readmit-able patients), (2) patients whose 30-day readmissions were detected and prevented by the day-2 and the day-12 checkups, (3) patients whose 30-day readmissions could have been prevented if the checkups were placed on days other than day-2 or day-12, and (4) patients who were going to be readmitted regardless of checkups and intervention (unavoidable readmissions).

To include all four types of patients, we went back to the chart review data set, which contained 327 cystectomy patients who underwent cystectomy at our collaborating hospital. We believe that the cohort of 327 patients included the four types of patients. Out of the 327 patients, 63 developed post-surgical conditions that lead to a 30-day readmission. The 63 patients included in our original analyses included type 3 and type 4 patients. The remaining $327-63 = 264$ patients included type-1 and type 2 patients, which were not readmitted and therefore not included in our original analysis.

Of the remaining 264 patients, 236 of them developed a condition at some point in their post-discharge recovery. The $264 - 236 = 28$ patients that never developed a condition were considered to be type 1 (not going to be readmitted regardless of monitoring policy). Of the 236 patients that developed a condition at some point, 24 patients were found to have had a condition detected on either the day-2 or the day-12 checkups as recorded on the medical chart. These 24 patients could have either (1) developed a non-readmission causing condition (reason 1) or (2) could have developed a readmission-causing condition that was mitigated by the checkups (reason 2). However, we do not have sufficient data to distinguish between the two reasons. Let q denote the proportion of reason 2 patients among the 24 patients. These patient could be considered as type 2 patients. We could estimate this proportion by looking at those $236 - 24 = 212$ patients who developed a condition but were not detected on day 2 or day 12 by the current follow-up protocol. Out of those 212 patients, 63 patients ($63/212 = 30\%$) were readmitted. This means that if we assume that the characteristics of those 24 patients are same as the population (212 patients), and that the checkups are perfect inspections that can prevent readmissions with probability one, then q can be estimated to be 30%. In reality, q may be smaller than 30% (if the checkups are not perfect) and it may not prevent readmissions with probability one; or q could be greater than 30% if patients who are found sick on day-2 and day-12 are more likely to be readmitted than the population average is. Another way to estimate q is to use the national average readmission rate of cystectomy (which was observed to be 24% in the SID database). We conducted sensitivity analyses around the proportion of type 2 patients at $q = 25\%, 50\%, 75\%$, and 100%. Gamma distributions were fitted to these cohorts with $q = 25\%, 50\%, 75\%$, and 100% of the 24 patients added.

Finally, we took the checkup policies obtained using the original gamma distribution (types 3 and 4 only) and computed their objective values (suboptimal) by plugging the computed policies into the gamma distributions that included patients (simulated by adding $q = 25\%, 50\%, 75\%$, and 100% of the 24 patients) type 2 patients. We then computed the difference in objective values between the suboptimal objective values and the optimal objective values (using the distribution that included type 2 patients as our testbed) for checkup policies consisting of 1 to 3 office visits and 1 to 7 phone calls. As seen in the following table, by ignoring types 1 and 2 patients, the detection probabilities degraded by at most 3.5%. The most likely value of q, according to our estimation, would be around 24%, which shows at worst a very small difference of 0.54% between the original checkup policy (from our simpler model containing only types 3 and 4 patients) and the true optimal. We believe that the small observed differences are sufficient to demonstrate that the results from our simpler analysis with only types 3 and 4 patients should still be valid.

To the best of our knowledge, this is the first study in the clinical or operational literature to attempt to characterize these two variables using actual data. This is because existing available datasets do not capture delay-time or time when a readmission-causing condition developed. Due to data scarcity, we conducted our numerical analysis using population-based distribution curves. Given sufficient delay-time data, our approach can be tailored to individual patients by applying transfer learning techniques for personalized readmission forecasting (Helm et al. (2016)). We

	q	1 Phone Call	2 Phone Calls	3 Phone Calls	4 Phone Calls	5 Phone Calls	6 Phone Calls	7 Phone Calls
1 Office Visit	25%	−0.42%	−0.46%	−0.48%	−0.49%	−0.49%	−0.49%	−0.50%
	50%	−1.16%	−1.25%	−1.24%	−1.24%	−1.22%	−1.18%	−1.24%
	75%	−1.96%	−2.15%	−2.20%	−2.24%	−2.24%	−2.22%	−2.28%
	100%	−2.64%	−2.90%	−2.98%	−3.05%	−3.04%	−3.02%	−3.09%
2 Office Visits	25%	−0.51%	−0.54%	−0.54%	−0.51%	−0.51%	−0.53%	−0.52%
	50%	−1.34%	−1.45%	−1.40%	−1.27%	−1.24%	−1.34%	−1.29%
	75%	−2.34%	−2.51%	−2.47%	−2.33%	−2.31%	−2.42%	−2.38%
	100%	−3.17%	−3.39%	−3.33%	−3.16%	−3.14%	−3.27%	−3.23%
3 Office Visits	25%	−0.49%	−0.53%	−0.52%	−0.53%	−0.49%	−0.49%	−0.54%
	50%	−1.17%	−1.32%	−1.26%	−1.26%	−1.11%	−1.14%	−1.38%
	75%	−2.22%	−2.41%	−2.37%	−2.39%	−2.20%	−2.22%	−2.49%
	100%	−3.04%	−3.28%	−3.22%	−3.25%	−3.02%	−3.04%	−3.37%

Table continued

Table 2.2: Difference in Detection Probabilities

demonstrate robustness of our optimal policies to distribution in Table 2.3 and the analytical results from Section 2.4 are not dependent on the form of the delay-time distribution. Further, the mean of the delay-time distribution observed in the data (2.35 days) is very close to delay-time estimates for common readmission-causing conditions in a survey given to an independent group of five practicing surgeons (average of 2 days). These cross-checks should help mitigate some concerns about the accuracy of the estimation. We also tested the dependency between delay-time and the time-to-develop the condition using the 63 readmitted patients from this new data set. The correlation between the two variables is 0.14, and they are independent (p-value < 0.05) using the Hilbert-Schmidt independence criterion (Gretton et al., 2007). While data for this study was collected manually as a proof of concept, this process could be appropriately scaled with IT support due to the proliferation of electronic health records. This type of analysis, however, is left to future work.

The second dataset comes from the the State Inpatient Databases (SID), which was gathered as part of the Healthcare Cost and Utilization Project sponsored by the Agency for Healthcare Research and Quality. From the SID dataset, we identified 717 cystectomy patients (ICD-9 code 577, 5771, and 5779) from the states of Florida, Iowa, North Carolina, New York, and Washington that were readmitted within 30 days of discharge in 2009 and 2010. As mentioned in Section 1, we choose cystectomy patients as a proof of concept given that our clinical collaborator is an expert in this type of surgery and that it has one of the highest readmission rates in the U.S. Note that subsequent work by our collaborator's surgical research group indicates the dynamics of cystectomy are similar to many other surgeries, particularly lower torso/abdomen surgeries (Jacobs et al., 2017), and our clinical collaborator believes this approach would be broadly applicable in the surgery domain; this includes surgeries targeted for inclusion in Medicare's readmission penalty program (HRRP). To further verify that the unimodality assumption holds for other surgery cohorts, we extracted the readmission records of patients who had some of the most common abdominal and chest surgeries in 2009 and 2010: Abdominal Aortic Aneurysm Repair (AAA), Esophagectomy, Pancreatectomy, Aortic Valve Replacement (AVR), Coronary Artery Bypass Grafting (CABG), and Lung Resection. In all six cases, the time-to-readmission and the estimated time-to-develop the condition curves (estimated using readmitted patients) appeared to be unimodal (see Figure 2.5.1).

We excluded patients who had ICD-9 code 4411, 4412, 4413, 4415 or 4416, patients who were 18 years old or younger, and patients who died during cystectomy or during their inpatient stay. The SID database captures the length of time between

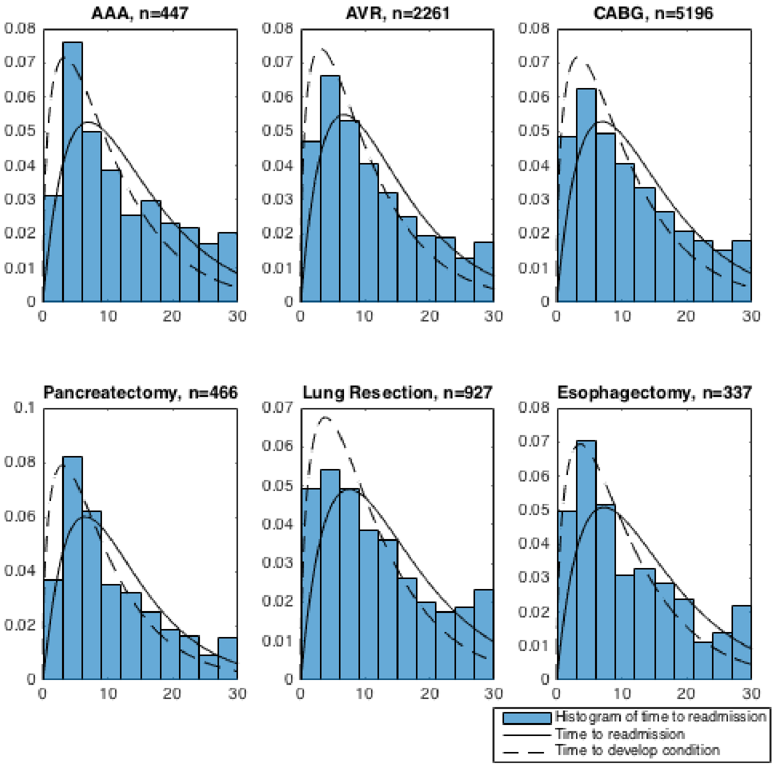


Figure 2.6: Time-to-Readmission and Time-to-Develop the Condition Distributions for Six Major Abdominal and Chest Surgeries

each patient's initial discharge and his/her subsequent readmission. Among the 717 patients, 385 patients from 2010 were used for parametrization and optimization of the models, and 332 patients from 2009 were used to test the optimal policies. We used the first dataset to estimate the delay-time distribution and to validate the efficacy of recovering the time-to-develop the condition distribution. To do that, we started by fitting distributions to the observed time-to-readmission (shown in Figure 2.7(a)) and to the observed delay-time (shown in Figure 2.7(b)). Gamma and exponential distributions worked well to model the time-to-readmission and the delay-time, respectively.

Given the time-to-readmission and the delay-time distributions, we recovered the time-to-develop the condition distribution through a numerical inverse Laplace transform. Next, we describe the inverse Laplace transform in detail.

Clinical data used to parametrize the delay-time models is limited in the fact that time-to-develop the condition is currently not recorded in any databases known to the authors. The historical data most readily available is the time-to-readmission. To obtain data on the delay-time, which is not recorded in any major clinical databases,

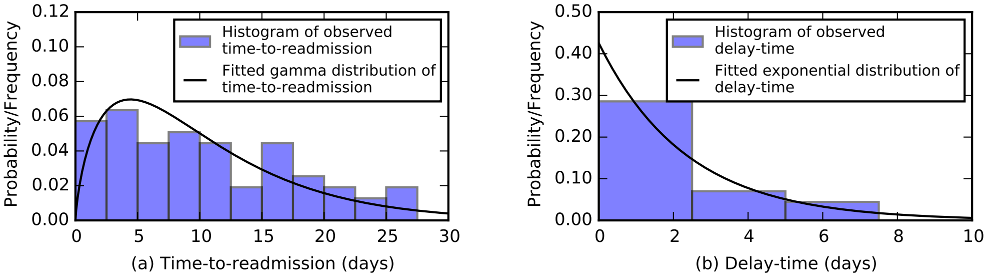


Figure 2.7: Time-to-Readmission and Delay-Time Distribution Fitted from Medical Charts

Note. Time-to-readmission $\rho \sim \text{gamma}(1.74, 5.98)$, delay-time $D \sim \text{exponential}(2.35)$

we conducted a study of 327 medical records and extracted data on how long the patient had been feeling ill before returning to the hospital based on triage notes upon readmission. However, given the time-to-readmission distribution and the delay-time distribution, we can obtain the time-to-develop the condition probability density on larger databases by applying the inverse Laplace transform.

Recall that the time-to-readmission, ρ, is the summation of the time-to-develop the condition δ and the delay-time D, i.e., $\rho = \delta + D$. Since δ and D are assumed to be independent, the Laplace transform of ρ, $\mathcal{L}\{g_\rho(x)\}(s)$, is equivalent to the product of the Laplace transforms of δ and D, i.e., $\mathcal{L}\{g_\delta(x)\}(s)$ and $\mathcal{L}\{f(x)\}(s)$.

$$\mathcal{L}\{g_\delta(x)\}(s)\mathcal{L}\{f(x)\}(s) = \mathcal{L}\{g_\rho(x)\}(s) \tag{2.69}$$

Dividing both sides by $\mathcal{L}\{f(x)\}(s)$, we get the following expression for the Laplace transform of the time-to-develop the condition, denoted by $\mathcal{G}(s)$:

$$\mathcal{L}\{g_\delta(x)\}(s) = \frac{\mathcal{L}\{g_\rho(x)\}(s)}{\mathcal{L}\{f(x)\}(s)} =: \mathcal{G}(s) \tag{2.70}$$

Applying the inverse Laplace transform $\mathcal{L}^{-1}\{\cdot\}$ to both sides of Eq. (2.70), we obtain the probability density function of the time-to-develop the condition:

$$g_\delta(x) = \mathcal{L}^{-1}\{\mathcal{G}(s)\}(x) \tag{2.71}$$

The inverse Laplace transform yields closed-form solutions for certain $g_\rho(\cdot)$-$f(\cdot)$ pairs such as Erlang-exponential and normal-normal. Given arbitrary $g_\rho(\cdot)$ and $f(\cdot)$, a closed-form solution may not exist. In such cases, numerical algorithms for inverse Laplace transform (Avdis and Whitt, 2007; Rizzardi, 1995; Lyness and Giunta, 1986) can be implemented.

The numerical Laplace inversion fitted the true time-to-develop the condition well with a Pearson χ^2 p-value $= 0.36$. This validates the efficacy of recovering the distribution of the time-to-develop the condition using inverse Laplace transform.

With an effective approach to recover the time-to-develop the condition, we expanded our analysis to the SID database (which includes patients from many hospitals across five states). Using the 2010 SID patients, we fitted a gamma distribution to the time-to-readmission as shown in Figure 2.8. Since the delay-time information was not recorded on the SID database, we assumed that the delay-time for the SID patients followed the same distribution as the delay-time observed on patients at our partner hospital (exponential(2.35)). We used the inverse Laplace transform to estimate the time-to-develop the condition distribution (see Figure 2.8).

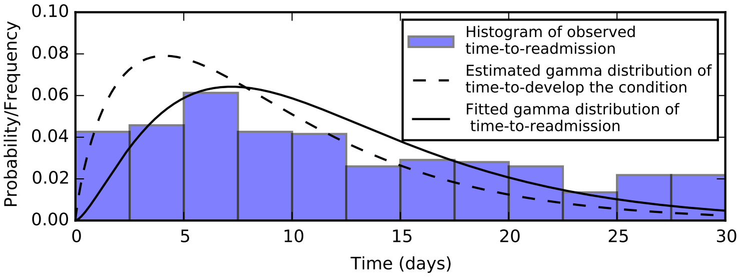


Figure 2.8: Fitted Time-to-Readmission and Recovered Time-to-Develop the Condition for 2010 SID Patients

Note. Time-to-readmission $\rho \sim \text{gamma}(2.50, 4.80)$, time-to-develop the condition $\delta \sim \text{gamma}(1.81, 5.08)$

2.5.2 Comparison of Policies Against Current Practice

With the model parameterized on the 2010 SID patients, we evaluated how our policy improves upon the current practice at our partner hospital. We also examined the robustness of our model by fitting various exponential and gamma delay-time distributions (see Table 2.3). The distributions tested in Table 2.3 satisfy the delayed readmission log-likelihood inequality defined in Section 2.4.3.

The current practice for post-discharge monitoring at our partner hospitals is to place a phone call on the 2nd day after discharge and an office visit on the 12th day after discharge. Throughout our numerical analyses, we assume that an office visit is a perfect inspection with detection rate $r = 1$; and a phone call is an imperfect inspection with detection rate $r = 0.6$ (given the patient has developed a condition, a phone call will detect the condition successfully with probability 0.6). These values were estimated by our clinical collaborators. In Section 2.5.4, we perform a sensitivity analysis on the detection rate.

Applying the algorithm described in Section 2.3.2, we solve for the optimal 2-checkup policies with one phone call and one office visit (for fair comparison with current practice) using the 2010 SID patients. We tested seven delay-time distributions (see Table 2.3) with the same mean and different variance as a sensitivity analysis, since the delay-time distribution is estimated based on a small sample of 63 patients and no other publicly available data set captured delay-time information. Table 2.3 shows how our policy outperforms current practice by significantly increasing the probability that ill patients are detected before readmission (defined as the detection probability). The relative improvement of the detection probability ranges from 23.9% to 65.3% (average = 49%) for the exponential and gamma delay-time distributions tested. This improvement is achieved solely by optimizing the timing and sequencing of the two checkups. As we shall see in the following sections, the detection probability further increases if we adopt more aggressive post-discharge monitoring policies by increasing the number of checkups. However, we would like to point out that the improvement is computed using readmitted patients only, which represent 19% of the entire cohort. Hence, when taking both readmitted and non-readmitted patients into account, the improvement might be smaller. As a sanity check, we conducted simulations and verified that, under current practice, the simulated readmission rates predicted by our model were very close to the readmission rates that were actually observed in the data (both around 20%).

Distribution			Time of	Time	Detection Probability		
Delay-time	E[D]	Var[D]	First	between	Optimal	Current	Relative
Distribution			Checkup	Checkups	2-Checkup	Practice	Improvement
exponential($\mu/2$)	1.2	1.4	4.9	3.1	0.13	0.08	56.8%
exponential(μ)*	2.4	5.5	5.9	4.4	0.23	0.16	43.7%*
exponential(2μ)	4.7	22.1	7.4	6.1	0.35	0.29	23.9%
gamma($1/2, 2\mu$)	2.4	11.0	6.5	5.0	0.20	0.15	30.8%
gamma($2, \mu/2$)	2.4	2.8	5.5	3.9	0.25	0.16	56.4%
gamma($3, \mu/3$)	2.4	1.9	5.3	3.6	0.26	0.16	62.1%
gamma($4, \mu/4$)	2.4	1.4	5.2	3.5	0.26	0.16	65.3%

Table 2.3: Optimal 2-Checkup for Exponential/Gamma Delay-Time Distributions)

Note. * marks the estimated delay-time distribution using our chart review data set. The timing of checkup (rounded to the first decimal place) is in days. In our numerical studies, we observed that the solutions are insensitive to rounding of the checkup timing.

In Table 2.3, where the mean of the gamma distribution is held constant and the variance is increased, we see that increased (gamma-distributed) delay-time variance leads to greater spacing between checkups. The performance of the optimal policy also degrades as the (gamma-distributed) delay-time variance increases. This implies that efforts at standardizing patients' behavior at home could have benefits for readmission reduction because it reduces the delay-time variance. This variance effect is offset in the exponential case by the concurrent increase in mean delay-time, which indicates that efforts to keep patient conditions from degrading too fast (e.g. compliance with physician orders and adherence to medication), can also provide significant benefit by allowing the healthcare provider time to detect the condition before it becomes too severe. Note that our approach can be tailored to each patient's time to readmission characteristics, but because of data scarcity, it is difficult tailor the delay-time. If there were sufficient data, the delay-time could also be personalized using the same method used to personalize time to readmission predictions (Helm et al., 2016).

2.5.3 Optimal Timing and Sequencing of Checkups: Timing outweighs sequencing

Next, we explore the delay-time-spaced block structure shown by Proposition II.4 and the optimal sequencing of checkups in a more generalized scheme involving four to ten checkups in total with three office visits. Though conducting ten checkups within a 30-day period could be burdensome for both clinicians and patients, the purpose here is to study 10-checkup policies as the extreme upper bound for the sake of comparison and completeness, and further investigate the structure of checkup policies and their timing and sequencing.

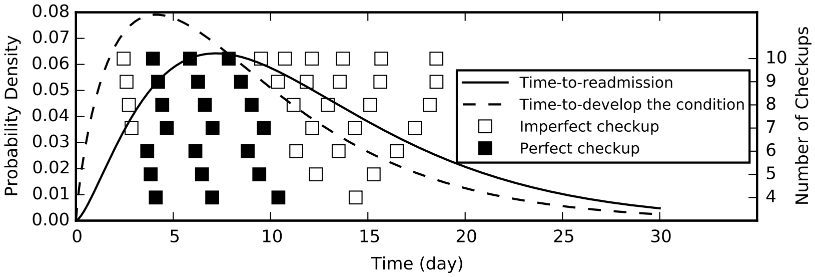


Figure 2.9: Optimal n-Checkup Sequencing and Timing, $n \in \{4, ..., 10\}$: Consecutive Perfect Checkups Appear around the Mode of $g_\delta(\cdot)$

Note. Assumptions: $D \sim$ exponential(2.35)); r of perfect checkups = 1, r of imperfect checkups = 0.6; the left axis denotes the probability density; the right axis denotes the number of checkups. The detection probabilities are 0.40, 0.43, 0.46, 0.48, 0.50, 0.52, and 0.54 respectively (from bottom to top).

From Figure 2.9, we draw the following insights: (1) checkups are scheduled in a contiguous block surrounding the mode of the time-to-develop the condition distribution with spacing approximately equal to the mean delay-time. Slightly wider spacing is observed around the perfect checkups and the spacing increases as the probability of developing the condition decreases; and (2) consecutive perfect checkups are placed surrounding the mode of the time-to-develop the condition curve (i.e. put the best checkups in the most hazardous period).

Although optimal policies favor consecutive perfect checkups around the mode, it is sometimes impractical to schedule them consecutively in a short period of time; particularly because many patients may live far from the hospital where their initial treatment occurred, making frequent travel to the hospital difficult or impossible. Fortunately, we find that, as long as the timing is optimal, the policies are robust to sequencing. That is, the gaps between the worst-case and the best-case sequences for all policies in our test suite (1 to 10 checkups consisting of 0 to 3 office visits and phone calls) ranged between 0.2% and 0.5%, indicating that the timing of checkups is much more important than the sequencing. One way to explain why sequencing is less important is that the optimization will mimic a perfect checkup by scheduling multiple imperfect checkups closer together. For example, three phone calls of detection rate 0.6 (made at once) have an equivalent detection rate of $1 - 0.4^3 = 0.94$. We conjecture that, by striking a balance between the spacing of checkups and the effective detection rate, the sub-optimal sequencing can mimic the behavior of the optimal sequencing. The robustness to sequencing is a valuable property: as the number of checkups increases, the number of permutations of checkup sequences becomes large (e.g. the 10-checkup policy in Figure 2.9 has $\binom{10}{3} = 120$ sequences), requiring a significant amount of computational power to obtain an optimal solution. Results from the sequencing analysis, however, generate near-optimal policies by fixing the checkup sequence that is convenient for the physician and the patient and then optimizing the timing of checkups. This also allows for accommodating physician and patient preferences with little degradation in performance.

Remark II.13 (Multi-modal time-to-develop the condition distributions)*.* We test our model numerically using a multi-modal time-to-develop the condition distribution estimated using a Gaussian Kernel Density Estimator. We show that checkup policies can still be solved numerically to optimality and the differences in optimal detection probabilities are within 2%.

To test the robustness of our model under a multi-modal distribution. We created a counter-factual time-to-develop the condition distribution by simulating the time-to-develop the condition of 63 patients according to the fitted gamma distribution presented in Section 2.5.1. Then, patients that may have been readmitted on day-12 but were not (possibly due to the current practice of following up with patients on day

12) were added to the cohort. We simulated 24 patients who developed a condition prior to day-12 based on the exponential delay-time distribution. A Gaussian Kernel Density Estimator (KDE) with bandwidth 0.8 was used to fit the time-to-develop the condition distribution curve. The KDE distribution is shown in Figure 2.10.

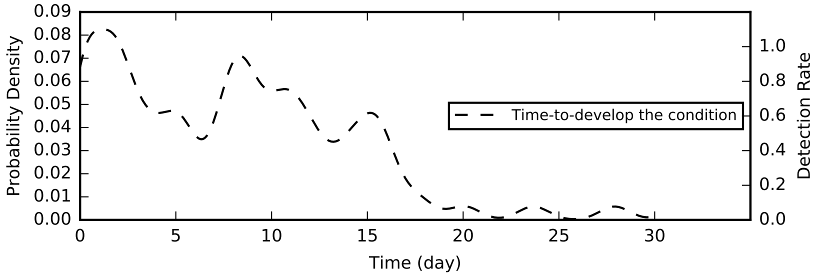


Figure 2.10: The Counter-Factual Multi-Modal Distribution Created by a Gaussian Kernel Density Estimator

We studied the sequencing and timing of checkups. Solving for optimal n-checkup ($n = 4, ..., 10$) sequencing and timing under a multi-modal distribution, we found that the insights on sequencing and timing developed under the unimodality assumption of Proposition II.4 did not hold in the multi-modal case as perfect checkups were no longer placed consecutively. However, the policies were still robust to the sequencing of checkups: the gaps between the worst-case and the best-case sequences for all policies in this test suite (4 to 10 checkups consisting of 3 office visits and 1 to 7 phone calls) were between 0.9% and 1.5%. Moreover, as can be seen in Table 2.5.1, the optimal detection probabilities of these policies were close to the ones obtained using the original gamma distribution.

	Checkup Policy	1P3O	2P3O	3P3O	4P3O	5P3O	6P3O	7P3O
Optimal Detection Probabilities	Original Gamma	0.40	0.43	0.46	0.48	0.50	0.52	0.54
	KDE	0.39	0.42	0.45	0.47	0.50	0.51	0.53
Gaps b/t Worst and Best Cases	Original Gamma	0.2%	0.3%	0.4%	0.5%	0.4%	0.3%	0.4%
	KDE	1.2%	1.2%	1.3 %	1.5%	1.1%	0.9%	1.2%

Table 2.4: Comparison of Optimal Detection Probabilities (P=Phone Call, O=Office Visit)

As the detection rate increased, we noticed that imperfect checkups were centered around each mode and placed closer together. However, increasing the detection rate did not necessarily widen the overall coverage area. Since the checkups were scattered to cover the prominent modes, the overall coverage area was dictated by the separation of the modes.

2.5.4 Impact of Detection Rate on Timing: Greater Coverage with Better Checkups

In this section, we study the impact of varying the detection probability of an imperfect checkup, r, and extend the insight drawn from Theorem II.10 using a realistic potential monitoring schedule (according to our clinical collaborator) of one office visit and nine phone calls. While this scenario is more aggressive than current practice, it is still reasonable because phone calls can be done cost-effectively using nurses, trained technicians, or even automated call systems (see www.cloud9hcs.com and Tagliente et al. (2016)). Results are presented in Figure 2.11.

In Figure 2.11, the spacing between checkups increases as the detection rate improves. This aligns with Theorem II.10 and our intuition: more accurate checkups can be spread farther apart; whereas less accurate checkups should be placed closer together to account for the higher probability that the condition is missed by previous checkups. With more accurate checkups, the associated larger spacings will cover a longer time period. Since checkups are scheduled less frequently, patients and family members are less likely to be inconvenienced. For example, too much contact may lead patients to become irritated, ignore phone calls, or not consider questions as attentively. Another benefit is that by covering a longer time period, there is an increased ability to detect potentially developing conditions. Finally, the extended monitoring period may help patients feel that they are receiving better attention/care, which can build trust between the patient and clinician, thereby improving patient satisfaction.

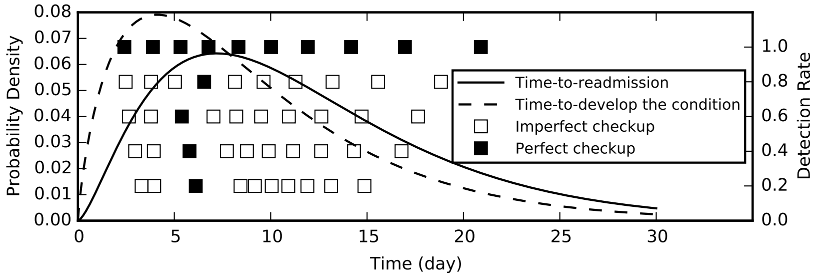


Figure 2.11: Optimal Checkup Timings under Different Detection Rates

Note. Assumptions: $D \sim$ exponential(2.35), number of checkups = 10. The detection probabilities are 0.31, 0.41, 0.50, 0.57, and 0.64 respectively (from bottom to top).

2.5.5 Marginal Benefits of Increasing Checkup Quantity vs. Improving Checkup Quality: Quantity Outweighs Quality

Since scheduling frequent follow-up office visits will increase the burden on frequently heavily loaded clinician schedules (Baron, 2010) , in this section, we consider the value of doing more phone calls as a substitute for office visits. Importantly for the clinical community, we find that checkup quantity is more important than quality; i.e. multiple phone calls function as a good substitute for office visits.

In our first experiment, we study optimal checkup policies that have a total number of checkups between one and ten with zero to three office visits.

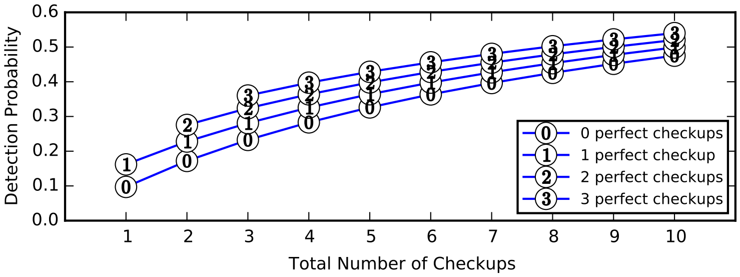


Figure 2.12: Detection Probability of n-Checkup Policies with 0–3 Perfect Checkups

Note. Assumptions: $D \sim$ exponential(2.35), r of perfect checkups = 1, r of imperfect checkups = 0.6

As shown in Figure 2.12, both increasing the number of checkups and increasing the number of perfect checkups improves the detection probability. However, we find that adding one additional phone call is nearly as effective as switching one phone call to an office visit. In our test suite where 1 to 10 checkups consisting of 0 to 3 office visits and phone calls were optimized, scheduling one additional phone call increases the detection probability by an average of 3.35% whereas replacing a phone call with an office visit (and rerunning the optimization) increases the detection probability by 3.45%. We also calculated the minimum number of additional phone calls needed to outperform replacing a phone call with an office visit. Across our test suite, on average, an office visit ($r = 1$) can be replaced with 2.57 phone calls ($r = 0.6$). Further, when the total number of checkups is less than five, an office visit can be replaced with 2 phone calls.

This result is highly valuable from the practical perspective, as phone calls are significantly less resource-intensive than office visits for both patients and physicians. Notice that phone calls have numerous benefits over office visits: (1) patients may be located far from the clinic and may have limited mobility and transportation options; (2) making an office visit is burdensome as the capacity of the clinic and physicians' time are limited; and (3) making phone calls can be done efficiently through specialized call centers or physicians' nursing or auxiliary staff in their spare time. The key finding is that an effective checkup policy can leverage these inexpensive phone calls to achieve similar results as those obtained with the more expensive and inconvenient office visits.

2.5.6 The Benefit of Improving the Efficacy of Phone Calls

One strong interest in efforts at readmission reduction lies in designing effective questionnaires for phone and telemedicine checkups (see, for example, readmission reduction startup company Cloud9, which has developed detailed questionnaires for many conditions, www.cloud9hcs.com) and testing predictive models based on historical responses to survey questions. Design of such questionnaires to effectively target the main causes of readmission (as an example for cystectomy, five main conditions account for almost all of the readmissions) can increase the detection probability of a phone call. These questionnaires are particularly easy to implement if the call is being conducted by someone who is not the physician or, or if it is conducted by an automated call system. To determine the importance of such improvements and subsequently the amount of effort that should be expended to perfect such surveys, we analyzed the impact of the detection probability, r, on the efficacy of a monitoring schedule.

Figure 2.13 shows that, as might be expected, the benefit of replacing a phone call with an office visit diminishes as the detection rate improves. To analyze the overall impact, we developed a test suite, where policies consisting of 10 checkups with 0-3 office visits were optimized. We incremented the detection rate from 0.2 to 0.8 (with a step size of 0.2), with 0.2 and 0.8 functioning as extreme lower/upper bounds for the sake of comparison and completeness. We started by computing the detection probability as a function of the detection rate of the phone calls. We then estimated (1) the improvement in detection probability achieved by upgrading an existing phone call to an office visit; and (2) the improvement in detection probability achieved by increasing the detection rate of the phone calls. Finally, we computed the relative effectiveness of increasing the phone call detection rate by 20% (compared to upgrading an existing phone call to an office visit). A relative effectiveness of 100% means that increasing the phone call detection rate by 20% is as effective as upgrading a phone call to an office visit.

Across this test suite, on average, increasing the detection rate by 20% absolutely (e.g. 0.2 → 0.4) achieves 29% to 70% (average = 47%) of the benefit achieved by replacing a phone call with an office visit. The following table shows the relative effectiveness.

The relative marginal benefit of increasing the detection rate is greater when the

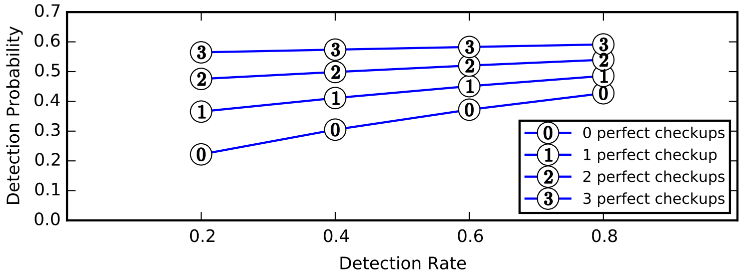


Figure 2.13: Detection Probability as a Function of the Detection Rate

Note. Assumptions: $D \sim$ exponential(2.35) with 10 checkups

Phone Call Detection Rate # of Office Visits	$0.2 \rightarrow 0.4$	$0.4 \rightarrow 0.6$	$0.6 \rightarrow 0.8$
$0 \rightarrow 1$	58%	63%	70%
$1 \rightarrow 2$	41%	46%	50%
$2 \rightarrow 3$	35%	29%	31%

Table 2.5: Relative Effectiveness of Increasing Phone Call Detection Rate with respect to Replacement of a Phone Call with an Office Visit

detection rate is low and the number of office visits is few (see Table 2.5). Notice that the relationship is not linear (plausibly concave as shown in Figure 2.13). The intuition is that the effort required to improve checkup policies increases as the policies get more aggressive and effective.

The practical implication suggests that physicians can benefit significantly by designing more effective phone call questionnaires, which may be used to help replace excessive or burdensome office visits. Increasing the detection rate of phone calls may be achieved by providing patient education upon hospital discharge (e.g. informing patients of symptoms that indicate worsening conditions), ensuring that the content of post-discharge questionnaires are tailored as much possible to individual patients and their personal characteristics (which can be identified with readmission risk models at the time of discharge), and targeting high risk conditions (e.g. infection, dehydration, kidney failure, failure to thrive) with focused questions.

2.5.7 Out-of-Sample Testing on a Separate Dataset and Solution Robustness

To validate and test our models, we parameterized our delay-time random variable using the first dataset from our partner hospital for radical cystectomy patients. We then estimated the time-to-readmission by fitting a gamma distribution (best fit) to the 2010 SID dataset, also for radical cystectomy patients. Using our inverse Laplace transform method, we were able to recover the distribution on the time-to-develop the condition. Finally, we generated an optimal monitoring schedule based on the dynamics obtained from combining our partner hospital delay-time data with the 2010 SID readmission data. We then tested this policy on a new dataset, 2009 SID data, comparing our checkup times to the readmissions for cystectomy patients across five states in 2009. To do so, we consider two methods. In both methods, we begin by determining the optimal policy with parameters estimated from 2010 SID data. *Method 1:* We compare the performance of the optimal policy from 2010 data when applied to a time-to-readmission curve estimated from the 2009 data versus the policy that optimizes according to the true 2009 time-to-readmission curve. We can then compare the optimality gap caused by errors in estimation of the time-to-readmission curve. *Method 2:* We apply the 2010 optimal policy to all the cystectomy patients from 2009 SID data and estimate the performance using each patient's actual readmission time and calculating the probability that his/her delay-time was long enough such that one of the inspections from our optimal policy would have caught the condition before it caused a readmission.

Method 1 is shown in Figure 2.17. The detection probabilities range from 0.1 to 0.5 and are very close to the detection probabilities using a time-to-readmission curve estimated with the 2009 data itself (in-sample). The absolute optimality gaps were less than 5% (see Table 2.6).

Total # of Checkups # of Office Visits	1	2	3	4	5	6	7	8	9	10
0	1.4%	2.1%	2.4%	2.7%	2.9%	2.6%	2.9%	2.9%	2.7%	2.9%
1	2.4%	2.7%	2.9%	3.0%	2.8%	2.9%	2.8%	2.9%	2.8%	2.8%
2	N/A	3.0%	3.1%	3.0%	3.1%	3.1%	3.1%	3.1%	3.0%	2.6%
3	N/A	N/A	3.2%	3.0%	3.2%	3.1%	2.9%	2.9%	2.8%	2.6%

Table 2.6: Absolute Optimality Gap for 2009 SID Patients Using n-Checkup Policies with 0-3 Perfect Checkups that Were Parameterized Using 2010 SID Patients

We also calculated the relative optimality gaps and switched the testing and training sets to further validate the findings (see Figures 2.5.7, 2.5.7, 2.5.7, 2.5.7. The largest relative optimality gaps were observed in one-checkup models, which are not advisable in practice. As the number of checkups increases, the relative optimality gap diminishes. This indicates that the model becomes more robust as the number of checkups increases, providing more support for the idea that quantity of checkups is highly important. Practically speaking, not only does larger quantity eliminate the need for excessive office visits, it also increases the robustness of the solution to errors in estimation.

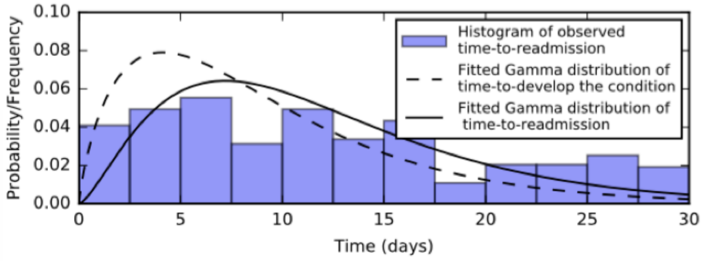


Figure 2.14: Fitted Time-to-Readmission and Recovered Time-to-Develop the Condition for 2009 SID Patients

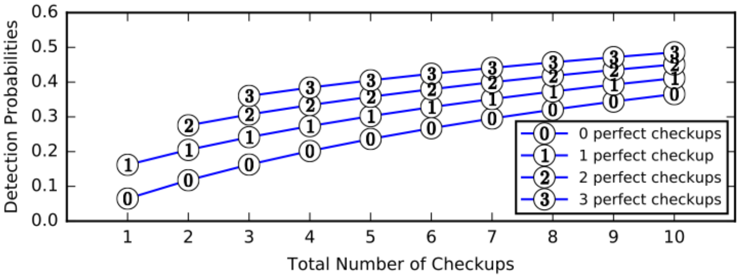


Figure 2.15: Detection Probability of Checkup Policies with 0-3 Perfect Checkups for 2009 SID Patients.

Method 2 assumes independence between delay-time and time-to-readmission. Let T be the time that the patient was actually readmitted (in the data). The detection probability, $\hat{\mathcal{D}}$, can then be calculated using the following formula. Let

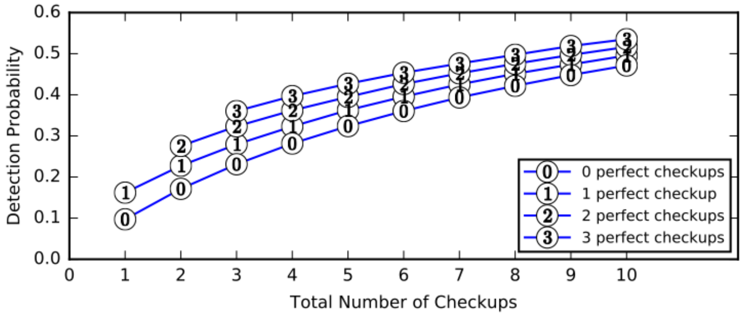


Figure 2.16: Detection Probability of Checkup Policies with 0-3 Perfect Checkups (Developed Using 2009 SID Patients) Tested on 2010 SID Patients (Method 1)

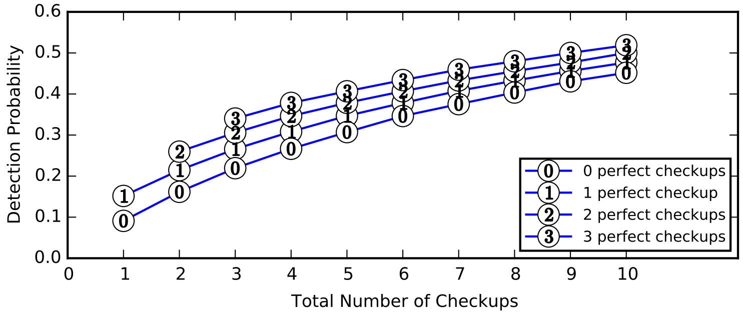


Figure 2.17: Detection Probability of Checkup Policies with 0–3 Perfect Checkups Tested on 2009 SID Patients (Method 1)

$N = \text{argmax}_n : t_n \leq T.$

$$\hat{\mathcal{D}} = r_1 \cdot (1 - F(T - t_1)) \tag{2.72}$$

$$+ \sum_{\alpha=2}^{N} \left(r_\alpha \cdot \left(\prod_{\beta=1}^{\alpha-1} (1 - r_\beta) \right) \cdot (1 - F(T - t_1)) \right) \tag{2.73}$$

$$+ \sum_{\alpha=3}^{N} \left(\sum_{\beta=2}^{\alpha-1} \left(r_\alpha \cdot \prod_{\gamma=\beta}^{\alpha-1} (1 - r_\gamma) \cdot (F(T - t_{\beta-1}) - F(T - t_\beta)) \right) \right) \tag{2.74}$$

$$+ \sum_{\alpha=2}^{N} (r_\alpha \cdot (F(T - t_{\alpha-1}) - F(T - t_\alpha))) \tag{2.75}$$

where the four summands collectively represent the total probability of detecting the patient as ill during every scheduled checkup. In particular, the first summand (2.72) represents the probability that the patient enters the ill state before the first checkup and is successfully identified as ill during the first checkup. The second summand (2.73) represents the probability that the patient becomes ill before the first checkup, but checkups 1 through $(\alpha - 1)$ fail to properly detect the patient condition and checkup $\alpha \in \{2, ..., N\}$ successfully identifies the patient as ill. The third summand (2.74) represents the probability that the patient becomes ill between checkups $(\beta - 1)$ and β for $\beta \in \{2, ..., (\alpha - 1)\}$, but is not properly identified as being ill until checkup $\alpha \in \{3, ..., N\}$. The fourth summand (2.75) represents the probability that the patient enters the ill state between checkups $(\alpha - 1)$ and α, and is immediately identified as being ill during checkup $\alpha \in \{2, ..., N\}$.

Method 2 evaluates how well the optimal policy would have performed in practice if implemented on the radical cystectomy patients from the 2009 SID data. Figure 2.18 presents the results of this study, indicating that the optimal policies estimated from 2010 SID data would in fact have been highly effective if put into practice on the patients of the out-of-sample dataset. In particular, the estimated detection probabilities (based on actual readmission times) are greater than 60% using one or more perfect checkups on the 2009 SID patients. It is worth highlighting the difference between Methods 1 and 2 (i.e. Figure 2.12 vs Figure 2.18): in Figure 2.12, we were plotting the objective function, which is parameterized with gamma and exponential distribution curves fitted from the training data set. Whereas in Figure 2.18, we were plotting a different objective, which uses the actual time to readmission observations combined with the delay-time distribution function plus the discrete observations.

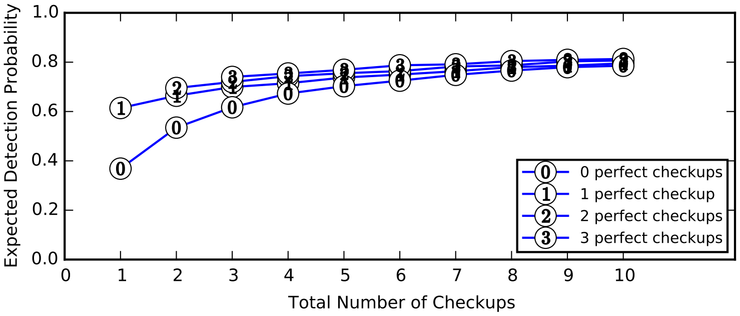


Figure 2.18: Proportion of Conditions Captured by the Optimal Policy with 0-3 Perfect Checkups Obtained Using 2010 SID Patients and Tested on 2009 SID Patients (Method 2)

This improved performance seems to stem from the fact that the true time-to-readmission for cystectomy patients tends to be more heavily front-loaded in the first 7-8 days than the fitted gamma distribution. Another fact that contributed to this higher performance is that the time to readmissions we used are in days (discrete) rather than time (continuous). Using discrete data created a lumping effect and lead to improved performance. We are unable to use the exact time of readmission (continuous data) to validate our model as it is protected information and could be used to identify patients. Moreover, since the optimal policies tend to also bunch a number of checkups soon after patient discharge, this policy ends up actually detecting more conditions in practice that would have been estimated by the fitted gamma distribution for time-to-readmission.

As a further benefit revealed by this study (seen in Figure 2.18), it appears that one office visit along with a few phone calls is sufficient to capture much of the value of post-discharge checkups. This is good news for busy clinicians concerned about the added burden of increased checkups.

2.5.8 Design of Practical Post-discharge Checkup Policy

Combining the insights drawn from our analytical and numerical analyses, we provide the following rules of thumb to facilitate the design of post-discharge checkup policies.

- **Timing of checkups outweighs sequencing:** (1) schedule checkups in a block surrounding the most-likely time (mode) of developing a condition; (2) keep the time between checkups close to the expected delay-time; (3) office visits should be scheduled near the time of highest risk of readmission for the patient.
- **Cover a longer time period and reduce office visits with better checkups:** Improving the quality of phone call checkups (e.g. better questionnaires, patient education) allows the checkup team to (1) cover a longer time period with less frequent calls (better for patients and detects more potential conditions), (2) reduce the number of office visits without reducing readmission detection (better for patients, clinicians, and healthcare organizations). Further, helping to standardize patient behavior at home, thereby reducing delay-time variance, has added detection benefits.
- **Quantity of checkups outweighs quality:** Multiple imperfect checkups serve as a good substitute for office visits; i.e., making more phone calls can be nearly as effective as replacing a few phone calls with office visits. Further, the larger the quantity of checkups, the more robust the solution is to errors in estimation/optimization.

In practice, hospitals could use the following steps to design better post-discharge monitoring policies: (1) estimate the time-to-develop the condition and the delay-time; (2) design an effective phone call questionnaire; (3) schedule checkups in a block with spacings approximately equal to the mean delay-time; (4) schedule office visits (perfect checkups) close to the time at which patients are at the highest risk of readmission; (5) spread phone calls farther apart from each other to cover a longer time period with improvements on the phone call questionnaires.

2.6 Discussion and Conclusion

In this study, we address the prevalent issue of hospital readmissions that concerns healthcare professionals, hospital patients, and policy-makers. We propose an analytical model based on delay-time analysis to design more effective post-discharge checkup policies for individual patients. Key results from our model not only provide theoretical extensions of the traditional delay-time analysis framework, but also important insights for healthcare decision-makers designing post-discharge checkup policies. By simultaneously optimizing with respect to multiple factors such as the number of checkups, the timing of checkups, and the types of checkup methods used, our model demonstrates significant improvements over current practice. Using the same number of checkups, current practice (which detects only 16% of the conditions experienced by readmission-bound patients) can be improved up to 23%, a relative improvement of 43.7%.

Future extensions upon this research may involve examining the benefit of detecting illnesses as early as possible. The current model assumes equal benefit from all early illness detections, however, it may be valuable to assign more benefit to earlier detections as they may result in less burden on the patients. Similarly, the current model also assumes that checkups have constant detection rate over the duration of a patient's readmission-causing condition. It may be valuable to examine the effect of time-dependent detection rate of phone calls, for example, the detection rate becomes higher as the patient has had the condition for a longer time. Another extension is to jointly optimize discharge (inpatient) and post-discharge (outpatient) decisions as the timing of discharge can affect readmission risk (Kelly et al., 2015; Rosen et al., 2017). While parameterizing our model with real data, we realized that empirical estimation could be challenging as our model requires two distributions (time-to-develop the condition and delay-time) as the input. One of the key empirical challenges is the issue of censoring, as we only utilized data within the finite 30-day readmission penalty window. In addition, patients have different intrinsic readmission risk: while some patients would not be readmitted, other patients would be readmitted regardless of post-discharge monitoring and interventions. Though the two distributions (and data beyond 30-day follow-up) are not widely available currently, we believe that our analysis will motivate the documentation and utilization of the delay-time and time-to-develop the condition information. We leave to future work the personalized delay-time and time-to-develop the condition forecast as well as more robust empirical estimation that considers the censorship of data.

The application of our model and findings has the potential for broad impacts including reduced hospital readmissions, improved quality of patient care, improved patient satisfaction, and reduced healthcare costs, all without overburdening clinicians (as clinician burden is often a major barrier to implementation of new healthcare practices). This is achievable by aligning checkup policy design with a number of key insights, namely: timing of checkups is the most important factor, checkup timing should be adjusted according to checkup detection rates, and checkup quantity is more important than checkup quality. At the same time, our model presents unique extensions to the traditional delay-time analysis framework by allowing for a time-varying failure rate and inhomogeneous detection rate. Thus, our model extends the scope of delay-time modeling and provides new insights into the structure of these types of problems. This ultimately broadens the scope of problems in which delay-time analysis can be applied.

Tested on an out-of-sample dataset containing 332 patients from the states of Florida, Iowa, North Carolina, New York, and Washington, our results demonstrate robustness, with absolute optimality gaps within 5%. As the number of checkups increases, the robustness further increases as the optimality gaps diminish. Our clinical collaborators have shown great interest in implementing our models and look forward to putting them through clinical testing. Going beyond cystectomy patients, the new framework developed has the potential to significantly reduce readmissions from a variety of surgical procedures, thereby improving the quality of patient care and decreasing healthcare costs.

CHAPTER III

Balancing Pre- and Post-Discharge Efforts

ABSTRACT: We developed a two-stage, Strengthen-then-Maintain framework for reducing hospital readmissions in the continuum of care spanning pre- and post-discharge. For the Markov Decision Process in the maintain stage, we develop a closed-form approximation for the cost-to-go with a theoretical accuracy bound which can then be used to understand the structure of the integrated two-stage framework. We apply this model to study 1) how a hospital balances readmission reduction efforts between pre- and post-discharge to minimize the cost of an episode of care and 2) how a healthcare funder designs a bundled payment and readmission penalty policy to incentivize readmission reduction. We specifically consider three policy levers: readmission penalty, treatment cost, and the length of the window of time that hospitals are responsible for readmissions (episode/penalty window). We provide a simple closed-form sufficient condition that captures the impact of these levers on the scope and efficacy of hospital readmission reduction programs. We find the episode/penalty window to be a major driver, which has yet to be purposefully employed by payers. Specifically, the length of the 30-day penalty window and 90-day bundled payment episode may be too long to incentivize readmission reduction for any but the low risk patients (which have little impact on overall readmission rate), even under additional penalty or subsidy. This may be an explanation for stalled readmission reduction after the implementation of the 30-day Medicare readmission penalty program, HRRP. Though payers want long windows to ensure hospitals cover as many readmission candidates as possible under reduction programs, long windows lead to smaller programs as hospitals "give up" on risky patients, whereas shortening this window can in fact encourage hospitals to expand readmission programs to include more and riskier patients.

3.1 Introduction

Hospital readmissions are burdensome and costly to the U.S. healthcare system. One in five Medicare Fee-For-Service patients were readmitted within 30 days of discharge (Jencks et al., 2009), and as much as 75% of readmissions have been determined to be preventable (Benbassat and Taragin, 2000). Reducing these preventable readmissions can save up to $25 billion for the U.S. healthcare system (PwC Health Research Institute, 2010). To incentivize hospitals to reduce readmissions, the Centers for Medicare and Medicaid Services (CMS) have established the Hospital Readmissions Reduction Program (HRRP) to penalized hospitals with excessive readmission rates. Moreover, the CMS has been experimenting with different reimbursement schemes, such as Pay-For-Performance (P4P) and Bundled Payments (BP), to provide further financial incentives for hospitals to reduce readmissions. Nonetheless, avoiding readmissions is still challenging both clinically and financially, and it is unclear whether the new payment structures offered by BP, P4P and penalty programs will be sufficient to incentivize further progress.

The battle against readmissions requires hospitals to exert readmission reduction effort in two phases of care: pre-discharge (during the inpatient stay) and post-discharge follow-up (after the patient has left the hospital). For instance, a hospital can extend the length of stay (LOS) to further stabilize a patient's condition and/or can perform follow-up checkups and treatments after the patient has been discharged. Though proven effective, these readmission reduction efforts can be overly costly. For example, the Reengineered Hospital Discharge Program (Project RED) conducted a randomized clinical trial and found that "the cost of the (readmission reduction) intervention (...) involved 0.5 full-time equivalent for a nurse and 0.15 full-time equivalent for a clinical pharmacist. If adopted broadly, this intervention could produce substantial effects on health care financing." (Jack et al., 2009). Due to the financial burden of the required efforts and investments, the momentum of readmission reduction has stalled since the implementation of the HRRP, as reported in a JAMA study (Desai et al., 2016).

To provide stronger readmission reduction incentives, the CMS is gradually shifting its reimbursement schemes from Fee-For-Service to Pay-For-Performance and Bundled Payment. Among these reimbursement schemes, the Bundled Payment is believed to be most effective at providing incentives to reduce readmissions (Andritsos and Tang, 2018; Guo et al., 2016). In 2013, the CMS established the Bundled Payment for Care Improvement (BPCI) Initiative. Under BPCI, a hospital receives a bundled payment for all costs incurred during an episode of care. Specifically, Model 2 of BPCI defines an episode of care to be “the inpatient stay in an acute care hospital plus the post-acute care and all related services up to 90 days after hospital discharge” (CMS, 2018a), including readmissions. Although a BP policy would provide stronger incentives for readmission reduction, the design of an effective BP policy still requires further investigation.

As a policymaker, the CMS faces challenging decisions when designing the BP policy and readmission penalty programs to properly incentivize readmission reduction. If the cost (plus penalty) of a readmission is too little, hospitals may be unmotivated to take action. If the costs of readmission reduction measures (e.g., extended LOS and intensive post-discharge follow-up care) are too expensive, hospitals may give up. In addition to the cost and penalty structures, the length of the readmission penalty window and length of an episode of care also play an important role. A New England Journal of Medicine article (Joynt and Jha, 2012) argued that hospitals have little control over readmissions that occur more than seven days after discharge, therefore policymakers should consider limiting the readmission penalty window. Besides the readmission time frame, further complicating the matter are factors such as the variation in the baseline readmission risk among surgical cohorts (e.g., total joint arthroplasty has a 4% 30-day readmission rate while cystectomy has a 25% one), the pathological nature of the complications that cause readmissions (infection, organ failure, heart attack, etc.), and the various effectiveness of the post-discharge treatments in preventing readmissions.

This chapter studies the policy-level decisions driven by the operational factors that are critical to a hospital’s readmission reduction. We study key policy-level decisions such as designing readmission penalty programs, subsidizing post-discharge follow-up treatments, and shortening/extending the length of an episode and the readmission penalty window length. We consider these factors in the context of the CMS’s BPCI program – “Model 2 and Model 3 involve a retrospective bundled payment arrangement where actual expenditures are reconciled against a target price for an episode of care” (CMS, 2018a). To study policy-level decisions, we study how a hospital may allocate readmission reduction efforts between the pre- and post-discharge stages of care. In the pre-discharge stage, the hospital exerts effort to reduce the readmission risk of a patient cohort. In the post-discharge stage, the hospital provides post-discharge follow-up care to the patient cohort whose readmission risk is determined by the pre-discharge efforts. This integrated two-stage framework enables us to analytically study the design of an effective bundled payment policy to align incentives.

This two-stage framework, which we call the *Strengthen Then Maintain* (STM) framework, is applicable to a set of machine maintenance problems beyond readmission management. In the strengthening stage, a machine is strengthened so that its failure rate is reduced. In the maintaining stage, the failure rate (determined by the strengthening efforts) is exogenous and the machine is (ad-hoc) repaired if defects appeared. If not repaired promptly, the machine could result a catastrophic failure. An example of such problems can be an aircraft maintenance problem – in the strengthening stage, an airline company performs (downtime) maintenance for an aircraft, which reduces the failure rate. In the maintaining stage, the aircraft is in operation, ad-hoc repairs are provided at the destination airports. Our model uncovers valuable managerial insights into this type of maintenance problems.

3.1.1 Contributions

- **Theoretical.** This chapter makes theoretical contributions to the reliability literature. We develop a novel Strengthen Then Maintain (STM) framework and study how a decision maker balances efforts between the strengthening stage and the maintaining stage. The maintaining stage is modeled as a discrete-time finite-horizon Markov Decision Process (MDP). We provide a closed-form expression for the cost-to-go for the machine maintenance MDP. Our approach does not require a parametric failure rate distribution – for an arbitrary (non-stationary) failure rate, we prove a theoretical accuracy bound for the closed-form cost-to-go under an arbitrary optimal policy. By studying the closed-form expression analytically, we show that the cost-to-go is affected by the failure rate at an asymptotically linear rate. By integrating the strengthening and maintaining stages, we provide a simple closed-form sufficient condition for a policy to incentivize the reduction of the failure rate (i.e., readmission rate). The analytical results of the STM framework can be generalized to machine maintenance problems with two distinct stages.

- **Practical.** We also uncover novel insights for designing an effective Bundled Payment policy for readmission reduction. Our analytical results suggest that hospitals have more financial incentives if the readmission penalty window is shortened, the cost of post-discharge follow-up treatment is reduced, the cost/penalty of readmission is increased, and the post-discharge treatment efficacy is improved. For patients that are likely to experience acute events leading to a swift readmission, a more effective mechanism is to shorten the penalty window. For other cohorts, subsidizing inpatient and outpatient efforts may be effective. We believe that our model is the first study that analytically addresses how the penalty window length impacts the incentives for readmission reduction. Unlike many game-theoretical studies which use stylized functional forms, our model is less restrictive with minimal assumptions imposed on the functional form. At the core of our model is a Markov Decision Process model, which directly captures the patient's pathological deterioration and the cost/penalty structures.

The practical implications of this chapter add valuable insights to Bundled Payment and readmission penalty policy design. As a proof of concept, we develop and validate our models using clinical data collected from bladder patients undergo cystectomy. We found that 90-day and 30-day windows may not provide readmission reduction incentives sufficiently for high-risk surgery cohorts. Our study suggests that 14-day window would be robust in incentivizing readmission reduction, even if post-discharge treatments are inefficient and the efficacy is low.

The rest of the chapter is organized as follows. Section 3.2 gives a brief review of the relevant literature. In Section 3.3, we give an overview of the modeling framework, which consists of a pre-discharge stage and a post-discharge stage. The post-discharge stage model is introduced in Section 3.4 and the pre-discharge stage model is introduced in Section 3.5. Section 3.6 studies the balance of efforts between the two stages. The policy implications are discussed in Section 3.7. In Section 3.8, we conduct numerical studies using institutional data from patient undergo cystectomy. Section 3.9 discusses some practical considerations and limitations of our policy recommendations and potential future works. Finally, Section 3.10 concludes the chapter.

3.2 Literature Review

This work lies at the junction of disease screening/monitoring and healthcare payment/contract policy. Four streams of literature are relevant to this chapter: 1) reliability; 2) disease monitoring and screening; 3) readmission management studies in the operations management literature; and 4) the healthcare payment/contract studies that aim to align incentives for better health outcomes. We provide a brief overview of these four streams of literature.

1) Reliability. In the reliability literature, many machine maintenance models assume that once a machine fails, it can be repaired or replaced and it will resume its normal operations. Hence, these models focus on minimizing the long-run cost and downtime in a steady-state (infinite horizon) setting (Wang, 2002). These models are thus not suitable for our problem as we consider a finite period of time (due to the finite nature of an episode of care, by definition). Our problem also considers failures to be catastrophic, since readmission is a major adverse health event and may require intensive treatments. Few papers, such as Özekici and Pliska (1991) and Milioni and Pliska (1988), study maintenance policies with catastrophic failures. However, most models focus on "maintenance" only – they do not study the "strengthening" stage where the failure rate can be reduced.

A relevant stream of literature studies maintenance under warranties. These studies are close to our work, as the manufacturer is responsible for all repair costs within a warranty (see Shafiee and Chukova (2013a) for a literature survey). In particular, models have been created to study used item upgrades/refurbishments and warranties – before selling the item, the manufacturer can spend money to upgrade/refurbish the item to reduce the failure rate. Among these models, Shafiee and Chukova (2013b) is one of the first that studies how the seller of a used item can spend efforts to upgrade an item, so that the failure rate is reduced and the warranty service cost is consequently reduced. Our model is different from this stream of literature in the following ways: 1) We aim to study the policy-level decisions where a policymaker (the CMS) makes a policy and an agent (the hospital) responds to the policy; while reliability models typically have only one decision maker – the seller, whose goal is to maximize profits. 2) Our model does not force the seller to provide warranty maintenance when it is economically not viable to do so (due to high costs). In particular, we allow a hospital to adopt an engaged policy or a disengaged policy, whichever is cheaper, depending on the cost and penalty structure and the patient's risk characteristics. The goal of our model is to incentivize (not to force) "upgrade/refurbish" behaviors via policy levers. 3) Reliability models often assume specific failure distributions (e.g., Weibull) but our model can handle arbitrary distributions with a proven accuracy bound.

2) Disease Monitoring and Screening. Many models have been studied for the prevention and treatment of chronic diseases such as cancer, hepatitis, and glaucoma (Ayer et al., 2012; He et al., 2016; Helm et al., 2015). These models optimize prevention and treatment decisions based on continuous and/or categorical variables that are well-established clinical studies of these diseases. Examples include the

Intra-ocular Pressure (continuous) of a glaucomatous eye and the METAVIR fibrosis score (categorical) of a hepatitis patient. Unlike these models, our problem focuses on modeling the post-discharge deterioration and readmission of a patient after the patient leaves the hospital, where little to no clinical information is observed while the patient is away from the hospital. Hence, we model the disease deterioration using three discrete states (healthy, sick, and readmitted), similar to many studies in existing literature (Helm et al., 2016; Liu et al., 2018a).

Some papers in the healthcare management literature consider the "strengthening" effect of inpatient care. For example, Andritsos and Tang (2018) models the health outcome as a co-production process. Zhang et al. (2016) and Adida et al. (2016) also model the readmission risk reduction effect of a hospitalization. We follow the literature and model the cost of "strengthening" efforts in a stylized way, without specific functional form assumptions. The novelty of our model is that we link the strengthening stage to a more structured MDP model that captures key operational factors in the post-discharge monitoring practice.

3) Readmission Management. Many clinical studies have found effective methods to reduce hospital readmissions. These methods include patient education, pre-discharge assessment, domiciliary aftercare, and post-discharge follow-up care. (Benbassat and Taragin, 2000; Wong et al., 2013). In the operations management literature, scholars have studied readmissions empirically and analytically. Kim et al. (2014) and Chan et al. (2012) studied how admission and discharge strategies of ICU admission affects readmission rates respectively. Chen et al. (2018) builds a readmission prediction model that incorporates latent heterogeneity using claims data. Bayati et al. (2014) developed a readmission prediction model and analyzed post-discharge intervention decisions. Helm et al. (2016) built a readmission prediction model and optimized staffing decisions for post-discharge follow-up appointments. Related to our work is Zhang et al. (2016), where the HRRP program was analyzed from a game theoretic approach. The authors studied single-year, multi-year, and two-hospital games to analyze the financial incentives under HRRP. Moreover, their model considers readmission reduction effort at a high level of abstraction (not detailing the operational tactics) and does not focus on the operational interplay between pre- and post-discharge efforts. Our model is different from these models as these models do not capture the effect of post-discharge follow-up care. Moreover, we believe we are the first to analytically study the effect of the readmission penalty window length.

4) Healthcare Payments and Contracts. Many papers studied the design of effective healthcare payment policies and contracts to align incentives for better care and health outcome. Some recent works include Adida et al. (2016); Andritsos and Tang (2018); Guo et al. (2016); Aswani et al. (2017); Bastani et al. (2016); Adida and Bravo (2018). Within this domain, a stream of literature takes a game-theoretic approach. Guo et al. (2016) is at the intersection between payment policy and hospital operations management. They studied a three-level game with an M/M/1 queue embedded to analyze incentives. Andritsos and Tang (2018) models the health outcome (readmissions) as a co-production process by the hospital and the patient. They studied the equilibria under FFS, P4P, and BP schemes. In their model, the hospital exerts effort in the inpatient stay stage only – it does not consider post-discharge outpatient treatments that can potentially avert readmissions. Close to our work is Adida et al. (2016), where the model consists of a first-stage (inpatient-stay stage) cost and a second-stage (post-discharge stage) cost. They assumed that the second-stage cost of follow-up is independent of the first stage cost. In our model, the two costs are interlinked by the readmission risk as a surrogate for the effort. This is an important distinction, since the operations management literature indicates that the efficacy of follow-up programs is strongly dependent on the patient risk at time of discharge (Helm et al., 2016), which is what is controlled in the inpatient stage.

3.3 Modeling Framework Overview

The overarching goal of this chapter is to study the design of an effective BP policy. To achieve this, we study the behavior of a hospital under a BP policy to infer policy-level implications for BP policy design. An overview of the framework is illustrated in Figure 3.1.

At the provider level, we study how a hospital behaves when it minimizes the cost of a T-day episode of care under a BP policy. In particular, we seek to understand what would entice a hospital to engage in readmission reduction in both the pre- and post-discharge stages. In the pre-discharge stage, the hospital invests $C_S(\rho)$ to reduce a patient's readmission risk to ρ from a baseline ρ_0 (see Definition III.2). A lower readmission risk ρ requires greater effort and thus a greater $C_S(\rho)$. This may involve clinical interventions such as administering advanced treatments and extending the LOS for further observation and stabilization. In the post-discharge stage, the hospital provides outpatient follow-up care for the patient whose readmission risk is ρ. The two stages are interlinked by the readmission risk ρ as we assume that the post-discharge readmission risk is a result of the pre-discharge efforts.[1] The

[1] We assume that the readmission risk can only be reduced during the index hospitalization. This is not a

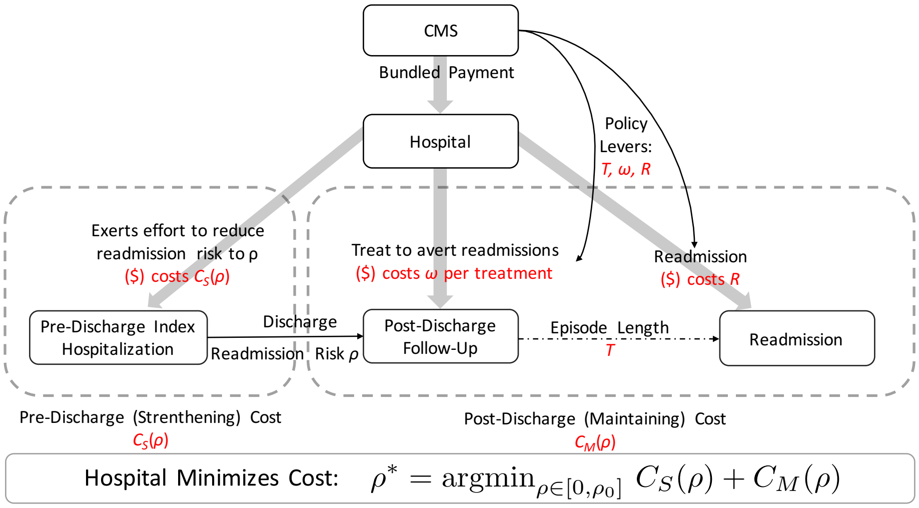

CMS
Bundled Payment
Hospital
Policy Levers: T, ω, R
Exerts effort to reduce readmission risk to ρ
($) costs $C_S(\rho)$
Treat to avert readmissions
($) costs ω per treatment
Readmission
($) costs R
Pre-Discharge Index Hospitalization
Discharge
Readmission Risk ρ
Post-Discharge Follow-Up
Episode Length
T
Readmission
Pre-Discharge (Strenthening) Cost
$C_S(\rho)$
Post-Discharge (Maintaining) Cost
$C_M(\rho)$
Hospital Minimizes Cost: $\rho^* = \text{argmin}_{\rho\in[0,\rho_0]} C_S(\rho) + C_M(\rho)$

Figure 3.1: Overview of the Framework

cost of follow-up is denoted by $C_M(\rho)$, which includes the cost of conducting post-discharge follow-up treatment (which costs ω per treatment) to avert readmissions (e.g., treating an infection outpatient) and potentially the cost a readmission (R) if readmitted. The hospital chooses the optimal readmission risk level ρ^* (as a surrogate for the optimal effort level) to minimize the cost of the entire episode of care $Z(\rho) = C_S(\rho) + C_M(\rho)$.

We would like to emphasize that our model is generalizable to a set of production and maintenance problems consisting of a strengthening stage – where a machine/product is strengthened, so that its failure rate is reduced to ρ at a cost of $C_S(\rho)$ – and a maintaining stage – where $C_M(\rho)$ is incurred by the maintenance and repair/replacement of the machine/product over a finite period. A table of notation is provided in Table 3.1.

restrictive assumption because, in clinically practice, physicians have more control over a patient's health via inpatient interventions. In the post-discharge stage, we assume that the outpatient treatments do not alter the nature of a patient's readmission pathology. This is a critical assumption to ensure the Markovian property and therefore the model's analytical tractability.

Notation	Description
$\rho, \rho_t, \vec{\rho}$	Readmission risk
ρ_0	Baseline readmission risk
$\bar{\rho}$	Engagement threshold
$\underline{\rho}$	A lower bound for $\bar{\rho}$
d	Delay-time failure rate
f	Treatment efficacy
$C_S(\rho)$	The pre-discharge (strengthening) stage cost of reducing readmission risk to ρ
$C_M(\rho)$	The post-discharge (maintaining) stage cost of follow-up with a patient cohort at readmission risk ρ
$Z(\rho)$	The total cost $Z(\rho) = C_S(\rho) + C_M(\rho)$
R	Cost of a readmission
ω	Cost of a post-discharge outpatient treatment
RA, S, H	Readmitted, Sick, Healthy states in the Markov Decision Process
V_t^*	The optimal cost-to-go
V_t^π	The cost-to-go if policy π is implemented

Table 3.1: Table of Notation

3.3.1 BP Policy and Balance Criterion

At the policy level, we study how the CMS shall leverage critical policy levers to incentivize hospitals to engage in readmission reduction. First, we define the notion of a BP policy and the key policy levers therein.

Definition III.1 (A Bundled Payment Policy)**.** Let the tuple (T, ω, R) denote a Bundled Payment (BP) policy. It has the following policy levers:

- T, the episode length: this is a key policy-level decision we consider for the CMS. The length of a BP episode (and the HRRP penalty window) is a nontrivial and challenging policy decision as argued by Joynt and Jha (2012).
- ω, the (average) cost of a post-discharge outpatient treatment: although this cost is determined by the types of treatment and procedures administered in each post-discharge outpatient encounter, we argue that the CMS can subsidize post-discharge monitoring and treatments (e.g., PCP visit, home care, and etc.) to effectively lower the cost of treatment ω to achieve a better continuum of care between pre- and post-discharge.
- R, the (average) cost (plus potential penalties) of a readmission: this cost is determined by the pathological nature of each readmission case. However, we argue that the CMS can have some control over this cost by penalizing readmissions and/or decreasing the Bundled Payment amount (so that readmission costs take up a greater portion of the BP budget).[2]

[2] Currently, BPCI and HRRP do not overlap – hospitals do not get penalized for excessive readmissions if they are reimbursed under a BP scheme. However, if stronger incentives are needed, it might be viable for the CMS to impose HRRP penalties on top of the BP scheme.

In this chapter, we do not consider the decision of the Bundled Payment amount (a.k.a. the target price as defined by BPCI). The CMS has a well-established method for calculating the target price (CMS, 2018b); therefore this is out of our research scope.

To determine whether the hospital is effectively reducing readmissions, we introduce the following notion of the baseline readmission risk. We follow Zhang et al. (2016) and define the baseline readmission risk as follows:[3]

Definition III.2 (Baseline Readmission Risk)**.** The *baseline readmission risk* of a patient cohort, ρ_0, is defined as the readmission risk observed without exerting any additional (costly) readmission reduction effort. Ideally, this would be the natural readmission rate intrinsic to the pathology of the surgery and its recovery. However, it is difficult to observe such a natural readmission risk since patients are rarely free of interventions. For the policy-level purpose of incentivizing readmission reduction, in practice, the CMS can set ρ_0 to the national average or the pre-HRRP historical readmission risk. This serves as a baseline for evaluating whether a hospital is effectively reducing readmissions.

We say a hospital or a BP policy is balancing pre- and post-discharge readmission reduction efforts if it manages to reduce the readmission risk to below the baseline risk. The following definition formalizes this criterion.

Definition III.3 (Balance)**.** Suppose the hospital minimizes the cost of the entire episode of care by choosing the optimal readmission risk level (as a surrogate for the effort level):

$$\rho^* = \arg\min_{\rho \in [0, \rho_0]} C_S(\rho) + C_M(\rho). \tag{3.1}$$

If $\rho^* < \rho_0$, the hospital and the BP policy is *balancing* efforts.

As we shall see later in Section 3.6, this implies that the hospital is engaged in post-discharge care so that the follow-up cost $C_M(\rho^*) < R$ is reduced and the hospital is exerting pre-discharge readmission reduction effort so that $C_S(\rho) > 0$.

3.4 The Post-Discharge ("Maintaining") Stage

In this section, we develop the post-discharge stage model. We first introduce the post-discharge MDP model that captures the key cost/penalty structure and the

[3] Zhang et al. (2016) defines the "cost of process improvements" to be $C(r_0, r)$, which captures the cost of reducing the readmissions rate from the baseline r_0 to r.

readmission risk characteristics. Then, we study a hospital's behavior if it minimizes the cost according to this MDP. As such, we obtain key insights from the MDP to guide policy-level decisions.

3.4.1 The Post-Discharge Follow-Up MDP

Under a BP policy, a hospital aims to minimize the expected cost of the entire episode of care. As such, in the post-discharge stage, the hospital seeks to find the optimal follow-up policy to minimize the follow-up cost. We assume that the hospital's optimal post-discharge policy can be obtained by solving a discrete-time, finite-horizon MDP. Figure 3.2 summarizes the immediate costs, the terminal costs, and the transition probabilities of the MDP. The notations of the post-discharge MDP are as follows.

$c_t(s,a)$	$s = RA$	$s = S$	$s = H$
$a = Treat$	0	ω	ω
$a = Wait$	0	0	0
Terminal Costs	R	0	0

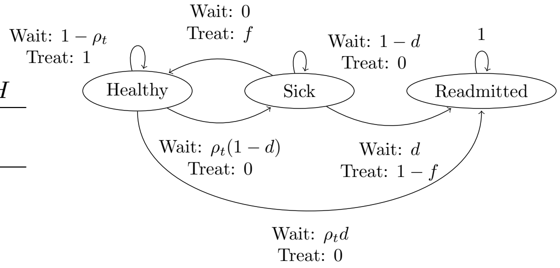


Figure 3.2: Costs and Transition Probabilities of the Post-Discharge Follow-Up MDP.

- $t \in \{0, 1, ..., T\}$: we assume that decisions are made on a daily basis after discharge. The day of discharge is denoted by $t = 0$. The episode length is denoted by T.

- $s, s_t \in \mathcal{S} = \{H, S, RA\}$: the state space $\mathcal{S}$ consists of *Healthy* (H) – the patient is free of readmission-causing conditions such as infections and failure to thrive; *Sick* (S) – the patient has developed some readmission-causing condition but has not yet readmitted; and *Readmitted* (RA) – the patient has been readmitted to the hospital.

- $a, a_t \in \mathcal{A} = \{Treat, Wait\}$ in each decision epoch, the healthcare provider decides whether to *treat* the patient or to *wait*.

- $c_t(s,a)$ represents the immediate cost of applying action a on a patient in state s in period $t \in \{0, 1, ..., T-1\}$. We assume that the cost of treating a patient who is not yet readmitted $(s \in \{H, S\})$ is ω. We set the cost of treatment on a readmitted patient zero, since readmission is assumed to be the absorbing state. The immediate cost of waiting (doing nothing) is zero in all states. $c_T(s)$ represents the terminal cost. We assume that the cost of a readmission is R. The cost of terminating in the healthy state or the sick state is zero.

- $P_t(s_{t+1}|s_t, a_t)$ denotes the transition probability from s_t to s_{t+1} when action a_t is applied in period $t \in \{0, 1, ..., T-1\}$. The developing and worsening of a readmission-causing complication is governed by two probabilities. The *readmission risk*, denoted by $\rho_t \in [0, 1]$, is the probability of a patient developing such a complication in period t (transitioning from H to S). The *delay-time failure rate*, denoted by $d \in [0, 1]$, is the probability of a sick patient worsening to the point of requiring a readmission in period t (transitioning from S to RA). This concept stems from the delay-time analysis framework (Liu et al., 2018a), which models the time it takes for a condition to worsen to a readmission. The *treatment efficacy*, denoted by $f \in [0, 1]$, is the probability of a treatment successfully averting a readmission. We assume that treating a healthy patient will maintain the patient's healthy state. Finally, we assume RA is an absorbing state.[4]

- $\pi(s, t)$ denotes a follow-up policy that maps $\mathcal{S} \times \{0, 1, ..., T-1\}$ to $\mathcal{A}$. An optimal follow-up policy is denoted by π^*.

The MDP model makes the following key assumptions. 1) The hospital minimizes the cost under a BP policy. Although hospitals are typically not-for-profit organizations, they more or less act like cost-minimizing business entities as widely modeled in the literature (Zhang et al., 2016; Adida et al., 2016). 2) An episode of care consists of an index hospitalization and at most one readmission due to the finite nature of an episode of care defined by BPCI. This is consistent with many Markov disease models in the literature – typically the absorbing state involves a major adverse event and/or a major pathological change (e.g., heart attack, death, readmission, cancer diagnosis). Moreover, among surgery patients, more than one readmissions within 90 days are rare. Even if there are repeated readmissions, the pathology after the index hospitalization and each re-hospitalization is drastically different.

To minimize the expected cost in the post-discharge stage, the hospital solves

[4] Note that from a modeling stand-point, it is possible to model death as an absorbing state. However, we do not model death since we assumed hospitals minimize costs and it is difficult to assign a monetary cost to death. Furthermore, the chance of death within the bundle payment window is rare compared to the probability of readmissions, and mortality rate is not penalized by the CMS. For these considerations, we do not model mortality.

the Bellman equations of the MDP. Let $V_t^*(s) = V^{\pi^*}(s)$ be the minimum expected post-discharge cost of following-up with a patient in state $s \in \mathcal{S}$ on the t^{th} day after discharge. The Bellman equations are given by:

If $t = T$: $V_t^*(s) = c_T(s), \forall s \in \mathcal{S}$. For $t \in \{0, 1, ..., T-1\} : V_t^*(RA) = R$, and

$$V_t^*(H) = \min \begin{cases} \omega + V_{t+1}^*(H) & (a_t = Treat) \\ \\ (1-\rho_t)V_{t+1}^*(H) + \rho_t(1-d)V_{t+1}^*(S) + \rho_t dR & (a_t = Wait) \end{cases} \tag{3.2}$$

$$V_t^*(S) = \min \begin{cases} \omega + fV_{t+1}^*(H) + (1-f)R & (a_t = Treat) \\ \\ (1-d)V_{t+1}^*(S) + dR & (a_t = Wait) \end{cases} \tag{3.3}$$

We are interested in analyzing hospitals' behavior as they reduce the readmission risk ρ_t. Hence, we introduce the following notation for the purpose of evaluating the post-discharge cost as a function of the risk ρ_t. When an arbitrary follow-up policy π (not necessarily optimal) is applied to a patient whose readmission risk is ρ_t, the expected follow-up cost is given by:

If $t = T$: $V_t^\pi(s, \rho_t) = c_T(s), \forall s \in \mathcal{S}$. For $t \in \{0, 1, ..., T-1\} : V_t^\pi(RA, \rho_t) = R$, and

$$V_t^\pi(H, \rho_t) = \begin{cases} \omega + V_{t+1}^\pi(H, \rho_{t+1}) & \text{if } \pi(H,t) = Treat \\ (1-\rho_t)V_{t+1}^\pi(H, \rho_{t+1}) + \rho_t(1-d)V_{t+1}^\pi(S, \rho_{t+1}) + \rho_t dR & \text{if } \pi(H,t) = Wait \end{cases}$$

$$V_t^\pi(S, \rho_t) = \begin{cases} \omega + fV_{t+1}^\pi(H, \rho_{t+1}) + (1-f)R & \text{if } \pi(S,t) = Treat \\ (1-d)V_{t+1}^\pi(S, \rho_{t+1}) + dR & \text{if } \pi(S,t) = Wait \end{cases}$$

Without loss of generality, the patient is assumed to be in the healthy state upon discharge in period $t = 0$. Under policy π, the expected T-period follow-up cost of a patient whose readmission risk is $\vec{\rho} = (\rho_0, \rho_1, \ldots, \rho_T)$ is denoted by

$$V_0^\pi(\vec{\rho}) = V_0^\pi(H, \rho_0). \tag{3.4}$$

If the readmission risk is stationary (i.e. $\rho_t = \rho, \forall t$), then we write $V_0^\pi(\rho) = V_0^\pi(H, \rho)$ for simplicity. Next, we shall analyze the behavior of a hospital when the minimize the expected cost according to the optimal policy obtained by solving the post-discharge follow-up MDP.

3.4.2 Technical Preliminaries

To facilitate our discussion, we present some technical preliminaries, including some assumptions we impose and some structural properties of the MDP. First, we make the following two mild assumptions so that the follow-up MDP model is realistic and the hospital's follow-up behavior is sensible.

Assumption III.4 (Treatments are Effective)**.** *We assume that* $1 - f < d$*, i.e., the probability of a sick patient being treated and readmitted is smaller than the (no-treatment) delay-time failure rate.*

In reality, this is very likely to hold since post-discharge care is clinically proven to be effective at reducing readmissions.

Assumption III.5 (Following-Up Every Day is Expensive)**.** *We assume that* $T\omega > R$*, i.e., the cost of providing treatments every single day over the* T*-day period is more expensive than a readmission.*

This assumption is likely to be true as verified in our numerical study (see Section 3.8.1). If this is violated, a hospital could provide post-discharge treatments every single day at a cost cheaper than a readmission. Under the current definition of an episode of care by BPCI, this may require the provider to see the patient every single day in the 90-day period, which will be unrealistic and much more expensive than a readmission.

In addition to these mild assumptions to ensure sensible behavior, we also make the following stationary readmission risk assumption for analytical tractability. We do acknowledge that a patient's readmission risk is not likely to be stationary in reality (Liu et al., 2018a). Nonetheless, this assumption is only used for tractability in the development of some initial structural properties.

Assumption III.6 (Stationary Readmission Risk)**.** *The readmission risk is stationary, i.e.,* $\rho_t = \rho, \forall t \in \{0, 1, ..., T-1\}$*.*

This assumption is only used in Propositions III.9 and Theorem III.13. When the readmission risk is nonstationary, Theorem III.16 provides a theoretical accuracy bound.

First, we present a lemma which will be used in later proofs. Lemma III.7 states that in any period, the cost-to-go of a healthy patient is no more than the cost-to-go of a sick patient. All costs-to-go are bounded by the cost of a readmission R since the worst case cost is R as it is assumed to be the absorbing state.

Lemma III.7. $V_t^*(H) \leq V_t^*(S) \leq V_t^*(RA) = R, \forall t \in \{0, 1, ..., T\}$.

Proof. Proof of Lemma III.7. Define a policy π such that $\pi(s,t) = W, \forall s \in \mathcal{S}, t \in \{0, 1, ..., T\}$. Note that this policy always applies "wait" to a patient, which incurs no immediate cost. Hence its maximum expected cost is no greater than the maximum terminal cost $c_T(RA) = R$. Then we have $V_t^*(s) \leq V_t^\pi(s) \leq R, \forall s \in \mathcal{S}, \forall t \in \{0, 1, ..., T\}$.

Next, we show $V_t^*(H) \leq V_t^*(S), \forall t \in \{0, 1, ..., T\}$ by induction. For $t = T$, $V_T^*(H) = 0 \leq V_T^*(S) = 0$ holds. Suppose the induction base assumption holds: $V_{t+1}^*(H) \leq V_{t+1}^*(S)$.

Recall the Bellman equations in t:

$$V_t^*(H) = \min \begin{cases} \omega + V_{t+1}^*(H) & (a_t = Treat) \\ (1-\rho_t)V_{t+1}^*(H) + \rho_t(1-d)V_{t+1}^*(S) + \rho_t dR & (a_t = Wait) \end{cases} \tag{3.5}$$

$$V_t^*(S) = \min \begin{cases} \omega + fV_{t+1}^*(H) + (1-f)R & (a_t = Treat) \\ (1-d)V_{t+1}^*(S) + dR & (a_t = Wait) \end{cases} \tag{3.6}$$

We first argue that $V_t(H, a_t = Treat) \leq V_t(S, a_t = Treat)$. $\omega + V_{t+1}^*(H) \leq \omega + V_{t+1}^*(S) \leq \omega + fV_{t+1}^*(S) + (1-f)R$ since $V_{t+1}^*(S) \leq R$ and $V_{t+1}^*(H) \leq R$. We next argue that $V_t(H, a_t = Wait) \leq V_t(S, a_t = Wait)$. It follows that $(1-\rho_t)V_{t+1}^*(H)+\rho_t(1-d)V_{t+1}^*(S)+\rho_t dR = (1-\rho_t)V_{t+1}^*(H)+\rho_t V_{t+1}^*(S)+\rho_t d(R-V_{t+1}^*(S)) \leq V_{t+1}^*(S) + d(R - V_{t+1}^*(S)) = (1-d)V_{t+1}^*(S) + dR$.

Finally, we argue that $V_t^*(H) = \min(V_t(H, a_t = Treat), V_t(H, a_t = Wait)) \leq V_t^*(S) = \min(V_t(S, a_t = Treat), V_t(S, a_t = Wait))$. We show this by contradiction. Suppose (for contradiction) that $\min(V_t(H, a_t = Treat), V_t(H, a_t = Wait)) > \min(V_t(S, a_t = Treat), V_t(S, a_t = Wait))$. At least one of the following must hold:

- $V_t(H, a_t = Treat) > V_t(S, a_t = Treat)$ and $V_t(H, a_t = Wait) > V_t(S, a_t = Treat)$
- $V_t(H, a_t = Treat) > V_t(S, a_t = Wait)$ and $V_t(H, a_t = Wait) > V_t(S, a_t = Wait)$

However, either one leads to contradiction. Hence, by contradiction, the induction holds.

□

Next, we develop some structural properties of the post-discharge MDP. The following proposition establishes the optimality of a control limit policy. In practice, a hospital is very likely to adopt such a control limit policy because it is insensible for a hospital to treat a sick patient and not to treat a healthy patient.

Proposition III.8 (Control Limit Policy Optimality). *If $f = 1$ (treatments are perfect), then there exists an optimal control limit policy π^* such that if $\pi^*(H, t) = Treat$ then $\pi^*(S, t) = Treat$, and if $\pi^*(S, t) = Wait$ then $\pi^*(H, t) = Wait$ $\forall t \in$*

$\{0, 1, ..., T-1\}$. *If $f < 1$ (treatments are imperfect), then condition 4 in Puterman (2005) Theorem 4.7.4 is violated. The existence of optimal control limit policies is not necessary.*

Proof. Proof of Proposition III.8. To show the existence of an optimal control limit policy, we introduce a dummy action $\varnothing$ and modifies the costs and transition probabilities as follows:

Costs

	$s = RA$	$s = S$	$s = H$
$a = \varnothing$	0	0	0
$a = Treat$	M	ω	ω
$a = Wait$	$2M$	0	0
Terminal Costs	R	0	0

ω: cost of treatment

R: cost of readmission

M: a sufficiently large cost ($M > R + T\omega$)

Transition Probabilities

$a_t = \varnothing$	$s_{t+1} = RA$	$s_{t+1} = S$	$s_{t+1} = H$
$s_t = RA$	1	0	0
$s_t = S$	1	0	0
$s_t = H$	1	0	0

$a_t = Treat$	$s_{t+1} = RA$	$s_{t+1} = S$	$s_{t+1} = H$
$s_t = RA$	0	0	1
$s_t = S$	0	0	1
$s_t = H$	0	0	1

$a_t = Wait$	$s_{t+1} = RA$	$s_{t+1} = S$	$s_{t+1} = H$
$s_t = RA$	1	0	0
$s_t = S$	d	$1-d$	0
$s_t = H$	$\rho_t d$	$\rho_t(1-d)$	$1-\rho_t$

Note that it is always optimal to take the dummy action in the RA state. Hence, the new MDP with the dummy action is equivalent to the original MDP.

Let us order the state space as $\mathcal{S} = \{0 = RA, 1 = S, 2 = H\}$ and order the actions as $\mathcal{A} = \{0 = \varnothing, 1 = Treat, 2 = Wait\}$. Now we can verify the sufficient conditions for the existence of an optimal control limit policy (Theorem 4.7.4 in Puterman (2005)).

1. $c_t(s, a)$ is non-increasing in $s \in \mathcal{S}$ for all $a \in \mathcal{A}$.

2. The tail sum of transition probabilities $q_t = (k|s, a) = \sum_{j=k}^{2} P_t(j|s, a)$ is nondecreasing in s for all $k \in \mathcal{S}$ and $a \in \mathcal{A}$.

3. $c_t(s, a)$ is subadditive on $\mathcal{S} \times \mathcal{A}$.

4. The tail sum of transition probabilities $q_t = (k|s, a)$ is supermodular on $\mathcal{S} \times \mathcal{A}$ for all $k \in \mathcal{S}$.

5. $c_T(s)$ is non-increasing in s.

Conditions 1 and 5 can be easily verified. We verify conditions 2, 3, and 4.

Condition 2:

$$q_t(k|s,a) = \begin{cases} \begin{array}{c|ccc} a=0 & k=0 & k=1 & k=2 \\ \hline s=0 & 1 & 0 & 0 \\ s=1 & 1 & 0 & 0 \\ s=2 & 1 & 0 & 0 \\ a=1 & k=0 & k=1 & k=2 \\ \hline s=0 & 1 & 1 & 1 \\ s=1 & 1 & 1 & 1 \\ s=2 & 1 & 1 & 1 \\ a=2 & k=0 & k=1 & k=2 \\ \hline s=0 & 1 & 0 & 0 \\ s=1 & 1 & 1-d & 0 \\ s=2 & 1 & 1-\rho_t d & 1-\rho_t \end{array} \end{cases}$$

Condition 3 holds as $\omega - 0 \leq M - 0$ and $0 - \omega \leq 2M - M$.

Condition 4:

- $k = 0$ then $q_t(0|s,a) = 1, \forall a \in \mathcal{A}$ and $s \in \mathcal{S}$. Supermodularity holds trivially.

- $k = 1$, we have

$$q_t(1|s+1, a+1) - q_t(1|s+1, a) = \begin{array}{c|cc} & a=0 & a=1 \\ \hline s=0 & 1 & -d \\ s=1 & 1 & -\rho_t d \end{array} \tag{3.7}$$

$$q_t(1|s, a+1) - q_t(1|s, a) = \begin{array}{c|cc} & a=0 & a=1 \\ \hline s=0 & 1 & -1 \\ s=1 & 1 & -d \end{array} \tag{3.8}$$

(3.9)

  $q_t(1|s+1, a+1) - q_t(1|s+1, a) \geq q_t(1|s, a+1) - q_t(1|s, a)$ holds for $s \in \{0, 1\}, a \in \{0, 1\}$.

- $k = 2$, we have

$$q_t(2|s+1, a+1) - q_t(2|s+1, a) = \begin{array}{c|cc} & a=0 & a=1 \\ \hline s=0 & 1 & -1 \\ s=1 & 1 & -\rho_t \end{array} \tag{3.10}$$

$$q_t(2|s, a+1) - q_t(2|s, a) = \begin{array}{c|cc} & a=0 & a=1 \\ \hline s=0 & 1 & -1 \\ s=1 & 1 & -1 \end{array} \tag{3.11}$$

(3.12)

  $q_t(1|s+1, a+1) - q_t(1|s+1, a) \geq q_t(1|s, a+1) - q_t(1|s, a)$ holds for $s \in \{0, 1\}, a \in \{0, 1\}$.

$q_t(2|s+1, a+1) - q_t(2|s+1, a) \geq q_t(2|s, a+1) - q_t(2|s, a)$ holds for $s \in \{0, 1\}, a \in \{0, 1\}$. Therefore, all five conditions are satisfied. There exists an optimal control limit policy π^* that is non-decreasing in s. In other words, if $\pi^*(H, t) = Treat$ then $\pi^*(S, t) = Treat$ and if $\pi^*(S, t) = Wait$ then $\pi^*(S, t) = Wait$.

□

Note that when $f < 1$, the existence of an optimal control limit policy is necessary. However, in our numerical experiments, optimal control limit policies still exist if ω/R

is sufficiently small and the treatment efficacy f is sufficiently large. As we shall see later in Definition III.22, a sufficiently small ω/R quotient requires the treatments to be sufficiently efficient.

The next proposition shows that reducing the readmission risk reduces the follow-up cost. Moreover, under a mild condition, the relationship is concave, implying that reducing readmission risk generates a larger saving as the risk gets closer to zero.

Proposition III.9 (Follow-Up Cost Concave Non-Decreasing in Readmission Risk)**.** *If Assumption III.6 (stationary readmission risk) holds, then the follow-up cost $V_0^{\pi}(\rho)$ is non-decreasing in the readmission risk ρ. Furthermore, under a mild condition (supermodularity in the health state and the readmission risk $(s,\rho) \in \mathcal{S} \times [0,1]$), the follow-up cost is concave in ρ.*

Proof. Proof of Proposition III.9. We prove by induction. The base case $t = T$ holds trivially. Suppose in period $t+1$, $V_{t+1}^*(s), s \in \{H, S\}$ is non-decreasing in ρ.

Recall the Bellman equations in t:

$$V_t^*(H) = \min \begin{cases} \omega + V_{t+1}^*(H) & (a_t = Treat) \\ (1-\rho)V_{t+1}^*(H) + \rho(1-d)V_{t+1}^*(S) + \rho dR & (a_t = Wait) \end{cases} \tag{3.13}$$

$$V_t^*(S) = \min \begin{cases} \omega + fV_{t+1}^*(H) + (1-f)R & (a_t = Treat) \\ (1-d)V_{t+1}^*(S) + dR & (a_t = Wait) \end{cases} \tag{3.14}$$

It follows trivially that $V_t(H, a_t = Treat) = \omega + V_{t+1}^*(H)$, $V_t(S, a_t = Treat) = \omega + fV_{t+1}^*(H) + (1-f)R$, and $V_t(S, a_t = Wait) = (1-d)V_{t+1}^*(S) + dR$ are non-decreasing in ρ. Since taking the minimum preserves monotonicity, it suffices to show that $V_t(H, a_t = Wait) = (1-\rho)V_{t+1}^*(H) + \rho(1-d)V_{t+1}^*(S) + \rho dR$ is non-decreasing in ρ. To see the non-increasing monotonicity, we shall look at the first order derivative. Let us simplify the notation, let $V_H(\rho) := V_{t+1}^{\pi^*}(H,\rho)$ and $V_S(\rho) := V_{t+1}^{\pi^*}(S,\rho)$. Following from Lemma III.7, we have:

$$\frac{\mathrm{d}}{\mathrm{d}\rho}\left((1-\rho)V_H(\rho) + \rho(1-d)V_S(\rho) + \rho dR\right) \tag{3.15}$$

$$= d(R - V_s(\rho)) + (V_S(\rho) - V_H(\rho)) + (1-\rho)V_H'(\rho) + (1-d)\rho V_S'(\rho) \geq 0 \tag{3.16}$$

To show the concavity, consider any $0 \leq \rho_1 \leq \rho_2 \leq 1$ and $\alpha \in [0,1]$. We shall verify the following inequality:

$$\begin{aligned} &(1 - \alpha\rho_1 - (1-\alpha)\rho_2)V_H(\alpha\rho_1 + (1-\alpha)\rho_2) + (\alpha\rho_1 + (1-\alpha)\rho_2)(1-d)V_S(\alpha\rho_1 + (1-\alpha)\rho_2) \\ \geq &\alpha(1-\rho_1)V_H(\rho_1) + (1-\alpha)(1-\rho_2)V_H(\rho_2) + \alpha\rho_1(1-d)V_S(\rho_1) + (1-\alpha)\rho_2(1-d)V_S(\rho_2) \end{aligned} \tag{3.17}$$

Applying the induction hypothesis, Eq. (3.17) can be rewritten as

$$\alpha(1-\alpha)(\rho_2-\rho_1)[V_H(\rho_2)-V_H(\rho_1)+(1-d)(V_S(\rho_1)-V_S(\rho_2))] \geq 0 \tag{3.18}$$

This follows as $V_s(\rho) = V_t^\pi(s,\rho)$ is assumed to be supermodular in $\{s,\rho\} \in \mathcal{S} \times [0,1]$ for all $t \in \{0,1,...,T\}$.

The supermodularity assumption states that the benefit of reducing the readmission risk is greater on a healthy patient than on a sick patient. Note that the supermodularity does not hold in general. For example, if $d = 0$, then supermodularity is violated. However, it holds for large enough R and d.

□

It is intuitive that the cost of follow-up is non-decreasing in the readmission risk ρ since lower risk patients require fewer treatments and are less likely to be readmitted. For the cost of follow-up to be concave in ρ, the Bellman equations are required to be supermodular in the health state and the readmission risk $(s,\rho) \in \mathcal{S} \times [0,1]$. Supermodularity here means that the marginal benefit of having a lower readmission risk, ρ, in the healthy state is greater than the marginal benefit for a patient in the sick state. This condition is intuitive, since risk of getting sick is less important to future cost when a patient is already sick, because there is a chance that she will never return to the healthy state and thus the risk of getting sick becomes irrelevant. In contrast, for a healthy patient, this risk is more relevant to future outcomes. Note that supermodularity does not necessarily hold. For instance, in the pathological case where the delay-time failure rate d is zero, supermodularity does not hold. However, for sufficiently large R and d, which is the most likely case in practice, supermodularity should hold.

Next, we numerically verify concavity and demonstrate that even when the function is not technically concave, it still resembles a concave shape. To do this, we compute the second derivative of $V(\cdot)$ numerically.

We tested the following 144 combinations of parameters: $T \in \{7, 14, 30, 90\}$, $\omega/R \in \{0.01, 0.2, 5, 10\}, d \in \{0.1, 0.3, 0.5\}, f \in \{0.5, 0.75, 1\}$. Out of the 144 cases, 97 cases were found to be concave as desired. The rest 47 cases were non-concave. Figure 3.3 shows the 47 non-concave cases. The tuple above each subplot denotes $(T, f, d, \omega/R)$. The horizontal axis is the readmission risk ρ and the vertical axis is the cost $V(\rho)$. In these nonconcave cases, $V(\rho)$ still behaves very much like a concave function.

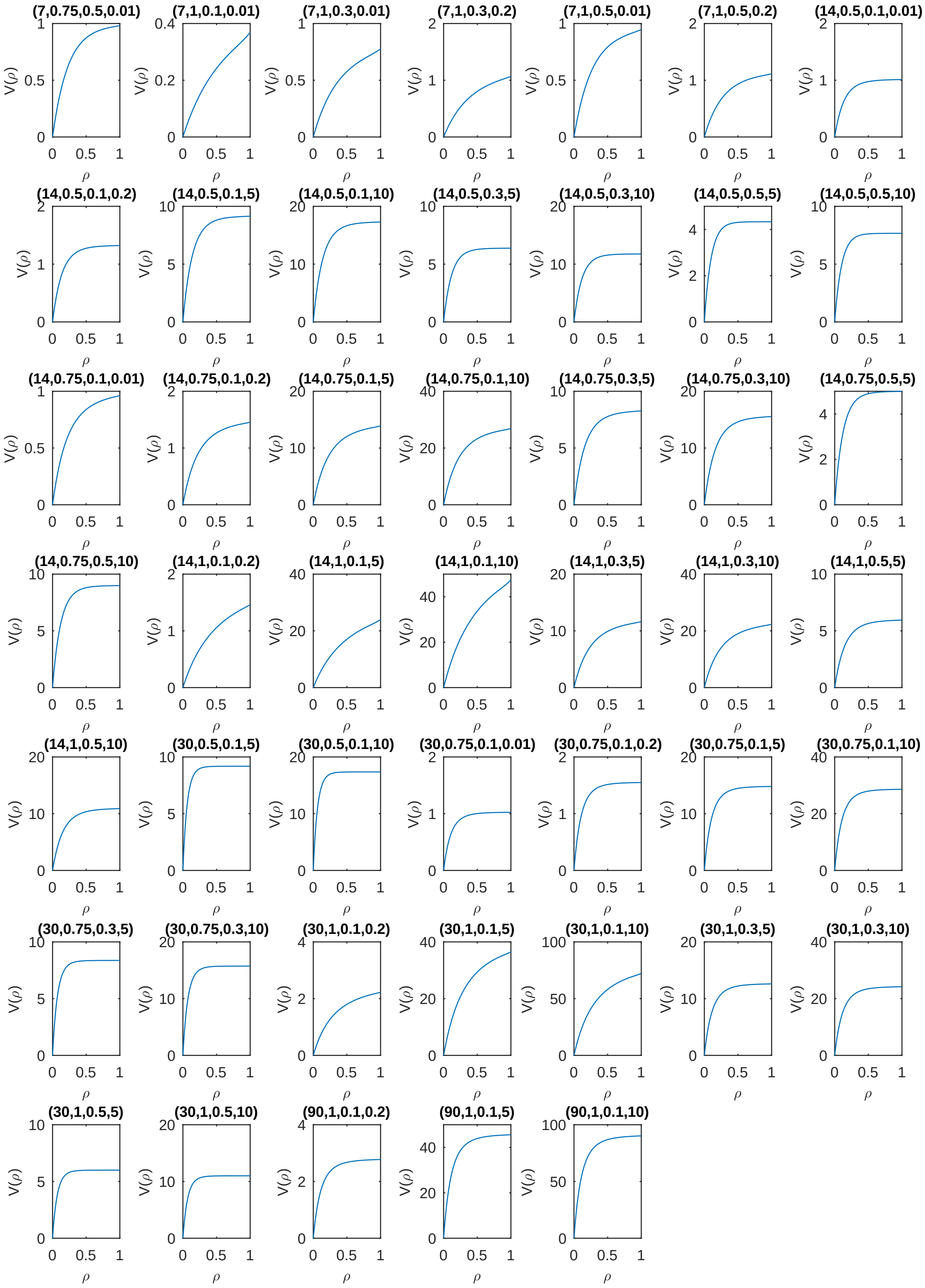


Figure 3.3: Examples of Non-concave V

3.4.3 Engaged and Disengaged Follow-Up Policies

In this section, we first identify two types of follow-up policies – the engaged policy and the disengaged policy (Proposition III.10). We show that disengaged policies are expensive and undesirable (Proposition III.11) and that the engaged policy is optimal when patients are at low risk and the treatments are effective and efficient (Proposition III.12). We argue that an effective BP policy should entice hospitals to adopt an engaged policy.

First, we identify two types of follow-up policies adopted by a hospital to minimize the expected follow-up cost. Under Assumption III.5 (following-up every day is expensive), a hospital will adopt either one of the following two policies.

Proposition III.10 (Engaged and Disengaged Follow-Up Policies)**.** *Suppose Assumption III.5 (following-up every day is expensive) holds, either one of the following two types of policies is optimal:*

- Engaged Follow-Up Policy: *a policy π_E is said to be an* engaged policy *if* $\forall t \in \{0, 1, \ldots, T-1\}$, $\pi_E(S,t) = Treat, \pi_E(H,t) = Wait, \forall t \in \{0,1,...,T-1\}$.
- Disengaged Follow-Up Policy: *a policy π_{DE} is said to be a* disengaged policy *if in some period $t \in \{0, 1, \ldots, T-1\}$, the action is $\pi_{DE}(H,t) = \pi_{DE}(S,t) = Wait$.*

Proof. Proof of Proposition III.10. Since $T\omega > R$, it is never optimal to treat the patient every single day. Hence, by definition, a hospital adopts either the engaged policy or a disengaged policy. □

Adopting an engaged follow-up policy, a hospital will provide treatment to a patient if and only if the patient is sick. This means that the hospital is responsive and fully engaged in providing timely post-discharge care. It also means that medical resources are not "wasted" on treating a healthy patient. In contrast, under a disengaged follow-up policy, a hospital is not fully engaged because it may prefer to wait and let a sick patient get readmitted rather than providing treatments to avert the readmission.

It is tempting to wonder why a hospital would ever adopt a disengaged policy to minimize the follow-up cost. Intuitively, one may conjecture that a hospital may adopt such a disengaged policy due to the following reasons. 1) The patient is at very high risk of readmission. As such, an engaged policy requires the hospital to provide intensive post-discharge interventions, which may be more costly than a readmission (this case was found to exist in the modeling efforts of Zhang et al. (2016) and found empirically by Desai et al. (2016)). 2) Treatments that can avert a readmission may be very expensive. In such cases, the hospital would rather readmit the patient. 3) Post-discharge treatment may be little effective in averting readmissions (f is very small). 4) Readmissions are very cheap (or not penalized sufficiently), therefore the hospital cares little about preventing readmissions. 5) The bundled payment episode length is too long, requiring too much effort. A hospital may give up on preventing readmissions for such a long time interval. As we shall see later in Section 3.6, our analytical results confirm our intuition.

As a policymaker, the CMS shall design a BP policy to eliminate the financial incentives for disengaged policies. Although it is not illegal for hospitals to deny non-emergent patients treatment (Adida et al., 2016), it is arguably irresponsible and unethical to do so. Moreover, from the provider's perspective, we argue that disengaged policies are also undesirable. The next proposition shows that a disengaged policy is undesirable due to its excessive cost.

Proposition III.11 (Disengaged Policies are Expensive)**.** *Suppose Assumption III.4 (treatments are effective) holds, if a disengaged policy is ever optimal, then over the T-day follow-up period, the hospitals incurs a follow-up cost $V_0^*(H) = V_0^{\pi_{DE}}(H) \geq R - \dfrac{\omega}{d-(1-f)}$, which equates to the cost of a readmission less a multiple of the cost of a treatment.*

Proof. Proof of Proposition III.11. Since in period t, the optimal action is to wait when a patient is sick, we have

$$(1-d)V_{t+1}^*(S) + dR \leq \omega + fV_{t+1}^*(H) + (1-f)R \tag{3.19}$$

By Lemma III.7, $V_{t+1}^*(H) \leq V_{t+1}^*(S)$ implies

$$(1-d)V_{t+1}^*(H) + dR \leq \omega + fV_{t+1}^*(H) + (1-f)R \tag{3.20}$$

$$V_{t+1}^*(H) \geq R - \frac{\omega}{f+d-1} \geq R - \frac{\omega}{f+d-1} \tag{3.21}$$

By Lemma III.7, we have $V_{t-1}^*(S) \geq V_{t-1}^*(H) \geq R - \frac{\omega}{f+d-1}$. Since all immediate costs are nonnegative, it follows that $V_0^*(S) \geq V_0^*(H) \geq R - \frac{\omega}{f+d-1}$ holds. □

Note that if $1-f \ll d$ (i.e., treatments are highly effective), then the disengaged policy follow-up cost becomes as expensive as the cost of a readmission. In this case, since treatments are very effective, not providing them (under a disengaged policy) will result in a very high cost. For instance, if treatments are perfect ($f = 1$) and the delay-time failure rate $d = 0.3$ (representing the baseline estimate, see Section

3.8.1), then the cost of follow-up is greater than $R - 3.3\omega$, which equates to the cost of a readmission less the cost of 3.3 treatments.

As disengaged policies are shown to be undesirable from both the policymaker's and the provider's perspectives, we now shift our focus to the engaged policy. The next proposition shows that a hospital will adopt an optimal engaged policy if 1) the patient is at very low risk of readmission, 2) the treatments are sufficiently effective (in this case a sufficient condition is $f = 1$), and 3) treatments are sufficiently efficient[5] ($\omega/R \leq d$, and this holds in our numerical studies, see Section 3.8.1).

Proposition III.12 (Engaged Policy Optimality)**.** *Suppose Assumption III.6 (stationary readmission risk) holds, if $\omega/R \leq d$ (this requires treatments to be sufficiently efficient).*[5] *There exists $f_E \geq 0$ and $\rho_E \geq 0$ such that for all $f \geq f_E$ and $\rho \leq \rho_E$, an engaged policy is optimal.*

Proof. Proof of Proposition III.12. Suppose $0 \leq \rho \leq \rho_E = 0$. Let $f = 1 \geq f_E = 1$. The Bellman equations become:

$$V_t^*(RA) = R \tag{3.22}$$

$$V_t^*(H) = \min \begin{cases} \omega + V_{t+1}^*(H) & (a_t = Treat) \\ V_{t+1}^*(H) & (a_t = Wait) \end{cases} \tag{3.23}$$

$$V_t^*(S) = \min \begin{cases} \omega + V_{t+1}^*(H) & (a_t = Treat) \\ (1-d)V_{t+1}^*(S) + dR & (a_t = Wait) \end{cases} \tag{3.24}$$

It follows that wait is the optimal action in the healthy state and $V_t^*(H) = 0, \forall t \in \{0, 1, ..., T-1\}$. So $V_t^*(S) = \min \begin{cases} \omega & (a_t = Treat) \\ (1-d)V_{t+1}^*(S) + dR & (a_t = Wait) \end{cases}$.

Since $\omega < dR$, the optimal action in the sick state is to treat the patient.

□

As the policymaker, we shall entice a hospital to adopt an engaged policy because it implies that the hospital is being responsive and engaged in providing post-discharge care.

[5] The quotient ω/R is defined as the treatment inefficiency in Definition III.22. Therefore, a smaller ω/R value implies higher efficiency.

3.4.4 Closed-Form Engaged Policy Follow-Up Cost

In this section, we develop crucial structural properties when an engaged follow-up policy is adopted by the hospital. We show that under an engaged policy, the follow-up cost can be closed-form expressed in terms of the MDP's costs, transition probabilities, and the episode length T (Theorem III.13). This closed-form expression reveals that reducing readmission risk generates savings (asymptotically) linear in the readmission risk (Corollary III.14). Moreover, Corollary III.14 uncovers the relationship between the follow-up cost and the policy-level levers (T, ω, R).

Theorem III.13 (Engaged Policy Follow-Up Cost in Closed-Form)**.** *Suppose Assumption III.6 (stationary readmission risk) holds. Given an engaged follow-up policy π_E, the cost of following-up with a patient (who is discharged healthy initially) for T days is*

$$V(\rho) := V_0^{\pi_E}(H, \rho) = \alpha_1 z_1^T + \alpha_2 z_2^T + \beta, \tag{3.25}$$

where $\alpha_1, \alpha_2, z_1, z_2$, and β are functions of $\rho, f, d, R,$ and ω. Note that ^{T} is the exponent not to be confused with the transpose operator in linear algebra. The closed-form expressions for these terms are defined in Eq. (3.27) – (3.30).

The function $V(\rho)$ is strictly increasing in ρ (following from the proof of Proposition III.9), though $V(\rho)$ is not necessarily concave.[6] *Under mild condition (supermodularity), $V(\rho)$ is concave (by Proposition III.9).*

Proof. Proof of Theorem III.13. We first list the variables:

$$\alpha_1 = \frac{(d-1)\omega\left(\sqrt{(1-\rho)^2+4\rho f(1-d)}-\rho-1\right) - R((d-1)f+1)\left(\sqrt{(1-\rho)^2+4\rho f(1-d)}+2d\rho-\rho-1\right)}{2((d-1)f+1)\sqrt{(1-\rho)^2+4\rho f(1-d)}} \tag{3.26}$$

$$\alpha_2 = \frac{(d-1)\omega\left(\sqrt{(1-\rho)^2+4\rho f(1-d)}+\rho+1\right) - R((d-1)f+1)\left(\sqrt{(1-\rho)^2+4\rho f(1-d)}-2d\rho+\rho+1\right)}{2((d-1)f+1)\sqrt{(1-\rho)^2+4\rho f(1-d)}} \tag{3.27}$$

$$z_1 = \frac{(1-\rho)-\sqrt{(1-\rho)^2+4\rho f(1-d)}}{2} \tag{3.28}$$

$$z_2 = \frac{(1-\rho)+\sqrt{(1-\rho)^2+4\rho f(1-d)}}{2} \tag{3.29}$$

$$\beta = \frac{(1-d)\omega + dR + (1-d)(1-f)R}{1-f+fd} \tag{3.30}$$

Since an engaged policy is implemented, we can rewrite the Bellman equations as follows:

[6] For example, let $T = 3, d = 0.4, \omega = 1, R = 2.9$, and $f = 1$, then $V''(0.99) = 0.12 > 0$ implies the function is not concave. However, we observe that, for large T and f, $V(\rho)$ is concave. Since $V(\rho)$ is obtained by solving a recurrence relation, it is difficult to derive a tractable and meaningful condition to ensure concavity. Nevertheless, one can numerically compute the second order derivative of $V(\rho)$ using the closed-form expression (Eq. (3.25)) to verify concavity.

If $t \in \{0, 1, ..., T-1\}$:

$$V_t^{\pi_E}(H) = (1-\rho)V_{t+1}^{\pi_E}(H) + \rho(1-d)V_{t+1}^{\pi_E}(S) + \rho dR \tag{3.31}$$

$$V_t^{\pi_E}(S) = \omega + fV_{t+1}^{\pi_E}(H) + (1-f)R \tag{3.32}$$

$$V_t^{\pi_E}(RA) = R \tag{3.33}$$

If $t = T$:

$$V_T^{\pi_E}(H) = 0 \tag{3.34}$$

$$V_T^{\pi_E}(S) = 0 \tag{3.35}$$

$$V_T^{\pi_E}(RA) = R \tag{3.36}$$

The Bellman equation in the healthy state can be expressed in the following recurrence relation:

$$V_t^{\pi_E}(H) = (1-\rho)V_{t+1}^{\pi_E}(H) + \rho(1-d)(\omega + fV_{t+2}^{\pi_E}(H) + (1-f)R) + \rho dR \tag{3.37}$$

To simplify our notation, define $V_t = V_{T-t}^{\pi_E}(H)$. Then we have

$$V_{t+2} = \begin{cases} (1-\rho)V_{t+1} + \rho(1-d)(\omega + fV_t + (1-f)R) + \rho dR & \text{if } t \in \{0, \ldots, T-2\} \\ 0 & \text{if } t = 0 \\ \rho dR & \text{if } t = 1 \end{cases} \tag{3.38}$$

We can solve this nonhomogeneous recurrence. First we solve for a homogeneous solution V_t^H. The characteristic equation is

$$z^2 - (1-\rho)z - \rho f(1-d) = 0 \tag{3.39}$$

The two roots are

$$z_1 = \frac{(1-\rho) - \sqrt{(1-\rho)^2 + 4\rho f(1-d)}}{2} \tag{3.40}$$

$$z_2 = \frac{(1-\rho) + \sqrt{(1-\rho)^2 + 4\rho f(1-d)}}{2} \tag{3.41}$$

So the homogeneous solution is

$$V_t^H = z_1^t \alpha_1 + z_2^t \alpha_2 \tag{3.42}$$

where α_1 and α_2 are constants that are determined later. For the particular solution, we guess a constant form $V_t^P = \beta$. Plugging in to the recurrence, $\beta - (1-\rho)\beta - \rho f(1-d)\beta = \rho(1-d)\omega + \rho d R + \rho(1-d)(1-f)R$. The solution is $V_t^P = \beta = \frac{(1-d)\omega + dR + (1-d)(1-f)R}{1-f+fd}$.

Now we can plug in the two boundary conditions to solve for α_1 and α_2:

$$V_0 = V_0^H + \beta = 0 \tag{3.43}$$

$$V_1 = V_1^H + \beta = \rho d R \tag{3.44}$$

The solutions are

$$\alpha_1 = \frac{(d-1)\omega\left(\sqrt{(1-\rho)^2 + 4\rho f(1-d)} - \rho - 1\right) - R((d-1)f+1)\left(\sqrt{(1-\rho)^2 + 4\rho f(1-d)} + 2d\rho - \rho - 1\right)}{2((d-1)f+1)\sqrt{(1-\rho)^2 + 4\rho f(1-d)}} \tag{3.45}$$

$$\alpha_2 = \frac{(d-1)\omega\left(\sqrt{(1-\rho)^2 + 4\rho f(1-d)} + \rho + 1\right) - R((d-1)f+1)\left(\sqrt{(1-\rho)^2 + 4\rho f(1-d)} - 2d\rho + \rho + 1\right)}{2((d-1)f+1)\sqrt{(1-\rho)^2 + 4\rho f(1-d)}} \tag{3.46}$$

□

Computing the follow-up cost in closed-form, instead of using backward induction, enables us to derive powerful analytical insights into the operational impact of bundled payment design features. Corollary III.14 reveals that if a hospital manages to reduce the readmission risk to a low level ($\rho \to 0$), then readmission reduction will generate savings linearly in the readmission risk. The marginal benefit is a function of the MDP's costs, transition probabilities, and the episode length.

Corollary III.14 (Marginal Benefit of Reducing Readmission Risk)**.** *For an engaged policy π_E with Bellman equation $V(\rho) := V_0^{\pi_E}(H, \rho)$, the marginal benefit of reducing the readmission risk is asymptotically linear for sufficiently small ρ. Formally, we have*

$$\lim_{\rho \to 0^+} V'(\rho) := \lim_{\rho \to 0^+} V_0^{\pi_E}(H, \rho) = \{d + [1 - (1-d)\, f]\,(T-1)\}\, R + (T-1)\,(1-d)\,\omega. \tag{3.47}$$

Proof. Proof of Corollary III.14. For the purpose of exposition, we prove for the case $f = 1$. For $f < 1$, the idea is similar.

Let $f = 1$: for notational convenience, we define the following function:

$$s(\rho) = \sqrt{(1-\rho^2) + 4\rho(1-d)} \tag{3.48}$$

$$k = -\frac{(1-d)\omega}{d} - R \tag{3.49}$$

$$g = 2Rd + k \tag{3.50}$$

Note that as $\rho \to 0^+$, we have

$$\lim_{\rho \to 0^+} s(\rho) = 1 \tag{3.51}$$

$$\lim_{\rho \to 0^+} s'(\rho) = \lim_{\rho \to 0^+} \frac{1+\rho-2d}{s(\rho)} = 1 - 2d \tag{3.52}$$

Then we can rewrite α_1, α_2, z_1, and z_2 as follows:

$$\alpha_1 = \frac{1}{2}\left(k - \frac{g\rho + k}{s(\rho)}\right), \alpha_2 = \frac{1}{2}\left(k + \frac{g\rho + k}{s(\rho)}\right) \tag{3.53}$$

$$z_1 = \frac{1-\rho-s(\rho)}{2}, z_2 = \frac{1-\rho+s(\rho)}{2} \tag{3.54}$$

Hence, we rewrite $V(\rho)$ as

$$V(\rho) = \frac{1}{2}\left(k - \frac{g\rho + k}{s(\rho)}\right)\left(\frac{1-\rho-s(\rho)}{2}\right)^T + \frac{1}{2}\left(k + \frac{g\rho + k}{s(\rho)}\right)\left(\frac{1-\rho+s(\rho)}{2}\right)^T - k \tag{3.55}$$

Take the derivative with respect to ρ, we have

$$V'(\rho) = \tag{3.56}$$

$$\begin{aligned} &-\frac{T}{4}\left(k - \frac{g\rho + k}{s(\rho)}\right)\left(\frac{1-\rho-s(\rho)}{2}\right)^{T-1}(1+s'(\rho)) + \frac{1}{2}\left(\frac{1-\rho-s(\rho)}{2}\right)^T\left(\frac{-gs(\rho)+(g\rho+k)s'(\rho)}{s(\rho)^2}\right) \\ &-\frac{T}{4}\left(k + \frac{g\rho + k}{s(\rho)}\right)\left(\frac{1-\rho+s(\rho)}{2}\right)^{T-1}(1-s'(\rho)) + \frac{1}{2}\left(\frac{1-\rho+s(\rho)}{2}\right)^T\left(\frac{gs(\rho)-(g\rho+k)s'(\rho)}{s(\rho)^2}\right) \end{aligned} \tag{3.57}$$

Taking the limit, we have

$$\begin{aligned} \lim_{\rho \to 0^+} V'(\rho) = &-\frac{T}{4}\left(k - \frac{0+k}{1}\right)\left(\frac{1-0-1}{2}\right)^{T-1}(1+s'(\rho)) + \frac{1}{2}\left(\frac{1-0-1}{2}\right)^T\left(\frac{-g+(0+k)s'(\rho)}{1^2}\right) \\ &-\frac{T}{4}\left(k + \frac{0+k}{1}\right)\left(\frac{1-0+1}{2}\right)^{T-1}(1-s'(\rho)) + \frac{1}{2}\left(\frac{1-0+1}{2}\right)^T\left(\frac{g-(0+k)s'(\rho)}{1^2}\right) \end{aligned} \tag{3.58}$$

$$=0 + 0 - \frac{Tk}{2}(1-s'(\rho)) + \frac{g-ks'(\rho)}{2} \tag{3.59}$$

$$=dRT + (T-1)(1-d)\omega \tag{3.60}$$

For $f < 1$, the idea behind the proof is the same. We omit the proof since the proof is too lengthy for exposition.

$\square$

3.4.5 Approximation of the Post-Discharge Cost

So far we have shown some very nice properties of the engaged follow-up policy. However, in reality, a hospital may not always adopt the engaged policy. In this section, we shall allow a hospital to adopt either the engaged policy or a disengaged policy, whichever is cheaper. We show that we can still leverage the closed-form engaged policy cost $V(\rho)$ to construct a piece-wise closed-form (PWCF) expression $C_M(\rho)$ that approximates the optimal post-discharge cost well (Eq. (3.63)). Even when the readmission risk is nonstationary, we can still use the PWCF expression to approximate the follow-up cost with a proven accuracy bound (Theorem III.16).

To construct the approximation, let us first compare the engaged policy follow-up cost $V(\rho)$ (given by Eq. (3.25)) with the optimal dynamic programming (DP) policy cost $V^*(\rho)$ (obtained by solving the Bellman equations). Under Assumption III.6 (stationary readmission risk), Figure 3.4 shows that the engaged policy is optimal in the light grey area, when the patient's readmission risk is below ρ_E (confirming Proposition III.12). Moreover, we see that the curve resembles a linear relationship as $\rho \to 0$, which confirms Corollary III.14. In this light grey area, the optimal DP policy cost (the dashed line) coincides with the engaged policy cost (the solid line) because the engaged policy is in fact optimal. However, as the readmission risk increases into the dark grey area, a disengaged policy is optimal.

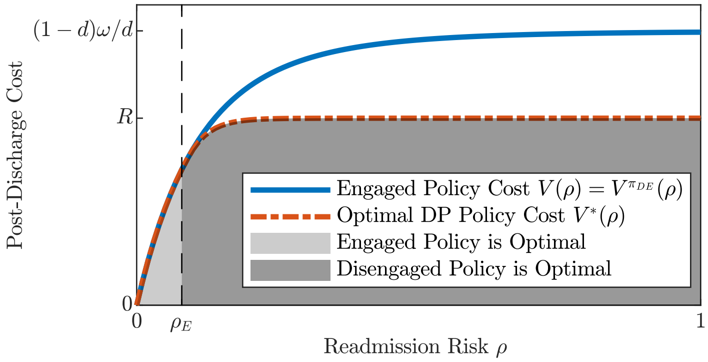


Figure 3.4: Engaged Policy Cost and Optimal DP Policy Cost ($T = 30, d = 0.3, f = 1$, $\omega = \$1,000, R = \$5,000$)

In the dark grey area, we also observe that the engaged policy cost can be more expensive than a readmission. For instance, in the extreme (pathological) case where

$\rho = 1$ and $f = 1$, we have $V(1) = R + (1-d)\omega/d \geq R$. This may seem counterintuitive because a trivial worst-case cost for the hospital is to provide no treatments at all in the post-discharge stage and get a readmission. This worst case costs at most R, which is lower than the engaged policy cost $R + (1-d)\omega/d \geq R$. In fact, the engaged policy may be suboptimal when the readmission risk is very high. In such cases, a patient being very likely to develop complications will require intensive and frequent post-discharge treatments. As such, the required intensive care can be more costly than a readmission, which makes a readmission more economical for the hospital. When $f = 1$, the difference between the engaged policy and the optimal dynamic programming policy is $(1-d)\omega/d$. As the delay-time failure rate d decreases and the cost of a treatment ω increases, the engaged policy performs worse since it "wastes" expensive treatments on a patient who is unlikely to be readmitted due to low failure rate.

Since the engaged policy might be more expensive than a readmission, we shall allow a hospital to switch to a disengaged policy if the engaged policy cost exceeds the cost of a readmission ($V(\rho)= V_0^{\pi_E}(H, \rho) \geq R$). As such, we define the piece-wise closed-form (PWCF) follow-up cost approximation function as

$$C_M(\rho) = \min\{V(\rho), R\}. \tag{3.61}$$

Since $V(\rho)$ is non-decreasing in ρ (Theorem III.13), we can define $\bar{\rho}$, the threshold below which an engaged policy is cheaper than the cost of a readmission, as

$$\bar{\rho} = \begin{cases} \arg\min_\rho V(\rho) \geq R & \text{if } \exists \rho \in [0,1] : V(\rho) \geq R \\ 1 & \text{otherwise} \end{cases} \tag{3.62}$$

Remark III.15 (Post-Discharge "Maintaining" Cost Function)*.* The post-discharge "Maintaining" cost is expressed as a PWCF function:

$$C_M(\rho) = \begin{cases} V(\rho) & \text{if } \rho \in [0, \bar{\rho}] \\ R & \text{if } \rho \in (\bar{\rho}, 1] \end{cases} \tag{3.63}$$

Next, we shall show that the PWCF function can approximate the optimal DP cost well. When the readmission risk is nonstationary, it can still be used to approximate the follow-up cost with a proven accuracy bound.

Theorem III.16 (PWCF Approximation is Δ-Accurate)**.** *Suppose Assumption III.4 (treatments are effective) holds. For a patient whose readmission risk ρ_t is non-stationary, let $V_0^{\pi^*}(H,\vec{\rho})$ denote the optimal DP policy cost of follow-up. We can approximate the cost of follow-up using $C_M(\rho)$ such that $|C_M(\rho) - V_0^{\pi^*}(H,\vec{\rho})| \leq \Delta$. The accuracy $\Delta = \max\left\{\dfrac{\omega}{f+d-1}, \sum_{t=0}^{T-1}\epsilon_t R\right\}$, where $\epsilon_t = |\rho_t - \rho|$ for all $t \in \{0,1,\ldots,T-1\}$. To achieve best accuracy, one can compute ρ using the Least Absolute Deviation (LAD) estimates of ρ_t, which is the median over the time index.*

Proof. Proof of Theorem III.16. To prove this, we first prove a lemma.

Lemma III.17 (Non-Stationary Readmission Risk Approximation)**.** *Suppose the true readmission risk $\vec{\rho} = (\rho_0,\ldots,\rho_T)$ is non-stationary. The engaged policy follow-up cost $V_0^{\pi_E}(H,\vec{\rho})$ can be approximated using $V(\rho)$ (which assumes stationary readmission risk ρ). The accuracy is given by $|V_0^{\pi_E}(H,\vec{\rho}) - V(\rho)| \leq \sum_{t=0}^{T-1}\epsilon_t R$, where $\epsilon_t = |\rho_t - \rho|$ for all $t \in \{0,1,\ldots,T-1\}$.*

Proof. Proof of Lemma III.17. To simplify the notation, let the engaged policy cost of follow-up for the **non-stationary** readmission risk model is $V_t^{\rho_t}(s) := V_t^{\pi_E}(s,\rho_t)$. Let the engaged policy cost of follow-up for the **stationary** readmission risk model is $V_t^{\rho}(s) := V_t^{\pi_E}(s,\rho)$. Define the loss in each period $g_t = \max_{s\in\mathcal{S}}|G_t(s)| = \max_{s\in\mathcal{S}}|V_t^{\rho}(s) - V_t^{\rho_t}(s)|$.

We prove by induction. Given $g_{t+1}, t \in \{1,\ldots,T\}$, we shall establish bounds for g_t. We focus on $G_t(S)$ first. Since the follow-up policy is to wait on sick patients, we have

$$G_t(S) = V_t^{\rho}(S) - V_t^{\rho_t}(S) \tag{3.64}$$

$$= \omega + fV_{t+1}^{\rho}(H) + (1-f)R - (\omega + fV_{t+1}^{\rho_t}(H) + (1-f)R) \tag{3.65}$$

$$= fG_{t+1}(H) \tag{3.66}$$

By the induction hypothesis, we have

$$-g_{t+1} \leq -fg_{t+1} \leq G_t(S) \leq fg_{t+1} \leq g_{t+1} \tag{3.67}$$

Next, we derive the bounds for the healthy state, in which the optimal action is to wait:

$$\begin{aligned} G_t(H) =& V_t^{\rho}(H) - V_t^{\rho_t}(H) && (3.68)\\ =&(1-\rho)V_{t+1}^{\rho}(H) + \rho(1-d)V_{t+1}^{\rho}(S) + \rho dR \\ & - (1-\rho_t)V_{t+1}^{\rho_t}(H) - \rho_t(1-d)V_{t+1}^{\rho_t}(S) - \rho_t dR && (3.69)\\ =&(1-\rho)G_{t+1}(H) + \rho G_{t+1}(S) + (\rho-\rho_t)(dR + (1-d)V_{t+1}^{\rho_t}(S) - V_{t+1}^{\rho_t}(H)) && (3.70) \end{aligned}$$

To obtain an upper bound, we construct a linear program (LP_H^{UB}):

$$\begin{aligned} G_H^{UB} = \max_{\rho_t, G_H, G_S, V_H, V_S} \quad & (1-\rho)G_H + \rho G_S + (\rho-\rho_t)(dR + (1-d)V_S - V_H) && (3.71)\\ \text{s.t.} \quad & \rho - \epsilon_t \le \rho_t \le \rho + \epsilon_t && (3.72)\\ & 0 \le V_H \le V_S \le R && (3.73)\\ & -g_{t+1} \le G_S, G_H \le g_{t+1} && (3.74)\\ & 0 \le \rho_t \le 1 && (3.75) \end{aligned}$$

This linear program attains its maximum $G_H^{UB} = g_{t+1} + \epsilon_t R$ when $\rho_t = \rho - \epsilon_t, V_S = R, V_H = 0$, and $G_S = G_H = g_{t+1}$.

To obtain a lower bound, we construct a linear program (LP_H^{LB}):

$$\begin{aligned} G_H^{LB} = \min_{\rho_t, G_H, G_S, V_H, V_S} \quad & (1-\rho)G_H + \rho G_S + (\rho-\rho_t)(dR + (1-d)V_S - V_H) && (3.76)\\ \text{s.t.} \quad & \rho - \epsilon_t \le \rho_t \le \rho + \epsilon_t && (3.77)\\ & 0 \le V_H \le V_S \le R && (3.78)\\ & -g_{t+1} \le G_S, G_H \le g_{t+1} && (3.79)\\ & 0 \le \rho_t \le 1 && (3.80) \end{aligned}$$

This linear program attains its maximum $G_H^{LB} = -g_{t+1} - \epsilon_t R$ when $\rho_t = \rho + \epsilon_t, V_S = R, V_H = 0$, and $G_S = G_H = -g_{t+1}$. Hence, $-g_{t+1} - \epsilon(t)R \leq G_t(H) \leq g_{t+1} + \epsilon_t R$. Thus, $g_t \leq g_{t+1} + \epsilon_t R$.

Observe that $G_T(S) = G_T(H) = 0$. Hence, it follows that $g_0 \leq \sum_{t=0}^{T-1} \epsilon_t R$ inductively. □

Now we leverage this lemma to prove the theorem.

Consider a T-period problem with non-stationary readmission risk $\vec{\rho}$. The state space $\vec{\rho} = (\rho_0, \rho_1, ..., \rho_{T-1}) \in [0,1]^T$ can be partitioned into two mutually exclusive sets: F and P. For any $\vec{\rho} \in F$, an engaged policy is optimal. For any $\vec{\rho} \in P$, a disengaged policy is optimal.

Let's focus on $\pi^* \in F$ first. Let the optimal engaged policy cost of follow-up for the non-stationary readmission risk model be $V_0^{\pi^*}(H, \vec{\rho})$. Since an engaged policy is optimal, i.e., $V_0^{\pi^*}(H, \vec{\rho}) = V_0^{\pi_E}(H, \vec{\rho})$, by Lemma III.17, we have

$$-\sum_{t=0}^{T-1} \epsilon_t R \leq V_0^{\pi^*}(H, \vec{\rho}) - V(\rho) \leq \sum_{t=0}^{T-1} \epsilon_t R \tag{3.81}$$

Since by Lemma III.7, $V_0^{\pi^*}(H, \vec{\rho}) \leq R$. We can pick ρ such that $V(\rho) = C_M(\rho) \leq R$ and ensure $|C_M(\rho) - V_0^{\pi^*}(H, \vec{\rho})| \leq \Delta = \sum_{t=0}^{T-1} \epsilon_t R$.

Next, let us consider $\pi^* \in P$. Take any $\vec{\rho} \in P$, by Proposition III.11 and Lemma III.7, we have $R - \frac{\omega}{f+d-1} \leq V_0^{\pi^*}(H, \vec{\rho}) \leq R$. We can pick ρ such that $C_M(\rho) = R$ to ensure $|C_M(\rho) - V_0^{\pi^*}(H, \vec{\rho})| \leq \Delta = \frac{\omega}{f+d-1}$.

□

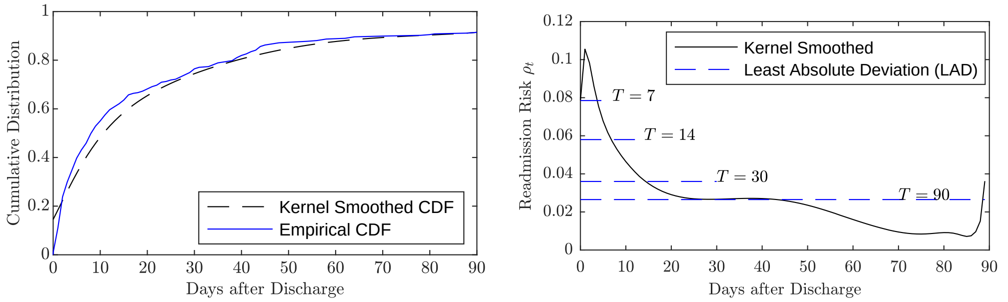
Theorem III.16 provides a theoretical accuracy bound for using the PWCF function $C_M(\rho)$ to approximate the optimal value function for a non-stationary readmission risk. This enables us to use the PWCF function to compute the cost of follow-up for patients with non-stationary readmission risks. To achieve the best accuracy, one can use the median readmission risk (over the time index) to approximate the optimal follow-up cost.

Note that in our numerical studies, the estimated ρ are small (between 0.01 and 0.08, see Table 3.2 in Section 3.8.1). So Theorem III.16 can potentially provide a reasonably tight theoretical accuracy bound. Numerically, the accuracy is observed to be reasonably tight for the purpose of policy-level decisions. For example, if $\omega/R \leq 1$ (which is likely to hold as it requires a treatment to be less expensive than a readmission), the accuracy is within 20% of the cost of a readmission (see Figure 3.6). Based on our baseline estimates (see Section 3.8.1), $20\% \times R \approx \$1,000$ is comparable to the cost of one outpatient treatment (e.g. one Primary Care Physicians (PCP) office visit and the medications and procedures administered). It is also comparable to the cost of a half day of hospitalization (Henry J Kaiser Family Foundation, 2015). When treatments are expensive (ω is large), the treatment efficacy f is low, or the delay-time failure rate d is small, the accuracy bound becomes large. Under these conditions, however, it is unlikely that a hospital would be interested in engaging in post-discharge effort since the cost is high and efficacy is low, and the patient is unlikely to be readmitted, hence these are unrealistic scenarios.

3.4.6 Analysis for Nonstationary Readmission Risk

In this section, we relax the assumption on stationary readmission risk. We fitted a kernel smoothed estimation of the readmission risk (Diehl and Stute, 1988) using the 327 patients from our partner hospital. Figure 3.5 shows the fitted and the empirical cumulative distributions as well as the readmission risk (failure rate/hazard). The readmission risk was non-stationary in this case, being noticeably high within the first two weeks after discharge. It then dropped and became steady after day-20. After day-50, the readmission risk dropped even more. Towards the end of the 90-day window, the readmission risk increased. One possible explanation for this slight increase near day-90 is the “bathtub” failure curve – increasing failure rate indicates “wear-out”. A more plausible cause of this bathtub shape was due to the censoring of our data and the finite support ($[0, 90]$) used in the kernel smoothed estimation.

For the 7-, 14-, 30-, and 90-day penalty windows, we computed the Least Absolute Deviation (LAD) estimates of the readmission risk: $\rho_{LAD}^{T=7} = 0.079, \rho_{LAD}^{T=14} = 0.059, \rho_{LAD}^{T=30} = 0.035$, and $\rho_{LAD}^{T=90} = 0.026$.



Panel A: CDF

Panel B: Readmission Risk

Figure 3.5: Fitted Kernel Smoothed Distribution (Panel A) and Least Absolute Deviation Estimates (Panel B) for the Readmission Risk

We first demonstrate that the PWCF approximation $C_M(\rho_{LAD}^T)$ is a good approximation. We normalize the readmission cost to $R = 1$ and let $d = 0.3$ and computed the difference between the LAD estimates $C_M(\rho^T_{LAD})$ and the optimal dynamic programming policy cost $(V_0^{\pi^*}(H, \vec{\rho}))$. Figure 3.6 shows the accuracy, which is defined as $|C_M(\rho^T_{LAD}) - V_0^{\pi^*}(H, \vec{\rho})|$. Since $R = 1$, the gaps can also be interpreted as a relative (percentage) gap with respect to the cost of a readmission (R). We varied the treatment inefficiency ω/R from 0.01 to 1 and plotted the accuracy versus the treatment inefficiency ω/R in a log scale. The cost-to-go using LAD approximation was within 20% of the optimal non-stationary cost for $\omega/R \leq 10^0 = 1$. This approximation error is comparable to the cost of one outpatient treatment (e.g. one PCP office visit) or half day of hospitalization (Henry J Kaiser Family Foundation, 2015). We also observed that the approximation gaps were smaller as the treatment efficacy increased.

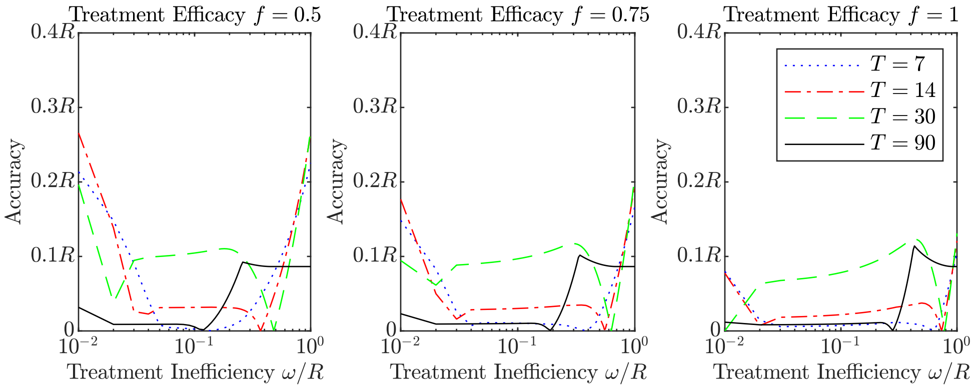


Figure 3.6: Gaps between the Optimal Dynamic Programming Cost with Non-stationary Readmission Risk and the LAD Estimations

Next, we consider reducing the non-stationary readmission risk and its effect on the cost of follow-up. Let $\vec{\rho}$ be the non-stationary readmission risk vector. We scale the readmission risk vector up/down element-wise $\rho^r_t \leftarrow \min\{\rho_t r, 1\}$ where $r > 0$ is the scaling factor. This scaling operation preserved the shape of the non-stationary risk. We varied r from a very small value $(10^{-5}/\max_t(\rho_t))$ to a very large value $(1/\min_t(\rho_t))$. As a result, the scaled readmission risk ranged from $(10^{-5}, ..., 10^{-5})$ to $(1, ..., 1)$ For each scaled readmission risk ρ^r_t, we computed the LAD estimate ρ_{LAD}.

Figure 3.7 shows that $C_M(\rho_{LAD})$ can approximate the optimal DP cost very well for $T \in \{7, 14, 30, 90\}$. Using Proposition III.21, the lower bound $\underline{\rho} = \frac{1}{d+[1-(1-d)f](T-1)+(1-d)(T-1)\omega/R}$ (plotted as dotted line) served as a reasonably tight lower bound on $\bar{\rho}$.

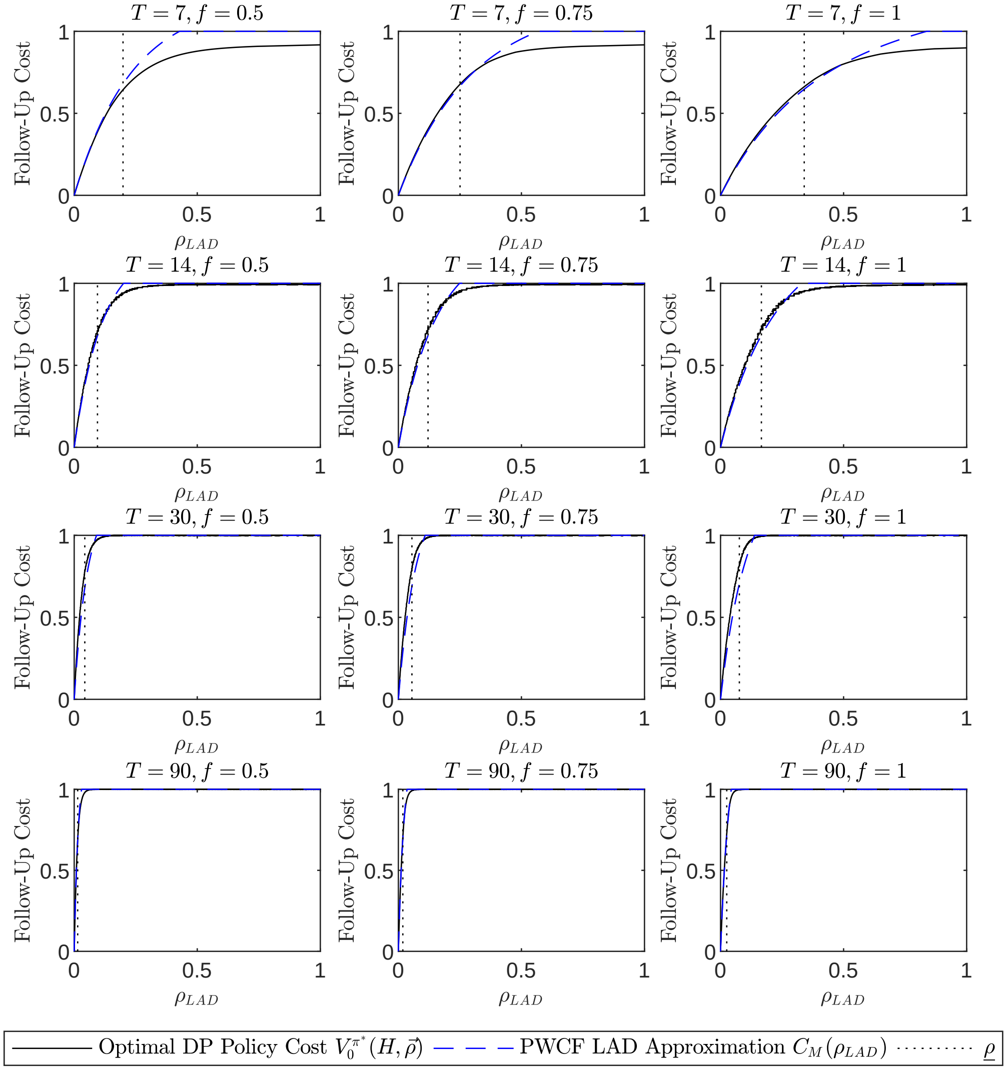


Figure 3.7: Effect of Reducing Non-stationary Readmission Risk Approximated Using LAD Estimations

3.4.7 Post-Discharge Care Engagement Threshold

Having established that the PWCF approximation $C_M(\rho) = \begin{cases} V(\rho) & \text{if } \rho \in [0, \bar{\rho}] \\ R & \text{if } \rho \in (\bar{\rho}, 1] \end{cases}$ is an accurate approximation, we now discuss the implications of the term $\bar{\rho}$.

Remark III.18 (Post-Discharge Care Engagement Threshold). The term $\bar{\rho}$ can be interpreted as an engagement threshold – for patients whose risk is at above this threshold, a hospital has no incentive to engage in exerting post-discharge readmission reduction efforts. If a hospital can manage to reduce the readmission risk to below the threshold $\bar{\rho}$, the hospital will engage in post-discharge readmission prevention and benefit from it at an (asymptotically) linear rate by Corollary III.14.

Note that $\bar{\rho}$ is uniquely determined by the MDP's costs, transition probabilities, and the episode length (ω, R, f, d, T). Since $V(\rho)$ is non-decreasing in ρ by Theorem III.13, $\bar{\rho}$ can be efficiently computed using a univariate binary search. However, a closed-form expression is not available for $\bar{\rho}$. We present Proposition III.19 to bound $\bar{\rho}$ from below using a closed-form expression. This lower bound is denoted by $\underline{\rho}$. For patients at risk below $\underline{\rho}$, a hospital would engage in post-discharge care since it could generate savings at a linear rate.

Proposition III.19. *If $V(\rho)$ is concave (see Theorem III.13, and Proposition III.9), then a lower bound for $\bar{\rho}$ is $\underline{\rho}$, such that*

$$\bar{\rho} \geq \underline{\rho} = \frac{1}{d + [1-(1-d)\,f]\,(T-1) + (T-1)(1-d)\omega/R}. \tag{3.82}$$

This means that for patients at risk below $\underline{\rho}$, the hospital would engage in post-discharge care because it could generate savings at a linear rate.

Proof. Proof of Proposition III.19. The result follows from the concavity of $V(\rho)$:

$$\begin{aligned}\bar{\rho} \geq \frac{R}{\lim_{\rho\to 0^+} V'(\rho)} &= \frac{R}{\{d + [1-(1-d)\,f]\,(T-1)\}\,R + (T-1)\,(1-d)\,\omega} \\ \Rightarrow \bar{\rho} \geq \underline{\rho} &= \frac{1}{d + [1-(1-d)\,f]\,(T-1) + (1-d)(T-1)\omega/R}\end{aligned} \tag{3.83}$$

□

This lower bound relies on the concavity of $V(\rho)$, as it implies the marginal benefit (i.e., the derivative) decreases in ρ. Although $V(\rho)$ is not necessarily concave (see footnote 6 for a counter-example), in our numerical studies, $V(\rho)$ behaves very much like a concave function even it is not concave.

3.5 The Pre-Discharge ("Strengthening") Stage

In this section, we provide a stylized way to capture the relationship between a patient's readmission risk and the required pre-discharge efforts and costs for achieving the readmission risk. We do acknowledge that it is very difficult to quantify the exact cost of a hospital exerting efforts that reduce the readmission risk to a specific value ρ. This chapter does not attempt to provide a structural pre-discharge model or to estimate $C_S(\rho)$ from data. Specifying and estimating such a structural model would require a separate empirical study like Bartel et al. (2014). Instead, we take a stylized approach and follow Zhang et al. (2016) to define this cost. Different from their work, our approach is more general as it does not assume a specific functional form. Note that the functional form used in Zhang et al. (2016) satisfies all of the assumptions we impose.[7]

Assumption III.20 (Pre-Discharge "Strengthening" Cost Function)**.** *Without specific functional form assumptions, the pre-discharge strengthening cost $C_S(\rho)$ should satisfy the following mild and intuitive conditions:*

1. $C_S(\rho) = 0$ *for* $\rho \in [\rho_0, 1]$ *– a hospital can achieve the baseline readmission risk (or worse) without any (additional) spending on readmission reduction (which is true by definition of* ρ_0*).*

2. $C_S(\rho)$ *is continuous on* $[0, 1]$*. This condition is for analytical tractability.*

3. $C_S(\rho)$ *is twice differentiable on* $[0, \rho_0)$ *and twice left differentiable at* ρ_0*. This is also for analytical tractability.*

4. $C_S(\rho)$ *is strictly convex and strictly decreasing on* $[0, \rho_0)$*. This requires that the cost of reducing the readmission risk becomes more costly as* ρ *gets closer to zero.*

5. $C_S(0) = H_0 > R$ *– eliminating the possibility of a readmission is more expensive than the cost of a readmission. This is likely to be true because otherwise, a hospital could spend* $C_S(0) < R$ *in the pre-discharge stage to eliminate readmissions completely.*

3.6 Balancing Pre- and Post-Discharge Efforts

Next, we integrate the pre- and post-discharge stages and study how a hospital balances efforts between the two stages to minimize the total expected cost of the entire episode of care.

To check whether a BP policy is balancing, we analyze how $Z(\rho)$ behaves and find the minimizer ρ^*. There are two important cases that must be considered: (1) $\bar{\rho} \leq \rho_0$ and (2) $\bar{\rho} > \rho_0$. Figure 3.8 provides an example of each case. Recall that $\bar{\rho}$ is a threshold above which the hospital does not engage in reducing readmissions and

[7] Following the functional form used in Zhang et al. (2016), we can define: $C_S(\rho) = \begin{cases} 0 & \text{if } \rho \geq \rho_0 \\ \frac{H_0}{\rho_0^\alpha}(\rho_0 - \rho)^\alpha & \text{if } \rho < \rho_0 \end{cases}$, where $1 < \alpha < \infty$. This satisfies Assumption III.20 for all $1 < \alpha < \infty$ if $H_0 > R$. The parameter α and H_0 characterize the convexity and the difficulty of reducing readmissions. A larger α indicates a more convex shape of $C_S(\rho)$ so that reducing readmissions become increasingly difficult as ρ approaches zero.

below which the hospital benefits from reducing readmissions at an (asymptotically) linear marginal rate. Theorem III.21, shows that a BP chosen such that $\bar{\rho} > \rho_0$ balances the efforts between both the pre- and post-discharge stages.

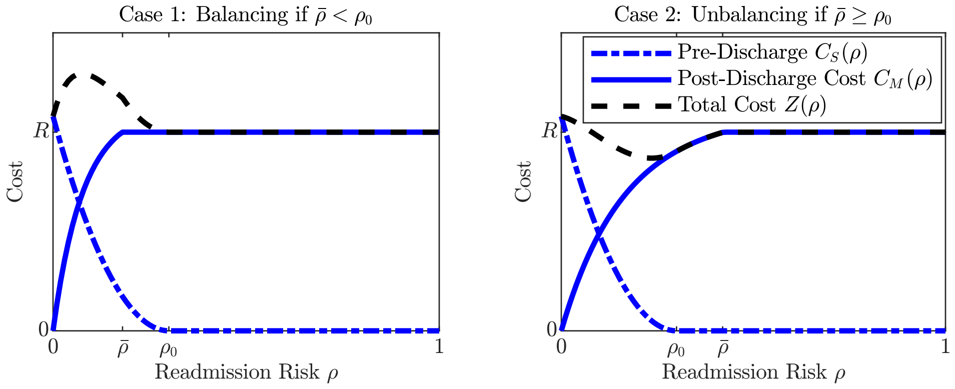


Figure 3.8: Schematic Sketch of the Costs in Cases 1 and 2

Theorem III.21 (Sufficient Condition For Balancing BP Policy)**.** *Suppose the following mild condition on the derivative of the pre-discharge strengthening cost function $C_S(\cdot)$ holds:*

$$C'_S(0) > \min\left\{-\frac{C_S(0) - R}{\rho_0}, -V'(\rho_0)\right\} \tag{3.84}$$

This condition requires that the first order derivative be sufficiently large (i.e., not too negative). Intuitively, this requires reducing readmissions in the pre-discharge stage to be not too difficult and expensive. Graphically, the C_S curve should not be too "steep." This condition holds if we use the specific functional form from Zhang et al. (2016).

If $V(\rho)$ is concave (see Theorem III.13, Proposition III.9), then a sufficient condition for a BP policy to be balancing is

$$\underline{\rho} = \frac{1}{d + [1-(1-d)\,f]\,(T-1) + (1-d)(T-1)\omega/R} \geq \rho_0. \tag{3.85}$$

If $V(\rho)$ is not concave, a sufficient condition is $\bar{\rho} > \rho_0$, where $\bar{\rho}$ can be efficiently computed using a univariate search for $V(\bar{\rho}) = R$ on $\bar{\rho} \in [0, 1]$.

Note that the cost of a treatment ω and the cost of a readmission R appear as a quotient in the denominator in Eq. (3.85). In fact, this quotient has a practical meaning – it quantifies how inefficient and expensive a treatment is, relative to the cost of a readmission. We shall now formally define this quotient as the treatment inefficiency.

Definition III.22 (Treatment Inefficiency)**.** The quotient ω/R denotes the *treatment inefficiency*. If $\omega/R \geq 1$, a treatment is considered very inefficient since it is as expensive as a readmission (if not more so). In this case, a hospital would rather readmit the patient and incur the readmission cost and avoid providing any treatments. If $\omega/R = 0$, the treatment is considered very efficient as it is virtually costless. Hence, a larger value of ω/R indicates that treatments are more inefficient.

Eq. (3.85) provides powerful insights into the design of a bundled payment policy. Next, we discuss each one of the moving parts and the policy implications of this equation.

3.7 Design of a Balancing Bundled Payment Policy

Theorem III.21 provides intuition into how the CMS can design a bundled payment policy that can incentivize readmission reduction. In this section, we discuss the policy implications. To provider stronger readmission reduction incentives, the following actions can be taken by the CMS.

- **Shortening the Episode Length ($T \searrow$).** Shortening the BP episode length provides stronger incentives for hospitals to exert more readmission reduction effort. As shown in Eq. (3.85), the episode length T being in the denominator suggests that reducing the length has a "convex" impact on the incentives – as the length gets shorter, the incentives get increasingly stronger (the convex relationship is shown numerically in Figure 3.13). This aligns with our intuition. A longer episode length means that the hospital is held accountable for a longer period of time, making preventing readmissions less feasible and more expensive.[8] Hence, extended episode length can strongly disincentivize hospitals to reduce readmissions, possibly explaining the observed lack of progress in national readmission reduction (Desai et al., 2016). Conversely, shortening the penalty window can incentivize hospitals to exert more readmission reduction effort, as the financial benefits of such efforts begin to outweigh the costs.

  Moreover, we find that shortening the episode length is more effective if the patient cohort experiences urgent and acute post-discharge complications that require immediate inpatient care (e.g., an organ failure). In such cases, patients will get readmitted very quickly once a complication develops. This is reflected

[8] For example, as T increases to ∞, Corollary III.14 implies that the post-discharge cost becomes a step function. This means that following-up will be as expensive as a readmission, unless the readmission risk is completely eliminated (i.e., $\rho = 0$). In this case, the hospital would not exert any readmission reduction effort in either pre- or post-discharge stage.

by a larger delay-time failure rate d. As d increases, the delay-time window in which post-discharge outpatient treatment is effective becomes shorter. This means that there is little room for error. The hospital must exert full effort to engage in preventing readmissions, because otherwise, if the patient developed a complication and was not treated timely, this complication would quickly trigger a readmission. Consider the extreme case in which complications require immediate inpatient care ($d = 1$). Despite perfect treatments ($f = 1$), subsidizing treatment and penalizing readmission would not work. In this case, Eq. (3.85) becomes $1/T \geq \rho_0$, from which ω/R disappears. The only way to increase incentives is to shorten the episode length.

In recent medical literature, how long a BP episode of care should be and how long the HRRP penalty window should be have been controversial. A New England Journal of Medicine article argued that "policymakers (...) could consider limiting the (HRRP) time window. The causes of readmissions occurring within 3 days after discharge or even 7 days after discharge are much more under the hospital's control, and these near-term readmissions are preventable far more often than later ones" (Joynt and Jha, 2012). The Healthcare Cost and Utilization Project (HCUP) (Fingar et al., 2017) found that over one-third of 30-day readmissions were 7-day readmissions and 7-day and 30-day readmissions were similar for many surgeries. Moreover, a recent study (Graham et al., 2018) showed that early readmissions (within 7 days of discharge) were more likely to be preventable than late ones (after day 7). Our findings support these medical hypotheses on the benefits of a shorter HRRP/BP window, and could provide a quantitative guideline for policy design.

- **Making Post-Discharge Treatment More Efficient ($\boldsymbol{\omega/R} \searrow$).** To further incentivize readmission reduction, the treatment inefficiency can be lowered by subsidizing post-discharge treatments and/or increasing the readmission penalty:

  - **Subsidizing Post-Discharge Treatments ($\boldsymbol{\omega} \searrow$).** To incentivize readmission reduction, the CMS should encourage post-discharge outpatient treatment by subsidizing and cost sharing. High cost of outpatient follow-up treatment can be burdensome thus can make reducing readmission costly for hospitals. Reducing this cost ω, without sacrificing the quality of care, can incentivize members of a healthcare bundle (e.g., Accountable Care Organizations) to work to reduce readmissions. Under current BP policy, the payment covers the costs of post-discharge outpatient treatments. However, the CMS does not have policies in place for post-discharge care subsidies. In some cases, post-discharge services and treatments are not even covered by Medicare or Medicaid. Our finding suggests that subsidizing post-discharge care could entice hospitals to reduce readmission. In fact, some private healthcare providers have already started to encourage and support their patients to seek better post-discharge care. For example, Tenet Healthcare, a private healthcare corporation, has put into effect a policy to fund patients for their spendings in post-discharge care such as "non-covered medically appropriate outpatient services at a hospital or another provider of the patients choice" (Tenet Healthcare, 2015). Our finding suggests that the CMS should encourage post-discharge outpatient treatment by subsidizing and cost sharing. In particular, for high-risk surgeries, such subsidies would play an important role to ensure hospitals exerts effort in the continuum of care pre- and post-discharge.

– **Penalizing Readmissions ($\boldsymbol{R} \nearrow$).** As readmissions become more costly or more penalized, the hospital will put more efforts into readmission reduction to avoid such a high cost. Although this cost is determined by the pathological nature of each readmission case, we believe that the CMS can still have some control over this cost by penalizing readmissions and/or decreasing the Bundled Payment amount (so that readmission costs take up a greater portion of the BP budget). Currently, BPCI and HRRP do not overlap – hospitals do not get penalized for excessive readmissions if they are reimbursed under a BP scheme. However, if stronger incentives are needed, it might be viable for the CMS to impose HRRP penalties on top of the BP scheme.

  In practice, the cost of a readmission exhibits large variation: a 30-day readmission after a major abdominal or chest surgery can cost as low as \$576 and as high as \$147,904 (Jacobs et al., 2017; Leow et al., 2018). As such, hospitals may not have enough incentives to reduce low-cost readmissions. In practice, a hospital can anticipate the cost of a future readmission – medical studies have found predictors and flags for high readmission costs and longer readmission LOS, such as blood transfusion, imaging, pathological stage, and comorbidity in the index hospitalization (Jacobs et al., 2017). Therefore, it is possible that a hospital decides to exert less effort on patients who exhibit low-cost readmission characteristics. To "equally" incentivize the reduction of both low-cost and high-cost readmissions, we argue that the CMS could consider penalizing readmissions differently to reduce the variance in costs. For instance, readmissions can be reviewed case by case by a panel of external clinicians, and accordingly, the CMS can determine a case-specific penalty based on the cause, cost, and intensity of each readmission. In fact, this fits well into the CMS's current policy and practice as the BPCI retrospectively reconciles with the hospital for the bundled payments (CMS, 2018a).

3.8 Numerical Case Study

In this section, we first estimate the parameters using two datasets of cystectomy patients. We then discuss how to design a BP policy and how each moving part impacts the policy.

3.8.1 Parameters Estimation

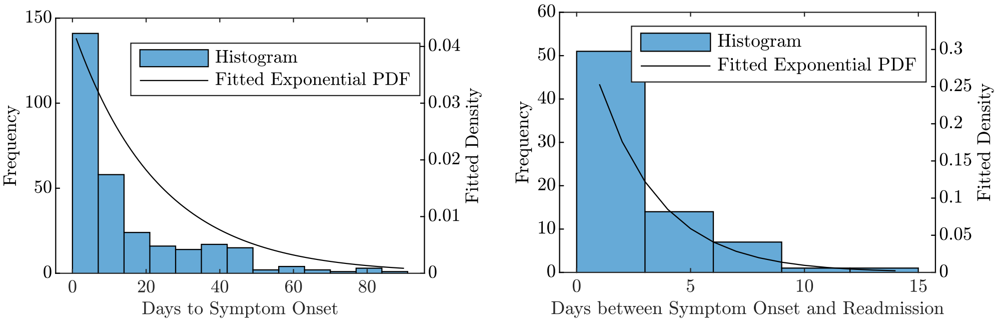
As a proof of concept, we use data collected from bladder patients who have undergone cystectomy surgery. Cystectomy, a surgery to remove the bladder, has one of the highest 30-day readmission rates among all major chest and abdominal surgeries. Using these patients as an initial testbed, we validate some of the key assumptions used in our analytical study.

We used two datasets to parameterize the model. The first dataset consists of 327 cystectomy patients discharged from our partner hospital between 2007 and 2012. The second dataset contains 717 cystectomy patients from the State Inpatient Database (SID) discharge in 2009 and 2010. For detailed description of the inclusion and exclusion criteria, we refer the readers to Liu et al. (2018a).

Baseline Readmission Risk ($\boldsymbol{\rho_0}$) and the Delay-Time Failure Rate ($\boldsymbol{d}$). To estimate the readmission risk ρ_0, we first fitted an exponential distribution with $\lambda = 0.044$ using right-censored data from the 327 cystectomy patients (see Meeker and Escobar (2014) for estimation method).[9]

We estimate the readmission risk $\rho_t = P(s_{t+1} = S | s_t = H, a_t = Wait) =$

[9] We would like to stress that the readmission risk is distinct from the probability of readmission. Recall that the readmission risk ρ is the probability of a healthy patient developing a post-operative complication (i.e. becomes sick). The reason why this rate may seem higher than expected is due to the fact that a post-operative complication may not necessarily lead to a readmission. Potential reasons include: 1) some complications were treated (thus a readmission was prevented); 2) some complications may have caused mortality (instead of a readmission); and 3) some complications were never treated and did not trigger a readmission. Nevertheless, in clinical practice, without proper diagnoses at a clinical encounter, it is difficult to tell whether a patient's condition would cause a readmission or not. Thus, as long as the patient is experiencing complications, our model treats them indifferently as a potential readmitable complication. Although this is a simplification, we believe that this is suitable for policy-level analyses.



Panel A: Readmission Risk ρ_0

Panel B: Failure rate d

Figure 3.9: Fitted Exponential Distribution for the Time to Symptom (Panel A) and Time between Symptom Onset and Readmission (Panel B)

$\frac{S(t)-S(t+1)}{S(t)}$, where $S(t)$ is the survival function (i.e. complementary cumulative distribution function) of the exponential distribution. For an exponential distribution with rate λ, the readmission risk can be computed using $\rho_t = 1 - \exp(-\lambda)$. Hence we estimate the baseline readmission risk $\rho_0 = 0.04$. This represents the 25% 30-day readmission rate of cystectomy patients. Since the overall 30-day readmission rate of Medicare beneficiaries are 20% (Jencks et al., 2009), we use $\rho_0 = 0.04$ as a baseline scenario and varied it between 0.01 and 0.08 to represent a low-risk case and a high-risk case respectively. This serves as a baseline for evaluating whether a hospital is effectively reducing readmissions. In this chapter, we do not attempt to estimate the natural readmission rate intrinsic to the pathology of the surgery and its recovery because it is difficult to observe such a natural readmission risk as patients are rarely free of interventions. Based on the data fitting shown in Figure 3.9, the delay-time failure rate d is estimated to be $d = 0.3$. We varied d between 0.1 and 0.5 for sensitivity analysis.

Cost of a Readmission (R) and Cost a Treatment (ω). There are huge variations in the cost of readmissions. A 30-day readmission can cost as low as \$576 and as high as \$147,904 (Jacobs et al., 2017; Leow et al., 2018). While this might be an issue for an operational model, for policy-level analyses, we use the population average to gain insights. The average cost of a readmission is estimated to be $R = \$5,000$ according to Leow et al. (2018). To estimate the cost of a treatment ω, we refer to Kilroy et al. (2013), which estimated that an average Emergency Department (ED) visit within 30 days of discharge costs \$1,900. Given the fact that ED visits are very expensive, if patients were treated at the Primary Care level, the cost will be much less. For a baseline treatment cost, we use the cost of half of an

ED visit $\omega = \$1,000$. The cost of post-discharge readmission reduction efforts is difficult to estimate in practice due to the varied methods that can be employed and the difficulty of quantifying the cost of those efforts in practice. To address this, we vary the cost of ω along a reasonable interval, with the average cost of a readmission \$5,000 as an upper bound and \$200 as a lower bound (which represents the average cost of a Primary Care visit).

Pre-Discharge Cost of Eliminating Readmissions ($C_S(0)$). We acknowledge that it is very difficult to estimate the cost of eliminating readmissions completely for a patient cohort. Here we propose a conservative estimate. Based on a study of 49,540 cystectomy patients in the U.S. between 2003 and 2010 (Leow et al., 2015), the 95% confidence upper bound for the 90-day hospital direct cost was \$27,269. The average LOS was 10.8 days. The average cost of hospitalization per day can be roughly estimated to be \$2,525. One conservative estimate for the cost of eliminating readmissions is the cost of keeping all patients hospitalized for 90 days after surgery. We estimate this cost to be $C_S(0) = \$230,000 \geq \$227,242 = 90 \times \$2,525$.

Treatment Efficacy (f). In the medical literature, studies have found varying levels of efficacy for post-discharge follow-up care to reduce readmissions. Jack et al. (2009) found that follow-up care, joined with other readmission reduction measures, reduced readmissions. Misky et al. (2010) reported that patients who had timely PCP follow-up after discharge had 2% readmission rate whereas patients lacking timely PCP visit had 21% readmission rate. According to Benbassat and Taragin (2000), up to 75% percent of readmissions are preventable. In line with the latter paper, we choose $f = 0.75$ to be the baseline scenario. We vary f between 0.5 and 1 to conduct sensitivity analyses.

For the purpose of exposition, we round the parameters and present them in Table 3.2 to summarize the estimation results, the uncertainties, and the bounds for sensitivity analyses. We acknowledge that many of these parameters are difficult to estimate, however we believe our ranges are reasonable based on available financial and clinical data for our stated purpose, which is to provide insight into system-level policy design, rather than operational-level decision support, which requires more accurate estimates of model parameters. There are many practical and theoretical challenges in the estimation of these parameters, which provide motivation for a separate empirical study as future work.

Parameter	Baseline	Uncertainty	Lower Bound	Upper Bound	Source
ρ_0	0.04	Moderate	0.01	0.08	Liu et al. (2018a)
d	0.3	Low	0.1	0.5	Liu et al. (2018a)
R	\$5,000	Very High	\$500	\$150,000	Leow et al. (2018)
ω	\$1,000	Very High	\$200	\$5,000	Kilroy et al. (2013)
$H(0)$	$\$230,000$	Moderate	\$230,000	$+\infty$	Leow et al. (2018)
f	0.75	High	0.5	1	Benbassat and Taragin (2000)

Table 3.2: Summary of Baseline Estimates, Uncertainties, and Bounds for Sensitivity Analyses

3.8.2 Assumption Justification and Validation

Assumption III.4 (treatments are effective) holds in the baseline scenario of $d = 0.3$ and $f = 0.75$. **Assumption III.5 (Following-Up Every Day is Expensive)** holds in the baseline scenario of $\omega = \$1,000$ and $R = \$5,000$ if $T \geq 5$. **Assumption III.6 (stationary readmission risk)** is used to derive a closed-form expression for analysis of the post-discharge follow-up cost. When this assumption violated, Theorem III.16 provides an accuracy bound for the PWCF expression. **Assumption III.20 (Pre-Discharge "Strengthening" Cost Function)** is quite general as it only specified the shape of the curve. As mentioned previously, if we follow the literature and use a similar functional form as in Zhang et al. (2016), $C_S(\rho) = \begin{cases} 0 & \text{if } \rho \geq \rho_0 \\ H_0(\rho_0 - \rho)^\alpha/\rho_0^\alpha & \text{if } \rho < \rho_0 \end{cases}$, then Assumption III.20 holds for $C_S(0) = H_0 > R$ and for all $\alpha \in (1, \infty)$.

3.8.3 Policy Recommendations

In this section, we first analyze the status quo and shed light on the recent trend indicating that readmission reduction has stalled under current HRRP penalty program. We then propose policy recommendations through policy levers of (1) shortening the episode length (or the HRRP penalty window) T. (2) subsidizing post-discharge treatments, and (3) penalizing readmissions.

Status Quo: lengthy episode/penalty window weakens readmission reduction incentives.

Much of the work on readmissions has focused on penalties and methods (costs) for preventing readmissions. In this section, we analyze these two methods in conjunction with a third dimension: the episode/penalty window length. Here, we define the status quo as $T = 30$ under the HRRP defined penalty window, and $T = 90$ under the BPCI defined length of an episode of care. We test whether the sufficient condition for balanced readmission reduction efforts is met by varying the treatment efficacy f and inefficiency ω/R for low ($\rho_0 = 0.01$), medium ($\rho_0 = 0.04$), and high ($\rho_0 = 0.08$) risk cohorts.

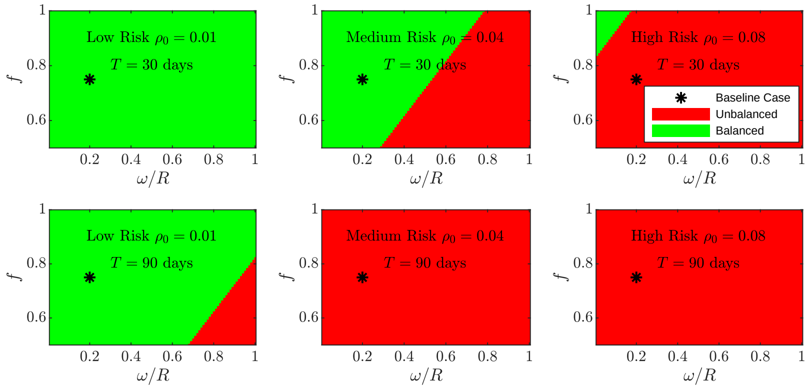


Figure 3.10: Status Quo: in Baseline Scenarios, 90-day BP Episode Length is Too Long for Medium and High-Risk Cohorts. HRRP 30-day Window is Too Long for High-Risk Cohorts

In Fig. 3.10, the green (light gray) region indicates that the optimal policy is balanced (i.e. the sufficient condition (Eq. (3.85)) is met) and red (dark gray) region is indicates the optimal policy is unbalanced and the hospital will not exert effort in both stages (i.e. the condition is not met). For the medium- and high-risk cohorts, it appears that a 90-day BP episode may be too long to incentivize readmission reduction; the effort of keeping these patients out of the hospital is too great *even if post-discharge treatments are very effective and efficient.* In this case, it is unlikely that a treatment subsidy or operational improvement to decrease cost of treatments or an increase in penalty can overcome the burden of preventing these riskier patients from returning to the hospital for up to 90-days after discharge. The only option for these cases is to *shorten the penalty and episode window.* This could be done on a case-by-case basis for diagnoses/procedures that are considered particularly at medium or high risk for readmission; e.g. radical cystectomy with a 25% readmission rate (Lee et al., 2019).

In the baseline scenario (marked with asterisks), a 30-day penalty window could incentivize readmission reduction only for low and medium risk cohorts. Unfortunately, the contribution of these patients to the overall readmission rate is dampened by the fact that their risk is low, hence targeting these patients may not have the desired magnitude of impact on readmissions. For the high-risk cohorts, treatment efficacy and efficiency must be high for readmission reduction incentives, even if there is only a 30-day penalty window. This observation supports one possible explanation for why the readmission reduction has plateaued under the HRRP's 30-day penalty program (Desai et al., 2016): the penalty window is arduously long, particularly for riskier patients that contribute significantly to the overall readmission rate, while the low hanging fruit (low and medium risk) provides insufficient improvement to move the needle.

In the next two subsections, we analyze each of these three levers in greater detail. To do so, we quantify what we mean by "incentivizing readmission reduction" as follows. For each BP policy, (ω, R, T), there is an upper threshold on the risk level of patients that would be included in a Readmission Reduction Program that we call the *RRP threshold*; e.g. only patients with baseline risk level $\rho_0 < 0.04$ would induce the hospital to employ a balanced policy (i.e. exert appropriate effort in a readmission reduction program). The higher the threshold, the more (and riskier) patients would be included in a readmission reduction program. Hence we quantify readmission reduction incentive in terms of the impact on this threshold: the higher the threshold, the more the incentive to expand readmission reduction programs and to target riskier (and hence more needy) patients. We refer to this quantitative measure as *X%-RRP (readmission reduction program) expansion* (or similarly *RRP expansion*), an increase in the risk threshold for program inclusion by $X\%$ (absolute).

Subsidizing Post-Discharge Treatments and Penalizing Readmissions.

In this section, we study how a policymaker (e.g., the CMS) may consider subsidizing post-discharge treatments and/or increasing the penalty of readmissions to incentivize readmission reduction.

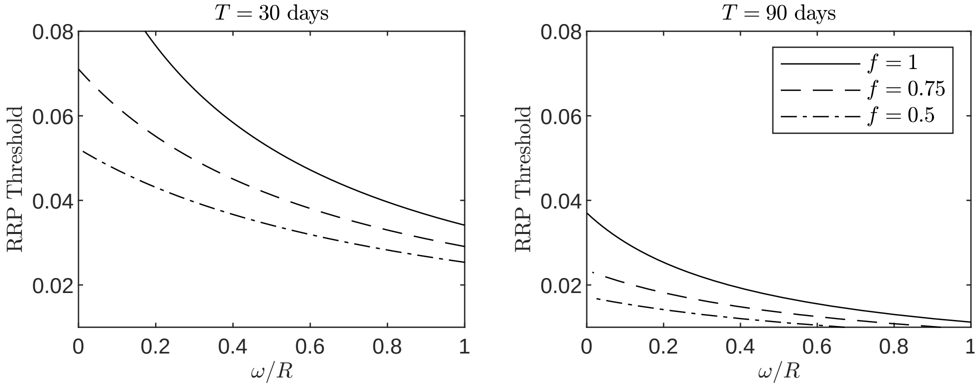


Figure 3.11: Readmission Reduction Program (RRP) Threshold is Convex Decreasing in Treatment Inefficiency ω/R ($d = 0.3$)

In Fig. 3.11, the lines show the RRP threshold for inclusion as a function of treatment inefficiency, ω/R . If a point $(\omega/R, \rho_0)$ falls below the line then the hospital will engage in readmission reduction and if the point lies above line, then the hospital will not. In the latter case, the CMS may subsidize the post-discharge treatments or increase the readmission penalty to incentivize readmission reduction. The curves can be interpreted as displaying how many extra patients (greater ρ_0) would be targeted for reduction efforts if treatments are made more efficient. Note, more efficient is to the left in this graph.

The convex shape of the impact of the subsidy/penalty inefficiency ratio indicates that an improvement in efficiency (move to the left on the curve) has greater impact when the treatments are already efficient. This means that, if the BP structure and readmission ecosystem is already inefficient, it will take significant efficiency improvement to see even a small change in RRP expansion (as a measure of readmission reduction effort). Conversely, if the system is already fairly efficient, small improvements in efficiency can result in increasingly more substantial RRP expansion. The managerial implication is that, it *may* take significant improvements in efficiency (e.g. like a setup cost) for incentives to begin having an impact on RRP programs (literature indicates current efforts may not be sufficient), but once we start seeing an impact a program may experience increasingly impactful gains from further increases in efficiency. This is with the caveat, however, that BP design may not linearly impact inefficiency as a function of their controllable inputs. We next study the impact of the subsidy/penalty levers individually to shed more insight on the policy design choices for insurers.

Fig. 3.12 plots the sensitivity (marginal) to subsidy/penalty for a wide range of ω and R (based on Table 3.2). On the left panel, we set the treatment cost to the baseline $\omega = \$1,000$ and varied the cost of readmission between the estimated lower and upper bounds (from \$500 to \$150,000). To do so, we compute the derivatives of the LHS of Eq. (3.85) with respect to ω and R. The absolute values of the derivative measure the sensitivities to subsidy and penalty ($|\,\mathrm{d}/\,\mathrm{d}\omega\ \underline{\rho}(\omega)|$ and $|\,\mathrm{d}/\,\mathrm{d}R\ \underline{\rho}(R)|$ respectively). Hence, a larger sensitivity indicates a greater change in terms of incentives (i.e., LHS of Eq. (3.85)) per dollar of subsidy/penalty.

We found that subsidizing is more effective when the cost of a readmission is relatively low. When the readmission penalty is too high, the additional impact of subsidy becomes muted because the high penalty is already sufficiently motivating. On the right panel, we set the readmission cost to the baseline $R = \$5,000$ and varied the cost of treatment between the estimated lower and upper bounds (from \$200 to \$5,000). We found that increasing the penalty is more robust to a variety of treatment costs. In contrast to subsidy, the general trend indicates that increasing

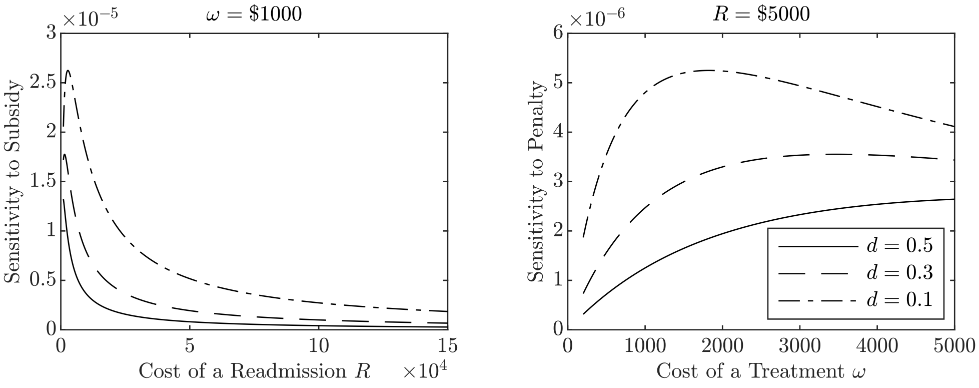


Figure 3.12: Sensitivities to Subsidy and Penalty $(T = 30, f = 0.75)$

readmission penalties is more effective when treatment costs are higher. Thus, if readmission prevention methods are expensive a penalty can be more incentivizing. Because penalty and subsidy move in opposite directions with respect to each other, it may be prudent to consider only engaging in one of these mechanisms, depending on the status quo of current costs.

Shortening the BPCI/HRRP Window Length.

As our analytical result suggests (see §3.7), reducing the episode/penalty window length T has a "convex" effect on the incentives, making the episode/penalty window length a highly effective mechanism to increase the incentives to reduce readmissions. In Fig. 3.13, we varied the window length T for two treatment inefficiency (ω/R) scenarios at 0.2 (baseline) and 0.8 (inefficient treatments). A larger area under the line in Fig. 3.13 indicates more robustness in readmission reduction incentives because the hospital will engage in readmission reduction for cohorts at or below the baseline readmission risk ρ_0. Due to the convex impact of T, shortening the window length is increasingly effective as the window shrinks. Fig. 3.13 indicates that a 14-day window would provide sufficient incentives for hospitals to reduce readmissions for patient cohorts at high risk. Moreover, shortening the window is also more effective than improving the inefficiency (as the great area under the curve covers a broader range of risks).

From Fig. 3.13, we also observed that when the treatment inefficiency increases (from 0.2 to 0.8), increasing the treatment efficacy becomes less effective in providing incentives. This is due to the fact that the high cost makes the post-discharge treatments less economically viable. As a result, under the same budgetary constraint (a hospital is willing to spend at most R in the post-discharge stage), the hospital

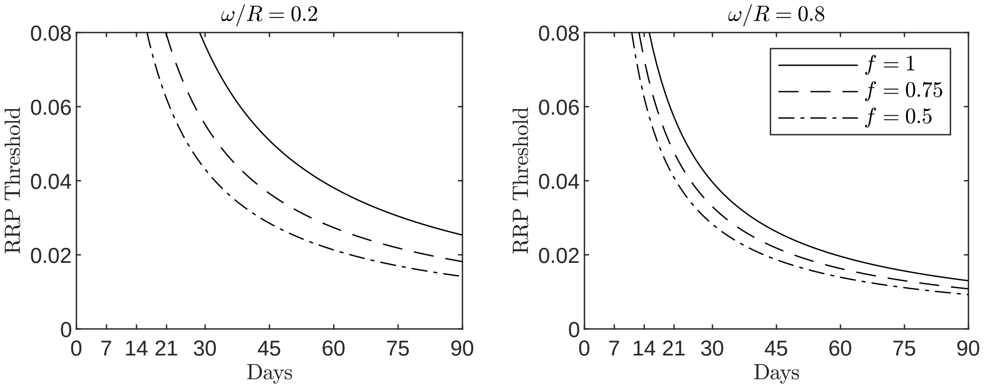


Figure 3.13: Readmission Reduction Program (RRP) Threshold is Convex Decreasing in the Window Length ($d = 0.3$)

will conduct fewer post-discharge treatments.

3.8.4 Summary of Numerical Insights

We summarize the insights from our numerical studies.

Status quo HRRP penalty window and BPCI episode are possibly too long. Specifically, current BPCI 90-day episode of care might be too long if the patient cohort is at medium- or high-risk of readmission. Under HRRP, if a hospital minimizes the cost (as it does under a BP policy), then 30-day HRRP penalty window might be too long for high-risk surgery cohorts. With too long a HRRP/BPCI window, there is no amount of subsidy or penalty that could incentivize hospitals to include medium or high risk patients in their readmission reduction program. This shed light on why readmission reduction has recently stalled under current HRRP scheme.

For high-risk patients, HRRP/BPCI window must be appropriately shortened first, before exploring other subsidy and penalty options. Shortening the BPCI episode and HRRP penalty window has a convex impact – as it shortens to 14-day, hospitals would gain sufficient incentives to reduce readmissions for patient cohorts at a broad range of readmission risks.

Subsidy of treatment and penalty of readmission can make treatments efficient enough to provide readmission incentives. In particular, the treatment efficiency has a convex impact – improving the efficiency has a greater impact when the system is already efficient. The more we work toward incentivizing readmission reduction, the more effective our efforts become. In conjunction with the shortening of the BPCI/HRRP window, subsidy and penalty in readmission reduction initiatives can become increasingly effective to overcome the stalling of the readmission reduction momentum.

Effectiveness of subsidy and penalty work in contrasting ways. When treatments are expensive, penalties are increasingly effective. In contrast, subsidy is only effective when penalties are small to moderate and rapidly becomes ineffective as the readmission cost/penalty increases. Depending on the readmission and treatment costs intrinsic to each patient cohort, the CMS should focus on the most effective lever between subsidy and penalty.

3.9 Practical Considerations, Limitations, and Future Work

Our analytical and numerical analyses have found the following policy levers could be used to increase the incentives: the readmission penalty could be increased; the post-discharge treatment that prevents readmissions can be subsidized; the HRRP penalty window and the BPCI episode of care length can be shortened to provide stronger incentives. Implementing such policy changes might raise some practical concerns. The policymakers should consider the following practical issues before actually implementing the recommended policy changes. These considerations also provide sensible directions for future research.

Shortening the BPCI Episode (or HRRP's Penalty Window) Length. Reducing the window T in certain instances may be very effective. However, there is an argument that this may lead hospitals to try to avoid readmissions by deferring the readmission to day $T+1$. For instance, if a patient presents him/herself at the ED on day T after discharge, the hospital can simply keep the patient in the ED or an observation unit till day $T+1$ and then admit the patient. However, this incentive exists regardless of the window, T, and there is anecdotal evidence that healthcare providers are already doing this with a 30-day readmission window. With a shorter window, the concern becomes that a larger fraction of the overall readmissions may be avoided in this manner. However, the choice of penalty window impacts this delay phenomenon only in terms of the number of potential readmissions that could occur within 1-2 days of the penalty cut-off. Hence, choosing an appropriate window based on historical readmission times (which would be not tampered with if the window were shortened below 30 days) could mitigate some of this behavior. A future research direction is to study and incorporate the this "defer" behavior of hospitals.

Post-discharge Treatment Subsidy. Subsidizing post-discharge treatment can help lessen the hospital's financial burden so that better post-discharge care can be provided. In practice, such subsidies can create unnecessary incentives for premature discharge and/or unnecessary post-discharge efforts exceeding what is economical to reduce readmissions. Earlier discharge, on one hand, could reduce the LOS and increase throughput. On the other hand, it could also lead to adverse health outcomes. Discharge timing and the LOS play important roles in readmission risk management. In this chapter, our inpatient stay model does not consider discharge timing. This is left to future work, where a more structured model for the inpatient stage could be developed with the support of empirical estimation.

Readmission Penalty. Currently, BPCI and HRRP do not overlap – hospitals do not get penalized for excessive readmissions if they are reimbursed under a BP scheme. However, if stronger incentives are needed, it might be viable for the CMS to impose HRRP penalties on top of the BP scheme. The HRRP readmission penalty is limited to 3% of the reimbursement amount Our study provides insights and a quantitative guideline for increasing this penalty, if stronger incentives are needed. However, if the readmission penalty is too high, it could lead hospitals to eschew patients from that payer (e.g. Medicare HRRP) or may cause significant financial hardship for the healthcare industry. Therefore, a sensible future extension is to incorporate patient selection.

While many aspects of our modeling framework require more investigation and empirical support, we believe that our analyses point to new directions for future research regarding incentives to reduce readmissions.

3.10 Conclusion

In this chapter, we study how a health funding policymaker can design an effective bundled payment and readmission penalty policy to incentivize hospitals to balance pre- and post-discharge efforts and thus reduce readmissions. To do so, we propose a novel Strengthen Then Maintain (STM) framework that models an episode of care consisting of an inpatient stay (strengthening) stage and a post-discharge follow-up (maintaining) stage. This framework is applicable to a set of machine maintenance problems where the failure rate of a machine can be reduced at a cost in the strengthening stage, and in the maintaining stage maintenance policies are optimized based on the failure rate produced by the strengthening stage. To study policy-level decisions, we first study how a hospital would behave under a BP policy. By analyzing the hospital's behaviors, we identified two possible follow-up monitoring regimes – in an engaged policy, the hospital treats the patient if and only if the patient is sick; whereas, in a disengaged policy, the hospital may not treat a sick patient due to the high costs associated with readmission prevention efforts. We show that a disengaged policy is undesirable due to its high costs and should be disincentivized at the policy level. We then develop a piece-wise closed-form expression for the post-discharge stage cost-to-go and provide a theoretical bound for the optimality gap when key assumptions are violated. Making minimal assumptions on the cost of the inpatient stay stage, we derive a sufficient condition to incentivize the hospital to reduce readmission while minimizing the cost of both the pre- and post-discharge stages.

Our analytical findings suggest that the following policy levers could be used to increase the incentives for reducing readmissions: 1) subsidizing post-discharge outpatient follow-up treatments; 2) penalizing readmissions; and 3) shortening the readmission penalty window and the length of an episode of care. We also provide a quantitative guideline for the CMS to decide the subsidization and penalty as well as the program window length. Parameterized with data from a cystectomy patient cohort, we found that the bundled payment policy and the readmission penalty program status quo may not provide sufficient incentives for hospitals to reduce readmissions. This sheds lights on the potential causes of the plateaued readmission reduction momentum since the implementation of the HRRP.

CHAPTER IV

Pre-Discharge Readmission Risk Prediction

ABSTRACT: Despite efforts to reduce their frequency and severity, complications and readmissions following radical cystectomy remain common. Leveraging readily available, dynamic information such as laboratory results may allow for improved prediction and targeted interventions for patients at risk of readmission. We used an institutional electronic medical records database to obtain demographic, clinical, and laboratory data for patients undergoing radical cystectomy. We characterized the trajectory of common postoperative laboratory values during the index hospital stay using support vector machine (SVM) learning techniques. We compared models with and without laboratory results to assess predictive ability for readmission. Among 996 patients who underwent radical cystectomy, 259 (26%) patients experienced a readmission within 30 days. During the first week after surgery, median daily values for white blood cell count, urea nitrogen, bicarbonate, and creatinine differentiated readmitted and non-readmitted patients. Inclusion of laboratory results greatly increased the ability of models to predict 30-day readmissions after cystectomy. Common postoperative laboratory values may have discriminatory power to help identify patients at higher risk of readmission after radical cystectomy. Dynamic sources of physiological data such as laboratory values could enable more accurate identification and targeting of patients at greatest readmission risk after cystectomy

4.1 Introduction

Radical cystectomy has one of the highest rates of complications and readmissions of any surgical procedure, with 25% of patients experiencing unplanned readmission within 30 days (Borza et al., 2017; Stimson et al., 2010; Hu et al., 2014; Skolarus et al., 2015). These high readmission rates, coupled with increasing policy focus on reducing readmissions, have motivated investigations into identification and optimization of patients at highest readmission risk. However, the ability to predict readmission using traditional administrative data is limited, making it unclear where and when to focus resources, leaving readmission rates largely unchanged (Minnillo et al., 2015; James et al., 2016).

There is increasing interest in incorporating dynamic data sources into readmission prediction models to better enable identification of high risk cohorts (Goldstein et al., 2017). While traditional administrative data are typically limited to static factors (e.g., demographics, comorbidities), widespread use of electronic health records has made dynamic sources of data, laboratory results for example, readily available. The degree to which readily available laboratory data used to guide day-to-day clinical decision-making might impact readmission risk prediction after cystectomy is unknown. Indeed, such variables can be successfully incorporated into prediction models to improve performance for other outcomes ranging from transfer to the intensive care unit to mortality (Escobar et al., 2008; Kipnis et al., 2016; Escobar et al., 2015; Lim et al., 2015). For cystectomy patients with frequent postoperative lab draws, models using dynamic laboratory data could allow for better risk stratification and postoperative planning.

In this context, we used data from our institutional electronic health record to examine whether incorporating dynamic laboratory data into readmission prediction models improved risk stratification after radical cystectomy. Specifically, we assessed daily post-operative values for commonly obtained laboratory tests, and used machine learning techniques to compare values between readmitted and non-readmitted patients. This study demonstrates the unique promise of readily available, dynamic data to inform risk stratification of patients most likely to be readmitted after cystectomy.

4.2 Data Source

We used the Michigan Medicine database containing records on all inpatient and outpatient visits at our tertiary care facilities. This dataset was queried for all inpatient encounters associated with a diagnosis of bladder cancer (International Classification of Diseases 9th Revision code 188.X) and procedural codes for radical cystectomy (57.71) for the period from 2006 to 2016.[1] This yielded a cohort of 996 patients who underwent radical cystectomy during the study period.

[1] The nature of this surgery, which is the removal of the entire bladder, has not changed over the study period.

4.3 Outcomes and Covariates

Our primary outcome for this study was unplanned readmission within 30 days of discharge from the index hospitalization. Among patients who did not have an additional inpatient record within 30 days of discharge, we manually reviewed their electronic charts to identify any patients who had documentation of readmission to an outside facility during the post-discharge time window. This revealed an additional 62 readmitted patients.

We extracted the following administrative data as covariates for this study including age, body mass index (BMI), Charlson comorbidity score (precalculated within the source data using ICD-9 codes), marital status, gender, race, and insurance. We also obtained daily laboratory result data for the index admission after radical cystectomy. In order to focus on the most clinically relevant and readily available laboratory data, we restricted our analysis to the most common laboratory tests during the postoperative period including: complete blood count (white blood cell count, hemoglobin, hematocrit, and platelet count), basic metabolic panel (sodium, potassium, chloride, bicarbonate, blood urea nitrogen (BUN), creatinine, and glucose), and coagulation studies (prothrombin time (PT), international normalized ratio (INR), partial thromboplastin time (PTT)). We used ICD-9 codes to determine postoperative complications for inclusion in our readmission prediction models using previously described methods (Tan et al., 2011).

4.4 Statistical Analysis

We tested for differences between readmitted and non-readmitted patient characteristics using chi square testing for categorical variables and t-tests or Wilcoxon rank-sum tests for continuous variables depending on distribution. We assessed changes in laboratory values in several ways including: minimum and maximum during hospitalization, mean across the entire hospitalization, proportion of measured laboratory values outside of the normal reference range, and binary indicator variables if values were ever outside reference ranges.

To assess for differences in laboratory results between readmitted and non-readmitted cohorts, we applied support vector machine (SVM) techniques. This is a machine learning method used to generate classifications for a group of observations. In this case, we used SVM to generate cut-off laboratory values between readmitted and non-readmitted patients across the postoperative time period, with unique thresholds generated for each laboratory value on each postoperative day. We also assessed whether variance in these laboratory values differed between readmitted and non-readmitted patients.

Lastly, we built multiple logistic regression models using a combination of the SVM laboratory value thresholds, complications data, and baseline patient demographic and clinical data to examine effects on readmission risk stratification. We halved our cohort into derivation and validation cohorts, and selected predictors for inclusion in the model on the basis of the Akaike information criterion, a statistic that estimates the relative quality of competing models (Akaike, 1974). For comparison we also built a model using the same variables and a random forest regression algorithm, a machine learning technique which captures interactions between variables.

All analyses were conducted using SAS software version 9.4 (SAS Institute, Cary, NC) and all testing was two-sided using an alpha of 0.05. This study was approved by our Institutional Review Board (HUM00128698).

4.5 Results

Demongraphics	Readmitted (N=259)	Non-readmitted (N=737)	p-value
Mean age, y. (SD)	67.6 (10.8)	66.3 (11.0)	0.09
Male gender, %	81.2	83.2	0.5
Mean BMI, kg/m2 (SD)	29.2 (7.1)	28.8 (7.5)	<0.01
Race, %			0.54
Caucasian	90.8	91.3	
Black	4.6	2.4	
Other/unknown	4.6	6.3	
Marital status, %			0.17
Married	74.1	68.1	
Unmarried	2.3	3.6	
Unknown	23.6	28.3	
Charlson comorbidity index, %			0.22
0	3.5	4.9	
1	0	0	
2+	96.5	95.1	
Primary Payer, %			0.41
Private	39	39.7	
Medicare	53.4	52.3	
Medicaid	5.7	5.9	
Other	1.9	2.1	
Robotic cystectomy, %	3.1	4.5	0.43

Table 4.1: Patient Characteristics Stratified by Readmission Status after Radical Cystectomy

Among the 996 patients included in this cohort 259 (26%) were readmitted within 30 days of discharge. Readmitted and non-readmitted patients were similar in their demographic and clinical characteristics, though readmitted patients had higher

Body Mass Index (BMI) values, on average (Table 4.1, $p < 0.01$). Most patients were older, married, Caucasian men, and the minority were treated with robotic cystectomy.

As illustrated in Figures 4.1, 4.2, 4.3, and 4.4, several of the laboratory tests in this study showed differences between readmitted and non-readmitted patients during the postoperative period. The common postoperative laboratory values demonstrating discriminatory ability according to readmission status included: white blood cell count, bicarbonate, blood urea nitrogen, and creatinine. On the top panels, the red (green) line shows the median lab values of (non-) readmitted patients. The red (green) shaded area shows the 25th and 75th percentile of the lab values (non-) readmitted patients. For each lab, a one-dimensional SVM is constructed for each postoperative day to separate readmitted and non-readmitted patients. The dashed line plots the SVM boundary. The SVM decision boundaries for these laboratory values are shown in Table 4.2. On the bottom panels, the red (green) line, which corresponds to the right axis, shows the number of (non-) readmitted patients whose lab results available on each day. Note that the number of available lab results dropped in the days after surgery (length of stay). On the 7th day after surgery, about 55% patients had lab taken. The bars (corresponding to the left axis) show the p-value of two sample t-tests for difference in mean and the black line shows the $\alpha = 0.1$ significance level. We included up to seven days of lab results in the model since the p-values were relatively small (indicating significant difference between readmitted and non-readmitted patients) within the first seven days after surgery.

Laboratory Value	Day 0	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7
CO2 (mmol/L, <)	24	26	26	26	26	26	25	25
Creatinine (mg/dL, >)	1.2	1.2	1.1	1	0.98	1	1	1
BUN (mg/dL, >)	18	19	18	17	16	18	19	20
WBC (billion cells/L, >)	12.9	11.4	10.5	8.8	8.6	8.3	8.6	9.2

Table 4.2: Daily Postoperative Laboratory Value Thresholds for Readmission Risk as Determined by Support Vector Machine Learning Techniques

To examine whether including postoperative laboratory data into readmission prediction models would increase predictive ability, we tested several multiple logistic regression models (Figure 4.5) (A) included baseline clinical and demographic values and achieved a c-statistic of 0.52. The value of this c-statistic increased to 0.54 with inclusion of laboratory value thresholds solely from the day of discharge (B). Next, we added daily postoperative laboratory thresholds (within 7 days of surgery) to the

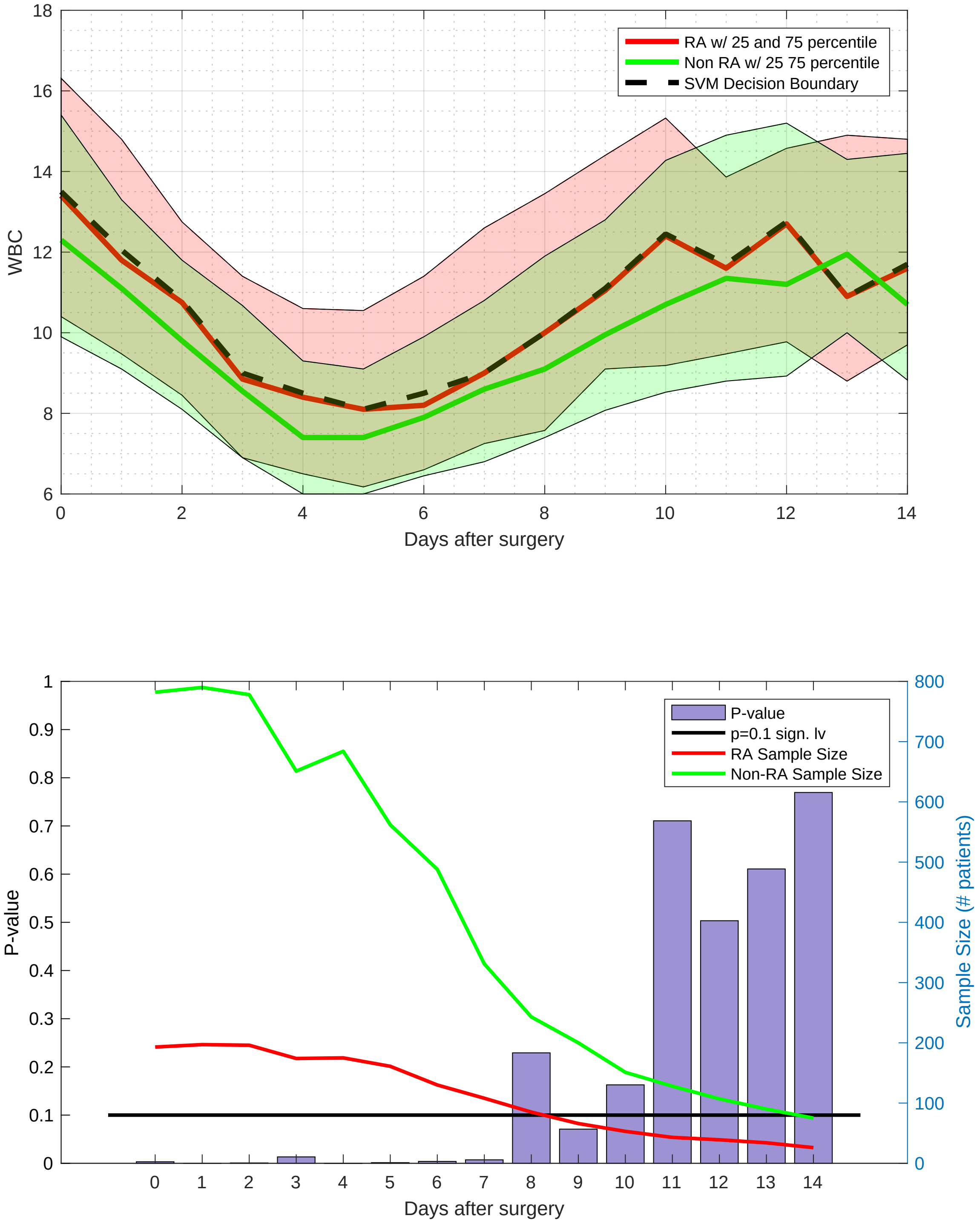


Figure 4.1: Daily WBC Values and Readmission Risk Thresholds during the Postoperative Period after Radical Cystectomy

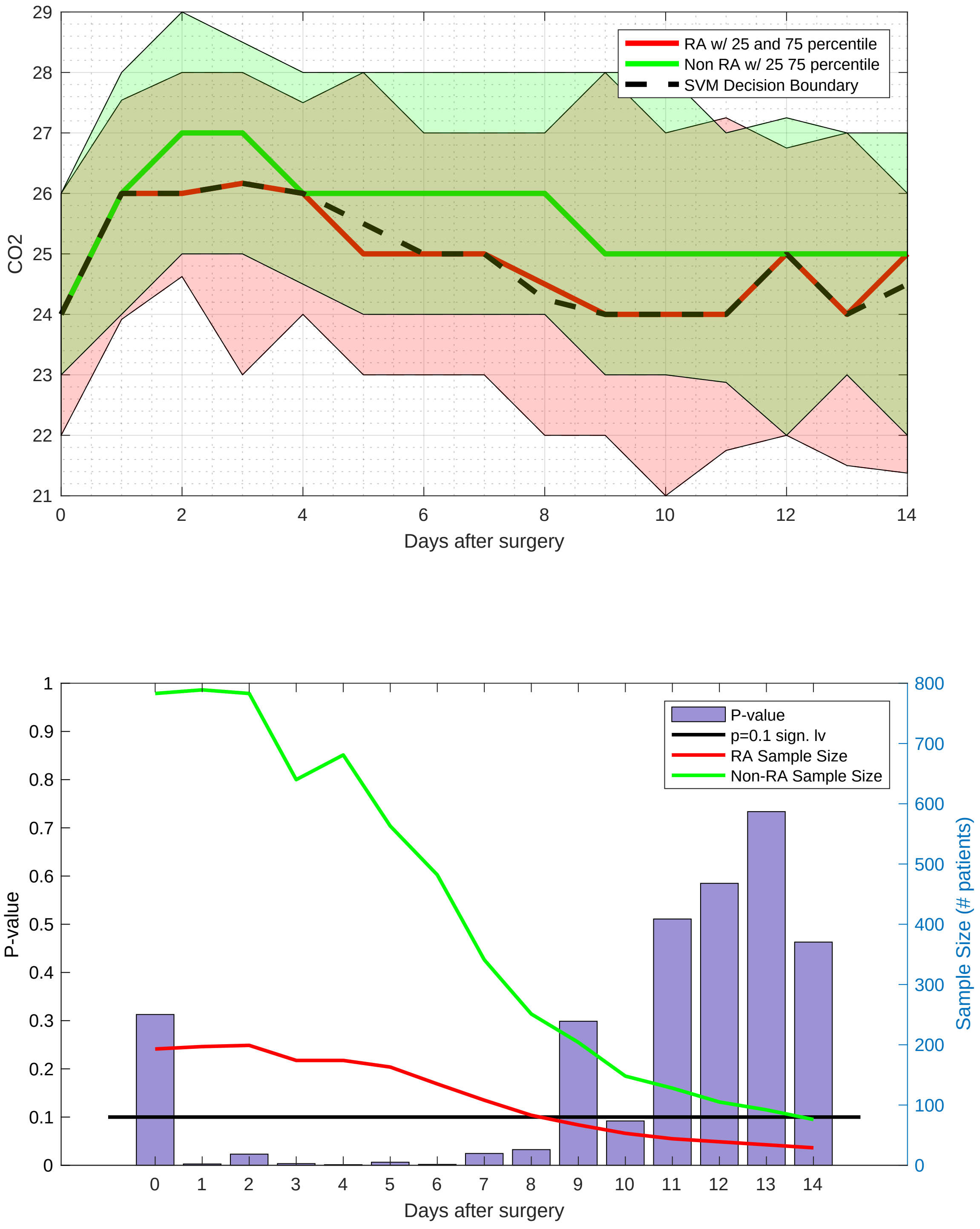


Figure 4.2: Daily CO2 and Readmission Risk Thresholds during the Postoperative Period after Radical Cystectomy

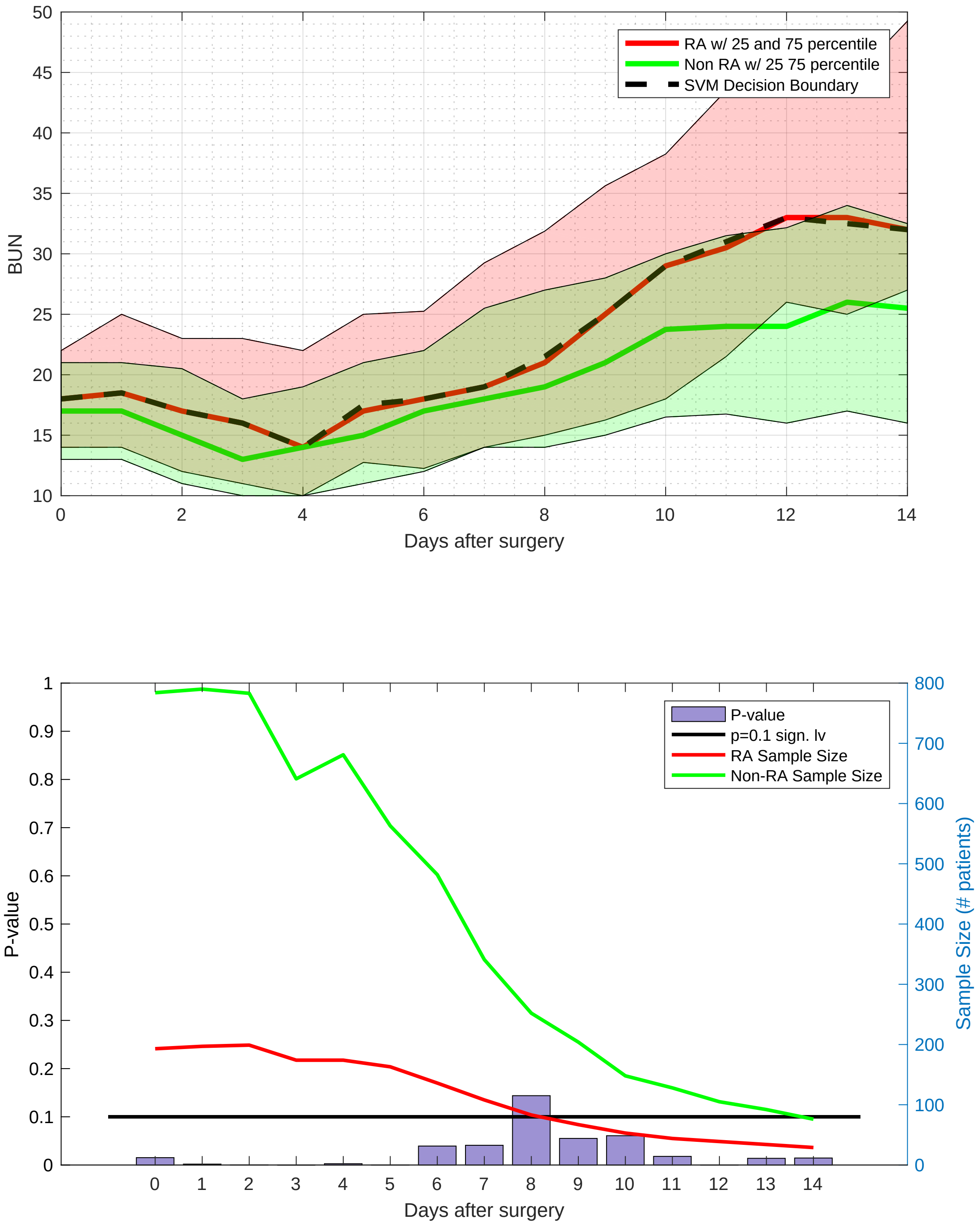


Figure 4.3: Daily BUN Values and Readmission Risk Thresholds during the Postoperative Period after Radical Cystectomy

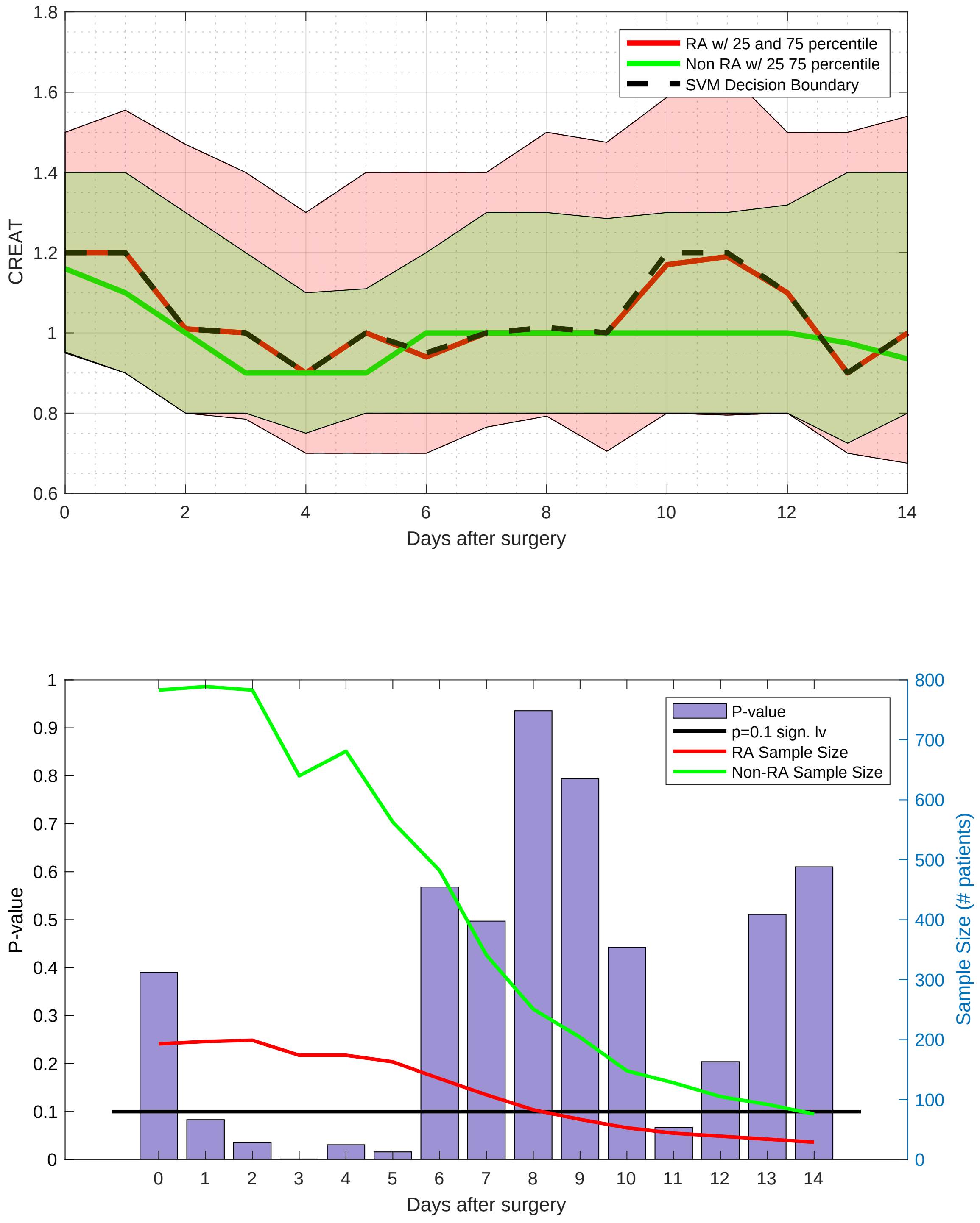


Figure 4.4: Daily CREAT Values and Readmission Risk Thresholds during the Postoperative Period after Radical Cystectomy

demographic and clinical variables to achieve a c-statistic of 0.59 when applied to the validation half of our sample (C). The inclusion of postoperative complications data further increased the c-statistic to 0.62 (D). Lastly, the random forest classification including baseline characteristics, laboratory values, and complications achieved a c-statistic of 0.68 (E). In between model (B) and (C), we included 3, 4, 5, 6, and 7 days of laboratory values incrementally. The c-statistics were 0.55, 0.55, 0.56, 0.56, and 0.59 respectively.

4.6 Discussion

Using machine learning techniques, we found differences in common postoperative laboratory values between readmitted and non-readmitted patients treated with cystectomy. We also calculated threshold values to help differentiate patients at high and low risk of readmission within 30 days of discharge. Moreover, incorporating daily postoperative laboratory value thresholds into our readmission prediction models greatly increased accuracy as measured by the c-statistic, when compared to models using only static demographic and clinical variables. Taken together these findings suggest that inclusion of dynamic sources of physiologic information such as laboratory values into prediction models may allow for important advancements in risk stratification and intervention targeting, especially in the seemingly refractory setting of readmissions after radical cystectomy.

While use of dynamic data points obtained from electronic health record sources is a new approach in the cystectomy literature, recognition and use of this information has been growing in other fields. The utility of EHR data such as laboratory values and vital signs has been established in prediction and risk adjustment for in-hospital mortality (Escobar et al., 2008; Liu et al., 2013; Escobar et al., 2013; Tabak et al., 2013). While these data are used to direct daily clinical decision-making, our cognitive capacity to incorporate subtle trends across multiple factors (e.g., laboratory data, vital signs, medications) and predict clinical decompensation and readmission is limited. For example, we found a trend for higher blood urea nitrogen over time among readmitted patients indicating potential subclinical postoperative dehydration placing patients at risk for readmission. In light of such human information processing limitations, similar methods have been explored in the projection of intensive care admission, extended length of stay, as well as condition-specific risk adjustment (Kipnis et al., 2016; Lim et al., 2015; Escobar et al., 2012; Liu et al., 2010; Smith et al., 2016). More recently, these techniques have used in predicting readmissions in large, generalized cohorts of patients (Escobar et al., 2015). As use

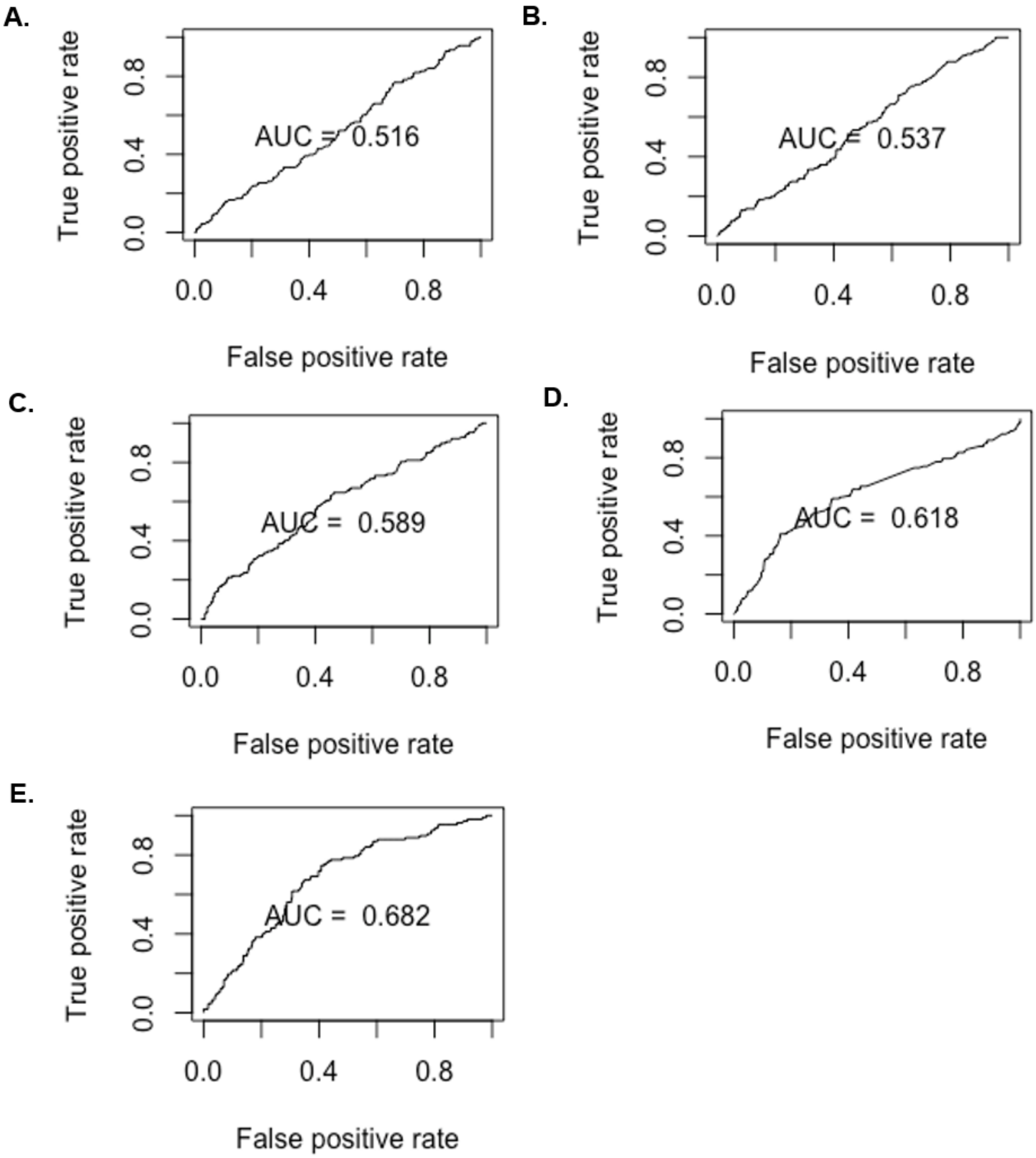


Figure 4.5: Area under the Receiver Operating Curve Showing the Performance of Multiple Logistic Regression Models Incorporating Daily Laboratory Values Tested in a Validation Cohort

of these methods continues to grow, they will likely become more standardized and broadly applicable (Goldstein et al., 2017).

The proximal opportunities for these approaches lie not only in their improved predictive ability, but also in the chance to directly build these algorithms into the electronic medical record systems from which they draw their data. This real-time approach has been piloted with promising results in the identification of patients at risk for death in the hospital (Khurana et al., 2016). More streamlined, albeit simplistic, rehospitalization prediction algorithms such as the LACE index already appear in some interfaces. Our findings suggest that these techniques may hold similar promise in the care of patients following radical cystectomy. Ongoing advances in predictive methodology and incorporation into medical information systems may significantly advance the ability of urological surgeons and their teams to more readily assess patient risk after cystectomy, consequently improving discharge planning and patient outcomes.

We note several important limitations to this study. First, using data from a single institution limits the external generalizability of our findings. However, our results are consistent with robust data from prior studies in different cohorts illustrating the predictive power of physiologic data contained in electronic health record systems. Next, while application of laboratory values in risk stratification for cystectomy patients is unique and improved prediction, we were unable to include vital signs into the models as the data were incomplete. Similarly, this analysis did not incorporate other granular data such as inpatient medications or in-hospital procedures. Nonetheless, existing studies have found the most significant improvements in model performance appear to be realized with inclusion of laboratory values. Last, while this retrospective study did include novel laboratory data, due to the nature of the data source we were unable to include more detailed disease-specific information (e.g., stage) that could impact postoperative care and readmission risk. However, our inclusion of daily postoperative laboratory testing could actually account for greater testing typically associated with more aggressive resections.

In spite of these limitations, this work highlights an opportunity to advance readmission risk assessment following radical cystectomy. If coupled with effective interventions, this could in turn have positive effects on bladder cancer patient outcomes regarding discharge timing decisions and after leaving the hospital. These improvements could also enable urologists to apply these techniques across the range of urological surgery, providing spillover benefits to an even larger patient cohort.

4.7 Conclusion

We found that readmission risk assessment following radical cystectomy is significantly improved by the addition of dynamic physiologic data collected in modern electronic health record systems. Future work should refine these algorithms and study their implementation into daily practice in order to help guide clinical decision making. While the problem of readmission after radical cystectomy appears refractory, innovative, dynamic approaches to existing data sources appear poised to enable significant progress towards risk stratification, ultimately helping ensure patients who are discharged home are able and ready to stay there.

CHAPTER V

Conclusion and Future Research

How to effectively and efficiently reduce hospital readmissions is one of the major challenges faced by healthcare systems. Readmissions burden patients (as well as hospitals and practitioners) and cause a significant amount of unnecessary healthcare spendings. While up to 75% of the readmissions are preventable, we still struggle to understand why unnecessary readmissions happen and how to prevent them from happening.

In this dissertation, we develop operations research models to reduce hospital readmissions. Our approach focuses on both the hospital operations level and the policymaker system level. We develop a delay-time optimization framework to maximize the detection of post-operative complications via post-discharge checkups. Then, we study how to design a bundled payment policy to balance and incentivize pre- and post-discharge readmission reduction efforts. We build a readmission prediction model using laboratory values observed during the index hospitalization. Ultimately, we provide novel methods for reducing readmissions between the pre- and post-discharge stages at the hospital and policymaker levels.

In the following sections, we discuss a potential future research direction for each of the three research works in this thesis. In the post-discharge stage, we ask questions regarding E-visits and their consequences on readmissions. In the pre-discharge stage, a stochastic programming model could be developed to incorporate the learning and reduction of a patient's readmission risk. Between the pre- and post-discharge stages, the management of a panel of patients at various readmission risks could be modeled and solved as a multi-armed bandit problem.

5.1 Future Research: E-Visits and Readmission

The analytical results in Chapter II demonstrate that E-visits and telemedicine visits could serve as replacements for traditional office visits. Nonetheless, the remote nature of E-visits requires more careful investigation.

The adoption of E-visits has significantly increased in recent years. Many believe that E-visits are cost-effective replacements for office visits. By replacing office visits with E-visits, clinics can increase the capacity and patients can receive health care at their convenience. Studies have found that E-visits are cheaper (in terms of the cost of the initial visit, subsequent medical care, and pharmacy) than many other patient-provider encounter modes such as PCP visits, ED visits, retail clinical visits, and urgent care center visits (Gordon et al., 2017). Moreover, the majority of patients (more than 60%) are willing to accept E-visits as shown in a survey of 1,378 ambulatory urology patients (Viers et al., 2015).

Although E-visits are more affordable and convenient, whether they can improve patient health outcomes remains unclear – studies in both the medical literature and the operations management literature have found mixed results (Schoenfeld et al., 2016; Bavafa et al., 2018). Moreover, a study empirically showed that E-visits may trigger 6% more follow-up visits (compared to office visits). This could undermine the premises of adopting E-visits and ultimately result in a 15% reduction in new patients acceptance each month (Bavafa et al., 2018). In the context of readmission reduction, a randomized trial found that phone calls after discharge did not decrease but increased the readmission rate (Auger et al., 2018). Before adopting E-visits widely to replace office visits, researchers need to further research and investigate the nature of E-visits and its consequences.

Specifically, we suggest the following research questions as future work that could be done in this field:

Who (Physicians and Patients) Should Adopt E-Visits? Shaw et al. (2018) pointed out that "whilst some clinicians are very keen to use this format (E-visits), others are reluctant or oppose." Aside from personal preferences (e.g., tech-savviness), what factors influence physicians' willingness to adopt E-visits? What patient cohorts (e.g., medical vs. surgical), what clinical settings (e.g., acute care vs. primary care), and what stages of care (e.g., chronic/routine care vs. post-acute care) should adopt E-visits?

Do Physicians Behave Differently in E-Visits? If So, Do E-Visits Impact Health Outcomes? Due to the remote communication nature of E-visits, physicians may behave more conservatively. A few medical papers (Mehrotra et al., 2013; Uscher-Pines et al., 2016, 2015) found that physicians prescribe more antibiotics in E-visits. Moreover, E-visits may trigger more follow-up tests and visits to resolve the uncertainties about the patient's condition that arise during the E-visit (Bavafa et al., 2018). It is unclear whether E-visits in a post-discharge setting will effectively reduce readmissions. Moreover, if E-visits can detect and intervene readmissions promptly, a potential future step would be to examine whether the readmission intensity (e.g., LOS and cost) is reduced by E-visits.

How do E-visits Affect Physician Productivity? How to Incorporate E-visits into a Physician's (and the Clinic's) Practice and Workflow? It is known that communications within an E-visit are different from a face-to-face encounter. According to Shaw et al. (2018), physicians are more likely to dominate the communication in E-visits. Moreover, both physicians and patients sometimes needed to state things explicitly in a remote consultation that remained implicit in a face-to-face encounter. Moreover, the cognitive tasks and resources that are required in an E-visit also differ. As such, switching between E-visit and office visit may incur a cognitive "switching" cost that degrades the productivity and burdens the physician's cognitively. While designing a physician's clinic schedule, one may want to schedule E-visits back-to-back to avoid such switching cost.

To address these questions, one may first use empirical methods (e.g., difference-in-difference and instrumental variables) to address these questions. One could use data from EHR and insurance claims to build empirical models.

5.2 Future Research: Incorporating Pre-Discharge Readmission Learning and Reduction Intervention

Chapter III focused on the policy-level analysis in a continuum of care spanning between pre- and post-discharge. In this section, we discuss a future research direction that focuses on the patient-level analysis.

As a future research, we propose a patient-level modeling framework for the trade-off faced between pre- and post-discharge readmission learning and reduction as a two-stage model. The problem of interest consists of two stages. The first stage is the inpatient stay stage. In clinical literature, there are mixed results on how the length of stay is associated with readmission risk Engelman et al. (1994); Lahey et al. (1998); Lazar et al. (2001); Bohmer et al. (2002); Cowper et al. (2007); Hannan et al. (2003). In the proposed future research, we could consider two actions in the inpatient stay, namely taking readmission reduction interventions and learning the patient's readmission risk characteristics. Upon discharge, the patient enters into the second stage, which is the post-discharge monitoring stage. Moreover, the post-discharge planning is dynamic, in the sense that we will update our belief on what time the patient will be readmitted dynamically using the information learned in each checkup. For example, if we checked up a patient and the patient was not sick, our belief on readmission timing (i.e., the time-to-readmission probability density curve) should change accordingly.

A partially observable Markovian decision process (POMDP) could be developed to jointly optimize both discharge decisions and post-discharge activities. One may incorporate a sequential learning component that enables the model to create personalized treatment and monitoring policies specific to each individual patient.

5.3 Future Research: Balancing Pre-Discharge Efforts for a Panel of Patients

Chapter IV focused on the prediction of readmission risk for an individual patient. In this section, we propose a model that utilizes the individual readmission risk prediction model for the management of a panel of patients at various risks of readmission.

Programs and initiatives for reducing readmissions have been developed and implemented at U.S. hospitals for almost a decade (Strunin et al., 2007; Mitchell et al., 2010). Particularly, the project called Re-Engineered Discharge (project RED) has been proven to be very effective at reducing hospital readmissions (Jack et al., 2009). However, implementing the project RED involves 12 components which requires the effort from both the physicians and nurses. Implementing project RED to every patient is very resource-intensive and impractical. Moreover, the probability of readmission, the time of the readmission, the cause of readmission, and the intensity of readmission vary across different types of patients (Jacobs et al., 2013; Skolarus et al., 2015). Given constrained resources, it is important to identify the patients that are at higher risk of readmission and focus the readmission reduction effort on the subset of high-risk patients.

There exists an extensive amount of literature in predicting the readmission risk for patients (Kansagara et al., 2011; Helm et al., 2016). However, many existing studies model the prediction and learning process as a static one. They do not utilize the longitudinal clinical data observed during the inpatient stay in a sequential fashion (online learning). Moreover, there does not exist a quantitative model for the readmission management of a panel of patients. We propose, as future work, to develop a sequential learning model to dynamically learn the readmission risk of the patients in the panel and optimally allocate readmission reduction resources across the inpatient stage and the post-discharge monitoring stage. While operations research models have been developed for screening a panel of patients at risk of developing conditions (Lee et al., 2015, 2018) and treating a panel of patients with chronic conditions (Jónasson et al., 2017), there does not exist a quantitative model addressing readmission management for a panel of patients at the patient-level.

5.4 Conclusion

In conclusion, this thesis provides operations research models that aims to reduce hospital readmissions, nonetheless, there is still much work to be done in this field. We propose three potential future research directions and hope that more investigation and research would be done to reduce hospital readmissions. In particular, we propose to study the effect of E-visits, and analyze whether they could help detect and avert readmissions, and how they could be incorporated into clinical practices. We also propose that one could build a patient-level model to personalize the readmission reduction interventions spanning between the pre- and post-discharge stages. As the medical data proliferates, one could incorporate dynamic readmission risk prediction models to manage the readmission risk for a panel of patients. As more attention is given to reducing readmission from both medical and operations research communities, it is our hope that hospital readmissions would be reduced and managed efficiently and effectively.

BIBLIOGRAPHY

BIBLIOGRAPHY

Adida, E., F. Bravo. 2018. Contracts for healthcare referral services: Coordination via outcome-based penalty contracts. *Management Science* doi:10.1287/mnsc.2017.3000.

Adida, E., H. Mamani, S. Nassiri. 2016. Bundled payment vs. fee-for-service: Impact of payment scheme on performance. *Management Science* **63**(5) 1606–1624.

Akaike, H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**(6) 716–723.

Andritsos, D. A., C. S. Tang. 2018. Incentive programs for reducing readmissions when patient care is co-produced. *Production and Operations Management* **27**(6) 999–1020.

Aswani, A., Z.-J. M. Shen, A. Siddiq. 2017. Data-driven incentive design in the medicare shared savings program. *Working Paper* doi:10.2139/ssrn.2808211.

Auger, K. A., S. S. Shah, H. L. Tubbs-Cooley, H. J. Sucharew, J. M. Gold, S. Wade-Murphy, A. M. Statile, K. D. Bell, J. C. Khoury, C. Mangeot, et al. 2018. Effects of a 1-time nurse-led telephone call after pediatric discharge: the H2O II randomized clinical trial. *JAMA Pediatrics* **172**(9) e181482–e181482.

Avdis, E., W. Whitt. 2007. Power algorithms for inverting Lapalce transforms. *INFORMS Journal on Computing* **19**(3) 341–355.

Ayer, T., O. Alagoz, N. K. Stout. 2012. OR Forum – A POMDP approach to personalize mammography screening decisions. *Operations Research* **60**(5) 1019–1034.

Ayer, T., O. Alagoz, N. K. Stout, E. Burnside. 2015. Heterogeneity in womens adherence and its role in optimal breast cancer screening policies. *Management Science* **62**(5) 1339–1362.

Barlow, R., F. Proschan. 1996. *Mathematical Theory of Reliability*, vol. 17. SIAM.

Baron, R. 2010. What's keeping us so busy in primary care? a snapshot from one practice. *New Engl. J. of Med.* **362**(17) 1632–1636.

Bartel, A. P., C. W. Chan, S.-H. H. Kim. 2014. Should hospitals keep their patients longer? the role of inpatient care in reducing post-discharge mortality. Tech. rep., National Bureau of Economic Research. doi:10.3386/w20499.

Bastani, H., M. Bayati, M. Braverman, R. Gummadi, R. Johari. 2016. Analysis of medicare pay-for-performance contracts. *Working Paper* doi:10.2139/ssrn.2839143.

Bavafa, H., L. M. Hitt, C. Terwiesch. 2018. The impact of e-visits on visit frequencies and patient health: Evidence from primary care. *Management Science* doi:10.1287/mnsc.2017.2900.

Bavafa, H., S. Savin, C. Terwiesch. 2017. Redesigning primary care delivery: Customized office revisit intervals and e-visits. *Working Paper* doi:10.2139/ssrn.2363685.

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

Bellone, J., J. Barner, D. Lopez. 2012. Postdischarge interventions by pharmacists and impact on hospital readmission rates. *J. Am. Pharm. Assoc.* **52**(3) 358.

Benbassat, J., M. Taragin. 2000. Hospital readmissions as a measure of quality of health care: advantages and limitations. *Archives of Internal Medicine* **160**(8) 1074–1081.

Bohmer, R. M., J. Newell, D. F. Torchiana. 2002. The effect of decreasing length of stay on discharge destination and readmission after coronary bypass operation. *Surgery* **132**(1) 10–15.

Borza, T., B. L. Jacobs, J. S. Montgomery, A. Z. Weizer, T. M. Morgan, K. S. Hafez, C. T. Lee, B. Y. Li, H. S. Min, C. He, et al. 2017. No differences in population-based readmissions after open and robotic-assisted radical cystectomy: implications for post-discharge care. *Urology* **104** 77–83.

Brailsford, S., P. Harper, J. Sykes. 2012. Incorporating human behaviour in simulation models of screening for breast cancer. *Eur. J. Oper. Res.* **219**(3) 491–507.

Brandeau, M., F. Sainfort, W. Pierskalla. 2004. *Operations research and health care: a handbook of methods and applications*, vol. 70. Springer.

Chan, C., V. Farias, N. Bambos, G. Escobar. 2012. Optimizing intensive care unit discharge decisions with patient readmissions. *Oper. Res.* **60**(6) 1323–1341.

Chen, S., N. Kong, X. Sun, H. Meng, M. Li. 2018. Claims data-driven modeling of hospital time-to-readmission risk with latent heterogeneity. *Health Care Management Science* doi:10.1007/s1072.

Christer, A. 1999. Developments in delay time analysis for modelling plant maintenance. *The Journal of the Operational Research Society* **50**(11) 1120–1137.

Christer, A., N. Jack. 1991. An integral-equation approach for replacement modelling over finite time horizons. *IMA Journal of Management Mathematics* **3**(1) pp. 31–44.

CMS. 2018a. Bundled payments for care improvement (BPCI) initiative: General information — center for medicare & medicaid innovation. goo.gl/4F4z72. (Accessed on 02/14/2018).

CMS. 2018b. Target price specifications model years 1 and 2. https://goo.gl/BDGtXT. (Accessed on 09/14/2018).

Costantino, M., B. Frey, B. Hall, P. Painter. 2013. The influence of a postdischarge intervention on reducing hospital readmissions in a medicare population. *Popul. Health Manag.* **16**(5) 310–316.

Cowper, P. A., E. R. DeLong, E. L. Hannan, L. H. Muhlbaier, B. L. Lytle, R. H. Jones, W. L. Holman, J. J. Pokorny, J. A. Stafford, D. B. Mark, et al. 2007. Is early too early? effect of shorter stays after bypass surgery. *The Annals of Thoracic Surgery* **83**(1) 100–107.

Dagpunar, J. 1994. Some necessary and sufficient conditions for age replacement with non-zero downtimes. *The Journal of the Operational Research Society* **45**(2) pp. 225–229.

D'Amore, J., J. Murray, H. Powers, C. Johnson. 2011. Does telephone follow-up predict patient satisfaction and readmission? *Popul. Health Manag.* **14**(5) 249–255.

Desai, N. R., J. S. Ross, J. Y. Kwon, J. Herrin, K. Dharmarajan, S. M. Bernheim, H. M. Krumholz, L. I. Horwitz. 2016. Association between hospital penalty status under the hospital readmission reduction program and readmission rates for target and nontarget conditions. *JAMA* **316**(24) 2647–2656.

Diehl, S., W. Stute. 1988. Kernel density and hazard function estimation in the presence of censoring. *Journal of Multivariate Analysis* **25**(2) 299–310.

Dudas, V., T. Bookwalter, K. Kerr, S. Pantilat. 2001. The impact of follow-up telephone calls to patients after hospitalization. *The American Journal of Medicine* **111**(9) 26–30.

Engelman, R. M., J. A. Rousou, J. E. Flack, D. W. Deaton, C. B. Humphrey, L. H. Ellison, P. D. Allmendinger, S. G. Owen, P. S. Pekow. 1994. Fast-track recovery of the coronary bypass patient. *The Annals of Thoracic Surgery* **58**(6) 1742–1746.

Erenay, F., O. Alagoz, A. Said. 2014. Optimizing colonoscopy screening for colorectal cancer prevention and surveillance. *Manufacturing & Service Operations Management* **16**(3) 381–400.

Escobar, G. J., M. N. Gardner, J. D. Greene, D. Draper, P. Kipnis. 2013. Risk-adjusting hospital mortality using a comprehensive electronic record in an integrated health care delivery system. *Medical Care* **51**(5) 446–453.

Escobar, G. J., J. D. Greene, P. Scheirer, M. N. Gardner, D. Draper, P. Kipnis. 2008. Risk-adjusting hospital inpatient mortality using automated inpatient, outpatient, and laboratory databases. *Medical Care* **46** 232–239.

Escobar, G. J., J. C. LaGuardia, B. J. Turk, A. Ragins, P. Kipnis, D. Draper. 2012. Early detection of impending physiologic deterioration among patients who are not in intensive care: development of predictive models using data from an automated electronic medical record. *Journal of Hospital Medicine* **7**(5) 388–395.

Escobar, G. J., A. Ragins, P. Scheirer, V. Liu, J. Robles, P. Kipnis. 2015. Nonelective rehospitalizations and postdischarge mortality: predictive models suitable for use in real time. *Medical Care* **53**(11) 916.

Fingar, K., M. Barrett, H. Jiang. 2017. A comparison of all-cause 7-day and 30-day readmissions, 2014. HCUP statistical brief no. 230 URL `goo.gl/wbnvY7`.

Fu, B., W. Wang, X. Shi. 2012. A risk analysis based on a two-stage delayed diagnosis regression model with application to chronic disease progression. *Eur. J. Oper. Res.* **218**(3) 847–855.

Goldstein, B. A., A. M. Navar, M. J. Pencina, J. Ioannidis. 2017. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *Journal of the American Medical Informatics Association* **24**(1) 198–208.

Gordon, A. S., W. C. Adamson, A. R. DeVries. 2017. Virtual visits for acute, nonurgent care: a claims analysis of episode-level utilization. *Journal of Medical Internet Research* **19**(2).

Graham, K. L., A. D. Auerbach, J. L. Schnipper, S. A. Flanders, C. S. Kim, E. J. Robinson, G. W. Ruhnke, L. R. Thomas, S. Kripalani, E. E. Vasilevskis, et al. 2018. Preventability of early versus late hospital readmissions in a national cohort of general medicine patients. *Annals of internal medicine* **168**(11) 766–774.

Green, L., S. Savin, Y. Lu. 2013. Primary care physician shortages could be eliminated through use of teams, nonphysicians, and electronic communication. *Health Affairs* **32**(1) 11–19.

Gretton, A., K. Fukumizu, C. Teo, L. Song, B. Schölkopf, A. Smola. 2007. A kernel statistical test of independence. *Advances in Neural Information Processing Systems*. 585–592.

Güneş, E., E. Örmeci, D. Kunduzcu. 2015. Preventing and diagnosing colorectal cancer with a limited colonoscopy resource. *Prod. Oper. Manag.* **24**(1) 1–20.

Guo, P., C. S. Tang, Y. Wang, M. Zhao. 2016. The impact of reimbursement policy on patient welfare, readmission rate and waiting time in a public healthcare system: Fee-for-service vs bundled payment. *UCLA Anderson School of Management Working Paper* URL `https://goo.gl/SdpiJj`.

Hannan, E. L., M. J. Racz, G. Walford, T. J. Ryan, O. W. Isom, E. Bennett, R. H. Jones. 2003. Predictors of readmission for complications of coronary artery bypass graft surgery. *JAMA* **290**(6) 773–780.

Harper, P., S. Jones. 2005. Mathematical models for the early detection and treatment of colorectal cancer. *Health Care Management Science* **8**(2) 101–109.

He, T., K. Li, M. S. Roberts, A. C. Spaulding, T. Ayer, J. J. Grefenstette, J. Chhatwal. 2016. Prevention of hepatitis c by screening and treatment in us prisons. *Annals of Internal Medicine* **164**(2) 84–92.

Helm, J. E., A. Alaeddini, J. M. Stauffer, K. M. Bretthauer, T. A. Skolarus. 2016. Reducing hospital readmissions by integrating empirical prediction with resource optimization. *Production and Operations Management* **25**(2) 233–257.

Helm, J. E., M. S. Lavieri, M. P. Van Oyen, J. D. Stein, D. C. Musch. 2015. Dynamic forecasting and control algorithms of glaucoma progression for clinician decision support. *Operations Research* **63**(5) 979–999.

Henry J Kaiser Family Foundation. 2015. Hospital adjusted expenses per inpatient day. `https://https://goo.gl/CfxTCe`. (Accessed on 07/06/2018).

Holland, R., J. Battersby, I. Harvey, E. Lenaghan, J. Smith, L. Hay. 2005. Systematic review of multidisciplinary interventions in heart failure. *Heart* **91**(7) 899–906.

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, et al. 2014. Sharpening the focus on causes and timing of readmission after radical cystectomy for bladder cancer. *Cancer* **120**(9) 1409–1416.

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. *Annals of Internal Medicine* **150**(3) 178–187.

Jacobs, B., Y. Zhang, J. Tan, H, Z. Ye, T. Skolarus, B. Hollenbeck. 2013. Hospitalization trends after prostate and bladder surgery: implications of potential payment reforms. *J. Urology* **189**(1) 59–65.

Jacobs, B. L., C. He, B. Y. Li, A. Helfand, N. Krishnan, T. Borza, A. A. Ghaferi, B. K. Hollenbeck, J. E. Helm, M. S. Lavieri, et al. 2017. Variation in readmission expenditures after high-risk surgery. *Journal of Surgical Research* **213** 60–68.

James, A. C., J. P. Izard, S. K. Holt, J. K. Calvert, J. L. Wright, M. P. Porter, J. L. Gore. 2016. Root causes and modifiability of 30-day hospital readmissions after radical cystectomy for bladder cancer. *The Journal of Urology* **195**(4) 894–899.

James, J. 2013. Health policy brief: Medicare hospital readmissions reduction program. *Health Affairs* doi:10.1377/hpb20131112.646839.

Jardine, A., A. Tsang. 2005. *Maintenance, Replacement, and Reliability: Theory and Applications*. Dekker Mechanical Engineering, Taylor & Francis.

Jencks, S. F., M. V. Williams, E. A. Coleman. 2009. Rehospitalizations among patients in the medicare fee-for-service program. *New England Journal of Medicine* **360**(14) 1418–1428.

Jónasson, J. O., S. Deo, J. Gallien. 2017. Improving hiv early infant diagnosis supply chains in sub-saharan africa: Models and application to mozambique. *Operations Research* **65**(6) 1479–1493.

Joynt, K. E., A. K. Jha. 2012. Thirty-day readmissions – truth and consequences. *New Engl. J. Med.* **366**(15) 1366–1369.

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.

Kelly, K. N., J. C. Iannuzzi, C. T. Aquina, C. P. Probst, K. Noyes, J. R. Monson, F. J. Fleming. 2015. Timing of discharge: a key to understanding the reason for readmission after colorectal surgery. *Journal of Gastrointestinal Surgery* **19**(3) 418–428.

Kent, D., R. Shachter, H. Sox, N. Hui, L. Shortliffe, S. Moynihan, F. Torti. 1989. Efficient scheduling of cystoscopies in monitoring for recurrent bladder cancer. *Med. Decis. Making* **9**(1) 26–37.

Khurana, H. S., R. H. Groves Jr, M. P. Simons, M. Martin, B. Stoffer, S. Kou, R. Gerkin, E. Reiman, S. Parthasarathy. 2016. Real-time automated sampling of electronic medical records predicts hospital mortality. *The American journal of medicine* **129**(7) 688–698.

Kilroy, C., D. Morgan-Solomon, P. Landrum. 2013. Preventing patient rebounds: Value-based care organizations should focus on more than just readmissions. `goo.gl/gcDCgb`. (Accessed on 03/06/2018).

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

Kipnis, P., B. J. Turk, D. A. Wulf, J. C. LaGuardia, V. Liu, M. M. Churpek, S. Romero-Brufau, G. J. Escobar. 2016. Development and validation of an electronic medical record-based alert score for detection of inpatient deterioration outside the icu. *Journal of Biomedical Informatics* **64** 10–19.

Kirk, P. S., X. Liu, T. Borza, B. Y. Li, M. Sessine, K. Zhu, Y. Qin, B. Jacobs, K. Urish, J. Helm, S. Gilbert, A. Weizer, J. Montgomery, B. K. Hollenbeck, M. Lavieri, T. A. Skolarus. 2018. Dynamic readmission prediction using routine post-operative laboratory results after radical cystectomy. *Working Paper* .

Kocher, K., B. Nallamothu, J. Birkmeyer, J. Dimick. 2013. Emergency department visits after surgery are common for medicare patients, suggesting opportunities to improve care. *Health Affairs* **32**(9) 1600–1607.

Koh, H., K. Sebelius. 2010. Promoting prevention through the affordable care act. *New Engl. J. Med.* **363**(14) 1296–1299.

Krishnan, N., X. Liu, M. S. Lavieri, M. Hu, A. Helfand, B. Li, J. E. Helm, C. He, B. K. Hollenbeck, T. A. Skolarus, et al. 2016. A model to optimize followup care and reduce hospital readmissions after radical cystectomy. *The Journal of Urology* **195**(5) 1362–1367.

Lahey, S. J., C. T. Campos, B. Jennings, P. Pawlow, T. Stokes, S. Levitsky. 1998. Hospital readmission after cardiac surgery. does "fast track" cardiac surgery result in cost saving or cost shifting? *Circulation* **98**(19 Suppl) II35–40.

Lazar, H. L., C. A. Fitzgerald, T. Ahmad, Y. Bao, T. Colton, O. M. Shapira, R. J. Shemin. 2001. Early discharge after coronary artery bypass graft surgery: Are patients really going home earlier? *The Journal of Thoracic and Cardiovascular Surgery* **121**(5) 943–950.

Lee, A. J., X. Liu, T. Borza, Y. Qin, B. Y. Li, K. L. Urish, P. S. Kirk, S. Gilbert, B. K. Hollenbeck, J. E. Helm, M. S. Lavieri, T. A. Skolarus, B. L. Jacobs. 2019. Role of postacute care on hospital readmission after high-risk surgery. *Journal of Surgical Research* **234** 116 – 122.

Lee, E., M. S. Lavieri, M. Volk. 2018. Optimal screening for hepatocellular carcinoma: A restless bandit model. *Manufacturing & Service Operations Management* doi: 10.1287/msom.2017.0697.

Lee, E., M. S. Lavieri, M. L. Volk, Y. Xu. 2015. Applying reinforcement learning techniques to detect hepatocellular carcinoma under limited screening capacity. *Health Care Management Science* **18**(3) 363–375.

Leeds, I., V. Sadiraj, J. Cox, S. Gao, T. Pawlik, K. Schnier, J. Sweeney. 2015. Discharge decision-making after complex surgery: Surgeon behavior compared to predictive modeling to reduce surgical readmissions. *J. Am. Coll. Surgeons* **221**(4) S123–S124.

Leow, J. J., A. P. Cole, T. Seisen, J. Bellmunt, M. Mossanen, M. Menon, M. A. Preston, T. K. Choueiri, A. S. Kibel, B. I. Chung, et al. 2018. Variations in the costs of radical cystectomy for bladder cancer in the usa. *European Urology* **73**(3) 374–382.

Leow, J. J., S. Reese, Q.-D. Trinh, J. Bellmunt, B. I. Chung, A. S. Kibel, S. L. Chang. 2015. Impact of surgeon volume on the morbidity and costs of radical cystectomy in the usa: a contemporary population-based analysis. *BJU International* **115**(5) 713–721.

Lim, E., Y. Cheng, C. Reuschel, O. Mbowe, H. J. Ahn, D. T. Juarez, J. Miyamura, T. B. Seto, J. J. Chen. 2015. Risk-adjusted in-hospital mortality models for congestive heart failure and acute myocardial infarction: Value of clinical laboratory data and race/ethnicity. *Health Services Research* **50**(S1) 1351–1371.

Liu, V., P. Kipnis, M. K. Gould, G. J. Escobar. 2010. Length of stay predictions: improvements through the use of automated laboratory and comorbidity variables. *Medical Care* **48** 739–744.

Liu, V., B. J. Turk, A. I. Ragins, P. Kipnis, G. J. Escobar. 2013. An electronic simplified acute physiology score-based risk adjustment score for critical illness in an integrated healthcare system. *Critical Care Medicine* **41**(1) 41–48.

Liu, X., M. Hu, J. E. Helm, M. S. Lavieri, T. A. Skolarus. 2018a. Missed opportunities in preventing hospital readmissions: Redesigning post-discharge checkup policies. *Production and Operations Management* doi:10.1111/poms.12858.

Liu, X., M. S. Lavieri, J. E. Helm, T. A. Skolarus. 2018b. Time for accountability: Are readmission responsibility windows too long? *Working Paper* URL `https://ssrn.com/abstract=3250443`.

Lyness, J., G. Giunta. 1986. A modification of the weeks method for numerical inversion of the Lapalce transform. *Mathematics of Computation* **47**(175) 313–322.

Maillart, L., J. Ivy, S. Ransom, K. Diehl. 2008. Assessing dynamic breast cancer screening policies. *Operations Research* **56**(6) 1411–1427.

Meeker, W. Q., L. A. Escobar. 2014. *Statistical methods for reliability data*. John Wiley & Sons.

Mehrotra, A., S. Paone, G. D. Martich, S. M. Albert, G. J. Shevchik. 2013. A comparison of care at e-visits and physician office visits for sinusitis and urinary tract infection. *JAMA Internal Medicine* **173**(1) 72–74.

Milioni, A. Z., S. R. Pliska. 1988. Optimal inspection under semi-markovian deterioration: The catastrophic case. *Naval Research Logistics* **35**(5) 393–411.

Minnillo, B. J., M. J. Maurice, N. Schiltz, A. C. Pillai, S. M. Koroukian, F. Daneshgari, S. P. Kim, R. Abouassaly. 2015. Few modifiable factors predict readmission following radical cystectomy. *Canadian Urological Association Journal* **9**(7-8) E439.

Misky, G. J., H. L. Wald, E. A. Coleman. 2010. Post-hospitalization transitions: examining the effects of timing of primary care provider follow-up. *Journal of Hospital Medicine* **5**(7) 392–397.

Mitchell, S., M. Paasche-Orlow, S. Forsythe, V. Chetty, J. O'Donnell, J. Greenwald, L. Culpepper, B. Jack. 2010. Post-discharge hospital utilization among adult medical inpatients with depressive symptoms. *Journal of Hospital Medicine* **5**(7) 378–384.

Myers, E., D. McCrory, K. Nanda, L. Bastian, D. Matchar. 2000. Mathematical model for the natural history of human papillomavirus infection and cervical carcinogenesis. *American Journal of Epidemiology* **151**(12) 1158–1171.

Özekici, S., S. R. Pliska. 1991. Optimal scheduling of inspections: A delayed markov model with false positives and negatives. *Operations Research* **39**(2) 261–273.

PerryUndem Research & Communications. 2013. The revolving door: a report on u.s. hospital readmissions. https://goo.gl/xn1rBX.

Pinsky, P. 2004. An early-and late-stage convolution model for disease natural history. *Biometrics* **60**(1) 191–198.

Puterman, M. L. 2005. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons.

PwC Health Research Institute. 2010. The price of excess: Identifying waste in healthcare spending. `https://goo.gl/iMbvod`. (Accessed on 10/07/2018).

Rizzardi, M. 1995. A modification of talbot's method for the simultaneous approximation of several values of the inverse Lapalce transform. *ACM T. Math. Software* **21**(4) 347–371.

Rosen, J. E., M. C. Salazar, K. Dharmarajan, A. W. Kim, F. C. Detterbeck, D. J. Boffa. 2017. Length of stay from the hospital perspective: practice of early discharge is not associated with increased readmission risk after lung cancer surgery. *Annals of Surgery* **266**(2) 383–388.

Sanders, G., A. Bayoumi, V. Sundaram, S. Bilir, C. Neukermans, C. Rydzak, L. Douglass, L. Lazzeroni, M. Holodniy, D. Owens. 2005. Cost-effectiveness of screening for hiv in the era of highly active antiretroviral therapy. *New Engl. J. Med.* **352**(6) 570–585.

Schoenfeld, A. J., J. M. Davies, B. J. Marafino, M. Dean, C. DeJong, N. S. Bardach, D. S. Kazi, W. J. Boscardin, G. A. Lin, R. Duseja, et al. 2016. Variation in quality of urgent health care provided during commercial virtual visits. *JAMA Internal Medicine* **176**(5) 635–642.

Shafiee, M., S. Chukova. 2013a. Maintenance models in warranty: A literature review. *European Journal of Operational Research* **229**(3) 561–572.

Shafiee, M., S. Chukova. 2013b. Optimal upgrade strategy, warranty policy and sale price for second-hand products. *Applied Stochastic Models in Business and Industry* **29**(2) 157–169.

Shaw, S. E., D. Cameron, J. Wherton, L. M. Seuren, S. Vijayaraghavan, S. Bhattacharya, C. ACourt, J. Morris, T. Greenhalgh. 2018. Technology-enhanced consultations in diabetes, cancer, and heart failure: Protocol for the qualitative analysis of remote consultations (QuARC) project. *JMIR Research Protocols* **7**(7) e10913–e10913.

Sim, S., J. Endrenyi. 1993. A failure-repair model with minimal and major maintenance. *IEEE Transactions on Reliability* **42**(1) 134–140.

Skolarus, T., B. Jacobs, F. Schroeck, C. He, A. Helfand, J. Helm, M. Hu, M. Lavieri, B. Hollenbeck. 2015. Understanding hospital readmission intensity after radical cystectomy. *The Journal of Urology* **193**(5) 1500–1506.

Smith, M. W., P. L. Owens, R. M. Andrews, C. A. Steiner, R. M. Coffey, H. G. Skinner, J. Miyamura, I. Popescu. 2016. Differences in severity at admission for heart failure between rural and urban patients: the value of adding laboratory results to administrative data. *BMC Health Services Research* **16**(1) 133.

Stimson, C., S. S. Chang, D. A. Barocas, J. E. Humphrey, S. G. Patel, P. E. Clark, J. A. Smith Jr, M. S. Cookson. 2010. Early and late perioperative outcomes following radical cystectomy: 90-day readmissions, morbidity and mortality in a contemporary series. *The Journal of Urology* **184**(4) 1296–1300.

Strunin, L., M. Stone, B. Jack. 2007. Understanding rehospitalization risk: can hospital discharge be modified to reduce recurrent hospitalization? *Journal of Hospital Medicine* **2**(5) 297–304.

Tabak, Y. P., X. Sun, C. M. Nunez, R. S. Johannes. 2013. Using electronic health record data to develop inpatient mortality predictive model: Acute laboratory risk of mortality score (ALaRMS). *Journal of the American Medical Informatics Association* **21**(3) 455–463.

Tagliente, I., L. Trieste, T. Solvoll, F. Murgia, S. Bella. 2016. Telemonitoring in cystic fibrosis: A 4-year assessment and simulation for the next 6 years. *Interactive Journal of Medical Research* **5**(2) e11–e11.

Tan, H.-J., J. S. Wolf Jr, Z. Ye, J. T. Wei, D. C. Miller. 2011. Complications and failure to rescue after laparoscopic versus open radical nephrectomy. *The Journal of Urology* **186**(4) 1254–1260.

Tenet Healthcare. 2015. Law department policy no.6: Hospital-provided post-discharge assistance to federal health care program beneficiaries. `goo.gl/3XXowv`. (Accessed on 03/08/2018).

Teng, Y., L. Han, W. Tu, N. Kong. 2011. Optimizing coverage for a chlamydia trachomatis screening program. *2011 IEEE International Conference on Automation Science and Engineering*. 531–536. doi:10.1109/CASE.2011.6042465.

Tsodikov, A., A. Szabo, J. Wegelin. 2006. A population model of prostate cancer incidence. *Statistics in Medicine* **25**(16) 2846–2866.

Uscher-Pines, L., A. Mulcahy, D. Cowling, G. Hunter, R. Burns, A. Mehrotra. 2015. Antibiotic prescribing for acute respiratory infections in direct-to-consumer telemedicine visits. *JAMA Internal Medicine* **175**(7) 1234–1235.

Uscher-Pines, L., A. Mulcahy, D. Cowling, G. Hunter, R. Burns, A. Mehrotra. 2016. Access and quality of care in direct-to-consumer telemedicine. *Telemedicine and e-Health* **22**(4) 282–287.

Viers, B. R., S. Pruthi, M. E. Rivera, D. A. O'Neil, M. R. Gardner, S. M. Jenkins, D. J. Lightner, M. T. Gettman. 2015. Are patients willing to engage in telemedicine for their care: a survey of preuse perceptions and acceptance of remote video visits in a urological patient population. *Urology* **85**(6) 1233–1240.

Wang, H. 2002. A survey of maintenance policies of deteriorating systems. *European Journal of Operational Research* **139**(3) 469–489.

Wang, W. 2012. An overview of the recent advances in delay-time-based maintenance modelling. *Reliability Engineering & System Safety* **106** 165–178.

Weinberger, M., E. Oddone, W. Henderson. 1996. Does increased access to primary care reduce hospital readmissions? *New Engl. J. Med.* **334**(22) 1441–1447.

White, C. 1977. A markov quality control process subject to partial observation. *Management Science* **23**(8) 843–852.

Wong, F. K. Y., S. K. Y. Chow, T. M. F. Chan, S. K. F. Tam. 2013. Comparison of effects between home visits with telephone calls and telephone calls only for transitional discharge support: a randomised controlled trial. *Age and Ageing* **43**(1) 91–97.

Yeh, R. 1997. Optimal inspection and replacement policies for multi-state deteriorating systems. *Eur. J. Oper. Res.* **96**(2) 248–259.

Zhang, D. J., I. Gurvich, J. A. Van Mieghem, E. Park, R. S. Young, M. V. Williams. 2016. Hospital readmissions reduction program: An economic and operational analysis. *Management Science* **62**(11) 3351–3371.

Zhang, J., B. Denton, H. Balasubramanian, N. Shah, B. Inman. 2012a. Optimization of prostate biopsy referral decisions. *Manufacturing & Service Operations Management* **14**(4) 529–547.

Zhang, S., P. Hanagal, P. Frazier, A. Meltzer, D. Schneider. 2012b. Optimal patient-specific post-operative surveillance for vascular surgery. *Proceedings of the 7th INFORMS Workshop on Data Mining and Health Informatics* URL `https://goo.gl/DYMWa5`.