Stochastic Control and Optimization Methods for Chronic Disease Monitoring and Control, Hospital Staffing, and Surgery Scheduling

by

Pooyan Kazemian

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
2016

Doctoral Committee:

Professor Mark P. Van Oyen, Chair
Associate Professor Brian T. Denton
Assistant Professor Jonathan E. Helm
Assistant Professor Mariel S. Lavieri
Professor Demosthenis Teneketzis

To my parents Azar and Mahmood with deepest gratitude

for their endless love and continuous support.

ACKNOWLEDGEMENTS

I would like to sincerely thank my advisor, Professor Mark Van Oyen for his incredible guidance, innumerous support and encouragement, and for everything he has taught me throughout the tenure of my Ph.D. study, without which this dissertation would not have been completed. I have also been greatly benefitted to work closely with Professor Mariel Lavieri, Professor Jonathan Helm, and Dr. Joshua Stein on the main chapter of my thesis related to monitoring and control of glaucoma. Their contribution has been essential in successfully completing this dissertation. I would also like to thank my other committee members Professor Demosthenis Teneketzis and Professor Biran Denton who have always been supportive of my work and provided valuable comments, suggestions and guidance during the course of this dissertation.

My special gratitude goes to Mrs. Merrill Bonder and the Seth Bonder Foundation for their wholehearted support, love, and inspiration. The support from Seth Bonder Foundation has been critical in successfully finishing this dissertation.

During the past five years, I have had the privilege to work with several collaborators from the Robert D. and Patricia E. Kern Center for the Science of Health Care Delivery at the Mayo Clinic in Rochester, MN. In particular, I would like to acknowledge the contribution and support of Professors Mustafa Sir, Kalyan Pasupathy, Susan Hallbeck, and Thomas Rohleder, Dr. Yue Dong, as well as Todd Huschka. I am really grateful for all of their support and guidance.

Further, I like to thank all of my previous instructors from the University of Tehran, Rutgers University, and the University of Michigan for their endless efforts inside and

outside the classroom. I am also thankful to Professor Melike Baykal-Gursoy for her advice and guidance during my two years at Rutgers.

I would also like to thank my brother Peyman Kazemian for his unconditional help and support, my friends Kamran Paynabar and Nima Salehi, and my officemates Maya Bam and Kayse Lee Maass for all the memories.

Finally, I dedicate this dissertation to my parents Azar Rahnavardi Azari and Mahmood Kazemian for their endless love, and unlimited supports, endurance and encouragement.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

CHAPTER I

Introduction

The United States has the most expensive healthcare system in the world with healthcare spending reaching $3.0 trillion or $9,523 per person in 2014 (*Centers for Medicare & Medicaid Services* 2014). However, reports consistently demonstrate that the U.S. underperforms on most dimensions of healthcare performance compared to other countries. Two critical challenges facing healthcare providers, and more broadly society, are controlling the cost of providing care to patients and improving the quality of the outcomes. This dissertation addresses these challenges at the patient level by developing stochastic control and optimization methods to better personalize the medical care and the delivery of care. The main unifying theme of this dissertation is the unprecedented use of data/information to design innovative decision support tools that keep the focus of care on the patients, their experience, and their outcomes and reduction of wasted resources. In other words, we leverage operations research techniques to develop patient-centered optimization models and decision methods to improve the quality of care, patient safety, and timeliness of access to care at lower cost, which all will result in better patient outcomes and therefore, a healthier society. This dissertation is presented in a multiple manuscript format. The results in Chapters II, III and IV have appeared as individual research papers *Kazemian et al.* (2016a), *Kazemian et al.* (2014), and *Kazemian et al.* (2016b).

In Chapter II, we take a step toward improving the quality of care and cost containment by integrating and personalizing the monitoring and treatment of chronic diseases. Chronic diseases affect almost one out of every two adults (*Ward et al.* 2014) and the treatment of them accounts for the expenditure of over 75% of direct healthcare costs in the U.S. (*Thrall* 2005). To effectively manage chronic disease patients, clinicians must know (1) how to monitor each patient, and (2) how to control the disease. In the treatment of chronic diseases, patients are quantitatively tested at prescribed intervals to monitor the degree of disease progression and subsequently whether a change in treatment is warranted. Hence, there are decisions on which tests to take and when to take them, as well as what treatments/interventions are needed. Tests are associated with costs to account for side-effects, discomfort, and inconvenience as well as the monetary cost of the test itself. Chapter II presents a dynamic personalized modeling framework that enables clinicians to (1) specify the optimal timing of each office visit and the appropriate subset of tests to perform at that visit considering the costs and value of each test and the uncertainty about the patient's disease state so as to achieve an accurate perception of disease progression without burdening patients and the healthcare system with over-testing [disease monitoring], and (2) identify optimal target levels for controllable disease risk factors to slow the rate of disease progression to an acceptable level without over-treating or under-treating the patient [treatment control]. We do so by providing the jointly optimal solution to a novel linear quadratic Gaussian state space model. For the new objective of minimizing the relative change in state over time (i.e., disease progression), which is necessary for management of irreversible chronic diseases, we show that the classical two-way separation of estimation and control holds, thereby making a previously intractable problem solvable by decomposition into two separate, tractable problems while maintaining optimality. The resulting optimization is applied to the management of glaucoma. Based on data from two large randomized clinical trials,

we validate our model and demonstrate how our decision support tool can provide actionable insights to the clinician caring for a patient with glaucoma. This methodology can be applied to a broad range of irreversible chronic diseases to optimally devise patient-specific monitoring and treatment plans. It can assist in managing the tradeoff between maximizing information about the disease and its control and the amount of healthcare resources (e.g., office visits and tests) to be provided. Further, this approach can broaden the quality of care to more people because it elevates the care even when provided by non-experts (e.g., optometrists and general ophthalmologists taking care of glaucoma patients). Therefore, the statistical and optimization framework developed in this work has the potential for broad impact on longitudinal patient care as well as cost containment.

Chapter III focuses on improving patient safety and therefore, improving health outcomes. This research was motivated by the new Accreditation Council for Graduate Medical Education (ACGME) duty-hour standards for residents and fellows that went into effect in 2011. These regulations were designed to reduce fatigue-related medical errors and improve patient safety. The new shift restrictions, however, have led to more frequent transitions in patient care (handoffs), resulting in greater opportunity for communication breakdowns between caregivers, which correlate with medical errors and adverse events. Recent research has focused on improving the quality of these transitions through standardization of the handoff protocols; however, no attention has been given to reducing the number of transitions in patient care. Chapter III leverages integer programming methods to design a work shift schedule for trainees that minimizes the number of error-prone patient handoffs which will result in fewer medical errors due to communication breakdowns. The new schedule complies with all ACGME duty-hour standards, provides required coverage, and maintains physician quality of life. We add constraints on physician's sleep hours, circadian rhythm and other human factors issues to reduce the fatigue-related medical errors and fur-

ther improve the patient safety. Moreover, this approach can reduce healthcare costs by optimizing the number of providers required. In a case study of redesigning the trainees' schedule for a Mayo Clinic Intensive Care Unit (ICU), we show that the number of patient handoffs can be reduced by 23% and still meet all required and most desired scheduling constraints. Furthermore, a 48% reduction in handoffs could be achieved if only the minimum required rules are satisfied.

Chapter IV addresses timely access to care and coordinated care delivery. Providing timely access to surgery is crucial for patients with high acuity diseases like cancer. In Chapter IV, we present a methodological framework to make efficient use of scarce resources including surgeons, operating rooms, and clinic time slots to meet the access delay service level for elective surgery using colorectal surgery (CRS) at the Mayo Clinic as a case study. We personalize the system to offer different tiers of access based on the acuity of patient's disease so that the more urgent cases are treated more promptly. Further, we increase the level of performance of the system to reduce waste of precious operating room time. In this chapter, we propose and evaluate 6 heuristic scheduling policies. All policies dramatically outperform the current scheduling protocol. The underlying idea behind these scheduling policies is the efficient and innovative use of patient information to tentatively book a surgery day at the same time the clinic appointment is set. We develop a logistic regression model to calculate for each clinic appointment order the probability that it will result in a surgery in CRS. Then, we creatively space out the clinic and surgery appointments such that if the patient does not need his/her surgery appointment, we have enough time to offer that to another patient. We develop a 2-stage stochastic and dynamic discrete-event simulation model to evaluate the 6 scheduling policies. In the first stage of the simulation, these policies are compared in terms of the average operating room overtime per day. The second stage involves fine-tuning the winning policy. This methodology is applied to historical patient data from Mayo CRS. Numerical results

Table 1.1: Overview of the dissertation topics and their main contributions.

	Topic	Main Contributions
Chapter II	Integrating and personalizing chronic disease monitoring and treatment (with application to glaucoma)	(1) Introduce a new objective for LQG state space modeling to minimize the relative change in state (i.e., disease progression) (2) Prove the two-way separation of optimal state estimation and control (separation principle) (3) Reduce measurement noise via Kalman filter and smoother (4) Provide the jointly optimal solution to both disease monitoring and treatment control problems in a feedback-driven control model (5) Personalize disease monitoring to avoid over/under-testing (i.e., when test and which subset of tests to take) (6) Personalize treatment to avoid over/under-treatment (i.e., what levels of controllable risk factors to target to sufficiently slow disease progression) (7) Develop a decision support tool for management of glaucoma
Chapter III	Designing new work shift schedules for healthcare providers to reduce the number of patient handoffs	(1) Integrate provider shift "design" and "assignment" (2) Minimize the number of error-prone patient handoffs, therefore, improve patient safety (3) Develop livable and equitable work shift schedules that provide the required coverage with minimum providers (4) Present application to a medical ICU
Chapter IV	Coordinating clinic and surgery appointments to reduce indirect wait time to elective surgery	(1) Proactively book a surgery appointment at the time a clinic appointment is booked (2) Creatively space out the clinic and surgery appointments to handle surgery cancelations (3) Propose six heuristic scheduling policies that outperform the current policy in terms of average operating room overtime (4) Evaluate and fine-tune the policies using a 2-stage stochastic simulation algorithm

demonstrate that the optimal policy performs 52% better than the current scheduling policy. Table 1.1 outlines the three main research topics of this dissertation and enumerates the main contributions of each chapter.

CHAPTER II

Dynamic Monitoring and Control of Irreversible Chronic Diseases with Application to Glaucoma

2.1 Introduction

Chronic diseases are the leading cause of death and disability and affect almost one out of every two adults in the United States (*Ward et al.*, 2014). In the management of chronic diseases, patients are quantitatively tested at prescribed intervals using a selected set of testing modalities to assess disease progression and decide whether a change in treatment is warranted. In this context, proper testing and treatment guidance is critical to both cost containment and patient outcomes in the management of chronic disease. In this chapter we develop a modeling framework for dynamic management of irreversible chronic diseases that enables us to (1) specify the optimal timing of each office visit and the appropriate suite of tests (i.e., the selection of testing modalities) to perform at that visit considering the costs and value of each test and the uncertainty about the patient's disease progression [disease monitoring], and (2) identify optimal target levels for controllable disease risk factors to slow the rate of disease progression without over-treating or under-treating the patient [treatment control].

To do so, we introduce and solve a new type of objective function for linear quadratic

Gaussian (LQG) systems that minimizes the relative change in state (i.e., disease progression) rather than the traditional objective of minimizing the cost of being in each state. We extend LQG theory by proving for this new objective that the classical two-way separation of optimal state estimation and control applies. This separation ensures computational tractability for the simultaneous optimization of disease monitoring and treatment control. This innovative modeling of dynamic disease monitoring and treatment control is developed generally to be applicable to many irreversible chronic diseases. As a proof of concept, we demonstrate the capabilities of this methodology by applying it to glaucoma, a chronic disease causing progressive blindness.

2.1.1 Scope of the Research

It is important to distinguish the disease monitoring problem from screening for a disease. The goal of disease screening is to determine whether or not a patient has a particular disease. A screening test is taken when the patient is considered to be at some risk of developing a condition but exhibits no symptoms of the illness. In contrast, for the disease monitoring problem the patient is already known to have the disease and the goal is to quickly detect the presence of disease progression and identify whether/how to adjust the treatment plan to slow/avert further disease progression.

For the treatment control portion of the problem, the goal is to determine the time-dependent intensity of treatment over a treatment cycle based on dynamically updating information on patient disease state from the monitoring portion. We emphasize that our model does not suggest a specific intervention. Rather, it provides patient-specific target levels for controllable/modifiable disease risk factors that help guide the doctor in selecting an appropriate treatment plan for the patient. Though one might try to model how each intervention affects the disease progression dynamics,

we feel it best to leave it to the clinician to employ his/her experience and expertise to decide what therapeutic interventions are most likely able to achieve the target levels suggested by our model.

2.1.2 Main Contributions

- **Theoretical:** (1) To the best of our knowledge, this is the first research to employ measurement adaptive systems theory to monitoring and control of chronic diseases (or even to any healthcare operations research problem), and this new application requires a major extension of the underlying theory. We extend the LQG state space modeling literature by introducing a new objective that minimizes the *relative change* in system state over time (i.e., the difference in estimated state elements between the current period and the previous period), rather than minimizing the cost of *current* state. In the previous literature, the goal of the controller has been to keep the system state on a *static* desired trajectory using costly control actions by minimizing the deviation of the current system position from the desired trajectory. However, in irreversible diseases such as glaucoma, once the disease has progressed, it is biologically impossible to recover the damage. In the context, the desired trajectory is to maintain the "current disease state position" (i.e. stop the disease from worsening), which the model *dynamically* updates as the disease progresses over time. This necessitates a new structure for the objective function (Eq. 2.3) not yet studied in LQG literature.

  (2) For LQG systems theory, the two-way separation of optimal state estimation and control (known as the *separation principle*) has been a critical foundation upon which to tractably and simultaneously optimize estimation and control of probabilistic systems (see *Witsenhausen* 1971). Our main theoretical results show that the two-way separation of optimal estimation and control extends

to this new objective of relative system state change, which involves two correlated state variables from the current and previous time periods. The treatment control can be optimized in closed form as a linear function of the best estimate of the patient's current disease state (i.e., filtered state mean given by the Kalman filter) via completion of squares. Furthermore, we show that the monitoring problem can be reduced to a continuous-state discrete-action Markov decision process (MDP) model with filtered and smoothed covariance matrices of the state serving as the information state and the Kalman filter and smoother equations acting as the system dynamics. The MDP can be solved via branch-and-bound dynamic programming to find the optimal monitoring schedule specific to each individual patient.

(3) Kalman filter and smoother are built into our modeling framework to extract noise from the raw measurements and to optimally estimate the disease state in each time period based on imperfect/noisy measurements. This is a key to accurately identifying genuine disease progression from testing artifacts. Kalman smoother is a new feature in our model (compared with the traditional LQG models), which is essential because of the new objective function we employ. State smoothing means using information gained at time t to update the prior estimate made at $t-1$ of the value of the state at $t-1$; filtering refers to estimating the current disease state based on new test results.

- **Practical:** (1) We develop an integrated, feedback-driven stochastic control model to provide the jointly optimal solution to both the disease monitoring and treatment control problems. It is worth noting that the disease control problem is affected by the monitoring regime because as new tests are performed, more information about the patient's disease state and disease dynamics become available. Such information can affect how the doctor controls/slows the progression of the disease. Therefore, it is critical to model and solve the disease

monitoring and control problems together to capture the interaction between them.

(2) The model explicitly determines which suite of tests to take at each time period. Some tests are significantly easier and cheaper to do than others. Different tests may provide less or more information about the patient's disease state. Therefore, it is important to be able to differentiate which tests to do at each time point in terms of improved monitoring and cost containment.

(3) We develop a data-driven decision support tool that provides a *menu of options* to the doctor based on how aggressively he/she wants to monitor and control the patient. The doctor can select an appropriate aggressiveness option depending on the patient's life expectancy, severity of disease, and other personal and clinical factors. For each aggressiveness option, the model incorporates new and past test results as well as clinically-believed and data-verified disease dynamics to predict and graph the future disease trajectory and recommend a patient specific monitoring regime and target level for controllable disease risk factors.

- **Data:** (1) We parametrize and validate our model using data from two landmark randomized clinical trials of patients with glaucoma. Our numerical results confirm that the model demonstrates low error in predicting the future disease trajectory.

  (2) Our model demonstrates potential for improving both patient outcomes and system cost when applied to patients from the clinical trial, which are already receiving a high level of care. This potential is likely greater for patients being treated by non-glaucoma specialists.

2.2 Literature Review

Papers relevant to this research are classified into three categories: (1) theoretical papers on measurement adaptive systems and sensor scheduling, (2) medical decision making papers on disease screening, diagnosis, and monitoring, and (3) optimization models on treatment planning and disease control. In this section we highlight some prominent papers in each category and briefly describe how our research methodologies and objectives are different.

Measurement Adaptive Systems and Sensor Scheduling: The closest paper to our work in terms of theory is *Meier et al.* (1967). This paper lays the foundations for measurement adaptive systems in which the controller tries to keep the system state on a *static* desired trajectory and simultaneously obtain information about the system state with minimum total cost over a finite horizon. They show that in the special case of discrete-time systems, linear system dynamics, quadratic cost of the *current state*, and Gaussian random noise processes, the problem of finding the optimal measurement policy reduces to the solution of a nonlinear, deterministic control problem. *Baron and Kleinman* (1969) extend their work to continuous-time measurements and investigate the optimal measurement duration for a human operator. *Bansal and Başar* (1989) provide an extension of this framework to the infinite-horizon setting with discounted costs. Our work differs in that it deals with a *dynamic* desired trajectory, minimizing the relative change in state in each time period (i.e., disease progression), which is essential for the management of irreversible chronic diseases as discussed in Subsection 2.1.2. For example, experiments using the model provided by *Meier et al.* (1967) lead to results that were considered clinically incorrect/unbelievable in the experience of our clinical co-author, a glaucoma specialist.

There is also extensive literature on sensor scheduling problems, in which a set of sensors is used to estimate a stochastic process, but because of cost or design con-

straints only one or a subset of them takes measurements at each time point. *Athans* (1972) considers the problem in which the controller has to select, at each time step, one measurement provided by one sensor out of many available sensors (with different measurement costs), such that a weighted combination of prediction accuracy and accumulated observation cost is minimized. Examples of other work in this area include *Gupta et al.* (2006), *Mehra* (1976), and *Vitus et al.* (2012). However, these papers differ from ours in that they do not consider the tradeoff between exploration vs. exploitation.

Disease Screening, Diagnosis, and Monitoring: While there is an extensive literature on disease screening and diagnosis problems, there is relatively little work on the disease monitoring problem that we defined in Section 2.1.1. *Helm et al.* (2015) and *Schell et al.* (2014) provide a heuristic approach for finding the time to next test based on patient's probability of progression. *Ayer et al.* (2012) provide a partially observable Markov decision process (POMDP) approach to personalize mammography screening decisions based on the prior screening history and personal risk characteristics of women. *Chhatwal et al.* (2010) develop a finite-horizon discrete-time Markov decision process (MDP) model to help radiologists determine the best time for biopsy based on the initial mammography findings and patient demographics to maximize a patient's total expected quality-adjusted life years. The works of *Yang et al.* (2013), *Mangasarian et al.* (1995), and *Saaty and Vargas* (1998) are other examples of disease screening models. These works differ from ours in that they focus on screening problem where the goal is to detect the presence of a particular disease with minimum delay. They do not provide any insights on how to monitor the patient if the presence of the disease is confirmed and progression trajectory can be monitored over time; nor do they consider treatment planning.

Treatment Planning and Disease Control: There has been a variety of works considering when to start treatment of a patient when the presence of disease is con-

firmed (also known as surveillance problems). *Lavieri et al.* (2012) develop a Kalman filter-based approach to help clinicians decide when to start radiation therapy in patients with prostate cancer based on predictions of the time when the patient's prostate specific antigen (PSA) level reaches its lowest point. *Shechter et al.* (2008) employ Markov decision processes (MDP) to optimize the time to initiate HIV treatment to maximize quality-adjusted life years of a patient. *Mason et al.* (2014) present an MDP model to determine the optimal timing of blood pressure and cholesterol medications. All of these works assume a measurement of the patient's health is taken periodically. Our work differs in that it solves the joint problem of optimal timing of each test and optimal treatment control.

Moreover, in most of the previous research mentioned, the patient's disease dynamics are assumed to be known or are estimated from population-based models. In our model, the population data is integrated with individual patient measurements that are gathered from sequential testing so that the predictions and decisions made are unique to each patient. Capturing the complex patient disease dynamics requires incorporating several health indices into the state vector. We employ a continuous state space that easily accommodates multivariate states (e.g. 9 dimensions in our model for glaucoma) and provide the jointly optimal solutions to both disease monitoring and control problems. Employing a continuous state space model is important as many quantitative tests for disease monitoring are in continuum. Problems with such a multivariate, continuous state-space often become intractable for MDP-based approaches due to the curse of dimensionality. Discretization of the state space and using approximate dynamic programming (ADP) to mitigate the curse of dimensionality of MDP models is an alternative approach when our modeling framework does not fit. For example, strongly discrete state variables, highly non-linear disease dynamics, and highly non-Gaussian random noises are features that are difficult for our model to handle.

2.3 The Modeling Framework

A continuous state space model is employed at the heart of our modeling framework with two key components: (1) a state transition process to model disease progression dynamics, and (2) a measurement/testing process to model how the true disease state is observed. Both processes (Eq.'s 2.1 and 2.2) are in the form of first order vector difference equations with additive Gaussian white noise (i.e., noise inputs at time t and t' are independent).

2.3.1 State Transition Process

The recursive state transition equation for our N-stage time horizon is given by

$$\alpha_{t+1} = \mathrm{T}_t\alpha_t + G_t\beta_t + \eta_t, \qquad t = 1, ..., N, \tag{2.1}$$

where α_t is the random variable representing the state of the disease at time t, β_t is the "disease control" variable administered at time t, η_t is the vector of Gaussian white noise that represents unmodeled disease process noise with $E[\eta_t] = 0$ and $Var(\eta_t) = Q_t$, T_t is the state transition matrix governing the underlying disease progression dynamics and G_t is a vector capturing the effect of disease control variable β_t on the next period state, α_{t+1}. β_t is one of the two optimization variables of the model. It determines how the modifiable disease risk factors should be adjusted at time t to optimally slow the progression of disease. Having such information will help clinicians select the appropriate treatment plan for the patient.

2.3.2 Measurement/Testing Process

The measurement equation gives the relationship between the true disease state, α_t, and the noisy raw reading/observation, z_t, as follows.

$$z_t = Z_t \alpha_t + \varepsilon_t, \qquad t = 1, ..., N, \tag{2.2}$$

where z_t is the observation vector (i.e., the result of test(s) performed on the patient), Z_t is the observation matrix and determines how components of the true state are observed, ε_t is the multi-variate Gaussian white test noise with $E[\varepsilon_t] = 0$ and $Var(\varepsilon_t) = H_t^{(\theta_t)}$. θ_t is the "test/measurement control" variable that determines which subset of tests to take in period t and is the other control variable the model optimizes. $H_t^{(\theta_t)}$ models the error associated with the tests and is directly affected by the decision on which test(s) to take at time t (which we highlight by adding (θ_t) to the superscript, i.e., $H_t^{(\theta_t)}$). It is worth noting that both α_{t+1} and z_t are Gaussian random vectors since they are a linear combination of independent Gaussian random variables. The initial state, α_0, is Gaussian with $E[\alpha_0] = \hat{\alpha}_0 = \hat{\alpha}_{1|0}$ and $Var(\alpha_0) = \hat{\Sigma}_0 = \hat{\Sigma}_{1|0}$. The random variables α_0, $\{\eta_t\}$, and $\{\varepsilon_t\}$ are mutually independent. Throughout the chapter, the notation $\hat{X}_{t|t'}$ means the estimated value of random variable X at time t with information up to time t'.

2.3.3 Objective Function

The novel objective function (performance criterion) we analyze is given by

$$J = E\left\{\sum_{t=1}^{N}\left[(\alpha_t - \alpha_{t-1})' A_t (\alpha_t - \alpha_{t-1}) + \beta_t' B_t \beta_t + l_t(\theta_t)\right] + (\alpha_{N+1} - \alpha_N)' A_{N+1} (\alpha_{N+1} - \alpha_N)\right\}, \tag{2.3}$$

in which A_t is the unit cost matrix of further worsening of the disease, B_t is the unit cost of administering disease control (i.e., further adjusting the modifiable disease risk factor), and the scalar $l_t(\theta_t)$ is the cost of taking tests/measurements in period t, which depends on the test control variable, θ_t.

The objective function consists of four terms: (1) $(\alpha_t - \alpha_{t-1})' A_t (\alpha_t - \alpha_{t-1})$ is the cost of relative change in the system state random variable (i.e., disease progression) between the previous period $t-1$ and the current period t (whereas the traditional LQG objective minimizes ${\alpha_t}' A_t \alpha_t$ as explained in Subsection 2.1.2), (2) ${\beta_t}' B_t \beta_t$ is the cost of controlling the disease risk factors including side effects and complications of medical or surgical treatments, (3) $l_t(\theta_t)$ is the cost of undergoing additional testing, and (4) $(\alpha_{N+1} - \alpha_N)' A_{N+1} (\alpha_{N+1} - \alpha_N)$ is the terminal cost of relative state change at the end of the treatment horizon. The quadratic form of the first part of the objective function ensures that a large disease worsening is penalized more aggressively than a small one. Furthermore, achieving a large adjustment in disease risk factors may require more aggressive treatments (e.g. surgery or laser therapy), which are associated with higher monetary costs and more side effects and discomfort than a smaller change in risk factors, which can often be achieved by simpler treatments like medications. Hence, the cost associated with a big relative change in patient's disease risk factors is much higher than a small one, and so the quadratic form of the second part of the objective function is a good choice for our application.

2.3.4 Kalman Filter and Kalman Smoother

When the state transition and measurement processes are both in the form of first order difference equations with Gaussian white noises, the optimal state estimation (to minimize the mean squared error of the estimate) is given by the Kalman filter (*Kalman* 1960). The Kalman filter obtains the prediction of state mean and covariance at time t with information up to time $t-1$, $\hat{\alpha}_{t|t-1}$ and $\hat{\Sigma}_{t|t-1}$ respectively, and the

current reading, z_t, as inputs to the algorithm and calculates the filtered state (i.e., optimal estimate of the true state) mean and covariance, $\hat{\alpha}_{t|t}$ and $\hat{\Sigma}_{t|t}$ respectively. The optimal state mean estimate at time t with information up to time t, $\hat{\alpha}_{t|t}$, is given by

$$\hat{\alpha}_{t|t} = \hat{\alpha}_{t|t-1} + K_t \tilde{y}_t, \tag{2.4}$$

where $\hat{\alpha}_{t|t-1}$ is the predicted state mean at time t given information up to time $t-1$ and $\tilde{y}_t$ is the measurement residual (error) given by

$$\hat{\alpha}_{t|t-1} = T_{t-1}\hat{\alpha}_{t-1|t-1} + G_{t-1}\beta_{t-1}, \tag{2.5}$$

$$\tilde{y}_t = z_t - Z_t(T_{t-1}\hat{\alpha}_{t-1|t-1} + G_{t-1}\beta_{t-1}), \tag{2.6}$$

and K_t is the Kalman gain given by

$$K_t = \hat{\Sigma}_{t|t-1} Z_t{}' S_t^{-1}, \tag{2.7}$$

in which S_t is the predicted covariance around the measurement given by $Z_t \hat{\Sigma}_{t|t-1} Z_t{}' + H_t^{(\theta_t)}$.

The predicted state covariance at time t given the information up to time $t-1$, $\hat{\Sigma}_{t|t-1}$, and the most recent state covariance estimate at time t with information up to time t, $\hat{\Sigma}_{t|t}$, satisfy

$$\hat{\Sigma}_{t|t} = \hat{\Sigma}_{t|t-1} - \hat{\Sigma}_{t|t-1} Z_t{}' \left(Z_t \hat{\Sigma}_{t|t-1} Z_t{}' + H_t^{(\theta_t)} \right)^{-1} Z_t \hat{\Sigma}_{t|t-1} = \left(I - K_t Z_t\right) \hat{\Sigma}_{t|t-1}, \tag{2.8}$$

$$\hat{\Sigma}_{t|t-1} = T_{t-1} \hat{\Sigma}_{t-1|t-1} T_{t-1}{}' + Q_{t-1}. \tag{2.9}$$

The initial state mean and covariance, $\hat{\alpha}_{1|0}$ and $\hat{\Sigma}_{1|0}$ respectively, are calculated based on population data from clinical trials. For more discussion on Kalman filter please

see *Bertsekas et al.* (1995) and *Harvey* (1990).

Because of the special form of the objective function that minimizes relative state change from time $t-1$ to time t (i.e., disease progression) we need to refine the estimation of previous state mean and covariance after a new measurement is taken at time t ($\hat{\alpha}_{t-1|t}$ and $\hat{\Sigma}_{t-1|t}$ respectively). This is called state smoothing and can be done via fixed-interval Kalman smoother as follows.

$$\hat{\alpha}_{t-1|t} = \hat{\alpha}_{t-1|t-1} + \hat{\Sigma}^*_{t-1}\left(\hat{\alpha}_{t|t} - \hat{\alpha}_{t|t-1}\right), \tag{2.10}$$

$$\hat{\Sigma}_{t-1|t} = \hat{\Sigma}_{t-1|t-1} + \hat{\Sigma}^*_{t-1}\left(\hat{\Sigma}_{t|t} - \hat{\Sigma}_{t|t-1}\right)\hat{\Sigma}^{*}_{t-1}{}', \tag{2.11}$$

in which $\hat{\Sigma}^*_{t-1} = \hat{\Sigma}_{t-1|t-1}T_{t-1}{}'\hat{\Sigma}^{-1}_{t|t-1}$. A derivation of the fixed-interval Kalman smoothing can be found in *Ansley and Kohn* (1982).

The control system block diagram is depicted in Figure 2.1. The dashed arrows indicate that the information is carried over from the current period, t, to the next period, $t+1$. The values in parentheses are not observable. Suppose the patient is in disease state α_t when visiting the doctor's office. Based on the optimal test control action θ^*_t (already determined in the previous time period), all or a subset of tests are performed on the patient. The noisy observation/reading, z_t, is then sent to the Kalman Filter. Based on the predicted and observed states, the Kalman Filter algorithm calculates the best estimate of the mean and covariance of the patient's disease state in period t, $\hat{\alpha}_{t|t}$ and $\hat{\Sigma}_{t|t}$ respectively, and sends the filtered values to both the Kalman Smoother and the controller (i.e. the decision support system itself for this analysis). The Kalman Smoother will then modify the best estimates of the state mean and covariance in period $t-1$, $\hat{\alpha}_{t-1|t}$ and $\hat{\Sigma}_{t-1|t}$ respectively, and send the smoothed values to the controller. Notice that this is a key departure from the traditional methodology. The controller receives both the filtered and smoothed values of the patient's disease state mean and covariance (the information state for the

optimization component of the model) and outputs the optimal treatment and test control actions, β_t^* and θ_{t+1}^*. Finally, the prediction of the state mean and covariance in period $t+1$, $\hat{\alpha}_{t+1|t}$ and $\hat{\Sigma}_{t+1|t}$, is sent to the Kalman filter and smoother to be used in the following time period.

In Section 2.4, we focus on the controller (decision support tool) and show how the optimal disease and test control actions (β_t^* and θ_{t+1}^*, respectively) can be calculated given the information state $\wp_t = \left(\hat{\alpha}_{t|t}, \hat{\Sigma}_{t|t}, \hat{\alpha}_{t-1|t}, \hat{\Sigma}_{t-1|t}\right)$.

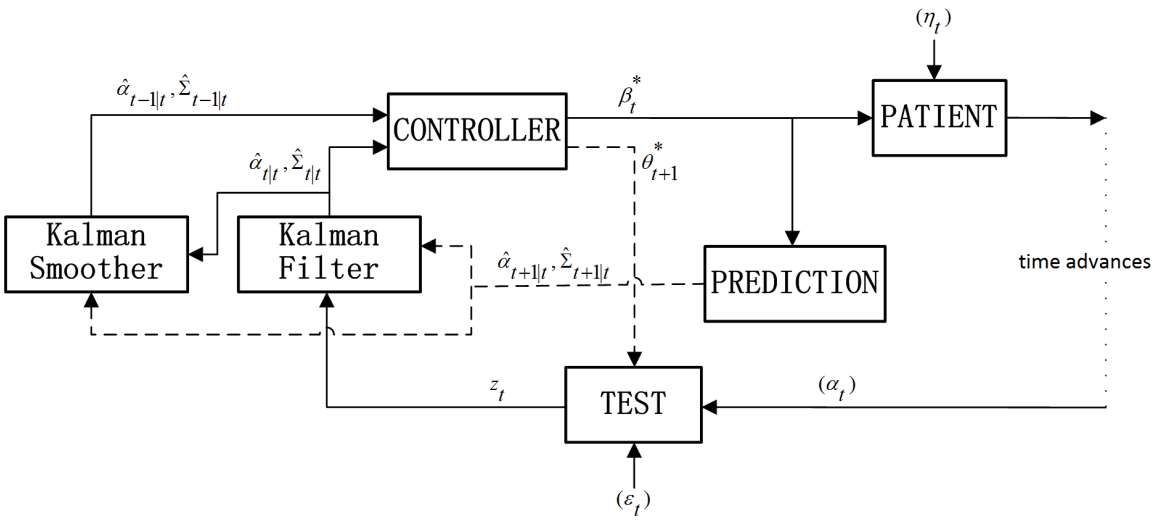

Figure 2.1: Control system block diagram.

2.3.5 Separation of Estimation and Control

One-way Separation of Estimation and Control: A control law is a function selected by the controller from the set of all admissible functions based on all data available at the time of the decision. This function generates a control action to be applied to the system. The problem is to make an optimal selection of such functions for all time steps that achieves the minimum expected cost (defined by the objective function) for the control horizon of the problem. For the general stochastic control problem with imperfect observations, given all the observations and previous control

actions, state estimation from noisy data is always independent of the control law. This is because the conditional density of the state given all the observations and the control actions is independent of the control law. This result is called the separation principle in control theory (see *Witsenhausen* 1971). The only underlying assumption for the separation principle to hold is to have one controller (i.e., centralized information) with perfect recall (i.e., the information on the previous observations and control actions do not get lost).

In general, the control law depends on the estimation of the system state; but, the estimation at time t is independent of all control laws given all observations up to time t and all the control actions up to time $t-1$. This is also known as the one-way separation of estimation and control. Since our LQG model is a special case of the general stochastic control problem with centralized information and perfect recall, the one-way separation principle holds. As seen in Section 2.3.4, the optimal state estimation at time t, described by $\hat{\alpha}_{t|t}$ and $\hat{\Sigma}_{t|t}$, is given by the Kalman filter (Eq.'s 2.4-2.9) and is independent of the control law given all the previous observations and control actions.

Two-way Separation of Estimation and Control: For LQG stochastic systems in which (1) the transition and measurement equations are linear in state and control action, (2) the objective function penalizes the quadratic cost of *current* state, and (3) the state and measurement noises are Gaussian, it has been shown that the control law is also independent of the state estimation (*Meier et al.* 1967). Therefore, for this traditional form of LQG models we have two-way separation of the estimation and the control; namely, the estimation is independent of the control law and the control law is independent of the estimation. In Section 2.4, we show for the new objective of minimizing the *relative change* in state, which involves two correlated state variables of current and previous time periods and requires smoothing in addition to filtering and prediction, that the optimal control law is still independent of state es-

timation. Thus, in this new and more complex environment, the two-way separation still holds. Furthermore, the optimal control action is linear in state estimation. This is extremely desirable for application because the control law is data independent and can be calculated offline (which greatly reduces the computational burden). The two-way separation of estimation and control for this special case of LQG models is a fundamental finding, which is critical to solution tractability.

2.4 Derivation of Optimal Disease and Test Controls

In this section we derive the optimal disease and test control actions given the information state at time t, $\wp_t$, which is defined as the filtered state mean and covariance at time t with information up to time t and the smoothed state mean and covariance at time $t-1$ with information up to time t, i.e. $\wp_t = \left(\hat{\alpha}_{t|t}, \hat{\Sigma}_{t|t}, \hat{\alpha}_{t-1|t}, \hat{\Sigma}_{t-1|t}\right)$.
In terms of $\wp_t$, a dynamic programming algorithm can be derived to find the optimum disease and test controls. The value function, $V_t(\wp_t)$, can be found recursively as follows.

$$V_t(\wp_t) = \min_{\beta_t, \theta_{t+1}} \left\{ L_t(\wp_t, \beta_t, \theta_{t+1}) + \underset{z_{t+1}}{E} \left[V_{t+1}(\wp_{t+1}) \right] \right\}, \quad t = 1, \ldots, N-1, \tag{2.12}$$

where $V_t(\wp_t)$ is the minimum expected cost from period t to N, the end of the control horizon, given the information state $\wp_t$, and $L_t(\wp_t, \beta_t, \theta_{t+1})$ is the expected instantaneous cost incurred in period t if the information state is $\wp_t$ and the control actions β_t and θ_{t+1} are chosen, given by

$$L_t(\wp_t, \beta_t, \theta_{t+1}) = E\left[(\alpha_t - \alpha_{t-1})' A_t (\alpha_t - \alpha_{t-1}) \right] + \beta_t' B_t \beta_t + l_{t+1}(\theta_{t+1}). \tag{2.13}$$

The boundary condition is given by

$$V_N\left(\wp_N\right) = \min_{\beta_N}\left\{\tilde{L}_N(\wp_N, \beta_N)\right\}, \tag{2.14}$$

where $\tilde{L}_N(\wp_N, \beta_N)$ is the expected cost incurred in the final period N if the information state is $\wp_N$ and the disease control action β_N is chosen.

The minimum cost during the entire control horizon can, therefore, be obtained by

$$J^* = l_1(\theta_1) + V_1(\wp_1), \tag{2.15}$$

in which $l_1(\theta_1)$ is the cost of initial tests during the patient's first office visit and $V_1(\wp_1)$ is the minimum cost to go from period 1 to the end of control horizon obtained recursively via Eq.'s 2.12 and 2.14. We assume all diagnostic tests are taken at the first visit.

In the remainder of this section we use an induction argument to prove the following theorems.

Theorem II.1. *For arbitrary time t ($t = 1, \ldots, N$), the control law is independent of the state estimation (i.e., we have two-way separation of optimal estimation and control). Moreover, the optimal disease control, β_t^*, is linear in the filtered state mean, $\hat{\alpha}_{t|t}$.*

Theorem II.2. *At an arbitrary time t ($t = 1, \ldots, N$), the optimal monitoring schedule, $\theta_{t+1}^*, \theta_{t+2}^*, \ldots$, can be found by solving a continuous-state discrete-action MDP model with filtered and smoothed covariance matrices of the state serving as the information state and the Kalman filter and smoother equations acting as the system dynamics.*

Theorem II.3. *For arbitrary time t ($t = 1, \ldots, N$), the value function with infor-*

mation up to time t has the following form.

$$V_t(\wp_t) = \left(\hat{\alpha}_{t|t} - \hat{\alpha}_{t-1|t}\right)' A_t \left(\hat{\alpha}_{t|t} - \hat{\alpha}_{t-1|t}\right) + tr\left[A_t \hat{\Sigma}_{t|t}\right] + \hat{\alpha}'_{t|t} P_t \hat{\alpha}_{t|t} + tr\left[P_t \hat{\Sigma}_{t|t}\right] + V_t^{\theta}\left(\wp_t^{\theta}\right) + b_t, \tag{2.16}$$

in which $V_t^{\theta}\left(\wp_t^{\theta}\right)$ represents the recursive terms that only depend on measurement control actions, i.e., when to take tests and which test(s) to take. They do not depend on the observations or on the disease control actions. Therefore, the measurement control problem can be solved separately from the treatment control problem. $\wp_t^{\theta}$ represents those elements of information state that are only affected by measurement control actions (i.e., $\wp_t^{\theta} = \left(\hat{\Sigma}_{t|t}, \hat{\Sigma}_{t-1|t}\right)$), and b_t is a constant. $V_t^{\theta}\left(\wp_t^{\theta}\right)$ and b_t will be obtained later in the proof.

Proof by induction: In Appendix 2.7.1 we prove that the value function in the final period is given by

$$V_N(\wp_N) = \left(\hat{\alpha}_{N|N} - \hat{\alpha}_{N-1|N}\right)' A_N \left(\hat{\alpha}_{N|N} - \hat{\alpha}_{N-1|N}\right) + tr\left[A_N \hat{\Sigma}_{N|N}\right] + \hat{\alpha}'_{N|N} P_N \hat{\alpha}_{N|N} + tr\left[P_N \hat{\Sigma}_{N|N}\right] + tr\left[A_N\left(\hat{\Sigma}_{N-1|N} - T_{N-1}\hat{\Sigma}_{N-1|N} - \hat{\Sigma}_{N-1|N}{T_{N-1}}'\right)\right] + tr\left[\tilde{P}_{N+1}\hat{\Sigma}_{N|N}\right] + tr\left[A_{N+1}Q_N\right], \tag{2.17}$$

where tr represents the trace of the matrix. By comparing Eq.'s 2.16 and 2.17 we see that (induction basis)

$$V_N^{\theta}\left(\wp_N^{\theta}\right) = tr\left[A_N\left(\hat{\Sigma}_{N-1|N} - T_{N-1}\hat{\Sigma}_{N-1|N} - \hat{\Sigma}_{N-1|N}{T_{N-1}}'\right)\right] + tr\left[\tilde{P}_{N+1}\hat{\Sigma}_{N|N}\right], \tag{2.18}$$

$$b_N = tr\left[A_{N+1}Q_N\right]. \tag{2.19}$$

Assume (induction hypothesis)

$$V_{t+1}\left(\wp_{t+1}\right) = \left(\hat{\alpha}_{t+1|t+1} - \hat{\alpha}_{t|t+1}\right)' A_{t+1} \left(\hat{\alpha}_{t+1|t+1} - \hat{\alpha}_{t|t+1}\right) + tr\left[A_{t+1}\hat{\Sigma}_{t+1|t+1}\right] + \hat{\alpha}'_{t+1|t+1} P_{t+1} \hat{\alpha}_{t+1|t+1} + tr\left[P_{t+1}\hat{\Sigma}_{t+1|t+1}\right] + V^{\theta}_{t+1}\left(\wp^{\theta}_{t+1}\right) + b_{t+1}. \quad (2.20)$$

We show that $V_t\left(\wp_t\right)$ follows the form given in Eq. 2.16 (induction step).

From Eq. 2.12 we know the general form of value function is

$$V_t\left(\wp_t\right) = \min_{\beta_t, \theta_{t+1}} \left\{ L_t(\wp_t, \beta_t, \theta_{t+1}) + \underset{z_{t+1}}{E}\left[V_{t+1}\left(\wp_{t+1}\right)\right] \right\}, \quad (2.21)$$

in which the information state at time $t+1$, $\wp_{t+1}$, is a function of $\wp_t$, β_t, θ_{t+1}, and z_{t+1}. The expected instantaneous cost in period t, $L_t(\wp_t, \beta_t, \theta_{t+1})$, is given in Eq. 2.13. Application of Lemma II.6 to the expectation in Eq. 2.13 results in

$$\begin{aligned} L_t(\wp_t, \beta_t, \theta_{t+1}) = {} & \left(\hat{\alpha}_{t|t} - \hat{\alpha}_{t-1|t}\right)' A_t \left(\hat{\alpha}_{t|t} - \hat{\alpha}_{t-1|t}\right) \\ & + tr\left[A_t\left(\hat{\Sigma}_{t|t} + \hat{\Sigma}_{t-1|t} - T_{t-1}\hat{\Sigma}_{t-1|t} - \hat{\Sigma}_{t-1|t}{T_{t-1}}'\right)\right] \\ & + {\beta'}_t B_t \beta_t + l_{t+1}(\theta_{t+1}). \end{aligned} \quad (2.22)$$

Replacing $L_t(\wp_t, \beta_t, \theta_{t+1})$ and $V_{t+1}\left(\wp_{t+1}\right)$ in Eq. 2.21 by their values given by Eq.'s 2.22 and 2.20 respectively, yields

$$\begin{aligned} V_t\left(\wp_t\right) = \min_{\beta_t, \theta_{t+1}} \{ & \left(\hat{\alpha}_{t|t} - \hat{\alpha}_{t-1|t}\right)' A_t \left(\hat{\alpha}_{t|t} - \hat{\alpha}_{t-1|t}\right) \\ & + tr\left[A_t\left(\hat{\Sigma}_{t|t} + \hat{\Sigma}_{t-1|t} - T_{t-1}\hat{\Sigma}_{t-1|t} - \hat{\Sigma}_{t-1|t}{T_{t-1}}'\right)\right] + {\beta'}_t B_t \beta_t + l_{t+1}(\theta_{t+1}) \\ & + \underset{z_{t+1}}{E}\left[\left(\hat{\alpha}_{t+1|t+1} - \hat{\alpha}_{t|t+1}\right)' A_{t+1}\left(\hat{\alpha}_{t+1|t+1} - \hat{\alpha}_{t|t+1}\right) + \hat{\alpha}'_{t+1|t+1} P_{t+1} \hat{\alpha}_{t+1|t+1}\right] \\ & + tr\left[A_{t+1}\hat{\Sigma}_{t+1|t+1}\right] + tr\left[P_{t+1}\hat{\Sigma}_{t+1|t+1}\right] + V^{\theta}_{t+1}\left(\wp^{\theta}_{t+1}\right) + b_{t+1}\}, \end{aligned} \quad (2.23)$$

Replacing $\underset{z_{t+1}}{E}\left[\left(\hat{\alpha}_{t+1|t+1} - \hat{\alpha}_{t|t+1}\right)' A_{t+1}\left(\hat{\alpha}_{t+1|t+1} - \hat{\alpha}_{t|t+1}\right)\right]$ and $\underset{z_{t+1}}{E}\left[\hat{\alpha}'_{t+1|t+1} P_{t+1} \hat{\alpha}_{t+1|t+1}\right]$

using Lemmas II.10 and II.11 in Appendix 2.7.2, respectively, yields

$$
\begin{aligned}
V_t\left(\wp_t\right) = \min_{\beta_t,\theta_{t+1}} \{ & \left(\hat{\alpha}_{t|t} - \hat{\alpha}_{t-1|t}\right)' A_t \left(\hat{\alpha}_{t|t} - \hat{\alpha}_{t-1|t}\right) \\
& + tr\left[A_t\left(\hat{\Sigma}_{t|t} + \hat{\Sigma}_{t-1|t} - T_{t-1}\hat{\Sigma}_{t-1|t} - \hat{\Sigma}_{t-1|t}{T_{t-1}}'\right)\right] + {\beta'}_t B_t \beta_t + l_{t+1}(\theta_{t+1}) \\
& + \left((T_t - I)\hat{\alpha}_{t|t} + G_t\beta_t\right)' A_{t+1}\left((T_t - I)\hat{\alpha}_{t|t} + G_t\beta_t\right) \\
& + tr\left[A_{t+1}\left(T_t\hat{\Sigma}_{t|t}{T_t}' + Q_t - \hat{\Sigma}_{t+1|t+1} - \hat{\Sigma}_{t+1|t+1}\hat{\Sigma}_t^{*\prime} - \hat{\Sigma}_t^*\hat{\Sigma}_{t+1|t+1} + \hat{\Sigma}_{t|t+1}\right)\right] \\
& + (T_t\hat{\alpha}_{t|t} + G_t\beta_t)' P_{t+1}(T_t\hat{\alpha}_{t|t} + G_t\beta_t) + tr\left[P_{t+1}\left(T_t\hat{\Sigma}_{t|t}{T_t}' + Q_t - \hat{\Sigma}_{t+1|t+1}\right)\right] \\
& + tr\left[A_{t+1}\hat{\Sigma}_{t+1|t+1}\right] + tr\left[P_{t+1}\hat{\Sigma}_{t+1|t+1}\right] + V_{t+1}^{\theta}\left(\wp_{t+1}^{\theta}\right) + b_{t+1}\}. \quad (2.24)
\end{aligned}
$$

Canceling terms results in

$$
\begin{aligned}
V_t\left(\wp_t\right) = \min_{\beta_t,\theta_{t+1}} \{ & \left(\hat{\alpha}_{t|t} - \hat{\alpha}_{t-1|t}\right)' A_t \left(\hat{\alpha}_{t|t} - \hat{\alpha}_{t-1|t}\right) \\
& + tr\left[A_t\left(\hat{\Sigma}_{t|t} + \hat{\Sigma}_{t-1|t} - T_{t-1}\hat{\Sigma}_{t-1|t} - \hat{\Sigma}_{t-1|t}{T_{t-1}}'\right)\right] + {\beta'}_t B_t \beta_t + l_{t+1}(\theta_{t+1}) \\
& + \left((T_t - I)\hat{\alpha}_{t|t} + G_t\beta_t\right)' A_{t+1}\left((T_t - I)\hat{\alpha}_{t|t} + G_t\beta_t\right) \\
& + (T_t\hat{\alpha}_{t|t} + G_t\beta_t)' P_{t+1}(T_t\hat{\alpha}_{t|t} + G_t\beta_t) \\
& + tr\left[\left(A_{t+1} + P_{t+1}\right)\left(T_t\hat{\Sigma}_{t|t}{T_t}'\right)\right] + tr\left[\left(A_{t+1} + P_{t+1}\right)Q_t\right] \\
& + tr\left[A_{t+1}\left(\hat{\Sigma}_{t|t+1} - \hat{\Sigma}_t^*\hat{\Sigma}_{t+1|t+1} - \hat{\Sigma}_{t+1|t+1}\hat{\Sigma}_t^{*\prime}\right)\right] + V_{t+1}^{\theta}\left(\wp_{t+1}^{\theta}\right) + b_{t+1}\}.
\end{aligned} \quad (2.25)
$$

The terms in Eq. 2.25 can be separated into three types: (1) those terms whose value are known at time t with information up to time t, (2) those that depend only on disease control action, β_t, and (3) those that depend only on test control action, θ_{t+1}. Hence, the minimization over β_t and θ_{t+1} can be separated as $V_t\left(\wp_t\right) =$

$V_t^{(1)}(\wp_t) + V_t^{(2)}(\wp_t) + V_t^{(3)}(\wp_t)$ in which

$$V_t^{(1)}(\wp_t) = \left(\hat{\alpha}_{t|t} - \hat{\alpha}_{t-1|t}\right)' A_t \left(\hat{\alpha}_{t|t} - \hat{\alpha}_{t-1|t}\right) + tr\left[A_t\left(\hat{\Sigma}_{t|t} + \hat{\Sigma}_{t-1|t} - T_{t-1}\hat{\Sigma}_{t-1|t} - \hat{\Sigma}_{t-1|t}{T_{t-1}}'\right)\right]$$
$$+ \hat{\alpha}'_{t|t}\left((T_t - I)' A_{t+1}(T_t - I) + {T_t}' P_{t+1} T_t\right)\hat{\alpha}_{t|t}, \tag{2.26}$$

$$V_t^{(2)}(\wp_t) = \min_{\beta_t}\left\{\beta_t'\left(B_t + {G_t}'(A_{t+1} + P_{t+1})G_t\right)\beta_t + \left(\hat{\alpha}_{t|t}((T_t - I)' A_{t+1} + {T_t}' P_{t+1})G_t\right)\beta_t\right.$$
$$\left. + \beta_t'\left({G_t}'(A_{t+1}(T_t - I) + P_{t+1}T_t)\hat{\alpha}_{t|t}\right)\right\}, \tag{2.27}$$

$$V_t^{(3)}(\wp_t) = tr\left[(A_{t+1} + P_{t+1})\left(T_t\hat{\Sigma}_{t|t}{T_t}'\right)\right] + tr\left[(A_{t+1} + P_{t+1})Q_t\right] + b_{t+1}$$
$$+ \min_{\theta_{t+1}}\left\{l_{t+1}(\theta_{t+1}) + tr\left[A_{t+1}\left(\hat{\Sigma}_{t|t+1} - \hat{\Sigma}_t^*\hat{\Sigma}_{t+1|t+1} - \hat{\Sigma}_{t+1|t+1}\hat{\Sigma}_t^{*\prime}\right)\right] + V_{t+1}^{\theta}\left(\wp_{t+1}^{\theta}\right)\right\}. \tag{2.28}$$

As before, the minimization over β_t is denoted by $\tilde{J}_t$, so

$$\tilde{J}_t = \min_{\beta_t}\{{\beta'}_t\left(B_t + {G_t}'(A_{t+1} + P_{t+1})G_t\right)\beta_t + \left(\hat{\alpha}'_{t|t}\left((T_t - I)' A_{t+1} + {T_t}' P_{t+1}\right)G_t\right)\beta_t$$
$$+ \beta_t'\left({G_t}'\left(A_{t+1}(T_t - I) + P_{t+1}T_t\right)\hat{\alpha}_{t|t}\right)\} \tag{2.29}$$

The minimization over β_t can be performed by completion of squares (similar to what is done in Lemma II.8 for the minimization over β_N) to yield Eq.'s 2.30 - 2.33. The optimal disease control at time t is given by

$$\beta_t^* = -U_t\hat{\alpha}_{t|t}, \tag{2.30}$$

in which the control law, U_t, is given by

$$U_t = \left(B_t + {G_t}'(A_{t+1} + P_{t+1})G_t\right)^{-1}\left({G_t}' A_{t+1}(T_t - I) + {G_t}' P_{t+1} T_t\right). \tag{2.31}$$

Moreover, the result of minimization over β_t is given by

$$\tilde{J}_t = -\hat{\alpha}'_{t|t}\tilde{P}_{t+1}\hat{\alpha}_{t|t}, \tag{2.32}$$

in which

$$\tilde{P}_{t+1} = \left((T_t - I)'A_{t+1}G_t + T_t{}'P_{t+1}G_t\right)\left(B_t + G_t{}'(A_{t+1} + P_{t+1})G_t\right)^{-1}\left(G_t{}'P_{t+1}T_t + G_t{}'A_{t+1}(T_t - I)\right). \tag{2.33}$$

As seen in Eq. 2.30 the optimal disease control β_t^* is a linear function of the filtered state mean $\hat{\alpha}_{t|t}$. It is worth noting that this function (more precisely, the control law U_t) depends only on parameters of the system dynamics and the objective function cost inputs. Hence, the control law is data independent and can be calculated off-line prior to solving the measurement and control problems.

As seen in Eq.'s 2.4 to 2.11, the optimal state estimation is independent of the control law. We also just showed that the optimal control law is independent of the state estimation. This completes proof of the Theorem II.1. ■

Replacing the minimization over β_t in Eq. 2.27 by its value given by Eq. 2.32 results in

$$\begin{aligned} V_t(\wp_t) = {} & \left(\hat{\alpha}_{t|t} - \hat{\alpha}_{t-1|t}\right)'A_t\left(\hat{\alpha}_{t|t} - \hat{\alpha}_{t-1|t}\right) + tr\left[A_t\left(\hat{\Sigma}_{t|t} + \hat{\Sigma}_{t-1|t} - T_{t-1}\hat{\Sigma}_{t-1|t} - \hat{\Sigma}_{t-1|t}T_{t-1}{}'\right)\right] \\ & + \hat{\alpha}'_{t|t}\left((T_t - I)'A_{t+1}(T_t - I) + T_t{}'P_{t+1}T_t - \tilde{P}_{t+1}\right)\hat{\alpha}_{t|t} + tr\left[(A_{t+1} + P_{t+1})\left(T_t\hat{\Sigma}_{t|t}T_t{}'\right)\right] \\ & + \min_{\theta_{t+1}}\{l_{t+1}(\theta_{t+1}) + tr\left[A_{t+1}\left(\hat{\Sigma}_{t|t+1} - \hat{\Sigma}_t^*\hat{\Sigma}_{t+1|t+1} - \hat{\Sigma}_{t+1|t+1}\hat{\Sigma}_t^{*\prime}\right)\right] + V_{t+1}^{\theta}\left(\wp_{t+1}^{\theta}\right)\} \\ & + tr\left[(A_{t+1} + P_{t+1})\,Q_t\right] + b_{t+1}. \end{aligned} \tag{2.34}$$

Letting

$$P_t = (T_t - I)'A_{t+1}(T_t - I) + T_t{}'P_{t+1}T_t - \tilde{P}_{t+1}, \tag{2.35}$$

and replacing $tr\left[(A_{t+1}+P_{t+1})\left(T_t\hat{\Sigma}_{t|t}T_t{}'\right)\right]$ in Eq. 2.34 by its other form given by Lemma II.12, Eq. 2.34 can be written as follows to match the form of value function we claimed in Eq. 2.16.

$$\begin{aligned}V_t(\wp_t) &= \left(\hat{\alpha}_{t|t}-\hat{\alpha}_{t-1|t}\right)'A_t\left(\hat{\alpha}_{t|t}-\hat{\alpha}_{t-1|t}\right)+tr\left[A_t\hat{\Sigma}_{t|t}\right]+\hat{\alpha}_{t|t}'P_t\hat{\alpha}_{t|t}+tr\left[P_t\hat{\Sigma}_{t|t}\right]\\ &+tr\left[A_t\left(\hat{\Sigma}_{t-1|t}-T_{t-1}\hat{\Sigma}_{t-1|t}-\hat{\Sigma}_{t-1|t}T_{t-1}{}'\right)\right]+tr\left[\left(\tilde{P}_{t+1}+A_{t+1}T_t+T_t{}'A_{t+1}-I\right)\hat{\Sigma}_{t|t}\right]\\ &+\min_{\theta_{t+1}}\{l_{t+1}(\theta_{t+1})+tr\left[A_{t+1}\left(\hat{\Sigma}_{t|t+1}-\hat{\Sigma}_t^*\hat{\Sigma}_{t+1|t+1}-\hat{\Sigma}_{t+1|t+1}\hat{\Sigma}_t^{*\prime}\right)\right]+V_{t+1}^{\theta}\left(\wp_{t+1}^{\theta}\right)\}\\ &+tr\left[\left(A_{t+1}+P_{t+1}\right)Q_t\right]+b_{t+1}. \end{aligned} \tag{2.36}$$

Now, by comparing Eq.'s 2.36 and 2.16, it can be easily seen that for $t=1,\ldots,N-1$

$$\begin{aligned}V_t^{\theta}\left(\wp_t^{\theta}\right) &= tr\left[A_t\left(\hat{\Sigma}_{t-1|t}-T_{t-1}\hat{\Sigma}_{t-1|t}-\hat{\Sigma}_{t-1|t}T_{t-1}{}'\right)\right]+tr\left[\left(\tilde{P}_{t+1}+A_{t+1}T_t+T_t{}'A_{t+1}-I\right)\hat{\Sigma}_{t|t}\right]\\ &+\min_{\theta_{t+1}}\{l_{t+1}(\theta_{t+1})+tr\left[A_{t+1}\left(\hat{\Sigma}_{t|t+1}-\hat{\Sigma}_t^*\hat{\Sigma}_{t+1|t+1}-\hat{\Sigma}_{t+1|t+1}\hat{\Sigma}_t^{*\prime}\right)\right]+V_{t+1}^{\theta}\left(\wp_{t+1}^{\theta}\right)\},\end{aligned} \tag{2.37}$$

and

$$b_t = tr\left[\left(A_{t+1}+P_{t+1}\right)Q_t\right]+b_{t+1}, \tag{2.38}$$

while from Eq.'s 2.18 and 2.19 we know for $t=N$

$$V_N^{\theta}\left(\wp_N^{\theta}\right) = tr\left[A_N\left(\hat{\Sigma}_{N-1|N}-T_{N-1}\hat{\Sigma}_{N-1|N}-\hat{\Sigma}_{N-1|N}T_{N-1}{}'\right)\right]+tr\left[\tilde{P}_{N+1}\hat{\Sigma}_{N|N}\right], \tag{2.39}$$

and

$$b_N = tr\left[A_{N+1}Q_N\right]. \tag{2.40}$$

Hence, proof of the Theorem II.3 (i.e., the value function we claimed in Eq. 2.16) is complete. ■

Note that the dynamic program defined by value function $V_t^\theta\left(\wp_t^\theta\right)$ for $t = 1, \ldots, N$ can be solved using a branch-and-bound-type algorithm to find the optimal monitoring schedule. The elements of this dynamic program, in a succinct form, are as follows.

***Information state:** $\wp_t^\theta = \left(\hat{\Sigma}_{t|t}, \hat{\Sigma}_{t-1|t}\right)$

***Action space:** $\theta \in \Theta$ where Θ is the set of all available tests for the corresponding disease.

***System dynamics** are given by Eq.'s 2.8, 2.9, and 2.11.

***Optimality equation** is given by Eq. 2.37.

***Boundary condition** is given by Eq. 2.39.

This completes proof of the Theorem II.2. ■

2.5 Case Study of Glaucoma

Thus far we have presented the modeling framework in its general form and derived the optimal disease and test control actions. In this section, we provide a proof of concept by applying our approach to glaucoma and demonstrating how it can help guide clinicians tailor their disease monitoring and treatment control.

Glaucoma is a major public health problem affecting almost 3 million patients in the United States (*Vajaranant et al.* 2012) and over 60 million patients worldwide (*Tham et al.* 2014). Glaucoma is the second leading cause of blindness in the US and a leading cause of visual impairment among Americans (*Stein et al.* 2011). In this section we show how the modeling framework and solution approaches described in Sections 2.3 and 2.4 can be applied to help clinicians in caring for patients with glaucoma. Further, we will elaborate on additional features of our approach designed specifically for glaucoma. Numerical results presented in this section are based on data from patients who were enrolled in two large glaucoma clinical trials.

2.5.1 Glaucoma

Glaucoma is a progressive eye disease which can cause irreversible vision loss and blindness if not adequately monitored and treated. From a societal perspective, the direct medical costs of managing glaucoma are estimated to total over 2.86 billion USD annually (*Rein et al.* 2006). Further, on a per patient basis, costs more than quadruple when patients progress from early to advanced glaucoma (*Lee et al.* 2006). The main risk factors associated with glaucoma and its progression include: non-white race, older age, elevated IOP, genetics and family history (*Tielsch et al.* 1990). It is worth noting that the patient's IOP (i.e., the pressure inside the eye) is the only known controllable/modifiable glaucoma risk factor. Therefore, the current management of glaucoma focuses on lowering the eye pressure by establishing a "target IOP", which is the level of IOP that the clinician feels is low enough to sufficiently slow disease progression (*Jampel* 1997).

Patients with glaucoma are monitored for disease progression using quantitative tests. Two primary methods to monitor a patient include: (1) tonometry (or measuring the IOP), and (2) perimetry (or visual field (VF) testing). Tonometry measures the patient's IOP, is relatively easy to perform, and is part of a standard eye examination. In most patients, vision loss occurs because elevated IOP damages the optic nerve, the structure that carries visual information to the brain for processing (*Sommer et al.* 1991). Lowering IOP reduces the chances of glaucoma progression. With glaucoma, patients often progressively lose peripheral vision and eventually central vision. The VF test quantifiably measures the extent and rate of peripheral vision loss by examining the sensitivity of the eye to light stimuli. It is more time-consuming than checking IOP, but provides important information on the status of the disease. VF testing can be anxiety provoking and challenging as it requires patient attention and cooperation. Two key global performance measures from VF testing include Mean Deviation (MD) and Pattern Standard Deviation (PSD), which estimate the deviation of peripheral

vision from a reference population who do not have glaucoma (*Choplin and Edwards* 1995). MD is usually a negative number; higher values of MD (i.e., values closer to zero) correspond to better vision quality (less vision loss). PSD is usually a positive number; lower values of PSD indicate less glaucomatous damage.

It is well established from prior work that both IOP and VF tests can be associated with noise. For example, patient performance on automated VF test can fluctuate considerably from one test to the next (*Choplin and Edwards* 1995). Likewise, IOP can fluctuate from hour to hour and day to day (*Wilensky et al.* 1993). To take such noise into consideration in deciding how to optimally monitor the patient and determine a target IOP, we harness the Kalman filter method (*Kalman* 1960) to extract noise from the raw measurements.

There are a number of treatments available to lower the IOP for a patient with glaucoma. Different eye drops, laser therapies, and incisional surgery can reduce the IOP to any number above 6 mmHg. However, glaucoma medications can be expensive and can have serious side effects including stinging, blurred vision, eye redness, itching, burning, low blood pressure, reduced pulse rate, fatigue, shortness of breath, headache, and depression. Therefore, it is important to find a target IOP level for each patient that sufficiently slows progression while avoiding unnecessary treatment. It is common in current practice to use fixed-interval monitoring regimes (e.g., annual exams) to test for disease progression. Furthermore, for each patient and eye, the ophthalmologist must currently make a gestalt-based estimate of a reasonable target IOP that considers the risk of disease progression and the side effects and costs associated with lowering the IOP. To our knowledge, no optimization-based approach presently exists to jointly determine how to monitor a patient with glaucoma and how to control the disease. Our approach considers the history of the patient (prior test performances) and her unique disease dynamics to provide clinicians with (1) a personalized monitoring regime to achieve an accurate assessment of whether there

is disease progression (exploration), and (2) a menu of target IOP options and how the glaucoma is likely to progress for different target IOP levels that the doctor can leverage to devise a treatment plan for the patient (exploitation). As a feature of our model, the clinician is able to select the desired aggressiveness level to monitor and treat the patient based on the unique characteristics/circumstances of each individual. We will elaborate on this menu of options in Subsection 2.5.5.

Figure 2.2 depicts a high-level overview of our dynamic monitoring and control decision support framework for patients with glaucoma. At each office visit, IOP and/or VF test(s) are performed. The raw data from these tests (which are known to be noisy) are fed into a Kalman filter to obtain an optimal estimation of the current disease state for a particular eye. Then, the estimate of the previous state is refined via a Kalman smoother given the new information. Each patient's label/type (fast, slow, or non-progressor) is determined to estimate how quickly the patient is likely to progress in the future. The decision support tool provides an optimal monitoring schedule (i.e., the timing of the next exam/test and which tests the patient should take) and a personalized target IOP for the patient for different aggressiveness levels/options (super-high, high, moderate, low, or super-low). Finally, the clinician chooses an aggressiveness option from the menu of choices that is appropriate for the individual patient.

2.5.2 Patient Disease State

We use a nine-dimensional state vector, α_t, to model the patient's disease state. The elements of the state vector include Mean Deviation (MD), Pattern Standard Deviation (PSD) and Intraocular Pressure (IOP) together with their discrete time first and second derivatives (i.e., velocity and acceleration, respectively);

$$\alpha_t = \begin{bmatrix} MD & MD' & MD'' & PSD & PSD' & PSD'' & IOP & IOP' & IOP'' \end{bmatrix}'. \quad (2.41)$$

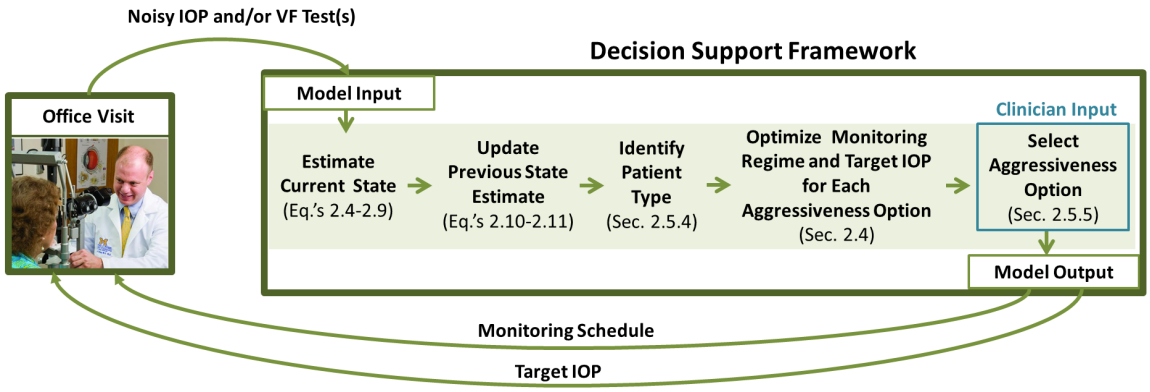

Figure 2.2: Big picture of the decision support framework illustrating the model inputs and outputs as well as the sequence of main steps of the disease monitoring and control algorithm.

The non-linear behavior of the disease dynamics is captured by including the velocity and acceleration of key disease state elements in the state vector. This is known to be an effective way to linearize a nonlinear model of state evolution (see *Bertsekas et al.* 1995). MD', PSD', and IOP' are the slope of a linear regression of the latest three MD, PSD, and IOP measurements, respectively. MD'', PSD'', and IOP'' are the difference of the latest two MD', PSD', and IOP' values divided by the time interval between them.

2.5.3 Data

To parameterize and validate our model, we use data from two multi-center randomized clinical trials: the Collaborative Initial Glaucoma Treatment Study (CIGTS) (*Musch et al.* 1999) and the Advanced Glaucoma Intervention Study (AGIS) (*Ederer et al.* 1994). These data sets are chosen because they include structured tonometry and perimetry data (IOP and VF readings) of glaucoma patients taken every 6 months during the course of the trials. We match the time step of our LQG model with these data sets to avoid the need for data interpolation (i.e., there is a 6-month time interval between periods t and $t+1$ in our model).

CIGTS followed newly-diagnosed glaucoma patients with mild to moderate disease who were randomized to medical or surgical therapy and followed for up to 11 years with IOP and VF tests taken every 6 months to assess for disease progression. In AGIS, patients with advanced glaucoma were randomized to laser therapy or incisional surgery and followed for up to 11 years with IOP and VF readings taken every 6 months.

For the purpose of the case study, we excluded patients from these trials with fewer than 5 readings. We also restricted our focus to the patients who received either medical or laser therapy, and excluded glaucoma patients who received incisional surgical interventions, because incisional surgery can abruptly change disease progression dynamics. We randomly divided all eligible participants from the trials (571 participants) into two sets of equal size: (1) a training set, and (2) a testing set. Both sets have approximately the same number of mild, moderate, and advanced glaucoma patients, with similar number of white and black patients, and equal numbers of patients contributing data from each trial. The training set is used for parametrization and calibration of our state space model and the testing set is used to evaluate the performance of our approach.

2.5.4 Patient Types (Fast/Slow/Non-progressor)

A fast-progressing patient is someone whose glaucoma is rapidly worsening and is part of the subset of patients at greatest risk of blindness. Although there is presently no gold standard for defining glaucoma fast-progressors, prior literature considers a loss of MD greater than 1 dB per year as a reasonable identifying feature of patients who are exhibiting fast-progression of glaucoma (see *Heijl et al.* 2013; *Gardiner and Crabb* 2002). We built our algorithms based on this definition of fast-progressors. To classify each patient, we calculated the slope obtained from a linear regression of their entire set of MD readings and labeled them as a:

- *fast-progressor* if the MD progression slope is declining by ≥ 1 dB/year,
- *slow-progressor* if the MD progression slope is declining between 0 and 1 dB/year, and
- *non-progressor* if the MD progression slope is not declining.

2.5.5 Aggressiveness Levels/Options/Policies

In clinical practice, the goals of care must be tailored to the needs of the individual patient. Rather than proposing one solution, a more powerful and useful approach is to provide the clinician with a range of options for how much effort (both from provider and the patient) will be put into monitoring and how aggressively IOP should be lowered such that future progression can be slowed. For instance, clinicians will likely see the need to monitor and treat a young patient who only has sight in one eye more aggressively than an older patient with good vision in both eyes who has multiple systemic medical comorbidities and is likely to expire before they go blind from glaucoma. As a useful and not overly complex approach, we develop optimization models tailored to three regimes, or "options," for monitoring and treatment (suggested by our clinical collaborator for ease of adoption into clinical practice). We refer to these three options as *low*, *moderate*, and *high levels of aggressiveness* to represent the level of intensity in care and monitoring. We also define two extreme levels of aggressiveness: *super-high* and *super-low*. Note that we choose these terms only for convenience in presenting the five options and to make it easier for the reader to remember the order of them in terms of how aggressively they test and treat patients; they are not meant to correspond to any existing terms or approaches currently used in clinical practice. These five options are useful not only for sensitivity analysis, but also suggest an effective way to implement a decision support system so that clinicians can pursue monitoring and treatment with the level of intensity that they, together

with the patient, determine to be the most appropriate for each individual. The five aggressiveness levels/policies/options follow:

1. *Super-high aggressiveness option*, which drops the IOP immediately to 6 mmHg (an ideal level of IOP for patients with any severity of glaucoma, but one that may be impractical for many patients due to limitations with the effectiveness of existing interventions, side effects, and/or complications),

2. *High aggressiveness option*, which tends to lower the IOP by 40%-60% compared to the patient's treated level of IOP that was achieved in the CIGTS/AGIS clinical trials after the initial intervention was given,

3. *Moderate aggressiveness option*, which tends to lower the IOP by 20% to 40% compared to the patient's treated level of IOP that was achieved in the CIGTS/AGIS clinical trials after the initial intervention was given,

4. *Low aggressiveness option*, which corresponds to the IOP achieved under no additional interventions beyond those employed in the CIGTS/AGIS trials,

5. *Super-low aggressiveness option*, which attempts to estimate progression of an untreated patient with glaucoma by removing the effect of existing interventions that were employed in CIGTS/AGIS on the patient's IOP.

It should be noted that the exact amount of IOP control suggested by high, moderate, and low aggressiveness policies is patient-specific and is optimized to yield the minimum total cost as defined by the objective function. However, the super-high and super-low aggressiveness policies are static policies that do not take the objective function into account. They are mainly added for purposes of analysis and comparison, but they can still provide valuable insight in a clinical setting by presenting the clinician the "best" and "worst" case options and their forecasted impact on disease

progression dynamics. These five options/policies also provide sensitivity analysis on the model cost parameters.

2.5.6 System Model Parameterization

The Expectation Maximization (EM) algorithm was employed for parameterization of our state space model. EM is an iterative algorithm for finding the maximum likelihood estimate of model parameters in the presence of missing or unobserved data. The EM algorithm alternates between expectation step (E-step) and maximization step (M-step). In the E-step, raw, noisy readings are filtered and missing data is estimated based on the observed data and the most recent estimate of system parameters. In the M-step, the log-likelihood function of all data points is maximized assuming the missing data is given by the estimates from the E-step. For more information about the EM algorithm please see *Dempster et al.* (1977); *Digalakis et al.* (1993); *Ghahramani and Hinton* (1996). While the model was presented in its general setting in Section 2.3, for the purpose of this case study we assume the model parameters are time-invariant. The output of the EM algorithm is the best estimate of system matrices $T_t = T$, $Q_t = Q$, $Z_t = Z$, $H_t^{(\theta_t)} = H^{(\theta_t)}$ for $t = 1, \ldots, N$, and initial state mean and covariance, $\hat{\alpha}_0$ and $\hat{\Sigma}_0$ (see Subsections 2.3.1 and 2.3.2 for the definition of these parameters). We further assume that $G_t = G = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}' T$ for $t = 1, \ldots, N$, because the control variable β_t is designed to control only the patient's IOP, which is the only controllable glaucoma risk factor. Furthermore, it is known that the intervention started or employed at time t has instantaneous effect on lowering the patient's IOP. For example, if patient's IOP is 20 mmHg at time period 7 and the control $\beta_7^* = -3$ mmHg, the expected value of IOP right after time period 7 is 17 mmHg. This IOP reduction affects other state elements and progression dynamics in the following time period through the transition equation. We use the EM algorithm to obtain 4 sets of system parameters. These sets of parameters are obtained from

(1) all patients in the training set, (2) only fast-progressors, (3) only slow-progressors, and (4) only non-progressors.

The model cost parameters were estimated based on the input from our glaucoma specialist collaborator so that the model outputs are reasonable from a clinical perspective. Note that it is the relative costs (rather than each absolute cost) that plays a key role in our analysis. It is beyond our scope to obtain definitive cost parameters; however, signicant sensitivity analysis was performed around those estimates to better understand the model's behavior and to ensure that the model in its entirety provides credible decision support. We confirmed with our glaucoma specialist collaborator that the model generates reasonable target IOPs and monitoring schedules for each patient. For instance, under the high aggressiveness option the model for fast-progressing patients (i.e., the most aggressive combination) should suggest taking both IOP and VF tests every six months and a target IOP of around 6 to 9 mmHg. Under the low aggressiveness option the model for non-progressing patients (i.e., the least aggressive combination) should suggest no further IOP reduction and taking IOP and VF tests every two years. This is in line with recommendations put forth by the *American Academy of Ophthalmology Glaucoma Panel* (2010). For all other combinations of aggressiveness level and patient type the model cost parameters are fine-tuned so that it suggests a monitoring regime and target IOP level that are reasonable in expert clinical opinion and are in between the two extreme combinations. The behavior of optimal policies is discussed in more detail in Section 2.5.8.2.

2.5.7 Model Usage for a Glaucoma Patient

For a patient who is newly diagnosed with glaucoma with no prior history of IOP and VF readings, both tests are taken in every period (i.e., every 6 months) for the first 5 periods. *Gardiner and Crabb* (2002) found that predictions based on 5 initial VF test results is a reasonable predictor of future vision loss in most patients. These 5

readings are used to (1) obtain baseline values for key disease state elements (i.e., MD, PSD, and IOP), (2) calculate the velocities and accelerations of key state elements, (3) warmup the Kalman filter and smoother, (4) reduce the initial uncertainty surrounding a given patient's disease state, (5) calculate the initial 5-period rate/slope of MD progression to label the patient as fast, slow, or non-progressor, and (6) differentiate the patient from the population mean and tailor the disease transition model to the specific patient. If the patient already has a history of IOP and VF readings, then these values can be used to create the initial state, or warmup, for our model. The system parameters obtained from all training patients are used in the Kalman filter and the Kalman smoother during the warmup period. At the end of the warmup period (i.e. after 5 readings), the patient is labeled as a fast-progressor, a slow-progressor, or a non-progressor based on her MD progression rate. Thereafter, only the type-specific system parameters (i.e., fast, slow, or non-progressor set of parameters) are used in the Kalman filter and smoother. Each time a test is taken, the MD progression slope is recalculated. We always consider the latest 5 filtered MD values to update the MD slope (i.e., a sliding window of length 5). Whenever the patient's latest MD slope indicates a label upgrade (i.e., the patient moves from non-progressor to slow/fast-progressor, or from slow-progressor to fast-progressor), the model (1) calls for a follow-up visit to take IOP and VF testing in the following time period, and (2) labels the patient as a suspect of the higher label/category (e.g., slow-progressor suspect or fast-progressor suspect). If the label change is confirmed at the next follow-up visit, the higher label is assigned to the patient; otherwise, the patient is returned to the previous lower label status. Note that in our analyses, we take a conservative approach and do not allow any label downgrading (on the recommendation of our glaucoma specialist collaborator). Once the label is upgraded for a patient, the model will recommend applying more IOP control (i.e., greater intensity of interventions) to slow glaucoma progression. Therefore, it can be expected that

the MD will tend to decline less rapidly once the amount of IOP control is increased, thus resulting in a lower classification/label at some point. However, if we were to downgrade the patient's label and let the model decrease the amount of IOP control, the patient would be at risk to start losing vision at the same rate as earlier in the disease course, which is not desirable. Therefore, we do not allow label downgrading for any patient once the label upgrade is confirmed. If a patient suspected to belong to a higher category does not get a confirmatory result at the very next follow-up visit, then, the patient remains at the original/lower label he/she was previously at. The glaucoma monitoring and treatment control algorithm steps are illustrated in Figure 2.3.

2.5.8 Numerical Results

In this subsection, we test the performance of our dynamic disease monitoring and control algorithm on glaucoma patients of CIGTS and AGIS clinical trials. We first validate our prediction model on the testing dataset, using the training dataset for parameterization. Then, we provide numerical results and examples on how the optimal policies behave. Lastly, we provide further results on the impact of optimal policies on patients with glaucoma.

2.5.8.1 Validation:

We first validate that the model is good at forecasting future disease progression trajectory and then validate that the results are consistent with clinical expectations. Our modeling approach efficiently captures the system and measurement noises using a set of stochastic first-order vector difference equations. To evaluate the performance of our prediction model, we used the first five data points of each patient in the testing dataset to warm up the Kalman filter and determine the patient type. Then, we predicted MD values for five periods into the future for each patient type (fast-

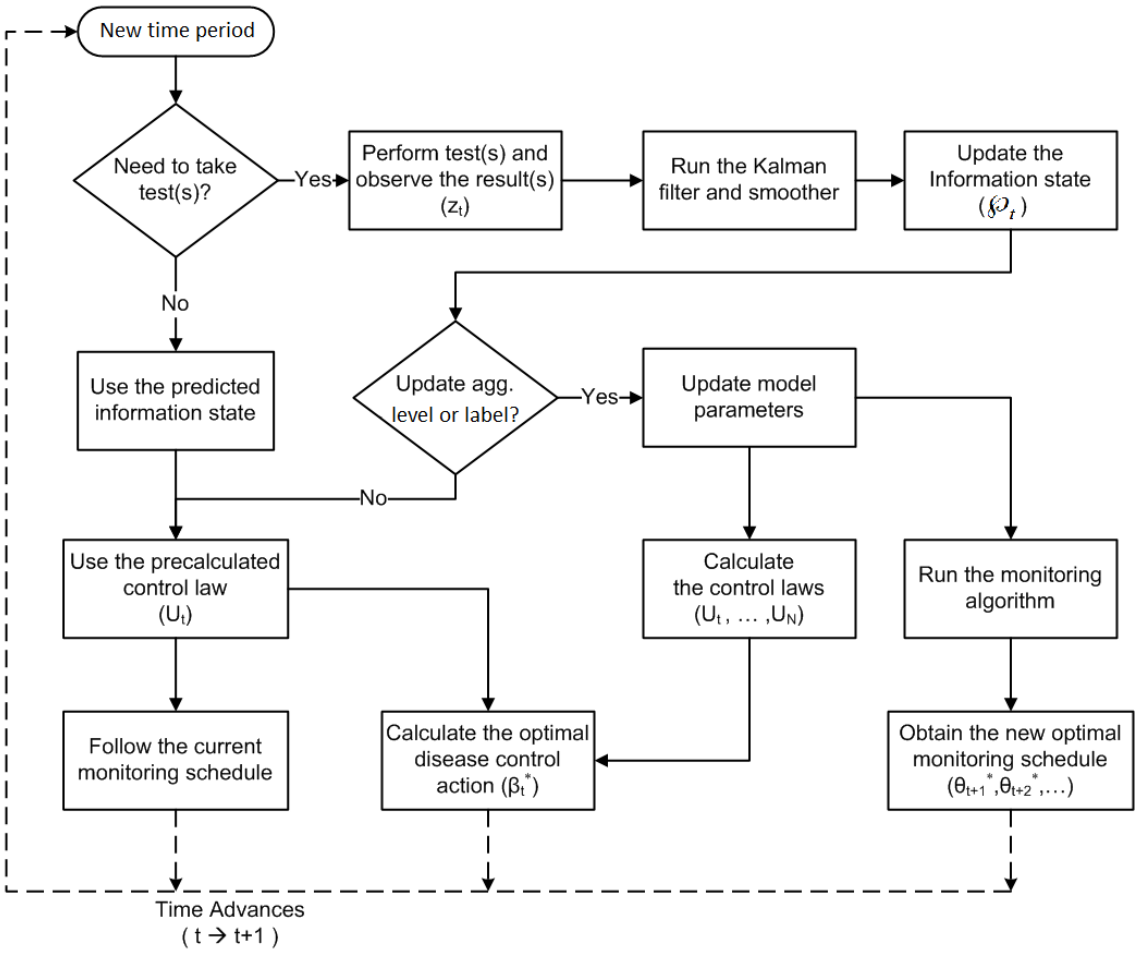

Figure 2.3: Glaucoma monitoring and treatment control flow diagram.

progressor, slow-progressor, and non-progressor) and calculated the prediction error (i.e., the predicted state mean minus the actual reading as obtained from the patients during their followup in the trial). Figure 2.4 shows the mean MD prediction error for up to five periods (2.5 years) into the future (each period represents a 6 month time interval). The dots correspond to the mean error and the bars represent the 95% confidence interval for the mean. These interval plots confirm that our prediction model has very little error in predicting the glaucoma state progression. One also sees that the fast-progressors (as defined in Subsection 2.5.4) vary the most in the datasets, and this is reflected in greater uncertainty and error. In here, we present

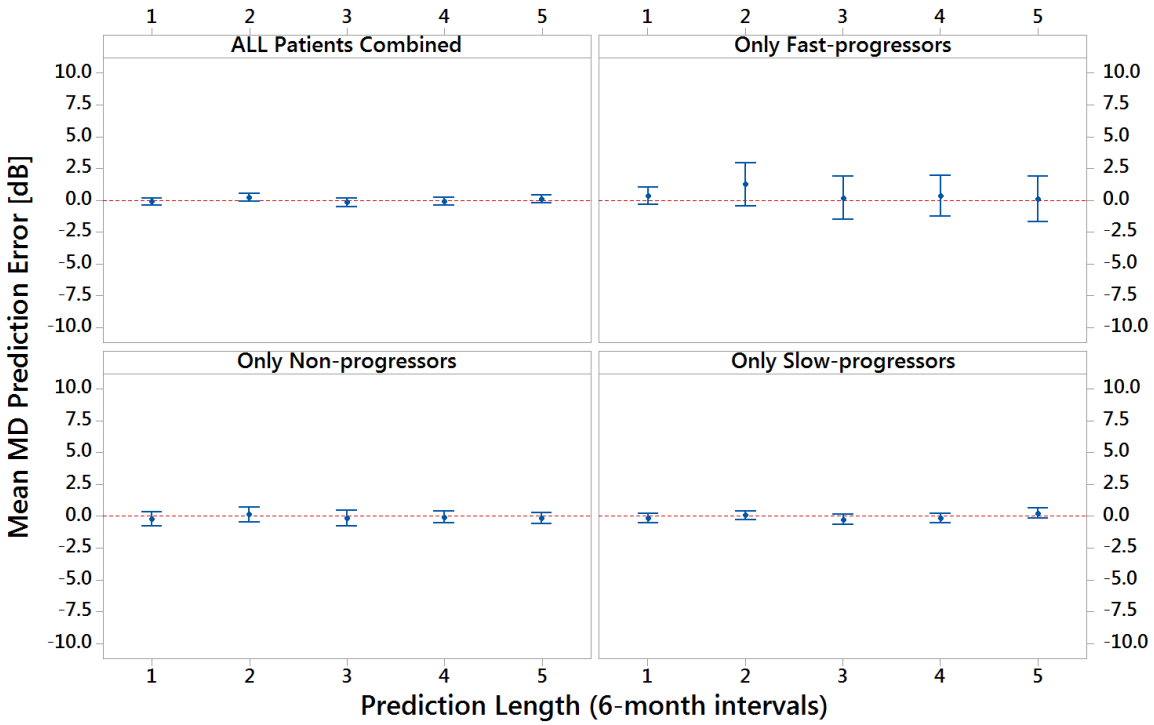

Figure 2.4: Interval plot of mean MD prediction error for different prediction lengths. The dots correspond to mean error and the bars represent 95% confidence interval for the mean.

the results for MD as it is the most clinically useful state variable. Similar results were obtained for other state elements (PSD and IOP).

2.5.8.2 Evaluation of the Optimal Policies:

Now that we have confirmed that the model creates an accurate forecast of disease progression, we next test that the output provides clinically reasonable results as confirmed by our clinical collaborator. The structure of optimal IOP control generated by our model under the moderate or high aggressiveness policy is a key to identifying the target IOP for each patient. We applied the high and moderate aggressiveness policies to all fast and slow-progressing patients in the test dataset to achieve a statistical characterization of how each policy behaves. For each group and each aggressiveness policy, we used the first five data points to warmup our model.

We then recorded the amount of optimal IOP control suggested by our model in the next 20 time periods (i.e., the following 10 years). Figure 2.5 depicts the results over all the patients in the testing dataset as box plots of optimal additional IOP control (β_t^*) applied in the current (i.e., period 0) and the following time periods. An IOP control of $-x$ mmHg corresponds to lowering the patient's IOP by x mmHg more than what was achieved in AGIS/CIGTS. The bottom and top of each box are the first and third quartiles, respectively. The lower and upper whiskers extend to the minimum and maximum data points within 1.5 box heights from the bottom and top of the box, respectively. As seen in the figure, our feedback-driven model recommends further lowering the IOP within the first few time periods. Afterwards, the optimal additional IOP control is close to zero. This results in the patient's IOP converging to "a number" over time. We call this number the "target IOP," per the common terminology used in the glaucoma community. As one would expect, the group of non-progressing patients do not get additional benefit from further lowering their IOP since they exhibit no signs of progression at the IOP levels they are maintaining from the treatments already employed in the trials. Therefore, they are not included in the graph.

We applied the five aggressiveness options to fast and slow-progressing patients in the test dataset and obtained the following IOP-related metrics for each combination of patient type and aggressiveness policy: (1) target IOP [mmHg] or the 10-year predicted IOP, (2) additional IOP control [mmHg] applied in 10 years (representing the amount/intensity of treatment), and (3) percentage of IOP change after 10 years. For each metric, we report the median and interquartile range (IQR), which are robust measures of location and scale respectively. The IQR is the difference between the upper and lower quartiles and provides a range that contains 50% of the data. As seen in Figure 2.5, the optimal additional IOP reduction is almost entirely applied during the first 6 periods (3 years) of employing the control policy. Hence, we evaluate

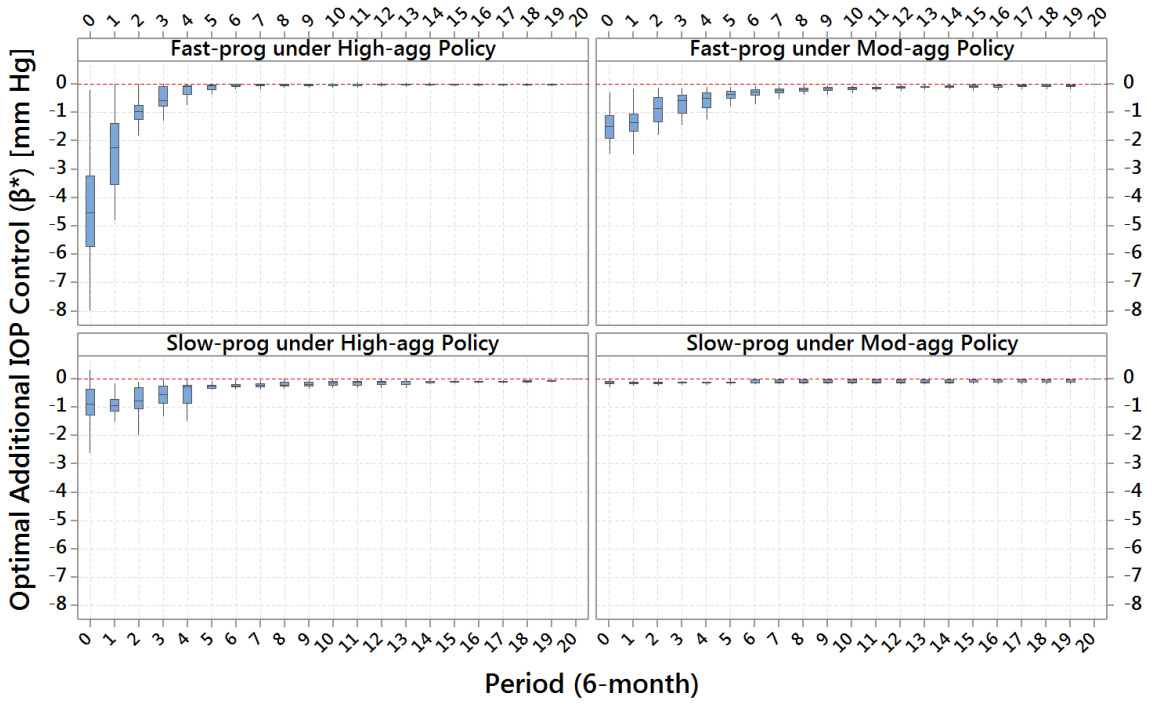

Figure 2.5: Optimal IOP controls suggested by our model for fast and slow-progressing patients under the high and moderate aggressiveness policies over 10 years. Period 0 is the current time period (i.e., the period at which the IOP control starts).

the patient's IOP after 10 years, which is a sufficiently long horizon for the IOP to become stable under treatment. We tested our IOP control model with longer time horizons and obtained similar results.

Table 2.1 summarizes the IOP-related results. For instance, applying the high aggressiveness policy to fast-progressors results in a median target IOP of 7.17 mmHg; this can be achieved by administering a median additional 9.36 mmHg IOP reduction from the baseline level of IOP attained in the trials. Such a target IOP is, on average, 55.24% lower than the baseline IOP (the IOP at the beginning of the 10-year prediction period) of fast-progressing patients in the trials. Since target IOP is an important metric that helps guide clinicians in selecting the appropriate treatment plan for the patient, the distribution of target IOPs is also given in Figure 2.9 in

Table 2.1: Comparison of the effect of different aggressiveness options on patient's IOP for fast and slow-progressing patients in CIGTS and AGIS.

		Target IOP [mm Hg] (10-year predicted IOP)		Cumulative/Additional IOP control [mm Hg] applied over 10 years		% of IOP change after 10 years	
		median	IQR	median	IQR	median	IQR
Fast-progressors	Super-high agg. Policy	6	0	-10.57	4.97	-64.17	9.50
	High agg. Policy	7.17	1.70	-9.36	4.17	-55.24	8.76
	Moderate agg. Policy	9.62	2.66	-7.43	3.91	-41.19	10.28
	Low agg. Policy	18.45	5.69	0	0	+3.47	29.32
	Super-low agg. Policy	22.98	6.37	+3.91	1.21	+26.53	39.23
Slow-progressors	Super-high agg. Policy	6	0	-12.00	3.09	-65.46	10.61
	High agg. Policy	9.52	2.07	-6.38	4.01	-44.11	8.51
	Moderate agg. Policy	13.08	2.02	-2.13	1.52	-22.69	23.29
	Low agg. Policy	14.95	2.66	0	0	-11.44	34.37
	Super-low agg. Policy	18.66	4.24	+4.93	1.27	+11.88	43.01

Appendix 2.7.3.

Table 2.2 summarizes the optimal monitoring regime for different combinations of patient type and aggressiveness level. For example, under the moderate aggressiveness level, the model for slow-progressing patients recommends measuring IOP every 6 months and checking the visual field every 12 months. It is worth noting that these protocols remain optimal as long as (1) the patient follows the monitoring schedule (i.e., does not miss a test), (2) the patient type/label remains unchanged, and (3) the doctor does not change the aggressiveness level. If any of the three criteria is not met, the model modifies the monitoring schedule to account for the missing information or the change in patient label/aggressiveness level. The monitoring regimes presented in Table 2.2 and the range and mean of target IOPs presented in Figure 2.9 are clinically appropriate in the professional opinion of our glaucoma specialist collaborator.

2.5.8.3 Menu of Options:

Now that we have validated our model and elaborated on the structure of the optimal policies, we provide an example of how our decision support tool can help guide

Table 2.2: Optimal monitoring regime for different combinations of patient type and aggressiveness level.

	Fast-progressors	**Slow-progressors**	**Non-progressors**
High aggressiveness	IOP+VF every 6 months	IOP+VF every 6 months	IOP+VF every 12 months
Moderate aggressiveness	IOP+VF every 6 months	IOP every 6 months VF every 12 months	IOP every 12 months VF every 24 months
Low aggressiveness	IOP every 6 months VF every 12 months	IOP 3x in 24 months VF every 24 months	IOP+VF every 24 months

clinicians in managing a patient with glaucoma. Figure 2.6 depicts the glaucoma progression trajectory (change in MD over time) for a randomly chosen fast-progressing patient from the AGIS trial. The figure depicts a sample output of the decision support tool (in regards to disease control) that compares how this patient is likely to progress over the following 10 periods (5 years) under different aggressiveness options defined in Subsection 2.5.5. As demonstrated in the figure, the patient progresses much slower and would have better MD value (i.e., vision quality) 5 years into the future as the aggressiveness of IOP control is increased. This graph demonstrates the type of insight our decision support tool can offer the clinicians. It provides a menu of options related to how aggressively the doctor wants to treat the patient, depicts the future disease progression trajectory, and provides the optimal target IOP and monitoring schedule for each aggressiveness option. The doctor is then able to select the right aggressiveness option based on evolving needs of the patient, adherence, health status, and other personal or clinical factors.

2.5.8.4 Insights into Treatment Effectiveness by Patient Type:

Figure 2.7 graphs the average MD loss per year against the total IOP reduction applied under different aggressiveness policies for all fast and slow-progressing patients in the test set of CIGTS and AGIS trials. This graph provides important insights for managing patients with glaucoma: (1) As seen in the graph, the curve for fast-

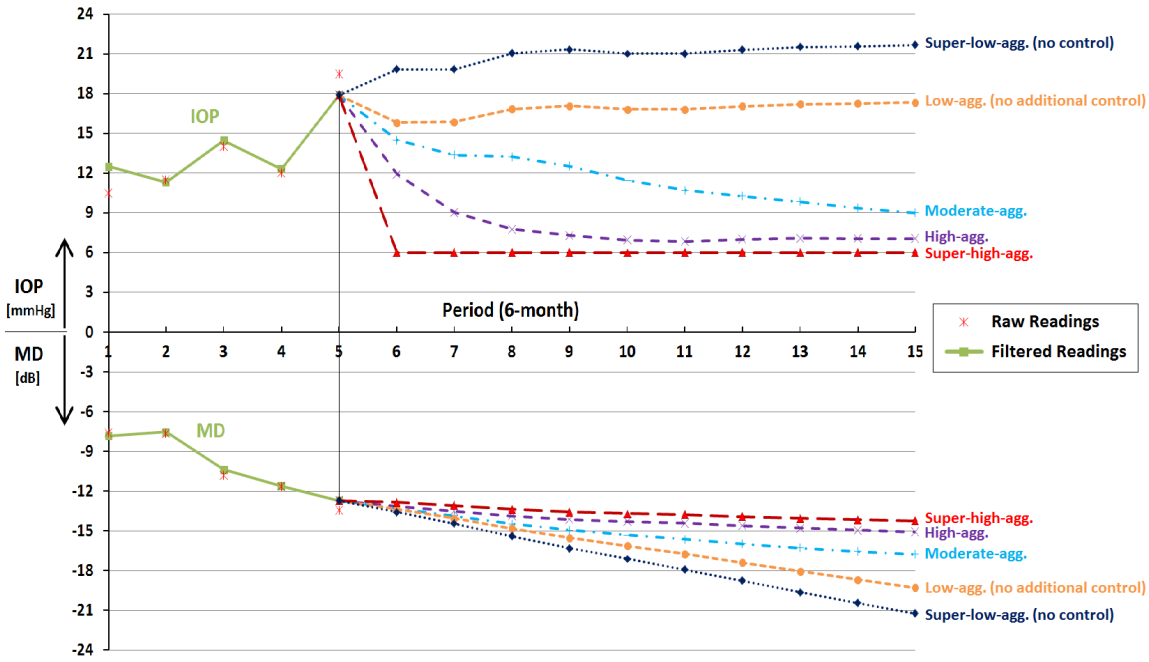

Figure 2.6: An example of the trajectory of glaucomatous progression as captured by changes to MD over time by employing each of the five different aggressiveness policies for a sample fast-progressing patient from the AGIS study (note: higher MD correlates with better vision quality).

progressors has a steeper slope, which indicates that this group of glaucoma patients benefits the most from further lowering of IOP. (2) It can be deduced from Figure 2.7 that the low aggressiveness policy (point B), which roughly speaking corresponds to treatment using eye drops, works well enough for most slow-progressors since the curve is fairly flat around point B on the slow-progressors curve. In other words, increasing the aggressiveness level from low to moderate/high has only minimal advantage for this group of patients. This highlights the importance of differentiating patients by progression type. Treating all patients the same risks over-treating for little gain or irreversible vision loss due to under-treating. It can also be seen that slow-progressors gain long-term benefit if treated under the super-high aggressiveness policy (point E), which roughly speaking corresponds to incisional surgery. There-

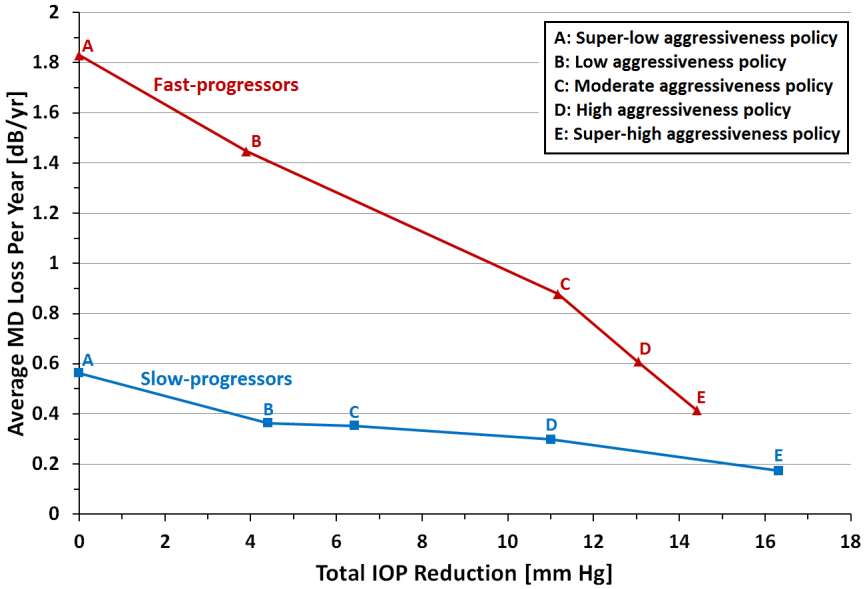

Figure 2.7: Average MD loss per year can be reduced by applying more IOP control. Fast-progressors get more benefit from lowering their eye pressure.

fore, for slow-progressing patients, the doctor may choose either the low or super-high aggressiveness policy, depending on the individual's life expectancy, severity of glaucoma, other personal and medical factors, and the preferences of the patient. (3) The steep slope of the fast-progressors' curve around point B implies that vision loss could be significantly averted (even in the short-term) by further reduction of their IOP. Hence, moderate, high, or super-high aggressiveness policies (points C, D, and E on the graph, respectively) may be more suitable for most fast-progressing glaucoma patients.

The same result is also verified by plotting the MD loss averted over 10 years by following the IOP controls suggested by our model. Please see Figure 2.10 in Appendix 2.7.3.

2.5.8.5 Sample Application of the Model in Practice:

In this subsection, we provide an illustration of how our modeling framework may be used to guide monitoring and control of a glaucoma patient. Figure 2.8 portrays

the disease trajectory of a sample patient. After the warmup period (i.e., the first 5 periods) the patient is initially identified as a non-progressor. In our example, the clinician chooses the low aggressiveness option to monitor and control the patient. Subsequently, the model suggests taking an IOP reading every year and a VF test every other year. We assume the clinician and the patient follow this protocol. The patient remains a non-progressor up to period 13, when she becomes classified as a slow-progressor suspect. This slow-progressor status is confirmed after obtaining IOP and VF testing at a follow-up visit in time period 14. When the patient becomes a confirmed slow-progressor, for the sake of this example, assume the doctor decides to increases the aggressiveness level to the moderate aggressiveness policy. Under this policy, the model recommends (1) lowering the IOP from 24 to about 21 mmHg, (2) measuring the IOP every six months, and (3) taking a VF test every year. After 1.5 years (i.e., at time period 17) the doctor and patient decide to further increase the aggressiveness level and continue care under the high aggressiveness policy. This policy suggests taking both IOP and VF tests every six months and recommends additional IOP reduction. Figure 2.8 also illustrates how the patient's glaucoma would likely progress after period 14 should the doctor have maintained the low aggressiveness IOP control policy during periods 5-14.

While this example relates to the management of a single patient, a few aspects should be highlighted. (1) As described in Subsection 2.5.7, glaucoma patients do not always maintain the same progression rate over time. Recall that each time a test is taken, the model updates the MD slope estimate; hence, it is possible that a patient moves from non-progressor status to slow/fast-progressor, or from slow-progressor to fast-progressor status. (2) As described in Subsection 2.5.8.2, whenever the patient's label is changed or the doctor decides to change the aggressiveness level, the model modifies the monitoring regime subsequently. (3) As described in Subsection 2.5.8.4, there is little gain (in terms of preventing vision loss) in increasing the aggressiveness level

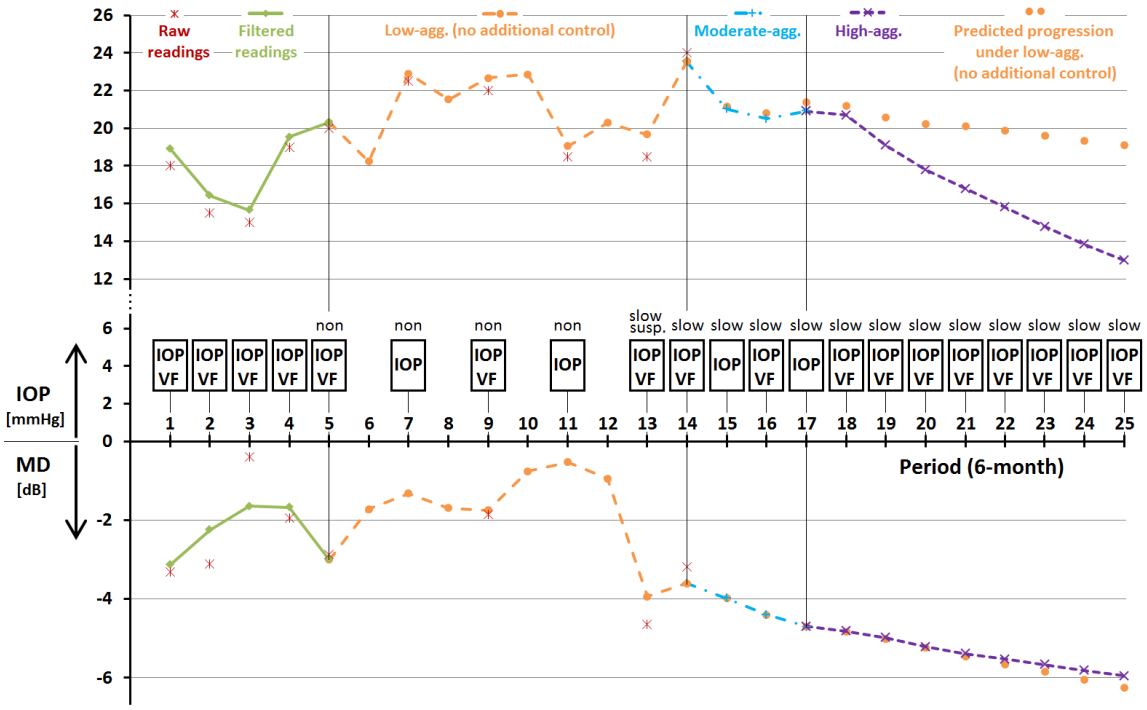

Figure 2.8: A non-progressing patient becomes slow-progressor. The clinician tailors care by increasing the aggressiveness level. The first 5 periods are warmup time. From period 5 to 13, the patient is a non-progrssor and the doctor selects the low aggressiveness policy. In period 13, the patient becomes a slow-progressor suspect and this label upgrade is confirmed in period 14. The doctor treats the patient under moderate aggressiveness policy from period 14 to 17. In period 17, the doctor increases the aggressiveness plociy to high in order to further slow the progresson rate. Periods 14-25 show forecasted values.

from low to high for slow progressing patients. Note the big gap in the optimal IOP under the low and high aggressiveness policies at time period 25. However, this gap results in a very small difference in the patient's MD values. Benefiting from this type of insight in a busy clinic can significantly enhance the ability of ophthalmologists and optometrists to appropriately take care of patients with glaucoma.

2.6 Conclusions

In this chapter we developed a dynamic personalized modeling paradigm for simultaneous monitoring and control of irreversible chronic diseases (e.g., glaucoma). Our model incorporates each patient's past and present readings in a feedback-driven control model to provide the jointly optimal solution to two critical questions facing clinicians: (1) when to schedule office visits and which suite of tests to perform to monitor for disease progression (exploration); and (2) what levels of controllable disease risk factors should be targeted to slow the rate of disease progression (exploitation).

Kalman filtering methodology is built into our modeling framework to extract noise from the raw measurements and to optimally estimate the disease state in each time period based on imperfect observations. This is a key to accurately identifying genuine disease progression from testing artifacts. We developed a multivariate continuous state space model of disease progression and model the state transition and the testing processes as first order vector difference equations with multivariate Gaussian random noises. For the new objective of minimizing the relative change in state (i.e., disease progression), which is imperative for management of irreversible chronic diseases, we proved the two-way separation of optimal estimation and control. This is a fundamental finding upon which solution tractability depends.

To demonstrate the effectiveness of our approach, we harnessed data from two landmark glaucoma randomized clinical trials to parametrize and validate our model. We showed that our Kalman filter-based model has low error in predicting the future disease progression trajectory. Further, we showed that our decision support tool provides a menu of options for the clinician based on how aggressively the doctor wants to manage the patient. For each aggressiveness option, the model provides for each glaucoma patient (1) future disease progression trajectory, (2) optimal monitoring schedule, (3) optimal target IOP. The doctor has the choice to select an appropriate

aggressiveness level depending on the patient's life expectancy, severity of glaucoma, and other personal and clinical factors. Our numerical results demonstrated that following the recommendations of our model not only results in patients with better vision quality over the treatment horizon, but also achieves significantly slower glaucoma progression rate, which means patients will keep their sight longer.

2.7 Appendix

2.7.1 Optimization of the Final Period Disease Control Action

The value function in the last period is given by

$$V_N(\wp_N) = \min_{\beta_N}\{E\left[(\alpha_N - \alpha_{N-1})' A_N (\alpha_N - \alpha_{N-1})\right] + \beta_N{}' B_N \beta_N + E\left[(\alpha_{N+1} - \alpha_N)' A_{N+1} (\alpha_{N+1} - \alpha_N)\right]\}. \quad (2.42)$$

Replacing $E\left[(\alpha_N - \alpha_{N-1})' A_N (\alpha_N - \alpha_{N-1})\right]$ and $E\left[(\alpha_{N+1} - \alpha_N)' A_{N+1} (\alpha_{N+1} - \alpha_N)\right]$ by their values given by Lemmas II.6 and II.7 in Appendix 2.7.2 respectively, and combining terms results in

$$\begin{aligned} V_N(\wp_N) = {} & \left(\hat{\alpha}_{N|N} - \hat{\alpha}_{N-1|N}\right)' A_N \left(\hat{\alpha}_{N|N} - \hat{\alpha}_{N-1|N}\right) + \hat{\alpha}'_{N|N}\left((T_N - I)' A_{N+1}(T_N - I)\right)\hat{\alpha}_{N|N} \\ & + tr\left[A_N\left(\hat{\Sigma}_{N|N} + \hat{\Sigma}_{N-1|N} - T_{N-1}\hat{\Sigma}_{N-1|N} - \hat{\Sigma}_{N-1|N}T_{N-1}{}'\right)\right] \\ & + tr\left[A_{N+1}\left((T_N - I)\hat{\Sigma}_{N|N}(T_N - I)' + Q_N\right)\right] \\ & + \min_{\beta_N}\{\beta_N{}'(G_N{}' A_{N+1} G_N + B_N)\beta_N + \beta_N{}'(G_N{}' A_{N+1}(T_N - I)\hat{\alpha}_{N|N}) \\ & + (\hat{\alpha}'_{N|N}(T_N - I)' A_{N+1} G_N)\beta_N\}, \end{aligned} \quad (2.43)$$

where tr represents the trace of the matrix. Let the minimization term in Eq. 2.43 be denoted by $\tilde{J}_N$. That is, let

$$\tilde{J}_N = \min_{\beta_N}\{\beta_N{}'(G_N{}'A_{N+1}G_N + B_N)\beta_N + \beta_N{}'(G_N{}'A_{N+1}(T_N - I)\hat{\alpha}_{N|N}) + (\hat{\alpha}'_{N|N}(T_N - I)'A_{N+1}G_N)\beta_N\}. \tag{2.44}$$

This minimization can be performed by completion of squares as described in Lemma II.8 in Appendix 2.7.2. Eq.'s 2.45 - 2.48 give the optimum disease control β^*_N and the result of minimization $\tilde{J}_N$. The optimum disease control at time N is given by

$$\beta^*_N = -U_N\hat{\alpha}_{N|N}, \tag{2.45}$$

where the control law of last time period, U_N, is given by

$$U_N = (G_N{}'A_{N+1}G_N + B_N)^{-1}G_N{}'A_{N+1}(T_N - I), \tag{2.46}$$

and the result of minimization over β_N is given by

$$\tilde{J}_N = -\hat{\alpha}'_{N|N}\tilde{P}_{N+1}\hat{\alpha}_{N|N}, \tag{2.47}$$

where

$$\tilde{P}_{N+1} = (T_N - I)'A_{N+1}G_N(G'_NA_{N+1}G_N + B_N)^{-1}G'_NA_{N+1}(T_N - I). \tag{2.48}$$

Substitution of Eq. 2.47 into Eq. 2.43 yields

$$\begin{aligned} V_N(\wp_N) = & \left(\hat{\alpha}_{N|N} - \hat{\alpha}_{N-1|N}\right)' A_N \left(\hat{\alpha}_{N|N} - \hat{\alpha}_{N-1|N}\right) \\ & + \hat{\alpha}'_{N|N} \left((T_N - I)' A_{N+1} (T_N - I) - \tilde{P}_{N+1}\right) \hat{\alpha}_{N|N} \\ & + tr\left[A_N \left(\hat{\Sigma}_{N|N} + \hat{\Sigma}_{N-1|N} - T_{N-1}\hat{\Sigma}_{N-1|N} - \hat{\Sigma}_{N-1|N}{T_{N-1}}'\right)\right] \\ & + tr\left[A_{N+1}\left((T_N - I)\hat{\Sigma}_{N|N}(T_N - I)'\right)\right] + tr\left[A_{N+1}Q_N\right]. \end{aligned} \tag{2.49}$$

Defining P_N as follows,

$$P_N = (T_N - I)' A_{N+1} (T_N - I) - \tilde{P}_{N+1}, \tag{2.50}$$

and also replacing $tr\left[A_{N+1}\left((T_N - I)\hat{\Sigma}_{N|N}(T_N - I)'\right)\right]$ by its simpler form as identified in Lemma II.9 in Appendix 2.7.2, we can further simplify Eq. 2.49 as follows.

$$\begin{aligned} V_N(\wp_N) = & \left(\hat{\alpha}_{N|N} - \hat{\alpha}_{N-1|N}\right)' A_N \left(\hat{\alpha}_{N|N} - \hat{\alpha}_{N-1|N}\right) + tr\left[A_N \hat{\Sigma}_{N|N}\right] \\ & + \hat{\alpha}'_{N|N} P_N \hat{\alpha}_{N|N} + tr\left[P_N \hat{\Sigma}_{N|N}\right] \\ & + tr\left[A_N \left(\hat{\Sigma}_{N-1|N} - T_{N-1}\hat{\Sigma}_{N-1|N} - \hat{\Sigma}_{N-1|N}{T_{N-1}}'\right)\right] \\ & + tr\left[\tilde{P}_{N+1}\hat{\Sigma}_{N|N}\right] + tr\left[A_{N+1}Q_N\right]. \end{aligned} \tag{2.51}$$

2.7.2 Lemmas

In the derivation of optimal control the following lemmas are needed.

Lemma II.4. *For any symmetric $n * n$ matrix A, the following holds.*

$$E\left[x' A y\right] = \bar{x}' A \bar{y} + tr\left[A V_{x,y}\right]. \tag{2.52}$$

Where

$$\bar{x} = E[x], \tag{2.53}$$

$$\bar{y} = E[y], \tag{2.54}$$

$$V_{x,y} = E\left[(x - \bar{x})(y - \bar{y})'\right]. \tag{2.55}$$

Proof. By writing the matrix operations in terms of summations we will have

$$x'Ay = \sum_{i=1}^{n}\sum_{j=1}^{n} x_i A_{ij} y_j. \tag{2.56}$$

Hence

$$E\left[x'Ay\right] = \sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij} E[x_i y_j], \tag{2.57}$$

but

$$E[x_i y_j] = E[x_i]E[y_j] + E\left[\left(x_i - E[x_i]\right)\left(y_j - E[y_j]\right)\right] = \bar{x}_i\bar{y}_j + V_{x_i,y_j}. \tag{2.58}$$

When Eq. 2.58 is substituted in Eq. 2.57:

$$E\left[x'Ay\right] = \bar{x}'A\bar{y} + \sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}V_{x_i,y_j} = \bar{x}'A\bar{y} + tr\left[AV_{x,y}\right], \tag{2.59}$$

where $tr\left[M\right]$ stands for trace of M (i.e., sum of diagonal terms).

□

Lemma II.5. *With information up to time t, we have the following covariance rela-*

tions.

$$Cov(\alpha_t, \alpha_{t+1}|\wp_t) = \hat{\Sigma}_{t|t}\mathrm{T}_t{}', \tag{2.60}$$

$$Cov(\alpha_{t+1}, \alpha_t|\wp_t) = \mathrm{T}_t\hat{\Sigma}_{t|t}, \tag{2.61}$$

$$Cov(\alpha_t, \alpha_{t-1}|\wp_t) = \mathrm{T}_{t-1}\hat{\Sigma}_{t-1|t}, \tag{2.62}$$

$$Cov(\alpha_{t-1}, \alpha_t|\wp_t) = \hat{\Sigma}_{t-1|t}\mathrm{T}_{t-1}{}'. \tag{2.63}$$

Proof. Before we start proving these equations note that

$$Cov\left(\alpha_t, \alpha_t|\wp_t\right) = E\left[\alpha_t\alpha_t{}'\right] - E\left[\alpha_t\right]E[\alpha_t]' \rightarrow E\left[\alpha_t\alpha_t{}'\right] = \hat{\Sigma}_{t|t} + \hat{\alpha}_{t|t}\hat{\alpha}'_{t|t}. \tag{2.64}$$

Therefore,

$$\begin{aligned} Cov(\alpha_t, \alpha_{t+1}|\wp_t) &= E\left[\alpha_t\alpha_{t+1}{}'\right] - E\left[\alpha_t\right]E[\alpha_{t+1}]' \\ &= E\left[\alpha_t(\mathrm{T}_t\alpha_t + G_t\beta_t + \eta_t)'\right] - E\left[\alpha_t\right]E[\mathrm{T}_t\alpha_t + G_t\beta_t + \eta_t]' \\ &= E\left[\alpha_t\alpha_t{}'\right]\mathrm{T}_t{}' + \hat{\alpha}_{t|t}\beta_t{}'G_t{}' - \hat{\alpha}_{t|t}\left(\hat{\alpha}'_{t|t}\mathrm{T}_t{}' + \beta_t{}'G_t{}'\right) \\ &= \left(\hat{\Sigma}_{t|t} + \hat{\alpha}_{t|t}\hat{\alpha}'_{t|t}\right)\mathrm{T}_t{}' - \hat{\alpha}_{t|t}\hat{\alpha}'_{t|t}\mathrm{T}_t{}' \\ &= \hat{\Sigma}_{t|t}\mathrm{T}_t{}', \end{aligned} \tag{2.65}$$

and similarly

$$Cov(\alpha_{t+1}, \alpha_t|\wp_t) = \mathrm{T}_t\hat{\Sigma}_{t|t}. \tag{2.66}$$

Furthermore,

$$\begin{aligned} Cov(\alpha_t, \alpha_{t-1}|\wp_t) &= E\left[\alpha_t {\alpha_{t-1}}'\right] - E\left[\alpha_t\right] E[\alpha_{t-1}]' \\ &= E\left[\left(\mathrm{T}_{t-1}\alpha_{t-1} + G_{t-1}\beta_{t-1} + \eta_{t-1}\right) {\alpha_{t-1}}'\right] \\ &- E\left[\mathrm{T}_{t-1}\alpha_{t-1} + G_{t-1}\beta_{t-1} + \eta_{t-1}\right] E[\alpha_{t-1}]' \\ &= \mathrm{T}_{t-1} E\left[\alpha_{t-1}{\alpha_{t-1}}'\right] + G_{t-1}\beta_{t-1}\hat{\alpha}'_{t-1|t} - \mathrm{T}_{t-1}\hat{\alpha}_{t-1|t}\hat{\alpha}'_{t-1|t} - G_{t-1}\beta_{t-1}\hat{\alpha}'_{t-1|t} \\ &= \mathrm{T}_{t-1}\left(\hat{\Sigma}_{t-1|t} + \hat{\alpha}_{t-1|t}\hat{\alpha}'_{t-1|t}\right) - \mathrm{T}_{t-1}\hat{\alpha}_{t-1|t}\hat{\alpha}'_{t-1|t} \\ &= \mathrm{T}_{t-1}\hat{\Sigma}_{t-1|t}, \end{aligned} \tag{2.67}$$

and similarly

$$Cov(\alpha_{t-1}, \alpha_t|\wp_t) = \hat{\Sigma}_{t-1|t}{\mathrm{T}_{t-1}}'. \tag{2.68}$$

□

Lemma II.6. *With information up to time t, the expected quadratic penalty of disease progression from period $t-1$ to t can be calculated as follows.*

$$\begin{aligned} E\left[(\alpha_t - \alpha_{t-1})' A_t \left(\alpha_t - \alpha_{t-1}\right)|\wp_t\right] &= \left(\hat{\alpha}_{t|t} - \hat{\alpha}_{t-1|t}\right)' A_t \left(\hat{\alpha}_{t|t} - \hat{\alpha}_{t-1|t}\right) \\ &+ tr\left[A_t\left(\hat{\Sigma}_{t|t} + \hat{\Sigma}_{t-1|t} - T_{t-1}\hat{\Sigma}_{t-1|t} - \hat{\Sigma}_{t-1|t}{T_{t-1}}'\right)\right]. \end{aligned} \tag{2.69}$$

Proof. Using Lemma II.4 and Lemma II.5,

$$
\begin{aligned}
&E\left[(\alpha_t-\alpha_{t-1})'A_t\left(\alpha_t-\alpha_{t-1}\right)|\wp_t\right] = E\left[{\alpha_t}'A_t\alpha_t|\wp_t\right] + E\left[{\alpha_{t-1}}'A_t\alpha_{t-1}|\wp_t\right] \\
&\quad - E\left[{\alpha_t}'A_t\alpha_{t-1}|\wp_t\right] - E\left[{\alpha_{t-1}}'A_t\alpha_t|\wp_t\right] \\
&\quad = \hat{\alpha}'_{t|t}A_t\hat{\alpha}_{t|t} + tr[A_t\hat{\Sigma}_{t|t}] + \hat{\alpha}'_{t-1|t}A_t\hat{\alpha}_{t-1|t} + tr[A_t\hat{\Sigma}_{t-1|t}] \\
&\quad - \hat{\alpha}'_{t|t}A_t\hat{\alpha}_{t-1|t} - tr[A_tT_{t-1}\hat{\Sigma}_{t-1|t}] - \hat{\alpha}'_{t-1|t}A_t\hat{\alpha}_{t|t} - tr[A_t\hat{\Sigma}_{t-1|t}{T_{t-1}}'] \\
&\quad = \left(\hat{\alpha}_{t|t}-\hat{\alpha}_{t-1|t}\right)'A_t\left(\hat{\alpha}_{t|t}-\hat{\alpha}_{t-1|t}\right) + tr\left[A_t\left(\hat{\Sigma}_{t|t}+\hat{\Sigma}_{t-1|t}-T_{t-1}\hat{\Sigma}_{t-1|t}-\hat{\Sigma}_{t-1|t}{T_{t-1}}'\right)\right].
\end{aligned}
\tag{2.70}
$$

□

Lemma II.7. *With information up to time t, the expected quadratic penalty of disease progression from period t to t + 1 can be calculated as follows.*

$$
\begin{aligned}
&E\left[(\alpha_{t+1}-\alpha_t)'A_{t+1}\left(\alpha_{t+1}-\alpha_t\right)|\wp_t\right] \\
&\quad = \hat{\alpha}'_{t|t}\left((T_t-I)'A_{t+1}(T_t-I)\right)\hat{\alpha}_{t|t} + {\beta_t}'({G_t}'A_{t+1}G_t)\beta_t + {\beta_t}'({G_t}'A_{t+1}(T_t-I)\hat{\alpha}_{t|t}) \\
&\quad + (\hat{\alpha}'_{t|t}(T_t-I)'A_{t+1}G_t)\beta_t + tr\left[A_{t+1}\left((T_t-I)\hat{\Sigma}_{t|t}(T_t-I)'+Q_t\right)\right].
\end{aligned}
\tag{2.71}
$$

Proof. Using Eq. 2.1, Lemma II.4 and Lemma II.5,

$$
\begin{aligned}
E\left[(\alpha_{t+1}-\alpha_t)'A_{t+1}\left(\alpha_{t+1}-\alpha_t\right)|\wp_t\right] &= E\left[{\alpha_{t+1}}'A_{t+1}\alpha_{t+1}|\wp_t\right] + E\left[{\alpha_t}'A_{t+1}\alpha_t|\wp_t\right] \\
&- E\left[{\alpha_{t+1}}'A_{t+1}\alpha_t|\wp_t\right] - E\left[{\alpha_t}'A_{t+1}\alpha_{t+1}|\wp_t\right] \\
&= (T_t\hat{\alpha}_{t|t}+G_t\beta_t)'A_{t+1}(T_t\hat{\alpha}_{t|t}+G_t\beta_t) + tr\left[A_{t+1}(T_t\hat{\Sigma}_{t|t}{T_t}'+Q_t)\right] \\
&+ \hat{\alpha}'_{t|t}A_{t+1}\hat{\alpha}_{t|t} + tr[A_{t+1}\hat{\Sigma}_{t|t}] \\
&- (T_t\hat{\alpha}_{t|t}+G_t\beta_t)'A_{t+1}\hat{\alpha}_{t|t} - tr\left[A_{t+1}T_t\hat{\Sigma}_{t|t}\right] - \hat{\alpha}'_{t|t}A_{t+1}(T_t\hat{\alpha}_{t|t}+G_t\beta_t) - tr\left[A_{t+1}\hat{\Sigma}_{t|t}{T_t}'\right] \\
&= \hat{\alpha}'_{t|t}\left((T_t-I)'A_{t+1}(T_t-I)\right)\hat{\alpha}_{t|t} + {\beta_t}'({G_t}'A_{t+1}G_t)\beta_t + {\beta_t}'({G_t}'A_{t+1}(T_t-I)\hat{\alpha}_{t|t}) \\
&+ (\hat{\alpha}'_{t|t}(T_t-I)'A_{t+1}G_t)\beta_t + tr\left[A_{t+1}\left((T_t-I)\hat{\Sigma}_{t|t}(T_t-I)'+Q_t\right)\right]. \qquad (2.72)
\end{aligned}
$$

□

Lemma II.8.

$$
\begin{aligned}
\tilde{J}_N = \min_{\beta_N}\{&{\beta_N}'({G_N}'A_{N+1}G_N+B_N)\beta_N + {\beta_N}'({G_N}'A_{N+1}(T_N-I)\hat{\alpha}_{N|N}) \\
&+ (\hat{\alpha}'_{N|N}(T_N-I)'A_{N+1}G_N)\beta_N\} = -\hat{\alpha}'_{N|N}\tilde{P}_{N+1}\hat{\alpha}_{N|N}. \qquad (2.73)
\end{aligned}
$$

Proof. The minimization over β_N can be performed by completion of squares. For a detailed discussion on how to take minimization by completion of squares please see Section 3.3 of *Sayed* (2011). In here, we provide a short proof.

$\tilde{J}_N$ can be expressed in matrix form as follows.

$$
\tilde{J}_N = \min_{\beta_N}\left\{\begin{bmatrix} 1 & {\beta_N}' \end{bmatrix}\begin{bmatrix} 0 & \hat{\alpha}'_{N|N}(T_N-I)'A_{N+1}G_N \\ {G_N}'A_{N+1}(T_N-I)\hat{\alpha}_{N|N} & {G_N}'A_{N+1}G_N+B_N \end{bmatrix}\begin{bmatrix} 1 \\ \beta_N \end{bmatrix}\right\}. \qquad (2.74)
$$

The center matrix in 2.74 can be factored into a product of upper-triangular, diagonal,

and lower-triangular matrices as follows.

$$\tilde{J}_N = \min_{\beta_N}\{ \begin{bmatrix} 1 & {\beta_N}' \end{bmatrix} \begin{bmatrix} 1 & {\omega_N}' \\ 0 & I \end{bmatrix} \begin{bmatrix} -\hat{\alpha}'_{N|N}(T_N - I)' A_{N+1} G_N \omega_N & 0 \\ 0 & {G_N}' A_{N+1} G_N + B_N \end{bmatrix} \begin{bmatrix} 1 & 0 \\ \omega_N & I \end{bmatrix} \begin{bmatrix} 1 \\ \beta_N \end{bmatrix} \}, \tag{2.75}$$

where

$$\omega_N = ({G_N}' A_{N+1} G_N + B_N)^{-1} {G_N}' A_{N+1} (T_N - I) \hat{\alpha}_{N|N}. \tag{2.76}$$

Expanding the right-hand-side of Eq. 2.75 yields

$$\tilde{J}_N = \min_{\beta_N} \left\{ -\hat{\alpha}'_{N|N}(T_N - I)' A_{N+1} G_N \omega_N + (\beta_N + \omega_N)' \left({G_N}' A_{N+1} G_N + B_N \right) (\beta_N + \omega_N) \right\}, \tag{2.77}$$

in which only the second term depends on the unknown β_N. Note that $({G_N}' A_{N+1} G_N + B_N)$ is positive semidefinite. This is because A_{N+1} and B_N are diagonal cost matrices with only positive terms on the main diagonal. So, the second term in Eq. 2.77 is always nonnegative and will be minimized by choosing $\beta_N = -\omega_N$.

Therefore, the optimum disease control β_N^* and the result of minimization, i.e. $\tilde{J}_N$, are given by the following equations respectively.

$$\beta_N^* = -U_N \hat{\alpha}_{N|N}, \tag{2.78}$$

where

$$U_N = ({G_N}' A_{N+1} G_N + B_N)^{-1} {G_N}' A_{N+1} (T_N - I), \tag{2.79}$$

and

$$\tilde{J}_N = -\hat{\alpha}'_{N|N}\tilde{P}_{N+1}\hat{\alpha}_{N|N}, \tag{2.80}$$

where

$$\tilde{P}_{N+1} = (T_N - I)'A_{N+1}G_N(G'_N A_{N+1}G_N + B_N)^{-1}G'_N A_{N+1}(T_N - I). \tag{2.81}$$

□

Lemma II.9.

$$tr\left[A_{N+1}\left((T_N - I)\hat{\Sigma}_{N|N}(T_N - I)'\right)\right] = tr\left[P_N\hat{\Sigma}_{N|N}\right] + tr\left[\tilde{P}_{N+1}\hat{\Sigma}_{N|N}\right] \tag{2.82}$$

Proof. We know $tr\left[XY\right] = tr\left[YX\right]$, $tr\left[X\left(YZ\right)\right] = tr\left[\left(XY\right)Z\right]$ and $tr\left[\left(X+Y\right)Z\right] = tr\left[XZ\right] + tr\left[YZ\right]$. Therefore,

$$tr\left[A_{N+1}\left((T_N - I)\hat{\Sigma}_{N|N}(T_N - I)'\right)\right] = tr\left[\left((T_N - I)'A_{N+1}(T_N - I)\right)\hat{\Sigma}_{N|N}\right] \tag{2.83}$$

From Eq. 2.50 we know $(T_N - I)'A_{N+1}(T_N - I) = P_N + \tilde{P}_{N+1}$. Hence,

$$\begin{aligned} tr\left[\left((T_N - I)'A_{N+1}(T_N - I)\right)\hat{\Sigma}_{N|N}\right] &= tr\left[\left(P_N + \tilde{P}_{N+1}\right)\hat{\Sigma}_{N|N}\right] \\ &= tr\left[P_N\hat{\Sigma}_{N|N}\right] + tr\left[\tilde{P}_{N+1}\hat{\Sigma}_{N|N}\right] \end{aligned} \tag{2.84}$$

□

Lemma II.10.

$$\underset{z_{t+1}}{E}\left[\hat{\alpha}'_{t+1|t+1}P_{t+1}\hat{\alpha}_{t+1|t+1}|\wp_t\right] = (T_t\hat{\alpha}_{t|t} + G_t\beta_t)'P_{t+1}(T_t\hat{\alpha}_{t|t} + G_t\beta_t) + tr\left[P_{t+1}\left(T_t\hat{\Sigma}_{t|t}T_t' + Q_t - \hat{\Sigma}_{t+1|t+1}\right)\right]. \quad (2.85)$$

Proof. From Lemma II.4 and Eq. 2.4

$$\begin{aligned}
&\underset{z_{t+1}}{E}\left[\hat{\alpha}'_{t+1|t+1}P_{t+1}\hat{\alpha}_{t+1|t+1}|\wp_t\right] = \underset{z_{t+1}}{E}\left[\hat{\alpha}_{t+1|t+1}|\wp_t\right]'P_{t+1}\underset{z_{t+1}}{E}\left[\hat{\alpha}_{t+1|t+1}|\wp_t\right] \\
&+ tr\left[P_{t+1}\underset{z_{t+1}}{E}\left[\left(\hat{\alpha}_{t+1|t+1} - \underset{z_{t+1}}{E}\left[\hat{\alpha}_{t+1|t+1}|\wp_t\right]\right)\left(\hat{\alpha}_{t+1|t+1} - \underset{z_{t+1}}{E}\left[\hat{\alpha}_{t+1|t+1}|\wp_t\right]\right)'\right]\right] \\
&= (T_t\hat{\alpha}_{t|t} + G_t\beta_t)'P_{t+1}(T_t\hat{\alpha}_{t|t} + G_t\beta_t) + tr\left[P_{t+1}K_{t+1}\underset{z_{t+1}}{E}\left[\tilde{y}_{t+1}\tilde{y}'_{t+1}\right]K_{t+1}{}'\right], \quad (2.86)
\end{aligned}$$

in which

$$\underset{z_{t+1}}{E}\left[\tilde{y}_{t+1}\tilde{y}'_{t+1}\right] = \underset{z_{t+1}}{E}\left[\left(z_{t+1} - Z_{t+1}(T_t\hat{\alpha}_{t|t} + G_t\beta_t)\right)\left(z_{t+1} - Z_{t+1}(T_t\hat{\alpha}_{t|t} + G_t\beta_t)\right)'\right]. \quad (2.87)$$

Replacing z_{t+1} by its value given by Eq. 2.2 yields

$$\begin{aligned}
\underset{z_{t+1}}{E}\left[\tilde{y}_{t+1}\tilde{y}'_{t+1}\right] &= \underset{z_{t+1}}{E}\left[\left(\varepsilon_{t+1} + Z_{t+1}(\alpha_{t+1} - T_t\hat{\alpha}_{t|t} - G_t\beta_t)\right)\left(\varepsilon_{t+1} + Z_{t+1}(\alpha_{t+1} - T_t\hat{\alpha}_{t|t} - G_t\beta_t)\right)'\right] \\
&= \underset{z_{t+1}}{E}\left[\left(\varepsilon_{t+1} + Z_{t+1}(\alpha_{t+1} - E[\alpha_{t+1}])\right)\left(\varepsilon_{t+1} + Z_{t+1}(\alpha_{t+1} - E[\alpha_{t+1}])\right)'\right] \\
&= H_{t+1}^{(\theta_{t+1})} + Z_{t+1}\left(T_t'\hat{\Sigma}_{t|t}T_t + Q_t\right)Z'_{t+1} \\
&= S_{t+1}. \quad (2.88)
\end{aligned}$$

Substitution of Eq. 2.88 into Eq. 2.86 results

$$\underset{z_{t+1}}{E}\left[\hat{\alpha}'_{t+1|t+1}P_{t+1}\hat{\alpha}_{t+1|t+1}|\wp_t\right] = (T_t\hat{\alpha}_{t|t}+G_t\beta_t)'P_{t+1}(T_t\hat{\alpha}_{t|t}+G_t\beta_t) + tr\left[P_{t+1}K_{t+1}S_{t+1}K_{t+1}'\right]. \tag{2.89}$$

Using Eq.'s 2.35, 2.7 - 2.9 yields

$$\begin{aligned}\underset{z_{t+1}}{E}\left[\hat{\alpha}'_{t+1|t+1}P_{t+1}\hat{\alpha}_{t+1|t+1}|\wp_t\right] &= (T_t\hat{\alpha}_{t|t}+G_t\beta_t)'P_{t+1}(T_t\hat{\alpha}_{t|t}+G_t\beta_t) \\ &\quad + tr\left[P_{t+1}\left(T_t\hat{\Sigma}_{t|t}T_t'+Q_t-\hat{\Sigma}_{t+1|t+1}\right)\right]. \end{aligned}\tag{2.90}$$

□

Lemma II.11.

$$\begin{aligned}&\underset{z_{t+1}}{E}\left[\left(\hat{\alpha}_{t+1|t+1}-\hat{\alpha}_{t|t+1}\right)'A_{t+1}\left(\hat{\alpha}_{t+1|t+1}-\hat{\alpha}_{t|t+1}\right)|\wp_t\right] \\ &\quad = \left((T_t-I)\hat{\alpha}_{t|t}+G_t\beta_t\right)'A_{t+1}\left((T_t-I)\hat{\alpha}_{t|t}+G_t\beta_t\right) \\ &\quad + tr\left[A_{t+1}\left(T_t\hat{\Sigma}_{t|t}T_t'+Q_t-\hat{\Sigma}_{t+1|t+1}-\hat{\Sigma}_{t+1|t+1}\hat{\Sigma}_t^{*\prime}-\hat{\Sigma}_t^*\hat{\Sigma}_{t+1|t+1}+\hat{\Sigma}_{t|t+1}\right)\right].\end{aligned}\tag{2.91}$$

Proof. This expectation can be divided into four parts as follows.

$$\begin{aligned}\underset{z_{t+1}}{E}\left[\left(\hat{\alpha}_{t+1|t+1}-\hat{\alpha}_{t|t+1}\right)'A_{t+1}\left(\hat{\alpha}_{t+1|t+1}-\hat{\alpha}_{t|t+1}\right)|\wp_t\right] &= \underset{z_{t+1}}{E}\left[\hat{\alpha}'_{t+1|t+1}A_{t+1}\hat{\alpha}_{t+1|t+1}|\wp_t\right] \\ - \underset{z_{t+1}}{E}\left[\hat{\alpha}'_{t+1|t+1}A_{t+1}\hat{\alpha}_{t|t+1}|\wp_t\right] - \underset{z_{t+1}}{E}\left[\hat{\alpha}'_{t|t+1}A_{t+1}\hat{\alpha}_{t+1|t+1}|\wp_t\right] &+ \underset{z_{t+1}}{E}\left[\hat{\alpha}'_{t|t+1}A_{t+1}\hat{\alpha}_{t|t+1}|\wp_t\right].\end{aligned}\tag{2.92}$$

The first expectation in 2.92 is similar to the expectation of Lemma II.10. Therefore,

$$\underset{z_{t+1}}{E}\left[\hat{\alpha}'_{t+1|t+1}A_{t+1}\hat{\alpha}_{t+1|t+1}|\wp_t\right] = (T_t\hat{\alpha}_{t|t}+G_t\beta_t)'A_{t+1}(T_t\hat{\alpha}_{t|t}+G_t\beta_t) + tr\left[A_{t+1}\left(T_t\hat{\Sigma}_{t|t}T_t'+Q_t-\hat{\Sigma}_{t+1|t+1}\right)\right]. \quad (2.93)$$

The second expectation in 2.92 can be simplified as follows.

$$\begin{aligned}\underset{z_{t+1}}{E}\left[\hat{\alpha}'_{t+1|t+1}A_{t+1}\hat{\alpha}_{t|t+1}|\wp_t\right] &= \underset{z_{t+1}}{E}\left[\hat{\alpha}_{t+1|t+1}|\wp_t\right]'A_{t+1}\underset{z_{t+1}}{E}\left[\hat{\alpha}_{t|t+1}|\wp_t\right] \\ &\quad + tr\left[A_{t+1}Cov\left(\hat{\alpha}_{t+1|t+1},\hat{\alpha}_{t|t+1}\right)\right] \\ &= (T_t\hat{\alpha}_{t|t}+G_t\beta_t)'A_{t+1}\hat{\alpha}_{t|t}+tr\left[A_{t+1}Cov\left(\hat{\alpha}_{t+1|t+1},\hat{\alpha}_{t|t+1}\right)\right], \end{aligned} \quad (2.94)$$

in which

$$\begin{aligned}Cov\left(\hat{\alpha}_{t+1|t+1},\hat{\alpha}_{t|t+1}|\wp_t\right) &= E\left[\hat{\alpha}_{t+1|t+1}\hat{\alpha}'_{t|t+1}|\wp_t\right]-E\left[\hat{\alpha}_{t+1|t+1}|\wp_t\right]E\left[\hat{\alpha}_{t|t+1}|\wp_t\right]' \\ &= E\left[\hat{\alpha}_{t+1|t+1}\left(\hat{\alpha}_{t|t}+\hat{\Sigma}^*_t(\hat{\alpha}_{t+1|t+1}-\hat{\alpha}_{t+1|t})\right)'|\wp_t\right] \\ &\quad - E\left[\hat{\alpha}_{t+1|t+1}|\wp_t\right]\hat{\alpha}'_{t|t} \\ &= E\left[\hat{\alpha}_{t+1|t+1}(\hat{\alpha}_{t+1|t+1}-\hat{\alpha}_{t+1|t})'|\wp_t\right]\hat{\Sigma}^{*\prime}_t \\ &= E\left[\hat{\alpha}_{t+1|t+1}\hat{\alpha}'_{t+1|t+1}|\wp_t\right]\hat{\Sigma}^{*\prime}_t - E\left[\hat{\alpha}_{t+1|t+1}|\wp_t\right](T_t\hat{\alpha}_{t|t}+G_t\beta_t)'\hat{\Sigma}^{*\prime}_t \\ &= \hat{\Sigma}_{t+1|t+1}\hat{\Sigma}^{*\prime}_t+(T_t\hat{\alpha}_{t|t}+G_t\beta_t)(T_t\hat{\alpha}_{t|t}+G_t\beta_t)'\hat{\Sigma}^{*\prime}_t \\ &\quad - (T_t\hat{\alpha}_{t|t}+G_t\beta_t)(T_t\hat{\alpha}_{t|t}+G_t\beta_t)'\hat{\Sigma}^{*\prime}_t \\ &= \hat{\Sigma}_{t+1|t+1}\hat{\Sigma}^{*\prime}_t. \end{aligned} \quad (2.95)$$

Therefore,

$$\underset{z_{t+1}}{E}\left[\hat{\alpha}'_{t+1|t+1}A_{t+1}\hat{\alpha}_{t|t+1}|\wp_t\right] = (T_t\hat{\alpha}_{t|t}+G_t\beta_t)'A_{t+1}\hat{\alpha}_{t|t}+tr\left[A_{t+1}\hat{\Sigma}_{t+1|t+1}\hat{\Sigma}^{*\prime}_t\right]. \quad (2.96)$$

In similar way, the third and fourth expectations in 2.92 can be simplified as follows.

$$\underset{z_{t+1}}{E}\left[\hat{\alpha}'_{t|t+1}A_{t+1}\hat{\alpha}_{t+1|t+1}|\wp_t\right] = \hat{\alpha}'_{t|t}A_{t+1}(T_t\hat{\alpha}_{t|t} + G_t\beta_t) + tr\left[A_{t+1}\hat{\Sigma}^*_t\hat{\Sigma}_{t+1|t+1}\right], \quad (2.97)$$

$$\underset{z_{t+1}}{E}\left[\hat{\alpha}'_{t|t+1}A_{t+1}\hat{\alpha}_{t|t+1}|\wp_t\right] = \hat{\alpha}'_{t|t}A_{t+1}\hat{\alpha}_{t|t} + tr\left[A_{t+1}\hat{\Sigma}_{t|t+1}\right]. \quad (2.98)$$

Replacing Eq.'s 2.93, 2.96, 2.97 and 2.98 into 2.92 yields

$$\begin{aligned}
\underset{z_{t+1}}{E}&\left[\left(\hat{\alpha}_{t+1|t+1} - \hat{\alpha}_{t|t+1}\right)'A_{t+1}\left(\hat{\alpha}_{t+1|t+1} - \hat{\alpha}_{t|t+1}\right)|\wp_t\right] = (T_t\hat{\alpha}_{t|t} + G_t\beta_t)'A_{t+1}(T_t\hat{\alpha}_{t|t} + G_t\beta_t) \\
&+ tr\left[A_{t+1}\left(T_t\hat{\Sigma}_{t|t}T_t' + Q_t - \hat{\Sigma}_{t+1|t+1}\right)\right] - (T_t\hat{\alpha}_{t|t} + G_t\beta_t)'A_{t+1}\hat{\alpha}_{t|t} - tr\left[A_{t+1}\hat{\Sigma}_{t+1|t+1}\hat{\Sigma}^{*\prime}_t\right] \\
&- \hat{\alpha}'_{t|t}A_{t+1}(T_t\hat{\alpha}_{t|t} + G_t\beta_t) - tr\left[A_{t+1}\hat{\Sigma}^*_t\hat{\Sigma}_{t+1|t+1}\right] + \hat{\alpha}'_{t|t}A_{t+1}\hat{\alpha}_{t|t} + tr\left[A_{t+1}\hat{\Sigma}_{t|t+1}\right] \\
&= \left((T_t - I)\hat{\alpha}_{t|t} + G_t\beta_t\right)'A_{t+1}\left((T_t - I)\hat{\alpha}_{t|t} + G_t\beta_t\right) \\
&+ tr\left[A_{t+1}\left(T_t\hat{\Sigma}_{t|t}T_t' + Q_t - \hat{\Sigma}_{t+1|t+1} - \hat{\Sigma}_{t+1|t+1}\hat{\Sigma}^{*\prime}_t - \hat{\Sigma}^*_t\hat{\Sigma}_{t+1|t+1} + \hat{\Sigma}_{t|t+1}\right)\right].
\end{aligned} \quad (2.99)$$

□

Lemma II.12.

$$tr\left[(A_{t+1} + P_{t+1})\left(T_t\hat{\Sigma}_{t|t}T_t'\right)\right] = tr\left[P_t\hat{\Sigma}_{t|t}\right] + tr\left[\left(\tilde{P}_{t+1} + A_{t+1}T_t + T_t'A_{t+1} - I\right)\hat{\Sigma}_{t|t}\right] \quad (2.100)$$

Proof. We know $tr[XY] = tr[YX]$, $tr[X(YZ)] = tr[(XY)Z]$ and $tr[(X+Y)Z] = tr[XZ] + tr[YZ]$. Therefore,

$$tr\left[(A_{t+1} + P_{t+1})\left(T_t\hat{\Sigma}_{t|t}T_t'\right)\right] = tr\left[\left(T_t'(A_{t+1} + P_{t+1})T_t\right)\hat{\Sigma}_{t|t}\right] \quad (2.101)$$

Using Eq. 2.35 to replace $T_t'\left(A_{t+1}+P_{t+1}\right)T_t$ we have

$$\begin{aligned} tr\left[\left(T_t'\left(A_{t+1}+P_{t+1}\right)T_t\right)\hat{\Sigma}_{t|t}\right] &= tr\left[\left(\tilde{P}_{t+1}+P_t+A_{t+1}T_t+T_t'A_{t+1}-I\right)\hat{\Sigma}_{t|t}\right] \\ &= tr\left[P_t\hat{\Sigma}_{t|t}\right]+tr\left[\left(\tilde{P}_{t+1}+A_{t+1}T_t+T_t'A_{t+1}-I\right)\hat{\Sigma}_{t|t}\right]. \end{aligned} \tag{2.102}$$

□

2.7.3 Results on Target IOP and MD Loss Averted

Since target IOP is an important metric that helps guide clinicians in selecting the appropriate treatment plan for the patient, the distribution of target IOPs is also of interest. Figure 2.9 shows the histogram of target IOPs for fast and slow-progressing patients under the high and moderate aggressiveness policies. The range and mean of each category is clinically appropriate in the professional opinion of our glaucoma specialist collaborator. Figure 2.10 graphs the MD loss averted in [dB]

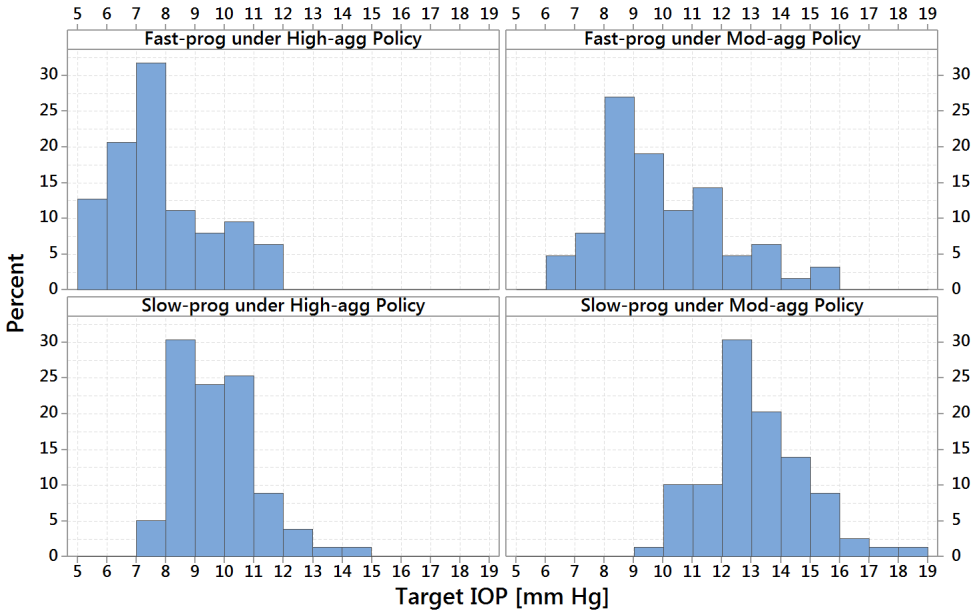

Figure 2.9: Histogram of target IOPs for CIGTS and AGIS patients under different aggressiveness policies.

for fast and slow-progressing patients under the high and moderate aggressiveness policies compared against the low aggressiveness policy over 10 years of following the IOP controls suggested by our model. As seen in the Figure, fast-progressing patients will lose fewer MD points (i.e., experience better vision quality) resulting from further lowering their eye pressure in short term, whether the doctor chooses moderate or high aggressiveness level. Slow-progressors, if treated under the high aggressiveness policy, could benefit from losing fewer MD points in the long term. However, this group of glaucoma patients does not gain evident benefit from employing the moderate aggressiveness policy even in the long term.

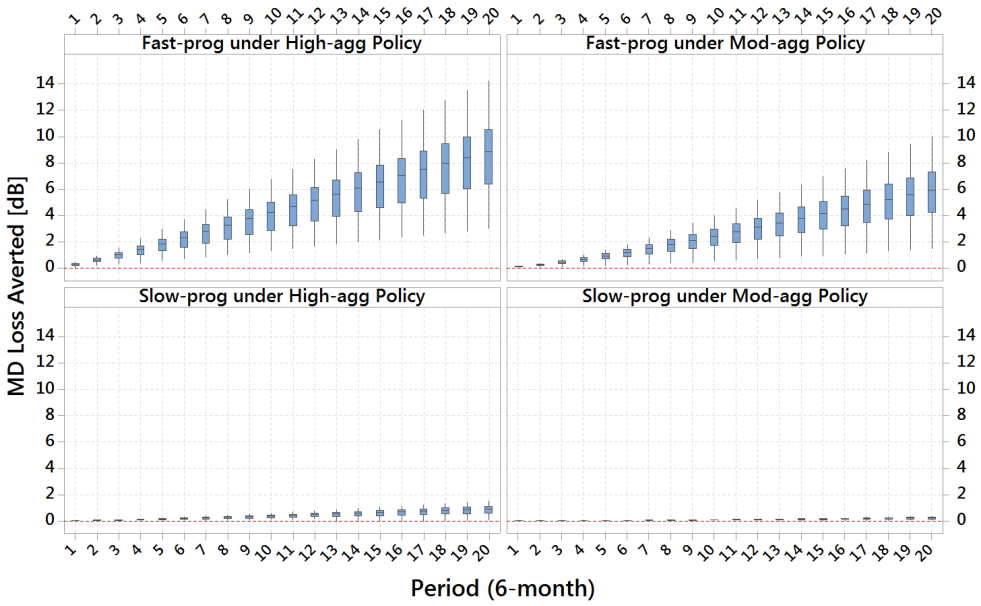

Figure 2.10: MD loss averted [dB] for fast and slow-progressing patients under the high and moderate aggressiveness policies compared against the low aggressiveness policy (i.e., no additional IOP reduction beyond those employed in trials) over 10 years of following the IOP control suggested by our model. Period 1 is six months into the future; period 20 is 10 years into the future.

CHAPTER III

Healthcare Provider Shift Design to Minimize Patient Handoffs

3.1 Introduction and Background

In November 1999, the U.S. Institute of Medicine (IOM) issued a report on medical errors estimating that nearly 100,000 patients die each year as a result of medical errors and another 15 million are harmed (*Kohn et al.* 2000). Root cause analysis of reported sentinel events from 1994 to 2004 revealed that two-thirds of these errors were due to communication failures (*Volpp and Grande* 2003). According to Dr. Lucien Leape, the number of deaths from medical errors in hospitals is equivalent to the death toll from three jumbo jet crashes every two days (*Leape et al.* 1999). In fact, more people die as a result of medical errors than from motor vehicle accidents, breast cancer, or AIDS. Recent reports continue to support the initial findings from IOM (*Institute of Medicine, Committee on Quality of Health Care in America* 2001) and connect fatigue-related medical errors with residents' duty hours (*Ulmer et al.* 2008).

In September 2010, the Accreditation Council for Graduate Medical Education (ACGME) enacted new duty-hour regulations for residents and fellows (see *Nasca et al.* 2010) that limited weekly work hours, length of duty periods, off time between shifts, and

frequency of consecutive on-call days and nights. These stricter duty-hour requirements went into effect on July 1, 2011. The goal was to reduce fatigue-related medical errors and improve patient safety by limiting residents/fellows (trainees) work hours. However, the more restrictive shifts have resulted in a significant increase in patient *handoffs* and communication failures (see *Vidyarthi et al.* 2006; *Horwitz et al.* 2006; *Philibert and Leach* 2005; *Horwitz et al.* 2007; *Hutter et al.* 2006; *Lockley et al.* 2006). Patient handoff is defined as "the process of transferring primary authority and responsibility for providing clinical care to a patient from one departing caregiver to one oncoming caregiver" (*Patterson and Wears* 2010). Some other terms that have commonly been used for handoff in the literature are *handover, sign-out, turnover, transition of care, transfer of care* and *shift change transfer*. It is worth noting that the working shifts of both residents and fellows must comply with the ACGME duty hour regulations. For ease of reference, we use the term "trainees" to refer to both residents and fellows in the rest of the chapter. We also refer to postgraduate year 2 (PGY-2) and above residents and all fellows as "senior trainees".

Several studies have correlated increased patient handoffs with more medical errors caused by communication breakdowns and therefore worse patient outcomes (see *Risser et al.* 1999; *Sutcliffe et al.* 2004; *Arora et al.* 2007; *Frankel et al.* 2006; *Jagsi et al.* 2005; *Greenberg et al.* 2007; *Landrigan et al.* 2004; *Petersen et al.* 1994; *Li et al.* 2011; *Kitch et al.* 2008; *Gandhi* 2005). It is believed that 20% - 30% of information conveyed during patient handoffs is not documented in the medical record (see *Sexton et al.* 2004; *Risser et al.* 1999). Figure 3.1 depicts the connection between the new ACGME duty-hour regulations and medical errors.

Several studies have focused on the communication aspects of handoffs and have provided recommendations to achieve high quality handoffs (see *Ye et al.* 2007; *Arora and Johnson* 2006; *Solet et al.* 2005, 2004; *Nemeth* 2012; *Arora et al.* 2005, 2008; *Henriksen et al.* 2008; *Cheung et al.* 2010; *Donchin et al.* 2003; *Kemp et al.* 2008). For

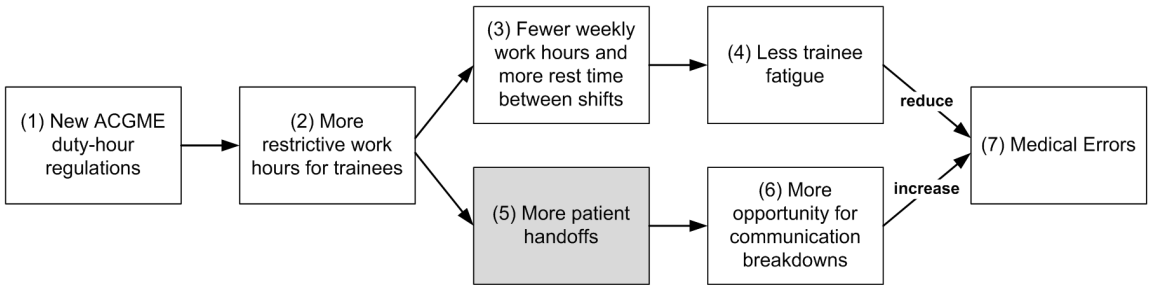

Figure 3.1: Connection between ACGME standards and medical errors - the net effect is uncertain.

example, *Kemp et al.* (2008) presented a methodology for conducting safe and effective sign-outs in a surgical service. *Clark et al.* (2011) designed a sign-out template to standardize the handoff process in a general surgery residency program. Other studies have focused on the overall handoff process. For instance, *Abraham et al.* (2012) proposed a clinician-centered approach that captures the entire clinician workflow prior to, during, and after handoff communication. Most of these studies have employed interviews, surveys, and observations to understand handoff failures and provide suggestions to enhance handoff fidelity.

While a high-quality, structured handoff process is important, decreasing number of patient handoffs is an additional and fundamental way to reduce opportunities for medical errors caused by communication breakdowns, supporting safer and more efficient patient care. The new ACGME duty-hour standards themselves specifically emphasize the importance of reducing handoffs:

"Programs must design clinical assignments to minimize the number of transitions in patient care."

Borman et al. (2012) recently surveyed surgery residents and identified that resident perceptions of causes of medical errors suggest that system changes are more likely to enhance patient safety than further hour limits. This research provides mathematical methods for effecting such system change, by redesigning schedules to reduce the

number of patient handoffs, hence reducing the opportunity for communication error. While recent research has focused on improving the quality of communication during the handoff process (i.e. improving box 6 in Figure 3.1) to reduce medical errors caused by communication breakdowns, to the best of our knowledge, no prior work leverages physician scheduling to reduce the quantity of handoffs (i.e. improving box 5 in Figure 3.1). This research contributes to the handoff literature by providing an Integer Programming (IP) approach to design trainees' schedules in a patient-centered manner that minimizes number of handoffs while respecting ACGME duty-hour standards. The methodology we develop is highly generalizable and, while the proof of concept is developed for an Intensive Care Unit (ICU), our approach can be applied to many different care units in hospitals, including different intensive care unit types (e.g. Medical ICU, Surgical ICU, Pediatric ICU, Critical Care Unit, etc.), the emergency department, and general floor care (internal medicine or surgery). This approach can also be employed for different provider levels, e.g. attending physicians, fellows, residents, nurses, etc.

Mathematical optimization techniques have been widely used to solve the physician and nurse scheduling problems in a provider-centered manner (see *Carter and Lapierre* 2001; *Ernst et al.* 2004; *Aickelin and Dowsland* 2004; *Gutjahr and Rauner* 2007). In the physician scheduling problem, given a set of doctors, a set of shifts and a planning period, one seeks to find fair schedules for all physicians (*Gendreau et al.* 2006). In the nurse scheduling problem, the cost of salaries should also be minimized.

Several studies have employed integer programming to formulate and solve the physician and nurse scheduling problems (see *Gascon et al.* 2000; *Bard and Purnomo* 2005; *Beaulieu et al.* 2000; *Sherali et al.* 2002; *Cohn et al.* 2009). These studies provide mathematical models to assign healthcare providers to pre-determined fixed shifts (i.e. shift assignment models). *Gascon et al.* (2000) studied the flying squad nurse scheduling problem. A multi-objective integer programming problem with binary

variables was employed to find a feasible schedule satisfying most of the constraints. The paper combined the sequential and the weighted method to obtain the best nurse schedule for minimizing the deviation measures in soft constraints. *Bard and Purnomo* (2005) developed an integer programming model to produce a revised schedule for regular and pool nurses to efficiently use them in the event of surge in demand for nursing services. The objective is to achieve sufficient coverage with the minimum cost of revising nurses' schedules. *Beaulieu et al.* (2000) addressed the problem of physician scheduling. This paper employed integer programming to make a schedule for physicians in the emergency room of a major hospital in Montreal, Canada. The model was able to generate a better schedule with smaller deviations from desired metrics in much shorter amount of time than the current method being used by hospital staff. *Sherali et al.* (2002) proposed mixed-integer programming models to address the resident scheduling problem concerned with prescribing work-nights for residents. Heuristic solution approaches were developed to solve the problem under different scenarios. *Cohn et al.* (2009) also combined IP-based techniques with user expertise and heuristic approaches to construct high-quality schedules for residents in the psychiatry program at Boston University School of Medicine based on their individual preferences.

Our IP-based shift *design and assignment* model differs in that it simultaneously (1) finds the best times for starting and ending the shifts to minimize the number of patient handoffs (this is the shift design part), and (2) assigns physicians to the shifts such that all ACGME duty-hour regulations are satisfied, required coverage is achieved, and livability rules are met (this is the shift assignment part). The shift design concept brings a new perspective to the problem of how best to incorporate the ACGME rules and provides a systematic, model-driven method for *designing* physicians' schedules compared to the conventional approach of selecting between either two 12-hour shifts or three 8-hour shifts per day. It further benefits patients by min-

imizing error-contributing handoffs while maintaining physicians' quality of life.

3.2 Model Development

In this section we present model assumptions, sets, parameters, variables, constraints and objective function. The parametric model in this section is based on the ICU setting for scheduling trainees at the Mayo Clinic; however, the same model (perhaps with slight modification) could be used for other hospital care units.

3.2.1 Assumptions

Because of the IP framework and also to ensure tractable and practical solutions, it is necessary to divide each day into discrete time blocks and to assume that shift change can happen only at the start/end of these time blocks. For example, if a day is divided evenly into 6 time blocks, each block would be 4 hours and shift changes can occur only at times 0, 4, 8, 12, 16 and 20. In other words, each physician can either work or not work in a full time block. We also approximate the number of patients handed off in each shift change based on historical data on ICU patient census by time of day and day of week.

3.2.2 Sets and Parameters

We use the following sets in our model.

- I: set of trainees,
- J: set of days within the planning horizon,
- T: set of weeks within the planning horizon,
- K: set of time blocks within a day,

- Kn: set of time blocks corresponding to night shift,
- Kr: set of time blocks that end during rounding time interval,
- $Kinc$: set of time blocks that end during inconvenient time interval for shift change (usually considered as late night and early morning).

The main model parameters are listed below.

- NbF: number of trainees (fellows or residents),
- NbD: number of days within the planning horizon,
- NbW: number of weeks within the planning horizon,
- NbB: number of time blocks within a day,
- ShL: maximum shift length allowed in hours,
- c_{jk}: approximate number of patient handoffs incurred by a shift change at the end of time block k in day j, calculated based on the average number of patients in the ICU at the time of shift change,
- d_{jk}: minimum number of trainees required to be on-call at time block k of day j.

We also use a few auxiliary parameters in our model to simplify the notation. They are directly calculated from the main parameters.

- $BL = \frac{24}{NbB}$: length of each time block in hours,
- $B^{ShL} = \lfloor \frac{ShL}{BL} \rfloor$: maximum number of consecutive time blocks which do not exceed ShL hours,
- $B_{10} = \lceil 10/BL \rceil$: minimum number of consecutive time blocks which exceed 10 hours.

3.2.3 Decision Variables

The following decision variables are used in the model.

- x_{ijk}: 1 if trainee i is assigned to time block k on day j, and 0 otherwise;
- y_{jk}: 1 if there is a shift change at the end of time block k on day j, and 0 otherwise;
- z_{ij}: 1 if trainee i is totally off-duty on day j, and 0 otherwise;
- w_{ij}: 1 if trainee i works at night on day j, and 0 otherwise.

3.2.4 Constraints

The model constraints can be classified into three categories: (1) required and (2) desirable constraints are associated with mandatory and optional scheduling rules respectively, while (3) linkage constraints enforce model dynamics.

3.2.4.1 Required Constraints

Certain constraints are required by regulation or organizational policy. The first five of these constraints are required by ACGME duty-hour regulations, while the last two are required by organizational policy.

1. Duty periods of postgraduate year 1 (PGY-1) residents must not exceed 16 hours in duration; however, senior trainees may be scheduled to a maximum of 24 hours of continuous duty. The following inequalities ensure that trainees do not work shifts longer than ShL hours. ShL is the maximum shift length allowed in hours so we use this value as an upper bound for shift length.

$$\sum_{k=s}^{B^{ShL}+s} x_{ijk} \leq \frac{ShL}{BL} \qquad \forall i \in I, \forall j \in J, \forall s \in \{1, ..., (NbB - B^{ShL})\}, \qquad (3.1)$$

$$\sum_{k=s}^{NbB} x_{ijk} + \sum_{k=1}^{s-(NbB-B^{ShL})} x_{i,j+1,k} \leq \frac{ShL}{BL} \qquad \forall i \in I, \forall j \in \{1, ..., NbD-1\},$$
$$\forall s \in \{(NbB - B^{ShL}) + 1, ..., NbB\}. \tag{3.2}$$

If $ShL = 24$ (i.e. the maximum allowed shift length is 24 hours), inequality 3.1 is not needed and only inequality 3.2 is kept.

2. Weekly duty hours must not exceed 80 hours:

$$\sum_{j=7(t-1)+1}^{7t} \sum_{k \in K} x_{ijk} \leq \frac{80}{BL} \qquad \forall i \in I, \forall t \in T. \tag{3.3}$$

3. Trainees must have a minimum of 10 hours free of duty between scheduled duty periods:

$$x_{i,j,k} - x_{i,j,k+1} + x_{i,j,k+s+2} \leq 1 \qquad \forall i \in I, \forall j \in J, \forall k \in \{1, ..., NbB - B_{10}\},$$
$$\forall s \in \{0, ..., B_{10} - 2\}, \tag{3.4}$$

$$x_{i,j,k} - x_{i,j,k+1} + x_{i,j,k+s+1} \leq 1 \qquad \forall i \in I, \forall j \in J,$$
$$\forall k \in \{NbB - B_{10} + 1, ..., NbB - 1\},$$
$$\forall s \in \{0, ..., (NbB - 1) - k\}, \tag{3.5}$$

$$x_{i,j,k} - x_{i,j,k+1} + x_{i,j+1,s+1} \leq 1 \qquad \forall i \in I, \forall j \in \{1, ..., NbD-1\},$$
$$\forall k \in \{NbB - B_{10} + 1, ..., NbB - 1\},$$
$$\forall s \in \{0, ..., k - (NbB - B_{10} + 1)\}, \tag{3.6}$$

$$x_{i,j,NbB} - x_{i,j+1,1} + x_{i,j+1,s+2} \leq 1 \qquad \forall i \in I, \forall j \in \{1, ..., NbD - 1\}, \forall s \in \{0, ..., B_{10} - 2\}. \tag{3.7}$$

If $NbB \geq 3$, inequalities 3.4 - 3.7 are required. Otherwise, this rule is automatically satisfied by other required constraints and the above inequalities are not needed to make sure trainees will get at least 10 hours off between shifts.

4. Trainees must get at least one day off per 7-day period (when averaged over 4 weeks):

$$\sum_{j=7(t-1)+1}^{7(t-1)+28} z_{ij} \geq 4 \qquad \forall i \in I, \forall t \in \{1, ..., NbW - 3\}. \tag{3.8}$$

5. Trainees must not be scheduled for more than 6 consecutive shifts of night duty (night float):

$$\sum_{s=0}^{6} w_{i,j+s} \leq 6 \qquad \forall i \in I, \forall j \in \{1, ..., NbD - 6\}. \tag{3.9}$$

6. The required coverage must be satisfied (coverage constraint):

$$\sum_{i \in I} x_{ijk} \geq d_{jk} \qquad \forall j \in J, \forall k \in K. \tag{3.10}$$

7. Shift change is not allowed during bedside multi-disciplinary rounds because this would disrupt the rounding process and impact the educational benefit to trainees:

$$y_{jk} = 0 \qquad \forall j \in J, \forall k \in Kr. \tag{3.11}$$

3.2.4.2 Linkage Constraints

The following inequalities serve as linkage constraints to connect x, y, z and w variables.

1. Inequalities 3.12 - 3.15 ensure that whenever there is a shift change at the end of time block k on day j, variable y_{jk} is assigned value 1.

$$y_{jk} \geq x_{ijk} - x_{i,j,k+1} \qquad \forall i \in I, \forall j \in J, \forall k \in \{1, ..., NbB - 1\}, \tag{3.12}$$

$$y_{jk} \geq x_{i,j,k+1} - x_{ijk} \qquad \forall i \in I, \forall j \in J, \forall k \in \{1, ..., NbB - 1\}, \tag{3.13}$$

$$y_{j,NbB} \geq x_{i,j,NbB} - x_{i,j+1,1} \qquad \forall i \in I, \forall j \in \{1, ..., NbD - 1\}, \tag{3.14}$$

$$y_{j,NbB} \geq x_{i,j+1,1} - x_{i,j,NbB} \qquad \forall i \in I, \forall j \in \{1, ..., NbD - 1\}. \tag{3.15}$$

2. The following inequality ensures that whenever z_{ij} is 1, trainee i is off-duty on day j.

$$\sum_{k \in K} x_{ijk} \leq NbB(1 - z_{ij}) \qquad \forall i \in I, \forall j \in J. \tag{3.16}$$

3. Inequalities 3.17 and 3.18 ensure that w_{ij} is 1 when trainee i works at night on day j, and 0 otherwise.

$$x_{ijk} \leq w_{ij} \qquad \forall i \in I, \forall j \in J, \forall k \in Kn, \tag{3.17}$$

$$w_{ij} \leq \sum_{k \in Kn} x_{ijk} \qquad \forall i \in I, \forall j \in J. \tag{3.18}$$

3.2.4.3 Desired Constraints

In addition to the required constraints, there are some other characteristics for a schedule which are not required, but are desirable to obtain more convenient and livable schedules. These could include vacation requests, sleep hours, circadian rhythm

or other human factors issues. In this part, we discuss the desired constraints in our model. These rules have been developed through several meetings and discussions with program directors, consultants, chief residents and fellows at Mayo Clinic.

1. To maintain regular sleep hours for trainees, we disallow shift changes at late night or early morning (inconvenient times).

$$y_{jk} = 0 \qquad \forall j \in J, \forall k \in Kinc. \tag{3.19}$$

2. ACGME rules only require one day off per 7-day period (when averaged over 4 weeks). However, working several days in a row could cause fatigue, irritability and reduced concentration for trainees. Hence, the desire is to provide at least one day off per any 7-day period (without averaging). This means that, trainees are permitted to work no more than six days in a row.

$$\sum_{s=0}^{6} z_{i,j+s} \geq 1 \qquad \forall i \in I, \forall j \in \{1, ..., NbD - 6\}. \tag{3.20}$$

3. Based on ACGME regulations, trainees are allowed to be on call for up to six consecutive night shifts. Nevertheless, it is believed that doing a lengthy run of night shifts might be associated with extreme fatigue, insomnia, and sleep deprivation. Hence, we limit the night float to a maximum of four consecutive night shifts.

$$\sum_{s=0}^{4} w_{i,j+s} \leq 4 \qquad \forall i \in I, \forall j \in \{1, ..., NbD - 4\}. \tag{3.21}$$

4. The minimum required off-time between scheduled duty periods is considered as 10 hours by ACGME. However, switching to day time work after doing a run of night shifts is hard for human brain. Preferably, trainees would have at least

Table 3.1: Scheduling constraints: required constraints (RC) and desired constraints (DC).

Scheduling Constraints	
RC1	Trainees must not work longer than *ShL* hours on a single shift.
RC2	Trainees must not work more than 80 hours per week.
RC3	Trainees must get at least 10 hours off-duty between shifts.
RC4	Trainees must get at least one day off per 7 days (averaged over 4 weeks).
RC5	Trainees must not be scheduled for more than six consecutive night shifts.
RC6	The required coverage must be satisfied.
RC7	Shift change is not allowed during bedside multi-disciplinary rounds.
DC1	Shift change is not allowed at late night or early morning (inconvenient times).
DC2	Trainees should not work more than six days in a row.
DC3	Trainees work no more than four night shifts in a row.
DC4	Trainees have at least one day off after a run of night shifts.
DC5	Shifts longer than 12 hours are not allowed.

one whole day off (in addition to the post-call day) after doing a run of night shifts to better adjust their sleep pattern. The following inequality ensures that after each night shift, either another night shift or a day off should be assigned to trainees.

$$w_{ij} - w_{i,j+1} - z_{i,j+1} \leq 0 \qquad \forall i \in I, \forall j \in \{1, ..., NbD - 1\}. \tag{3.22}$$

5. Although ACGME duty hour regulations allow a shift to last up to 16 hours for PGY-1 and 24 hours for PGY-2 and above, shifts longer than 12 hours are believed to be associated with fatigue, headaches, irritability and reduced concentration. Hence, we limit the shift length to 12 hours by setting the value of ShL parameter to 12 in inequalities 3.1 and 3.2.

To summarize our discussion in this section, all the Required Constraints (RCs) and Desired Constraints (DCs) are listed in Table 3.1.

3.2.5 Objective Function

The objective is to minimize the approximate number of patient handoffs during the scheduling horizon, calculated based on the average ICU patient census at the time of shift change. Figure 3.2 provides two examples of how the number of patient handoffs is affected by the number of patients in the ICU. In Fig. 3.2(a), there are two shift changes (provider transfers) at 7 am/pm. At the 7 am shift change there are two patients in ICU, so we incur two handoffs, while at 7 pm there is one patient in ICU and we incur one handoff. Fig. 3.2(b) illustrates the same scenario, but with only one shift change at noon, where 4 patients are handed off. Clearly, minimizing number of patient handoffs is not equivalent to minimizing number of shift changes. Furthermore, longer shifts do not guarantee fewer handoffs as seen in the example of Figure 3.2 (which also illustrates how the number of handoffs is calculated in our model). The challenge lies in designing a schedule that complies with all required constraints (and preferably most desired constraints) in a way that fewer patients have to be handed off. The objective function can be written as follows:

$$min \quad \sum_{j \in J} \sum_{k \in K} c_{jk} y_{jk}. \tag{3.23}$$

3.3 Case Study

This section presents a detailed discussion of how our model can be applied in a healthcare setting to help redesign trainees' on-call shifts to minimize the number of patient handoffs. We applied our model to the Medical Intensive Care Unit (MICU) at Saint Marys hospital in Rochester, Minnesota operated by Mayo Clinic. The MICU at Saint Marys hospital is a 24-bed unit. Our focus is on redesigning the fellows' shifts, as their service has the most impact on patient outcomes. Similar analysis can be applied to other provider levels (e.g., residents, attending consultants, etc.) and

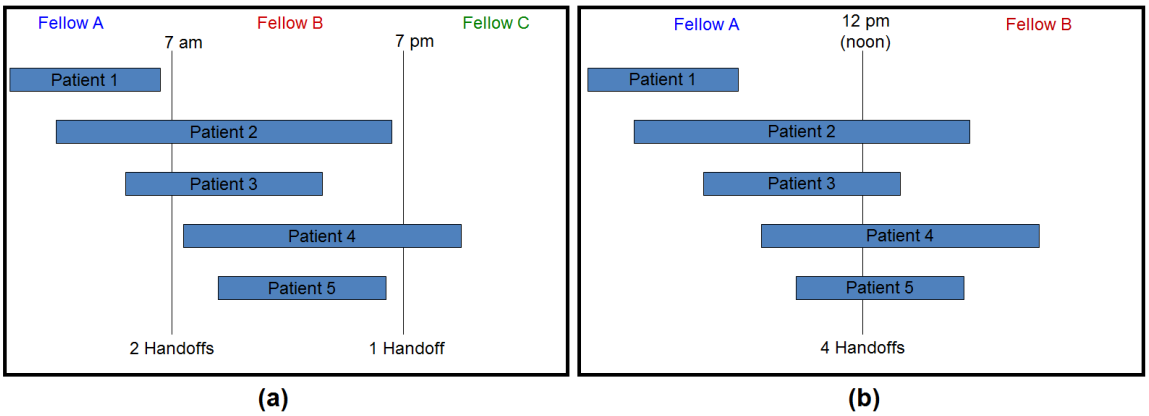

Figure 3.2: Examples of how the number of patient handoffs is calculated in the performance analysis model: (a) 2 shift changes and 3 patient handoffs, (b) 1 shift change and 4 patient handoffs.

other hospital units as well.

3.3.1 Assumptions

The detailed parameterization of the model for the case study was obtained through several meetings with residency and fellowship program directors, chief residents and fellows, as well as feedback from different medical providers at Mayo Clinic. For the case study, we consider a 4-week scheduling horizon which starts on a Saturday and ends on a Friday as trainees at Mayo Clinic rotate between different units every four weeks. Two years of MICU admission and discharge data were used to calculate the approximate MICU census for different days of week and times of day. Each day was divided into 12 time blocks. Hence, shift changes can happen at any of times 0, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20 and 22 in military time format. One nice property of the 2-hour time block is that, for this case study, the resulting schedule has a symmetric structure. The symmetric structure of our proposed schedule makes it easy to remember and much more appealing for implementation.

Based on ACGME rules, the maximum shift length is 16 hours for postgraduate year

1 (PGY-1) residents and 24 hours for senior trainees. Because the 24-hour shifts are believed to cause extreme tiredness and sleep deprivation contributing to more fatigue-related medical errors and poor patient outcomes, we limit the maximum shift length to 16 hours for fellows. Currently fellows work 12-hour shifts in the MICU. We start our analysis with a 16-hour limit on shift length, but will perform a sensitivity analysis on shorter and longer shifts later.

Some constraints deal with night shifts. In our study, we define night to be from 10:00 pm to 6:00 am. Hence, if a fellow is on call at any time in this interval, we assume he/she is on a night shift.

To provide 24/7 coverage, at least three fellows are required. This is because each fellow can work a maximum of 80 hours per week and we want to provide $24 * 7 = 168$ hours weekly coverage. Hence, we need at least $\lceil \frac{168}{80} \rceil = 3$ fellows.

For inequality 3.11, which ensures there is no shift change during the bedside multidisciplinary rounds, we need to determine the set of time blocks that end during this interval. Currently, the bedside rounds happen from 8:30 am to 11:00 am in the MICU.

Finally, a shift change is not allowed at inconvenient times (late night and early morning) through inequality 3.19. In this case study, we assume any time after 10:00 pm and before 4:00 am is inconvenient for a shift change.

3.3.2 Set and Parameter Values

Based on our previous discussion, model sets and parameters are assigned the following values.

Sets:

- $I = \{1, 2, 3\}$,
- $J = \{1, 2, ..., 28\}$,

- $T = \{1, 2, 3, 4\}$,
- $K = \{1, 2, ..., 12\}$,
- $Kn = \{1, 2, 3, 12\}$,
- $Kr = \{5\}$,
- $Kinc = \{1, 2, 12\}$.

Parameters:

- $NbF = 3$,
- $NbD = 28$,
- $NbW = 4$,
- $NbB = 12$,
- ShL is equal to 12 if DC5 is included in the scenario under consideration, and 16 otherwise,
- $d_{jk} = \begin{pmatrix} 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{pmatrix}_{28*12}$,
- $BL = \frac{24}{NbB} = 2$,
- $B^{ShL} = \lfloor \frac{ShL}{BL} \rfloor = 8$,
- $B_{10} = \lceil 10/BL \rceil = 5$,
- c_{jk} is equal to the average MICU patient census at the end of time block k in day j.

3.3.3 Data Collection

We used two years of MICU admission and discharge data to obtain patient census profiled by time of day and day of week. A computer program was developed to extract the required data from the dataset and to keep track of patient admissions and discharges for each time block of every day. Figure 3.3 shows the average MICU admission and discharge patterns during the day. As seen in this graph, there are almost no discharges at nights. Bedside rounds start at 8:30 am during which the discharge decisions are made by the team of residents, fellows and consultants. Patient discharges typically start around 9:00 am. The admission process is smoother with a higher average during the daytime. Figures 3.4 and 3.5 show the average MICU patient census versus different times of day and days of week. Mornings are more crowded than evenings since there is no discharge from the MICU at nights and before the rounds start in the morning. These results make sense intuitively and are in line with expert opinion which supports our data collection. Although the MICU census fluctuates from month to month, the pattern for different times of day and different days of week is similar. Since the census pattern is what matters for our shift design study (rather than the actual census numbers), we take the grand average census over months of year and use these numbers to approximate number of patient handoffs in our data-driven numerical analysis.

3.3.4 Experimental Scenarios

In this section, we solve the scheduling problem for different combinations of constraints to determine their effect on the objective function. The intent is to determine which desired constraints have the most impact on the number of patient handoffs. As discussed before, required constraints are those constraints that must be enforced in order to obtain valid or feasible schedules. Desired constraints are not required to be satisfied, but they make the resulting schedule more appealing. We perform

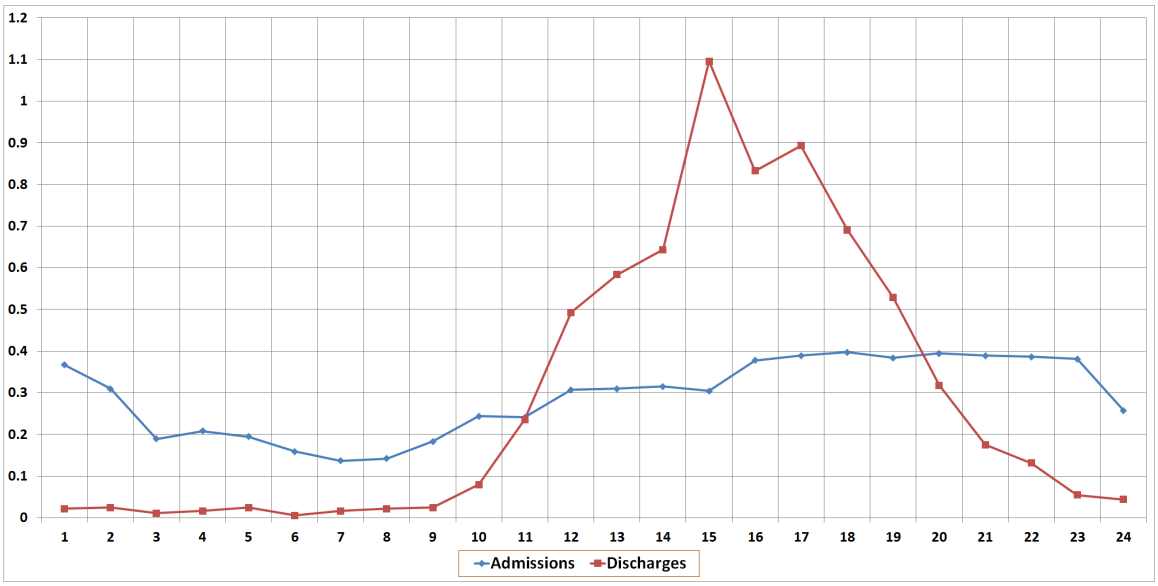

Figure 3.3: MICU admissions and discharges.

our analysis by adding one or a combination of desired constraints to the model and study their impact on the objective value (number of patient handoffs). If a desired constraint results in a great increase in the number of patient handoffs, loosening its bound or removing it will help avoid an increase in handoffs. This provides insight into the relative cost in terms of handoffs of a desired constraint.

There are 32 combinations of the five desired constraints. Those include having no desired constraints satisfied (1 case), having one desired constraint satisfied (5 cases), and so on. We will show the approximate number of patient handoffs for each case later in this section. First, we start with two extreme cases.

Scenario A - Only Required Constraints: The first scenario we investigate is the case in which only required constraints are satisfied. The resulting schedule will provide a lower bound on the minimum achievable number of handoffs. The number of patient handoffs from this case is used as the baseline for our comparison. The solution yields 635 patient handoffs over the 4-week scheduling horizon.

Scenario B - All Required Constraints and All Desired Constraints: The second sce-

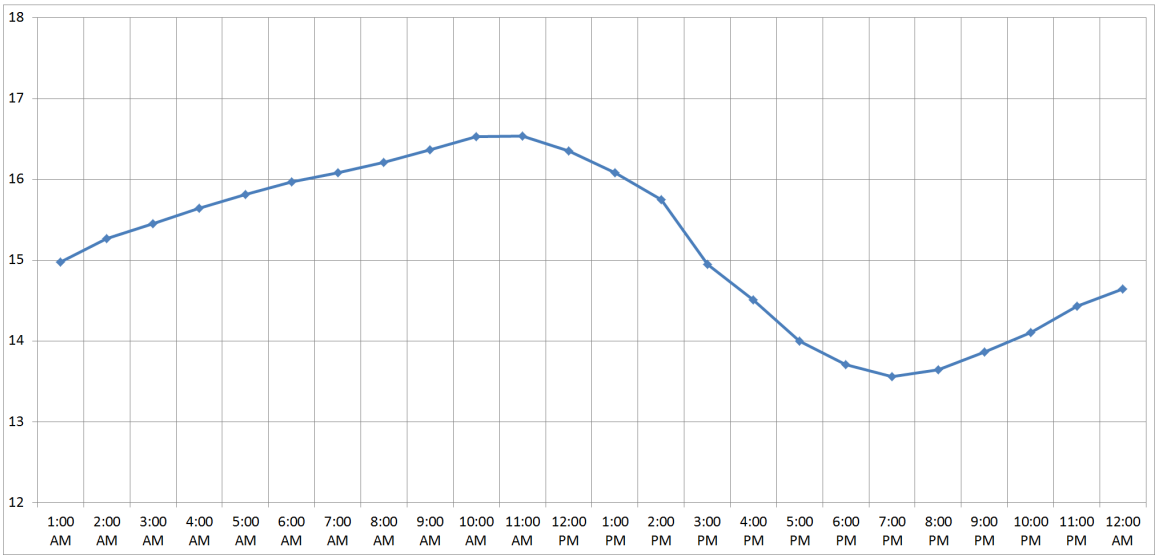

Figure 3.4: Average MICU patient census vs. time of day.

nario we study is the case in which all required and desired constraints are satisfied. This provides an upper bound on the number of patient handoffs. Interestingly, the resulting schedule was the same as current MICU schedule with 831 patients handed off during the 4-week horizon. The cost of having all desired constraints satisfied is a 31% increase in the number of patient handoffs.

Scenario C - All Required Constraints Together With One Desired Constraint: The previous scenarios provide a lower and an upper bound for the number of patient handoffs (635 and 831 respectively). In this scenario, we study the impact of each desired constraint on the number of patient handoffs by including them in the model individually. Scenarios C1, C2, C3, C4 and C5 are related to the cases in which DC1, DC2, DC3, DC4 and DC5 are added to the model respectively. The results show that adding DC2 or DC4 does not increase the number of patient handoffs, while adding DC1 or DC3 results in 641 patient handoffs (1% increase). On the other hand, adding DC5 (limiting shift length to 12 hours) results in 831 patient handoffs, the same result as the upper bound.

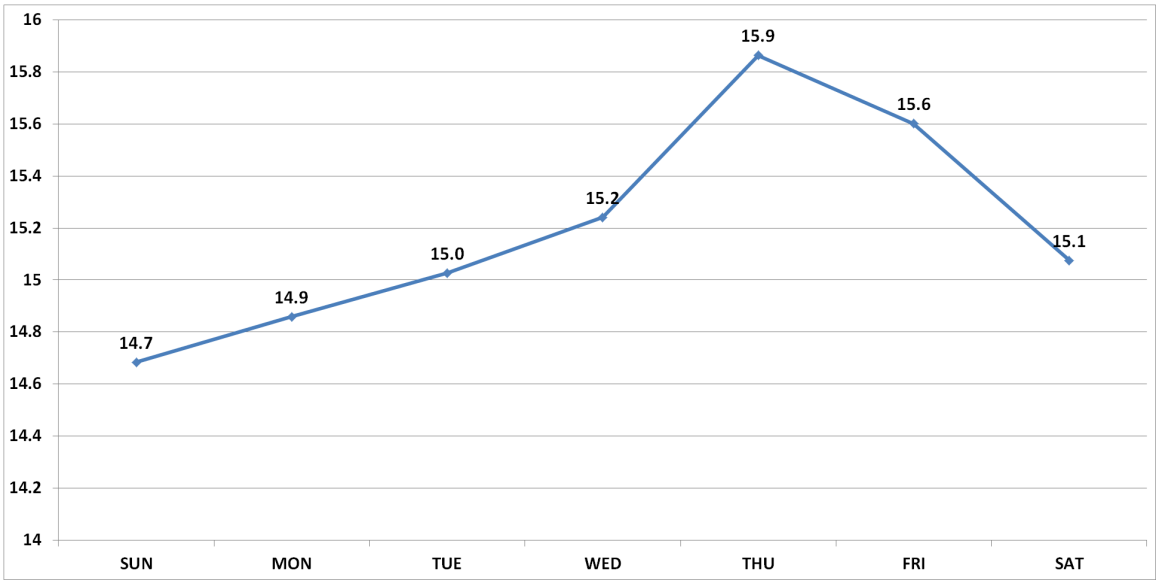

Figure 3.5: Average MICU patient census vs. day of week.

Scenario D - All Required Constraints Together With DC1-DC4: The results of previous scenarios revealed that each of these constraints individually does not significantly degrade the objective function. Including all of them simultaneously leads to a schedule with 641 patient handoffs on average over the 4-week horizon. This is only 1% greater than the lower bound (Scenario A), which includes none of the desired constraints.

Scenario E - All Required Constraints Together With DC2 and DC4: In scenario C, we saw that adding DC2 or DC4 to the model one at a time would not increase the number of patient handoffs. In this scenario, we investigate the effect of having both of them satisfied. The solution results in 641 handoffs, exactly the same as scenario D where DC1-DC4 are satisfied. Consequently, scenario E is dominated by scenario D.

Table 3.2 summarizes the results in this section. It shows all cases together with the scenarios we discussed in this section, the number of patient handoffs for each case, and whether each case is *efficient* or not. Those cases that are not dominated by any

other case are *efficient* schedules.

As seen in Table 3.2, scenarios B, C_2, C_4 and D are efficient schedules. However, from a practical standpoint, scenario D is preferred to scenarios C_2 and C_4, since DC1 - DC4 are satisfied with only 1% increase in the number of patient handoffs. Hence, from practical standpoint, only scenarios B and D are efficient schedules. The only difference between these two schedules is the limit on shift length. Figure 3.6 shows the resulting schedules corresponding to scenarios B and D (12 hours and 16 hours shift length limit, respectively).

Fellows at the Mayo MICU are currently working 12-hour shifts with shift changes happening at 6 am and 6 pm. Our analysis in this part showed that the current on-call schedule for fellows at Mayo's MICU is indeed optimal if we want to maintain 12-hour shifts; however, the 16-hour shift length of scenario D is very attractive due to the large number of handoffs saved (190 or 23% fewer).

Day\Time	0-2	2-4	4-6	6-8	8-10	10-12	12-14	14-16	16-18	18-20	20-22	22-24
1 (SAT)	1	1	1	2	2	2	2	2	2	1	1	1
2 (SUN)	1	1	1	2	2	2	2	2	2	3	3	3
3 (MON)	3	3	3	2	2	2	2	2	2	3	3	3
4 (TUE)	3	3	3	1	1	1	1	1	1	2	2	2
5 (WED)	2	2	2	1	1	1	1	1	1	2	2	2
6 (THU)	2	2	2	3	3	3	3	3	3	1	1	1
7 (FRI)	1	1	1	3	3	3	3	3	3	1	1	1
8 (SAT)	1	1	1	2	2	2	2	2	2	3	3	3
9 (SUN)	3	3	3	2	2	2	2	2	2	3	3	3
10 (MON)	3	3	3	1	1	1	1	1	1	2	2	2
11 (TUE)	2	2	2	1	1	1	1	1	1	2	2	2
12 (WED)	2	2	2	3	3	3	3	3	3	1	1	1
13 (THU)	1	1	1	3	3	3	3	3	3	1	1	1
14 (FRI)	1	1	1	2	2	2	2	2	2	3	3	3
15 (SAT)	3	3	3	2	2	2	2	2	2	3	3	3
16 (SUN)	3	3	3	1	1	1	1	1	1	2	2	2
17 (MON)	2	2	2	1	1	1	1	1	1	2	2	2
18 (TUE)	2	2	2	3	3	3	3	3	3	1	1	1
19 (WED)	1	1	1	3	3	3	3	3	3	1	1	1
20 (THU)	1	1	1	2	2	2	2	2	2	3	3	3
21 (FRI)	3	3	3	2	2	2	2	2	2	3	3	3
22 (SAT)	3	3	3	1	1	1	1	1	1	2	2	2
23 (SUN)	2	2	2	1	1	1	1	1	1	2	2	2
24 (MON)	2	2	2	3	3	3	3	3	3	1	1	1
25 (TUE)	1	1	1	3	3	3	3	3	3	1	1	1
26 (WED)	1	1	1	3	3	3	3	3	3	2	2	2
27 (THU)	2	2	2	3	3	3	3	3	3	2	2	2
28 (FRI)	2	2	2	3	3	3	3	3	3	1	1	1
Day\Time	0-2	2-4	4-6	6-8	8-10	10-12	12-14	14-16	16-18	18-20	20-22	22-24

a

Day\Time	0-2	2-4	4-6	6-8	8-10	10-12	12-14	14-16	16-18	18-20	20-22	22-24
1 (SAT)	1	1	1	2	2	2	2	2	2	2	2	3
2 (SUN)	3	3	3	3	3	3	3	2	2	2	2	2
3 (MON)	2	2	2	1	1	1	1	1	1	1	1	2
4 (TUE)	2	2	2	2	2	2	2	3	3	3	3	3
5 (WED)	3	3	3	1	1	1	1	1	1	1	1	3
6 (THU)	3	3	3	3	3	3	3	1	1	1	1	1
7 (FRI)	1	1	1	2	2	2	2	2	2	2	2	1
8 (SAT)	1	1	1	1	1	1	1	2	2	2	2	2
9 (SUN)	2	2	2	3	3	3	3	3	3	3	3	2
10 (MON)	2	2	2	2	2	2	2	3	3	3	3	3
11 (TUE)	3	3	3	1	1	1	1	1	1	1	1	3
12 (WED)	3	3	3	3	3	3	3	1	1	1	1	1
13 (THU)	1	1	1	2	2	2	2	2	2	2	2	1
14 (FRI)	1	1	1	1	1	1	1	3	3	3	3	3
15 (SAT)	3	3	3	2	2	2	2	2	2	2	2	3
16 (SUN)	3	3	3	3	3	3	3	1	1	1	1	1
17 (MON)	1	1	1	2	2	2	2	2	2	2	2	1
18 (TUE)	1	1	1	1	1	1	1	3	3	3	3	3
19 (WED)	3	3	3	2	2	2	2	2	2	2	2	3
20 (THU)	3	3	3	3	3	3	3	1	1	1	1	1
21 (FRI)	1	1	1	2	2	2	2	2	2	2	2	1
22 (SAT)	1	1	1	1	1	1	1	2	2	2	2	2
23 (SUN)	2	2	2	3	3	3	3	3	3	3	3	2
24 (MON)	2	2	2	2	2	2	2	1	1	1	1	1
25 (TUE)	1	1	1	3	3	3	3	3	3	3	3	1
26 (WED)	1	1	1	1	1	1	1	3	3	3	3	3
27 (THU)	3	3	3	2	2	2	2	2	2	2	2	3
28 (FRI)	3	3	3	3	3	3	3	1	1	1	1	1
Day\Time	0-2	2-4	4-6	6-8	8-10	10-12	12-14	14-16	16-18	18-20	20-22	22-24

b

Figure 3.6: Resulting schedules with (a) 12 hours per scenario B and (b) Scenario D with 16 hours shift length limit. Each number and color refers to one of the three fellows who is assigned to the corresponding time block.

Table 3.2: Summary of results of experimental scenarios.

Case #	Scenario	Desired Constraints	Number of Patient Handoffs	Dominated By Scenario
1	A	-	635 (baseline)	C_2 & C_4
2	C_1	1	641 (+1%)	D
3	C_2	2	635 (+0%)	
4	C_3	3	641 (+1%)	D
5	C_4	4	635 (+0%)	
6	C_5	5	831 (+31%)	B
7		1, 2	$\geq$ 641	D
8		1, 3	$\geq$ 641	D
9		1, 4	$\geq$ 641	D
10		1, 5	$\geq$ 831	B
11		2, 3	$\geq$ 641	D
12	E	2, 4	641 (+1%)	D
13		2, 5	$\geq$ 831	B
14		3, 4	$\geq$ 641	D
15		3, 5	$\geq$ 831	B
16		4, 5	$\geq$ 831	B
17		1, 2, 3	$\geq$ 641	D
18		1, 2, 4	$\geq$ 641	D
19		1, 2, 5	$\geq$ 831	B
20		1, 3, 4	$\geq$ 641	D
21		1, 3, 5	$\geq$ 831	B
22		1, 4, 5	$\geq$ 831	B
23		2, 3, 4	$\geq$ 641	D
24		2, 3, 5	$\geq$ 831	B
25		2, 4, 5	$\geq$ 831	B
26		3, 4, 5	$\geq$ 831	B
27	D	1, 2, 3, 4	641 (+1%)	
28		1, 2, 3, 5	$\geq$ 831	B
29		1, 2, 4, 5	$\geq$ 831	B
30		1, 3, 4, 5	$\geq$ 831	B
31		2, 3, 4, 5	$\geq$ 831	B
32	B	1, 2, 3, 4, 5	831 (+31%)	

3.3.5 Sensitivity Analysis and Discussion

In this section, we briefly review the main results from the previous section and then provide further analysis of the shift length constraint.

The previous section showed that most of the desired constraints can be accommodated without significantly increasing the number of patient handoffs. Those constraints include: no shift change at late night or early morning, at least one day off every seven days, no more than four night shifts in a row and a minimum one day off after a run of night shifts. The desired constraint that restricts shift length to 12 hours, however, increases the number of patient handoffs by more than 30%.

Based on ACGME duty-hour regulations, on-call shifts of residents in PGY-1 must not exceed 16 hours while senior trainees (PGY-2 and above residents and all fellows) may be scheduled for a maximum of 24 hours of continuous duty. To explore the effect of shift length, we run the model for different shift lengths from 12 hours to 24 hours (in 2-hour increments). We keep all other required and desired constraints active and only change the shift length bound (ShL parameter). Figure 3.7 shows the resulting change in number of patient handoffs.

As seen in Figure 3.7, increasing shift length limit results in fewer patient handoffs. The current 12-hour shifts in Mayo MICU cause an average of 831 patients to be handed off per month. Should the shift length be extended to 16 hours, this will result in nearly a 23% reduction in number of patient handoffs per month. Increasing the shift length to its maximum allowed limit, i.e. 24 hours, results in almost a 48% reduction in number of patient handoffs per month. On one hand, shorter shifts correlate with more frequent patient handoffs, which potentially results in more medical errors due to communication breakdown and loss of information during the handoff process. On the other hand, longer shifts are associated with extreme tiredness and sleep deprivation which can also contribute to fatigue-related medical errors.

A reasonable tradeoff between fatigue and handoffs should be established to mini-

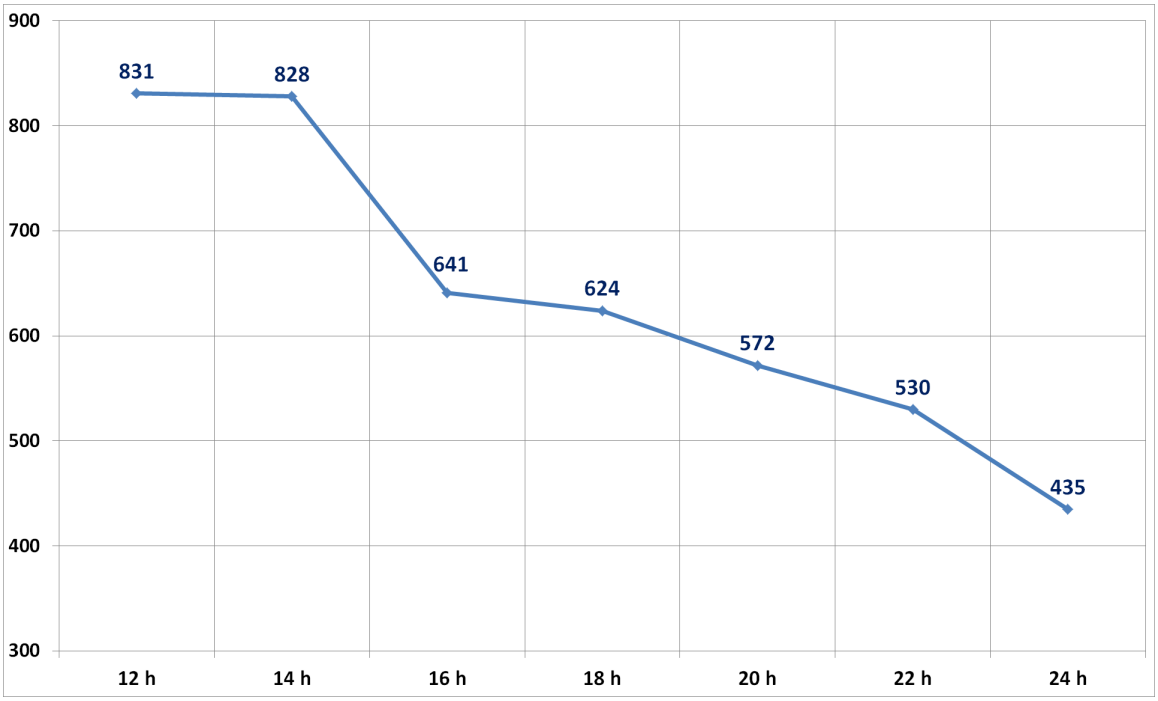

Figure 3.7: Number of patient handoffs for different shift length limits.

mize medical errors and achieve the best patient outcomes. However, to date there is no solid methodology to quantify physicians' fatigue and the effect of fatigue on quality of care and patient outcomes. The best we could do was to collect expert opinions. Several program directors and physicians that we have interviewed believe that 24-hour shifts are acceptable and worth the benefit of fewer patient handoffs. A majority of program directors, chief residents and fellows believe 16-hour shifts are very reasonable and worth the benefit of the 20%-25% reduction in patient handoffs (compared to 12-hour shifts). The 16-hour shift length limit is permitted by ACGME and could be applied to different trainee levels (i.e. PGY-1 residents, senior (PGY-2 and above) residents, and fellows). This appears to provide a good tradeoff between the adverse effects of physicians' fatigue and the adverse effects of more frequent patient handoffs.

One final point is the importance of maintaining fairness and balance among the

trainees' schedules. While this is not a mandated requirement, it is clearly important for implementation and trainee morale. Therefore, we added measures and associated constraints to ensure balance among the schedules for average duty hours, number of night shifts, and the number of days off. Figure 3.8 shows the resulting equitable schedule and associated *fairness* values that yields the same 641 patient handoffs (with the shift length limit set at 16 hours).

Average Duty-hours	1.Geen Fellow: 56 hrs/wk 2.Red Fellow: 56 hrs/wk 3.Yellow Fellow: 56 hrs/wk
# Night Shifts	1.Geen Fellow: 9 2.Red Fellow: 10 3.Yellow Fellow: 9
# Days Off	1.Geen Fellow: 6 2.Red Fellow: 7 3.Yellow Fellow: 7

Day\Time	0-2	2-4	4-6	6-8	8-10	10-12	12-14	14-16	16-18	18-20	20-22	22-24
1 (SAT)	1	1	1	1	1	1	1	3	3	3	3	3
2 (SUN)	3	3	3	2	2	2	2	2	2	2	2	1
3 (MON)	1	1	1	1	1	1	1	2	2	2	2	2
4 (TUE)	2	2	2	3	3	3	3	3	3	3	3	2
5 (WED)	2	2	2	2	2	2	2	3	3	3	3	3
6 (THU)	3	3	3	1	1	1	1	1	1	1	1	3
7 (FRI)	3	3	3	3	3	3	3	1	1	1	1	1
8 (SAT)	1	1	1	2	2	2	2	2	2	2	2	1
9 (SUN)	1	1	1	1	1	1	1	2	2	2	2	2
10 (MON)	2	2	2	3	3	3	3	3	3	3	3	2
11 (TUE)	2	2	2	2	2	2	2	3	3	3	3	3
12 (WED)	3	3	3	1	1	1	1	1	1	1	1	3
13 (THU)	3	3	3	3	3	3	3	2	2	2	2	2
14 (FRI)	2	2	2	1	1	1	1	1	1	1	1	2
15 (SAT)	2	2	2	2	2	2	2	1	1	1	1	1
16 (SUN)	1	1	1	3	3	3	3	3	3	3	3	1
17 (MON)	1	1	1	1	1	1	1	2	2	2	2	2
18 (TUE)	2	2	2	3	3	3	3	3	3	3	3	2
19 (WED)	2	2	2	2	2	2	2	3	3	3	3	3
20 (THU)	3	3	3	1	1	1	1	1	1	1	1	3
21 (FRI)	3	3	3	3	3	3	3	2	2	2	2	2
22 (SAT)	2	2	2	1	1	1	1	1	1	1	1	2
23 (SUN)	2	2	2	2	2	2	2	1	1	1	1	1
24 (MON)	1	1	1	3	3	3	3	3	3	3	3	1
25 (TUE)	1	1	1	1	1	1	1	3	3	3	3	3
26 (WED)	3	3	3	2	2	2	2	2	2	2	2	3
27 (THU)	3	3	3	3	3	3	3	1	1	1	1	1
28 (FRI)	1	1	1	2	2	2	2	2	2	2	2	1
Day\Time	0-2	2-4	4-6	6-8	8-10	10-12	12-14	14-16	16-18	18-20	20-22	22-24

Figure 3.8: Equitable schedule with all required and desired constraints and 16-hour shift length limits.

3.4 Conclusions

In this chapter, we developed a new patient-centered model for scheduling residents and fellows (trainees) to minimize number of patient handoffs, which have been linked with medical errors caused by communication breakdowns and adverse events. While previous literature focuses on the logistics of the handoff, we bring a new sys-

tems perspective to this problem by designing schedules that minimize the number of patients that are handed off, thereby reducing the opportunity for serious error. Our integer programming model designs on-call shifts such that all ACGME duty-hour regulations are satisfied, required coverage is achieved, livability rules are met, and patient handoffs are minimized. The general form of our model can be used by any healthcare operation that wants to reduce patient handoffs and that has duty-hour restrictions and similar livability constraints. Should the size of the model render the problem intractable for other healthcare units, heuristics approaches such as the Tabu Search and Ant Colony Optimization can be employed to solve the integer program (see, for example, *Balas and Martin* 1980; *Azadeh et al.* 2013; *Lokketangen and Glover* 1998).

In a case study of an ICU at an academic medical center (the Mayo Clinic in Rochester, Minnesota) we demonstrated how our model could be applied to reduce the number of patient handoffs. We found that most desired constraints (livability rules) can be satisfied with a negligible increase in number of patient handoffs. The desired constraint that had the largest impact on handoffs was the shift length. By increasing the shift length from 12 to 16 hours it was possible to reduce handoffs by 23% relative to the current MICU schedule. 24-hour shifts (the maximum allowable shift length) resulted in a 48% reduction in the number of patient handoffs. It is worth noting that in the proposed schedule no new trainees (fellow for the case study of Mayo MICU) need to be added beyond the minimum number needed to provide the required coverage specified by the 10th required constraint (require coverage is 24/7 for the case study). Therefore, in terms of financial costs, the as-is schedule and the proposed schedules are exactly the same.

Based on discussions with staff at Mayo Clinic, we found that 16-hour shifts provided a reasonable tradeoff between medical errors due to trainee fatigue and medical errors due to communication breakdowns as a result of more frequent patient handoffs. The

new shift design approach discussed in this chapter is under consideration for implementation at Mayo Clinic. A shift assignment approach based on the work discussed in this chapter was accepted by the practice and additional services are considering the use of the general methodology to assist in staff scheduling at Mayo Clinic.

CHAPTER IV

Coordinating Clinic and Surgery Appointments to Meet Access Delay Service Level for Elective Surgery

4.1 Introduction

The United States healthcare is facing a significant challenge in providing timely access to care (*Davis et al.* 2014), which not only affects patient satisfaction but also directly influences patient safety and health outcomes (*Murray and Berwick* 2003; *Koopmanschap et al.* 2005). Long wait times may also result in elevated healthcare costs because of additional treatments (*Somasekar et al.* 2002; *Hilkhuysen et al.* 2005). To ensure patients can get medical care when they need it, it is critical to make efficient use of the current healthcare resources to minimize access delay, which can lead to a healthier society.

In this paper, we leverage systems engineering and operations research to improve access to elective surgery in a highly specialized surgical unit. Our research is motivated by the issue of long waiting times for patients who need a surgery in the colorectal surgery (CRS) department at the Mayo Clinic in Rochester, MN. In CRS, like many surgical units across the country, the increase of patient demand is outpacing the growth of surgical capacity, leading to long wait times. We define wait time

as the number of business days between the day a patient is referred to CRS and the day he or she is operated in the operating room (OR). Previous research indicates that ineffective scheduling regimes are among the main factors contributing to underutilization of ORs (see *Weinbroum et al.* 2003; *Jonnalagadda et al.* 2005). In this paper, we propose and evaluate 6 carefully chosen scheduling policies that work better than the current scheduling protocol. The underlying idea behind these scheduling protocols is to efficiently use patient information to integrate and coordinate clinic and surgery appointment scheduling such that patients can get a clinic visit and a surgery appointment within a time period that is clinically safe for them while the OR overtime is minimized.

4.2 Problem Description

CRS at the Mayo Clinic includes 8 surgeons. They form 2 teams of orange and blue each consists of 4 surgeons. On any given day, one team sees patients in the clinic and the other team is operating in the OR. The teams switch places on the following business day. Therefore, each surgeon has a clinical and a surgical calendar. CRS receives patient appointment requests from two separate channels: (1) the surgeon's desk (internal referrals), and (2) the direct clinic (external referrals). About 75% of the requests come through the surgeon's desk, which are usually referred by the gastrointestinal (GI) department at the Mayo Clinic. The remaining 25% are referred from other hospitals. Each new request needs a clinic appointment with a CRS surgeon. The clinic visit may need to be followed by a surgery. Recently, CRS has experienced an increase in the earliest surgery time they can offer to their patients. The chief surgeon is worried that the long wait times can lead to adverse events and poor patient outcomes especially for the urgent patients (e.g. colon cancer patients). Figure 4.1 shows the two appointment request channels.

Current scheduling policy: Presently, the scheduling protocol/policy at CRS is

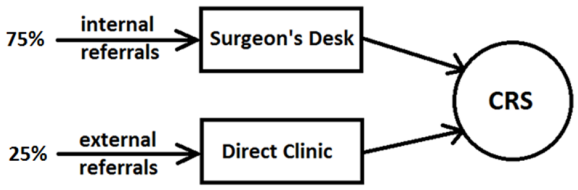

Figure 4.1: CRS appointment request diagram.

focused strictly on finding the surgeon with the first available clinic slot and scheduling a clinic visit for the patient with that surgeon. During the clinic visit, it is determined whether or not surgery is required and the details concerning surgery. If it is determined that the patient needs a surgery, the surgeon then looks into his/her surgical calendar and offers the earliest date he/she can perform the surgery.

The above protocol now used in practice often results in a long wait time for urgent patients to obtain surgery, because it waits until the day of the clinic visit to reserve a date for surgery. A common scenario that motivated this research is as follows. A cancer patient is assigned to surgeon X based on the clinical availability of surgeons. When surgeon X sees the patient in the clinic, the surgeon and the patient jointly determine to proceed with surgery. The surgery should be performed as soon as possible to prevent the tumor from further growing. Surgeon X looks into his surgical calendar and finds out that his next surgical availability is 3 weeks into the future; but, one of his colleagues, say surgeon Y, has surgical availability in 4 days. Surgeon X asks the patient if he/she prefers to wait 3 weeks or be operated by surgeon Y. The patient almost always prefers to wait and get the surgery with the same surgeon that they got to know during the clinic visit rather than a colleague they have never met. This is because the patient has already established some trust and a relationship with surgeon X. Therefore, in this paper we assume that the surgery must be performed by the same surgeon who has seen the patient in clinic.

4.3 Literature Review

Appointment scheduling has been a topic of interest for many researchers over the past few decades. *Gupta and Denton* (2008) provide a literature survey on appointment scheduling systems in 3 different healthcare settings: (1) primary care, (2) specialty clinic, and (3) elective surgery. They describe the goal of a well-designed appointment system as delivering timely access to health services for all patients. This is in line with our main objective in this paper. *Gupta and Denton* (2008) also indicate that most of the previous research has focused on minimizing the direct wait time, which is the time from the moment patient enters the clinic to the time he/she sees the doctor. Our focus in this research is to reduce the indirect wait time, which corresponds to the number of days from the day of appointment request until the day of appointment itself.

Two closely related papers to our work are *Patrick et al.* (2008) and *Saure et al.* (2012). *Patrick et al.* (2008) present a Markov decision process (MDP) model with rolling horizon to dynamically allocate the available capacity of a diagnostic resource (CT scanner) between different patient priority classes. Each patient requires one appointment slot. Surge capacity is diverted or delayed. Their objective is to minimize the number of patients who does not get an appointment by their maximum wait time target. They solve the equivalent linear program through approximate dynamic programming (ADP) to mitigate the curse of dimensionality of the MDP model. *Saure et al.* (2012) develop an MDP-based modeling framework to allocate available treatment capacity to incoming demand for scheduling cancer treatments in radiation therapy units while reducing wait times. They consider multiple identical radiation therapy machines and determine the total capacity by aggregating the individual capacities of machines. Each patient visit triggers a sequence of visits with different number of appointment slots for each visit. Surge capacity is delayed or accommodated using overtime. They use an ADP approach similar to *Patrick et al.*

(2008) in order to solve the model. There are a number of differences between these papers and our work. First, we consider multiple non-identical surgeons as scarce resources and directly book appointments with them considering the individual capacities as opposed to considering one single resource or multiple identical resources and aggregating the individual capacities. Second, they consider off-line scheduling in which the scheduling decisions are made once a day; whereas, we consider online scheduling in which the orders are scheduled one by one as they arrive to CRS. Third, they consider that each new request requires either one or multiple visits and each visit requires either one or multiple appointment slots; but, these are all assumed to be known at the time the request is received. However, in our problem, each request requires a clinic appointment and may or may not be followed by a surgery. The uncertainty about needing a surgery is realized during the patient's clinic visit. If it is deemed that the patient does need a surgery, the actual surgery duration is not realized until after the surgery is done, but can be estimated beforehand using historical data.

Gupta and Wang (2008) develop a model to decide whether to accept or reject an appointment request to maximize revenue while incorporating the patients' preferences in terms of which doctor to see and the time of appointment into their decision making. *Wang and Gupta* (2011) present a novel decision support tool that dynamically learns and updates patients' preferences and use this information to improve booking decisions. *Liu et al.* (2010) propose heuristic dynamic policies for scheduling patient appointments considering no-shows and cancelations and evaluate them using computer simulation. In this paper, we do not consider patient no-shows and cancelations as they rarely happen in a highly-specialized care unit like colorectal surgery. We also do not honor patients' surgeon preference per recommendation of the chief CRS surgeon. Clearly, in many healthcare environments, the patient can prioritize the selection of the provider with whom they feel most comfortable. Our scope is,

however, limited to the important class of environments in which the patients typically accept the surgeon offering the earliest access. It should be noted that it is not too difficult to manually intervene in cases in which the algorithmic solution needs to be tweaked to account for the occasional patient with strong provider preferences, if they are willing to accept the potential delay.

Many researchers have used computer simulation to study appointment and/or surgery scheduling and to test the performance of heuristic scheduling rules (*Erdogan and Denton* 2011). Simulation is a powerful tool for assessing the potential response of a health system to policy changes (*Fone et al.* 2003). For instance, *Ho and Lau* (1992) study the performance of over 50 scheduling rules using simulation and conclude that their performances are affected by the environmental conditions of the operating environment. *Vasilakis et al.* (2007) develop a discrete event simulation model to compare two methods of scheduling outpatient clinic appointments (individual surgeon vs. pooled lists) in terms of number of patients waiting for appointments and wait times to appointment and surgery. *LaGanga and Lawrence* (2007) leverage discrete event simulation to investigate appointment over-booking as one means of reducing the negative impact of no-shows and conclude that over-booking can improve patient access, provider productivity, and overall clinic performance. *Everett* (2002), *Gul et al.* (2011), *Berg et al.* (2014), *Liang et al.* (2015), and *Zhang and Xie* (2015) have also used computer simulation to study appointment/surgery scheduling. The most important difference between our work and the previous papers mentioned is that we expand our view to proactively coordinate the scheduling of clinic and surgery appointments such that patients with different levels of urgency can be served within their maximum wait time target windows with minimum OR overtime. The concept of coordinating clinic and surgery appointments is unique and novel in our work and has never been studied before in the context of surgery scheduling and planning. Unlike some surgical scheduling papers, our paper does not strive to determine how

many ORs to open on each day and how to allocate the available OR time to the surgeons. In other words, we assume each surgeon has rigidly allocated block time. We also do not specify the start time and the sequence of surgeries listed to be performed on each day. Our model is concerned with determining (1) a clinic appointment day, (2) a tentative surgery day, and (3) a surgeon to assign to both the clinic and the surgery appointments for each new patient referred to CRS.

4.4 Solution Approach

In this section, we describe our solution approach to the problem of long wait time to surgery in Mayo CRS that was described in Section 4.2. We tentatively book/reserve a surgery day for all patients at the same time we are scheduling the patient's clinic appointment. In other words, instead of waiting until the patient's clinic appointment to figure out if they need a surgery and then try to book a surgery day if they need one, we proactively schedule a tentative surgery day together with a clinic appointment when the patient referral is received by CRS. This guarantees timely access to surgery for all patients.

4.4.1 Data

We use 5-year historical data (2011 to 2015) on the CRS clinic and surgery appointments to perform our analysis. This includes data on about 12,000 patients. When CRS receives a new appointment request (either through the surgeon's desk or the direct clinic), the following patient information are among the available data:

1. indication (i.e., colon cancer, anal cancer, rectal cancer, Crohn's disease, chronic ulcerative colitis, diverticulitis, neoplasi/polyp, rectal prolapse, slow transit constipation, hernia, ostomy, or CRS other),

2. geographical location (i.e., local, regional, or national/international),

3. age

4. referral type (i.e., internal or external),

5. estimated surgery duration (predicted from historical surgery duration data).

In addition to the above data, we have the actual surgery durations for cases that have resulted in a surgery.

As mentioned in Section 4.2, on any given day, there are 4 surgeons in the OR and 4 in the clinic. On average, 2 of the surgeons get two ORs and 2 of them get one OR. Those surgeons who get two ORs can schedule parallel-staggered surgeries and let their team do the preparation, cutting, and closing of the patient. In this paper, we assume surgeons with one OR and those with two or more ORs can schedule 8 and 12 hours of surgery per day, respectively. Per the recommendation of our clinical collaborators, we also assume no limit for the clinical capacity of the surgeons. This is because the clinic visits are relatively short (each clinic visit takes only 15 to 30 minutes); so, they are never the bottleneck and can be accommodated easily. Further, the surgeons are willing to use clinic overtime to see a patient in clinic if it benefits OR utilization and patient outcomes.

4.4.2 Patient Priority Levels/Types

Mullen (2003) and *MacCormick et al.* (2003) provide a comprehensive survey on prioritizing patients in a waiting list. One common way to do so is based on the patients' acuity of disease. In this research, we use expert opinion to label each patient as either priority type 1, 2, 3, 4, or 5 based on a combination of (1) the indication of the patient, (2) the geographical zone, and (3) the referral type. Type 1 is for the most urgent cases, while on the other extreme, type type 5 is assigned to patients who only need a clinic appointment for follow-up/consult or those who don't need a surgery in near future. Each priority type corresponds to a maximum wait

time target (MWTT) to surgery, except that type 5 that does not require a surgery. MWTT is the maximum time that the patient's surgery can be safely delayed. We determined this based on the input from our clinical collaborators. This is a common approach in the literature for patient prioritization (see, for example, *Naylor* 1991). Our goal is to guarantee that all patients are offered at least one surgery day within their MWTT in the planning solution. It is worth noting that if a patient doesn't seize the offered surgery date (for any reason), we can offer them the next available surgery date but we cannot guarantee that the alternative date is within their MWTT. Table 4.1 summarizes the patient priority type for each combination of indication, referral type, and geographical zone. Table 4.2 provides the MWTT for each priority type in business days. For example, a colon cancer patient who has traveled from overseas and is referred to CRS internally is considered as priority 1. Consequently, we would like to offer a clinic and a surgery appointment to him/her within 3 business days. Note that rectal cancer patients are considered as type 5 because of the long required duration between clinic appointment and surgery due to chemotherapy. Also, throughout this paper, we only focus on business days since CRS is off during the weekends.

4.4.3 Logistic Regression

In this section, we develop a logistic regression model (also known as the logit model) in order to predict the probability of a new clinic appointment order will result in a surgery. We refer to this as the "probability of surgery" from now on. In the logit model the log odds of the outcome (response variable) is modeled as a linear combination of the independent variables (covariates). The patient's indication, age, and referral type are the covariates of the logit model. Eq. 4.1 presents the logit model in which y is the logit transform of the probability of surgery and $I(x)$ is an indicator function. For example $I(Crohn's)$ is equal to 1 if the patient's indication is

Table 4.1: Patient priority type for different combinations of patient indication, referral type, and zone. Smaller priority level numbers are associated with more urgent patients. Priorities 1, 2, 3, and 4 are assigned to patients who might need a surgery following their clinic visit to CRS. Priority 5 is assigned to patients who only need a clinic visit for follow-up/consult.

Indication	Internal Referral			External Referral		
	Local	Regional	National / International	Local	Regional	National / International
Colon cancer	3	2	1	3	2	2
Anal cancer	3	2	1	3	2	2
Rectal cancer	5	5	5	5	5	5
Crohn's disease	3	2	1	3	2	2
Chronic ulcerative colitis	3	2	1	3	2	2
Diverticulitis	3	2	1	3	2	2
Neoplasi/polyp	3	2	1	3	2	2
Rectal prolapse	4	3	2	5	4	3
Slow transit constipation	4	3	2	5	4	3
Hernia	4	3	2	5	4	3
Ostomy	4	3	2	5	4	3
CRS other	3	3	3	3	3	3
Follow-up/consult	5	5	5	5	5	5

Table 4.2: Maximum wait time target to surgery for different patient priority types.

Priority Type	Maximum Wait Time Target (MWTT) to Surgery [business days]
1	3
2	10
3	20
4	40
5	NA (only need a clinic visit)

Crohn's disease and 0 otherwise.

$$y = 0.72279 * I(\text{crohn's}) - 0.45483 * I(\text{neoplasi}) - 0.69236 * I(\text{STC}) - 0.65543 * I(\text{other}) - 0.00937 * age + 0.53546 * I(\text{internal referral}) + 0.39429 * I(\text{external referral}). \tag{4.1}$$

Therefore, the probability of surgery can be obtained via

$$p = \frac{e^y}{1 + e^y}. \tag{4.2}$$

4.4.4 Scheduling Policies/Protocols

In this subsection, we introduce 6 scheduling policies considered in our analysis. We evaluate the performance of these policies in terms of average OR overtime per day in the Section 4.5 using a 2-stage stochastic and dynamic simulation model.
In order to define the policies, the following terminology is needed:

- *Earliest Feasible Clinic (EFC)* is the earliest day we can schedule a clinic visit appointment for the patient. For a clinic appointment request that is received on day t, if it is an internally referred patient (meaning that the patient is physically at the Mayo Clinic), we assume EFC is the same day (i.e., day t). For external referrals, we assume EFC is 5 business days from the day the appointment request is received (i.e., $t + 5$) to give the patients at least one week to make travel arrangements to Mayo. In other words,

$$EFC = t + 0 \times I(\text{internal referral}) + 5 \times I(\text{external referral}). \tag{4.3}$$

- *Clinic to Surgery Gap (CSG)* is an arbitrary number that corresponds to the required number of days between the patient's clinic and surgery appointments. This number depends on the patient's priority type and the probability of

surgery obtained from the logistic regression model. Recall that (1) we tentatively reserve a surgery date for all patients before we know if the patient will require a surgery, and (2) the decision of whether or not to perform a surgery is made during the patient's clinic visit. To avoid the waste of OR time in cases where the patient doesn't end up pursuing a surgery, we space out the clinic and surgery appointments for patients with low probability of surgery. In general, the lower the probability of surgery is, the larger the CSG should be. This gives us some time to assign the OR time of a canceled tentatively booked surgery to a new patient and prevent the OR time from getting wasted.

- *Earliest Feasible Surgery (EFS)* is the earliest day we can schedule the surgery for the patient. Per definition, $EFS = EFC + CSG$ for all patients.

The above definitions and approach represent a new paradigm for setting appointments in advance while being sensitive to waiting times for service. The concept is intuitive and relatively simple to implement. There are, however, details that must be resolved to render the approach effective. Next, we are going to define the scheduling policies. These policies first find a tentative surgery day and a surgeon for the patient; then, they schedule a clinic appointment with the same surgeon prior to the tentative surgery day. Both are done at the time the order for clinic appointment is received by CRS. We first describe how each policy finds a tentative surgery day. Then, we elaborate on how the clinic appointment day is determined. Recall that the current scheduling policy of CRS was defined in Section 4.2.

4.4.4.1 Finding a tentative surgery day

Policy A: Assign the patient to the surgeon with the earliest OR availability in the interval $[EFS, MWTT]$ starting from EFS and moving forward in time. If there is no availability in this interval, assign her to the surgeon-day with the most remaining surgical availability in this interval and use OR overtime.

Policy B: If the patient is of priority 1, follow the rules of Policy A. If the patient is of priority 2, 3, or 4, assign her to a surgeon with sufficient OR availability in the interval $[EFS, MWTT]$ starting with day MWTT and moving backward in time. If there is no availability in this interval, assign her to the surgeon-day with the most remaining surgical availability in this interval and use OR overtime. This is similar to the policy derived in *Patrick et al.* (2008).

Policy C: If the patient is of priority 1, follow the rules of Policy A. If the patient is of priority 2, 3, or 4, sort the days in the interval $[EFS, MWTT]$ in descending order of aggregated surgical availability. Assign the patient to a surgeon with sufficient OR availability starting with the first day on this list (i.e., the least busy day) and moving down the list. If there is no availability in this interval, assign her to the surgeon-day with the most remaining surgical availability in this interval and use OR overtime.

Policy D: If the patient is of priority 1, follow the rules of Policy A. If the patient is of priority 2, 3, or 4, follow the procedure of Policy C in order to find a tentative surgery date for the patient using regular OR time. If there is no availability in the interval $[EFS, MWTT]$, calculate the "expected" surgical workload of each surgeon-day on this interval. Then, assign the patient to the surgeon-day with the least expected surgical workload (i.e., the most expected surgical availability) and use OR overtime. The expected surgical workload of surgeon i on day t is calculated via

$$E[W(i,t)] = \sum_{s \in S(i,t)} p_s E[D_s], \tag{4.4}$$

where $E[W(i,t)]$ is the expected workload of surgeon i on day t, $S(i,t)$ is the set of surgeries assigned to surgeon i on day t, p_s is the probability of surgery s (obtained via the logit model), and $E[D_s]$ is the expected duration of surgery s. The idea behind policy D is to take the uncertainty of the tentatively booked surgeries into consideration when we select a surgeon-day for using OR overtime. For example,

consider days t_1 and t_2 as the candidate surgery days for a new order with expected surgery duration of 3 hours. Further, assume that we have only one surgeon with surgical capacity of 8 hours per day. Day t_1 includes two tentatively booked surgeries of length 3 and 4 hours with probabilities of surgery of 0.3 and 0.2, respectively. Day t_2 includes two other tentatively booked surgeries of length 2 and 4 hours with 0.8 probability of surgery for both. Under policy C, day t_2 is selected for the new order because $2+4 < 3+4$. However, under Policy D, day t_1 is selected because the effective surgical workload of the surgeon on day t_1 $(0.3 * 3 + 0.2 * 4 = 1.7\ hrs)$ is less than that of day t_2 $(0.8 * 2 + 0.8 * 4 = 4.8\ hrs)$.

Policy E: If the patient is of priority 1, follow the rules of Policy A. If the patient is of priority 2, 3, or 4 (i.e., $p \in \{2, 3, 4\}$), follow the same procedure as Policy C but limit the search for a surgery day to the interval $[max(EFS, MWTT_{p-1}), MWTT_p]$. In here, $MWTT_p$ is the MWTT of the patient under consideration whose priority is p and $MWTT_{p-1}$ is the MWTT of priority $p-1$ patients.

Policy F: If the patient is of priority 1, follow the rules of Policy A. If the patient is of priority 2, 3, or 4 (i.e., $p \in \{2, 3, 4\}$), follow the same procedure as Policy D but limit the search for a surgery day to the interval $[max(EFS, MWTT_{p-1}), MWTT_p]$.

4.4.4.2 Finding a clinic appointment day

While the 6 scheduling policies differ in how they find a tentative surgery day for the patient, they all follow the same procedure in order to determine the clinic appointment day. If the patient is internally referred to CRS (i.e., he/she is physically located at the Mayo Clinic) on day t, the policies book a clinic appointment with the same surgeon with whom the tentative surgery is booked either on the same day or the following day (i.e., either on day t or $t+1$) depending on which day is a clinical day for the corresponding surgeon. However, if the order is an external referral (i.e., the patient is not physically located at the Mayo Clinic), the clinic appointment is

scheduled CSG days prior to the tentative surgery day. For example, suppose the tentative surgery day is set to day $t+10$ and $CSG = 3$ days. If the patient is already at Mayo (i.e., internal referral) we can offer a clinic appointment as soon as possible to remove the uncertainty regarding the surgery decision; hence, the clinic appointment is scheduled with the same surgeon on day $t+1$. If the patient is not physically at Mayo (i.e., external referral) the clinic appointment is scheduled on $t+7$, which gives the patient some time to make travel plans. Mathematically, we can obtain the clinic appointment day via Eq. 4.5,

$$\begin{aligned}\text{clinic app. day} = (EFC + \text{mod}(\text{tent. surgery day} - EFC + 1, 2)) \times I(\text{internal referral}) \\ + (\text{tent. surgery day} - CSG) \times I(\text{external referral}),\end{aligned} \tag{4.5}$$

in which $mod(x, 2)$ returns the remainder after division of x by 2.

4.4.5 Performance Criteria/Penalty Function

In order to evaluate and compare the scheduling policies, we define a piecewise-linear function that penalizes OR overtime. Per recommendation of our clinical collaborators, we consider 1 hour prior to and past 5 pm (i.e., 4 pm to 6 pm) as the *sweet spot* for ending the surgeon's day. Hence, we assign no penalty to the first 60 minutes of overtime for each surgeon. After that, a linear penalty is considered. Suppose $C(i,t)$ and $W(i,t)$ are the surgical capacity and the actual/effective workload of surgeon i on day t, respectively. The penalty of day t for surgeon i is obtained via

$$f(i,t) = (W(i,t) - C(i,t) - 60)^+, \tag{4.6}$$

in which $x^+ = \max\{x, 0\}$. It is worth noting that OR overtime happens because of (1) lack of accuracy in estimating the surgery duration, and (2) intentionally "overbooking" on some days to maintain the access delay service level.

4.5 Simulation Optimization Results

We begin this section by describing our 2-stage stochastic and dynamic discrete-event simulation and continue by presenting numerical results. In the first stage of our simulation model, we evaluate the performance of the 6 policies described in Subsection 4.4.4 for different order arrival rates under a fixed assumption on clinic to surgery gap (CSG). In the second stage, we fine-tune the winning policy of stage 1 by investigating the CSG function. We will call the fine-tuned version of the winning policy "the optimal policy" and further evaluate its performance.

We created a stochastic discrete-event simulation model of the CRS appointment system in MATLAB. It is assumed that requests/orders for clinic appointment arrive to CRS according to a Poisson process with rate λ. We consider online scheduling of orders. In other words, orders arrive one by one and get scheduled in the order they arrive. Each order is generated as a random sample from our large dataset. When a new order arrives (1) the patient's priority level is determined based on Table 4.1, (2) the maximum wait time target (MWTT) is determined based on Table 4.2, (3) the probability of surgery is calculated via Eq. 4.2, (4) the earliest feasible clinic (EFC) appointment day is determined via Eq. 4.3, (5) the clinic to surgery gap (CSG) is considered as 1 day if the priority level is 1 and 3 days otherwise (this will be further studied in the second stage of our simulation model), (6) a tentative surgery appointment is scheduled following the rules of the policy under consideration (see Section 4.4.4.1 for more details), and (7) a clinic appointment is scheduled according to the approach described in Section 4.4.4.2. At the end of an arbitrary day t, the simulation model either confirms or cancels the tentatively booked surgery appointments for patients whose clinic appointment was scheduled on day t. This is done by generating a random uniformly distributed number between 0 and 1 and comparing it against the patient's probability of surgery. Note that to better reflect practice we are taking a conservative approach and assume that the tentative

surgery appointments that are canceled throughout the day are realized at the end of the day (therefore, they cannot be rebooked until the following day).

Each replication of simulation is run for 350 days. Since we start with an empty system (i.e., all surgeons have full availability at the start of the simulation) we allow 100 days of warmup and use the data of days 101 to 350 (roughly one year since these are business days) to compare the policies. Note that orders may be scheduled beyond day 350; therefore, there is no end-of-horizon effect.

4.5.1 Simulation stage 1: comparing policies to find the winning policy

In the first stage of our stochastic simulation approach, we compare the scheduling policies for different rates of Poisson arrival process. We use the penalty function of Eq. 4.6 to calculate a penalty score for each policy. Mayo CRS performs about 200 surgeries per month. This corresponds to a rate of about 30 orders per day (remember that not every order results in a surgery). In this paper, we investigate arrival rates from 20 to 40 orders per day in increments of 2. We performed 200 replications of the simulation and calculated the daily overtime penalty of each replication (i.e., $\frac{1}{250}\sum_{t=101}^{350}\sum_{i=1}^{8} f(i,t)$, where $f(i,t)$ is given in Eq. 4.6), then took the average of these daily overtime penalties. Table 4.3 summarizes the results. As can be seen in Table 4.3, Policy D performs better than the other scheduling policies and results in smaller average overtime penalty per day. We also simulated the current scheduling policy described in Section 4.2. It is worth noting that all of the 6 scheduling policies, which tentatively book a surgery day at the same time the clinic appointment day is set, significantly outperform the current policy that waits until the clinic visit to book the surgery. Figure 4.2 portrays the variation of overtime penalty per day among simulation runs using jitter and box plots. Figure 4.3 depicts the average overtime penalty per day for Policy D and the current policy for different rates of the arrival process. At the rate of 30 orders per day (about 200 surgeries per month) which

Table 4.3: Average overtime penalty per day for the 6 scheduling policies and the policy of the current practice under different arrival rates. Policy D outperforms the rest for every rate in the range.

Arrival Rate (# Orders per Day)	Avg. # of Surgeries per Month	Average Overtime Penalty per Day						
		Policy A	Policy B	Policy C	Policy D	Policy E	Policy F	Current Policy
20	133.65	20.50	19.24	16.58	16.12	16.52	16.20	43.91
22	147.38	27.10	24.93	22.82	21.99	22.94	22.04	56.67
24	160.64	35.98	33.71	31.04	30.12	30.98	30.31	71.03
26	174.40	48.86	46.68	44.12	43.05	44.57	43.43	92.89
28	187.48	65.71	63.04	61.27	60.03	61.49	60.13	118.00
30	201.28	90.94	90.53	87.91	86.02	88.93	86.31	150.28
32	214.40	119.07	121.86	120.09	116.92	121.27	117.87	182.96
34	227.70	159.13	165.03	164.06	158.36	164.14	160.07	226.48
36	241.14	211.24	219.68	217.27	210.70	217.98	211.30	275.39
38	254.54	273.64	281.97	279.41	270.28	280.05	270.81	326.11
40	267.73	347.36	356.75	355.26	340.16	355.36	341.85	388.96

corresponds to the present business of Mayo CRS, Policy D performs 43% better than the current scheduling policy.

4.5.2 Simulation stage 2: fine-tuning the winning policy

Having found strong evidence that Policy D is the best policy among the 6 policies we investigated, we next fine-tune Policy D by evaluating its performance under different clinic to surgery gaps (CSGs). Recall that in the first stage of the simulation we assumed a CSG of 1 day for priority 1 and 3 days for other priority type patients. Now, we assume that the CSG can be 1, 3, or 5 days, and set the value based on the particular patient's probability of surgery, which is obtained from the logit model. The lower the probability of surgery is, the longer the gap between clinic and surgery appointments should be. This is because we are tentatively booking a surgery day for all patients before knowing whether the patient will actually need a surgery. If during the clinic visit it is deemed that the patient does not need a surgery, we will have $CSG - 1$ days to assign that OR time to another patient. Note that (1) we do not consider CSGs of 2 and 4 days because if the tentative surgery is scheduled on day t, the surgeon will not be in the clinic on days $t-2$ and $t-4$ to see the patient;

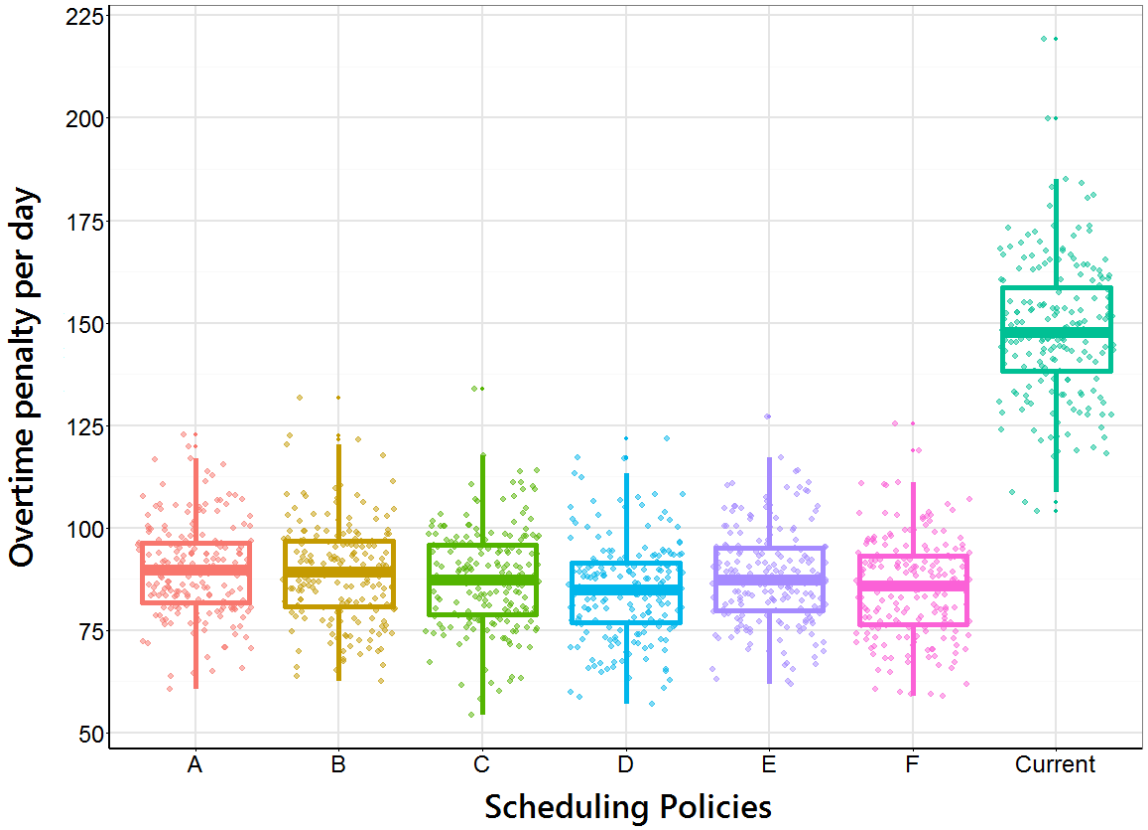

Figure 4.2: The overtime penalty per day of different scheduling policies and the current policy. All six policies outperform the current policy. Policy D performs the best, followed closely by Policy F.

and (2) the CSG cannot be 5 days for priority 1 patients because the MWTT for them is 3 days. Therefore, we consider the following CSG function:

$$CSG = \begin{cases} 1 \times I(p \geq \alpha) + 3 \times I(p < \alpha) & \text{if priority} = 1 \\ 1 \times I(p \geq \beta) + 3 \times I(\gamma \leq p < \beta) + 5 \times I(p < \gamma) & \text{if priority} \in \{2, 3, 4\} \end{cases} \tag{4.7}$$

in which p is the patient's probability of surgery and α, β, and γ are probability thresholds that determine which indicator functions are equal to 1. In this subsection, our goal is to find a good combination of α, β, and γ. We assume CSG is 3 days for

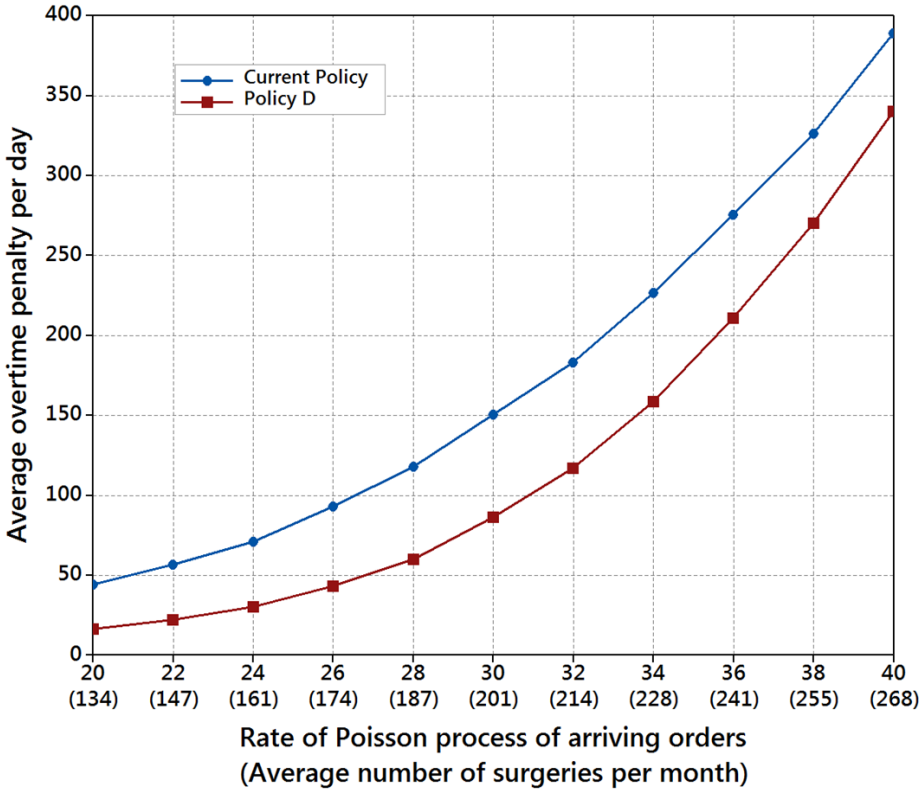

Figure 4.3: The average overtime penalty per day of Policy D and the current policy for different rates of the arrival process. The Policy D results in 43% less overtime penalty compared with the current scheduling policy of CRS at 200 surgeries per month.

all priority 2, 3, and 4 patients (i.e., $\beta = 1$ and $\gamma = 0$) and evaluate Policy D under different values of α. For each value we calculate the average overtime penalty per day based on 200 replications of the simulation model. Table 4.4 provides the results. It reveals that a threshold level of $\alpha = 0.4$ results in slightly better performance of this policy. Now that we have approximated the optimum value of α, we evaluate the performance of Policy D under different values of β and γ fixing α at 0.4. Table 4.5 presents the results. As can be seen in the table, Policy D performs better with threshold levels of $\beta = 0.8$ and $\gamma = 0.5$. From now on, we call the fine-tuned version of Policy D "the optimal policy" or "Policy D*".

Table 4.4: Average overtime penalty per day under Policy D for different values of α when $\beta = 1$ and $\gamma = 0$. The threshold level of $\alpha = 0.4$ results in slightly better performance.

α	Average Overtime Penalty per Day Under Policy D with β=1 and γ=0
0	85.74
0.1	85.74
0.2	85.74
0.3	85.74
0.35	85.64
0.4	85.35
0.45	86.35
0.5	93.28
0.55	110.41
0.6	121.32
0.7	143.49
0.8	194.36
0.9	194.36
1	194.36

4.5.3 The optimal policy (Policy D*)

Now that we found a good combination of the probability thresholds in the CSG function proposed in Eq. 4.7, below we summarize the steps of our methodological framework for coordinating clinic and surgery appointments to meet access delay service levels using the optimal policy (policy D*). For each new clinic appointment order/request take the following procedure:

Step 1: determine the patient's priority level, MWTT, probability of yielding a surgery, and EFC using Table 4.1, Table 4.2, and Eq. 4.2, and Eq. 4.3, respectively.

Step 2: Use Eq. 4.7 to calculate the CSG with $\alpha = 0.4$, $\beta = 0.8$ and $\gamma = 0.5$.

Step 3: Perform the following in order to schedule a tentative surgery day:

- If the patient is of priority 1, assign her to the surgeon with the earliest OR availability in the interval $[EFS, MWTT]$ starting from EFS and moving forward in time. If there is no availability in this interval, assign her to the surgeon-day

Table 4.5: Average overtime penalty per day under Policy D for different values of β and γ when $\alpha = 0.4$. The Policy D performs better with threshold levels of $\beta = 0.8$ and $\gamma = 0.5$.

β	γ	Average Overtime Penalty per Day Under Policy D with a=0.4
0.8	0.3	85.32
0.7	0.3	89.01
0.6	0.3	91.72
0.5	0.3	102.25
0.4	0.3	120.44
0.8	0.4	83.01
0.7	0.4	86.69
0.6	0.4	89.40
0.5	0.4	99.93
0.8	0.5	79.52
0.7	0.5	83.21
0.6	0.5	85.92
0.8	0.6	85.92
0.7	0.6	85.92
0.8	0.7	85.92

with the most remaining surgical availability (or the least amount of overtime) in this interval and use OR overtime.

- If the patient is of priority 2, 3, or 4, sort the days in the interval $[EFS, MWTT]$ in descending order of aggregated surgical availability. Assign the patient to a surgeon with sufficient OR availability for the expected surgery length of the case starting with the first day on this list (i.e., the least busy day) and moving down the list. If there is no availability in the interval $[EFS, MWTT]$, calculate the expected surgical workload of each surgeon-day on this interval using Eq. 4.4. Then, assign the patient to the surgeon-day with the least expected surgical workload (i.e., the most expected surgical availability) and use OR overtime.

Step 4: Use Eq. 4.5 in order to schedule a clinic appointment day for the patient with the same surgeon.

Step 5: During the clinic appointment, if the surgeon and the patient decide to

pursue a surgery, offer the tentatively booked surgery day (this is guaranteed to be within their MWTT). If that day does not work for the patient, go to step 6.

Step 6: Offer the surgeon's next surgical availability (this is not guaranteed to be within the patient's MWTT). If that day does not work for them, repeat step 6 until a surgery day is found.

4.5.4 The optimal vs. the current policy

In this subsection, we compare the average overtime penalty per day of Policy D* with that of Policy D and the current policy (i.e., the scheduling policy in current practice) using the previous simulation performance analysis approach. Figure 4.4 depicts the results. It can be seen in Figure 4.4 that at arrival rate of 30 orders per day (about 200 surgeries per month), Policy D* performs 17% better than Policy D and 52% better that the current policy. Figure 4.5 shows the box plots of the three policies at different arrival rates to depict the variation of overtime penalty per day among simulation runs. Note that in this figure the horizontal axis is a factorial vector of arrival rates with 11 levels (from 20 to 40 in increments of 2) and not a continuous scale.

4.6 Conclusions

Patients are referred to the colorectal surgery (CRS) department at the Mayo Clinic either internally or externally. They first see a surgeon in the clinic, and usually the clinic visit is followed by a surgery. Recently, CRS has been experiencing a rapid increase in the waiting time for the next available surgery day they can offer to the patients. The increased wait times can negatively impact patient safety and health outcomes.

In this paper, we presented a 2-stage stochastic and dynamic discrete-event simulation model that finds (1) a tentative surgery day with a surgeon in CRS, and (2) a clinic

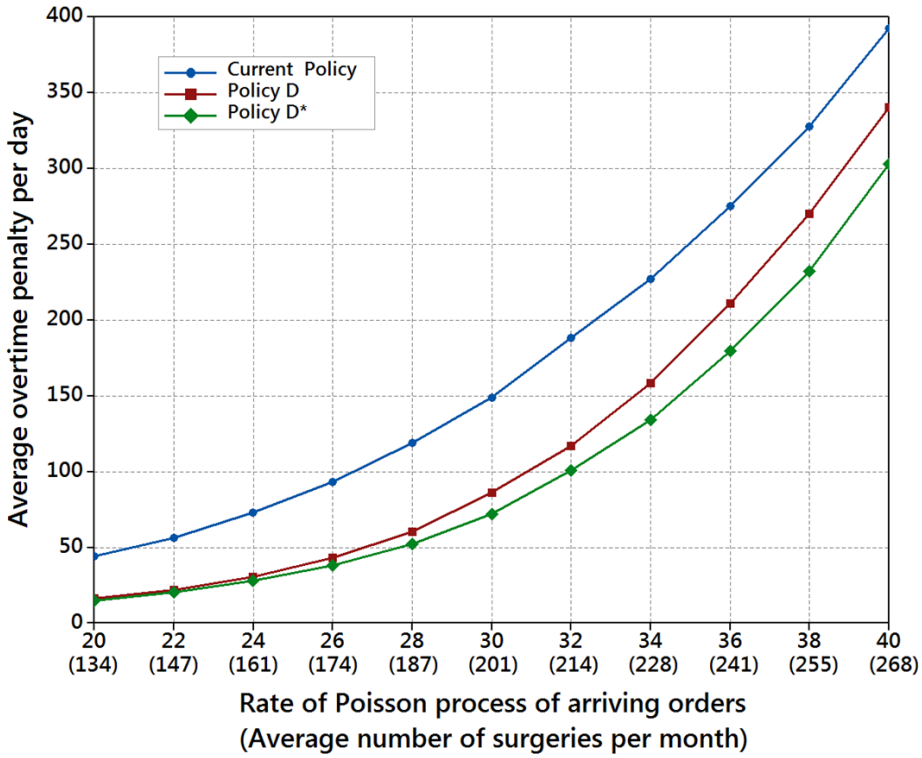

Figure 4.4: The average overtime penalty per day of Policies D*, D, and the current policy for different rates of the arrival process. The optimal policy (Policy D*) results in 52% less overtime penalty compared with the current scheduling policy of CRS at 200 surgeries per month.

appointment day with the same surgeon, for all new patents referred to CRS such that (1) all patients are offered at least one clinic and surgery day prior to their maximum wait time target (MWTT), and (2) OR overtime is minimized. To evaluate scheduling policies we developed a simulation model based on historical patient data from Mayo CRS. In the first stage of our approach, we investigated 6 different scheduling policies that all outperform the current scheduling policy and found the best/winning policy. In the second stage, the winning policy was fine-tuned through the investigation of different variations of the clinic to surgery gap. The fine-tuned version of the winning policy (i.e., the optimal policy) was compared against the current practice. Numerical results confirm that the optimal policy performs 52% better that the current policy in terms of the average overtime per day.

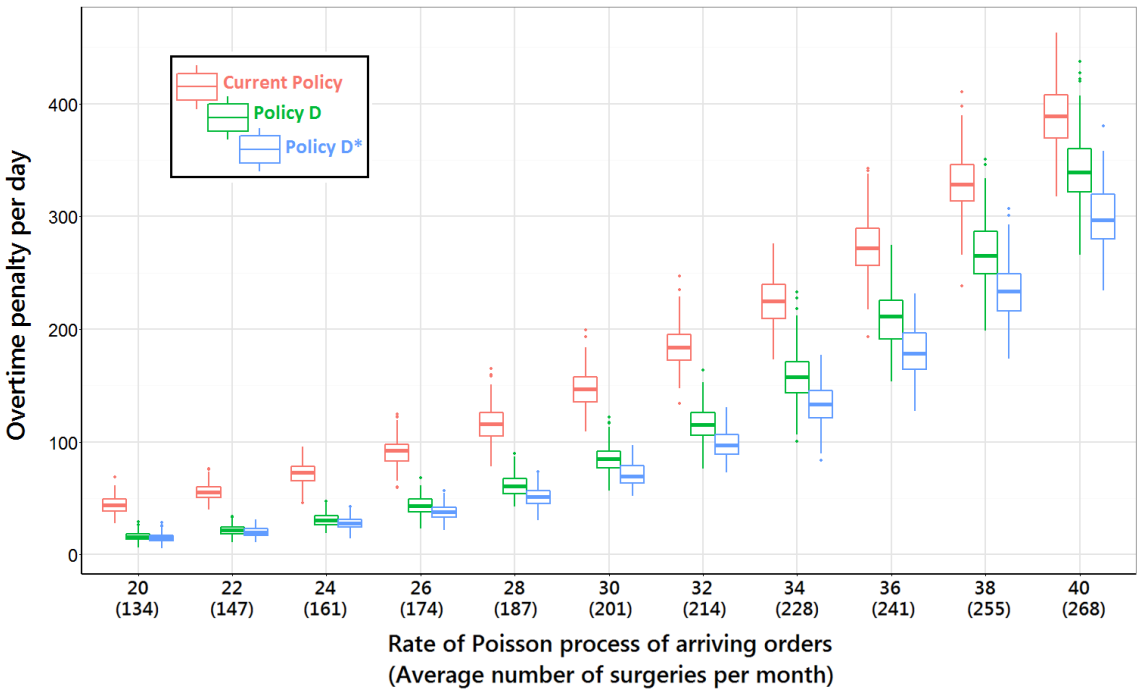

Figure 4.5: The box plots of overtime penalty per day of Policies D*, D, and the current policy for different rates of the arrival process. Policy D* clearly performs the best.

CHAPTER V

Conclusions and Future Research

This dissertation has focused on developing new stochastic control and optimization models to improve medical decision making and healthcare operations with the objective of improving health outcomes and cost containment. These objectives were achieved through better personalizing both the medical care itself and the operational aspects of the delivery of care in the work on medical decision making for chronic diseases, which focused on glaucoma. In other chapters, the operational/logistical aspect of care delivery was addressed through improving patient safety by reducing medical errors and providing timely access to care to the right patients based on the urgency needs of different types of patients.

Chapter II contributes to the medical decision making literature by developing a dynamic, personalized modeling paradigm for simultaneous monitoring and control of chronic diseases. Unlike previous research that solves the monitoring and treatment control problems in isolation, we provide a joint optimal solution to both problems in an integrated model. Our model incorporates each patient's past and present readings in a feedback-driven control model to provide the answer to two critical questions facing clinicians: (1) when to schedule office visits and which suite of test to perform to monitor disease progression (exploration); and (2) what levels of key disease risk factors should be targeted to slow the rate of disease progression (exploitation). For

glaucoma, a progressive eye disease that can lead to blindness, our model determines the best time to measure patient's intraocular pressure (IOP) and/or take a visual field (VF) test. Additionally, it provides a patient-specific target IOP to slow glaucoma progression. Since IOP is the only modifiable glaucoma risk factor, having such information will help the doctor select the appropriate treatment plan for the patient. Kalman filtering methodology is built into our modeling framework to extract noise from the raw measurements and to optimally estimate the disease state in each time period based on imperfect observations. This is a key to accurately identify genuine disease progression from testing artifact since both IOP and VF tests are associated with significant noise. Further, the algorithm can perform at different aggressiveness levels (low, moderate, and high-aggressiveness) based on individual patient characteristics (e.g. a 40-year-old patient with advanced disease vs. a 90-year-old patient with mild disease).

Capturing the complex patient disease dynamics requires incorporating several physiological indicators into the state vector. These state elements are best modeled with continuous state space model, reflecting continuous test measurements which have been found to have approximately Normal additive noise. A vector state space model of patient disease progression based on the linear quadratic Gaussian (LQG) dynamic control framework allowed us to create a data-driven model that separates process noise from measurement noise, and allows for incomplete/imperfect state observations. Our models can be described as possessing "measurement subsystems control," because the optimization and control must dynamically decide which measurements (tests of the patient) to take at each instant of time, thereby controlling the times at which tests are taken. Furthermore, the need to model testing process noise makes approximate dynamic programming (ADP), partially observable Markov decision processes (POMDP), and other approaches common in the medical decision making literature less effective in developing tractable solution methods for monitor-

ing and treatment control for chronic diseases. By developing a multivariate continuous state space model of disease progression and modeling the state transition and the testing processes as first order vector difference equations with multivariate Gaussian random noises, we surmount the computational hurdles of these other methods.

Chapter II also contributes to the literature on linear quadratic Gaussian (LQG) state space systems modeling and the theory of optimal control of measurement adaptive systems by introducing a new objective that minimizes the relative change in state (i.e., disease progression) rather than the traditional objective of simply minimizing the cost of being in each state. This new objective is important for our application since the goal is to prevent the patient from getting worse (in the setting of progressing diseases which are irreversible but can be slowed or arrested). Further, we proved an important result that the classical two-way separation of optimal estimation and control extends to this new objective of relative system state change. This is a fundamental finding upon which solution tractability depends. Leveraging this result, we were able to show that the optimal treatment control action at each time point is a linear function of filtered state mean, while the function itself (i.e., the control law) can be calculated offline. Moreover, we showed that the optimal monitoring schedule can be obtained by solving a recursive value function of filtered and smoothed covariance matrices of the state via branch and bound dynamic programming.

To demonstrate the effectiveness of our approach, we harnessed two 10+ year randomized clinical trials to parametrize and validate our model: the Collaborative Initial Glaucoma Treatment Study (CIGTS) and the Advanced Glaucoma Intervention Study (AGIS). Our numerical results demonstrated that our model not only results in patients with better vision quality over the treatment horizon, but also achieves significantly slower glaucoma progression rate, which means patients keep their sight longer.

Given the power of our approach, we expect that our new medical decision making

formulation and solution approach will be confirmed by future work from other researchers, and that a new research area for the healthcare operations research and medical community will flourish to advance quality of care and quality of life for patients with chronic diseases.

While our modeling framework has shown great potential in improving the monitoring and control of chronic disease patients with glaucoma, it comes with a few limitations and areas for improvement. First, we only considered the IOP and VF tests in developing a monitoring schedule for the patient, whereas in practice there are additional tests that can also be used to monitor glaucomatous progression. For instance, optical coherence tomography (OCT) is a non-invasive imaging test that measures the thickness of retinal nerve fiber layer (RNFL) (see *Schuman et al.* (2004)). This newer testing modality was not commercially available at the time of the CIGTS and AGIS clinical trials on which our analysis is based. Fortunately, the decision framework we have developed is scalable and can easily accommodate quantitative data of tests such as OCT. Should newer modalities for quantitatively assessing the status of a patient's glaucoma arise, data from such modalities can be incorporated into the model as well. In the future, we hope to acquire access to other data sources which contain OCT data and expand our state vector to accommodate data from this testing modality.

Second, the focus of our work is on patients who already have glaucoma (mild, moderate, or advance glaucoma). Our model provides scheduling regimes and IOP reduction suggestions to optimally monitor and control the progression of glaucoma. However, some patients have elevated eye pressures (i.e., ocular hypertension) with no signs of glaucoma. These patients are usually considered as "glaucoma suspects" and are at risk of developing glaucoma. Future research can enhance the model to help forecast which of these patients will go on to develop glaucoma and how often these patients should be monitored to see if they develop glaucoma.

Third, we used 6-month spaced time intervals because CIGTS and AGIS datasets

contain readings of patient's IOP and VF every six months. One can leverage the same modeling framework for data that is collected more or less frequently (e.g., monthly or every 3 months) without loss of generality. However, expanding our algorithm to automatically handle unequally spaced data is another potential path for future research.

Moreover, it is important to emphasize that our model does not suggest a specific medicine, laser therapy, or surgery. Rather, it provides a patient specific target IOP that helps guide the doctor in selecting an appropriate treatment plan. Though one might try to model how each glaucoma intervention affects the disease progression dynamics, we feel that it is best to leave it to the clinician to employ his/her experience and expertise to decide what therapeutic interventions are most likely able to achieve the target IOP suggested by our model.

Finally, future research can focus on further studying the model cost parameters. In this research, we relied on expert opinion and developed a set of cost parameters for each aggressiveness policy that performed on our extant data. Future research, however, can try to develop algorithms to optimize the balance between the cost of losing vision over time from glaucoma, cost of purchasing medications/undergoing surgery, cost of office visit and diagnostic testing, anxiety and stress of undergoing glaucoma tests, and side effects and complications of medical and surgical therapy to lower IOP further.

In Chapter III we aimed to reduce medical errors and therefore, improve patient safety, by designing a new work shift schedules for residents and fellows to minimize the number of error-prone patient handoffs. Patient handoffs are among the primary sources of medical errors in inpatient hospital care due to communication breakdowns. Several studies have focused on the fidelity of handoffs and provided recommendations and protocols to improve the communication aspects of handoffs. In this chapter, we contributed to the patient handoff literature by providing an

integer programming-based approach to design physician's work shift schedule in a patient-centered manner that minimizes the number of patient handoffs while respecting duty-hour standards. The ACGME rules allow shifts up to 24 hours. Based on several discussions with our medical collaborators at the Mayo Clinic, we found that limiting the shift length to 16-hours provided a reasonable tradeoff between medical errors due to physician fatigue and medical errors due to communication breakdowns as a result of more frequent patient handoffs. Therefore, we proposed to increase the current shift length from 12 to 16 hours, which reduces the number of handoffs while avoiding high fatigue-inducing shifts.

Using historical data from a medical ICU at the Mayo Clinic, we demonstrated that the computer-generated schedule with 16-hour shifts results in 23% reduction in the number of patient handoffs. It is worth noting that the proposed schedule satisfies all of the required scheduling rules, provide the required coverage, and maintains physician quality of life by satisfying a set of desired livability constraints in addition to the required constraints. Further, we added fairness constraints to ensure that all the physicians get almost the same number of hours on duty, night shifts, and days off.

There are, however, several directions for future research in this line. First, it is not currently known how much the extra weariness due to longer shifts contributes to fatigue-related medical errors. If physician fatigue and its effect on medical errors can be quantified in a systematic way, it will help in scientifically evaluating schedules that minimize the number of patient handoffs. Second, the connection between ICU rounding time and patient census pattern could be further investigated. In this study, we included a required constraint to keep bedside rounds at the current time to avoid the complex effect of rounding time on the patient discharge process, which directly influences ICU patient census.

Chapter IV presented novel ideas on coordinating clinic and surgery appointment scheduling with the objective of improving timeliness of access to surgery. In collab-

oration with colorectal surgery (CRS) at the Mayo Clinic, we showed how patient information can be used to proactively schedule a tentative surgery day for all patients at the same time that we schedule a clinic appointment for them. Patients are classified into different types based on their indication, geographical zone, and referral source. Each type is associated with a maximum wait time target (MWTT) that is determined based on expert opinion from the leadership of Mayo Clinic. This is the maximum time each patient's surgery can be safely delayed. Further, we developed a logit model to calculate the probability that an arriving patient of any type will yield a surgery in CRS. Surgery duration is also estimated for all incoming patient types based on historical patient data. We schedule a clinic appointment day and a tentative surgery day for all patients by their MWTT to guarantee that at least one surgery day is offered to all patients without jeopardizing their health status due to access delay. Overtime is used if it is the only way to achieve the MWTT access. However, the clinic appointment days and the tentative surgery days are determined by algorithms that carefully space them out to ensure that when the tentative surgery bookings are not needed (i.e., get canceled) they can be assigned to other patients. This tradeoff of utilization/overtime versus waiting time is a difficult one at the core of this chapter.

To solve this problem, we developed a 2-stage stochastic and dynamic discrete event simulation model to compare the performance of 6 heuristic scheduling policies. It is worth noting that all 6 policies significantly outperform the current scheduling policy. The current policy sets a surgery day for the patient at the time of his/her clinic visit. In the first stage of the simulation model, we found the policy that works best among the 6 policies (which we call the winning policy). In the second stage, we fine-tuned the winning policy by optimizing the rules that determine the distance/gap between clinic and surgery days. All of the policies will achieve the MWTT goals, using overtime when needed. Numerical results confirm that the optimal policy performs 52%

better that the current policy in terms of the average overtime per day. While the proof of concept study is given for Mayo CRS, this methodology can be applied to any specialty care unit that provides clinic and surgery appointments. Future research can study the effect of MWTTs on the optimal policy and/or develop a modeling framework to dynamically update the MWTTs based on the current workload of the surgeons. Moreover, in this research we assumed that (1) all patients must get at least one clinic and surgery day by their MWTT, and (2) overtime will be used as needed to accommodate the surge capacity. One can relax this assumption by trying to balance the tradeoff between using overtime and missing the MWTTs. In this case, the system should be penalized more for missing the MWTT of higher priority patients. Lastly, while we studied the performance of 6 carefully chosen scheduling policies, it is possible to investigate additional heuristic scheduling policies, in particular those that belong to the category of threshold policies. In this type of policies, a certain percentage of capacity is reserved for each priority type patients. The threshold limit of each day may be relaxed as we get close to the day to avoid the risk of ending up with unused clinic and operating room time. All of the research presented in this dissertation has been done in close collaboration with several medical collaborators from cutting-edge healthcare organizations (i.e., University of Michigan Health System and the Mayo Clinic) to identify real-world problems, obtain relevant datasets, and ensure practical significance. Combining the ideas and methods developed in this dissertation will result in a step toward bending the cost curve of healthcare spending and can contribute to a positive impact on health outcomes and improve the quality of life for many people.

BIBLIOGRAPHY

BIBLIOGRAPHY

Abraham, J., T. G. Kannampallil, and V. L. Patel (2012), Bridging gaps in handoffs: a continuity of care based approach, *Journal of biomedical informatics*, *45*(2), 240–254.

Aickelin, U., and K. A. Dowsland (2004), An indirect genetic algorithm for a nurse-scheduling problem, *Computers & Operations Research*, *31*(5), 761–778.

American Academy of Ophthalmology Glaucoma Panel (2010), Preferred practice pattern guidelines: Primary open-angle glaucoma, *San Francisco, CA: American Academy of Ophthalmology*, available at: www.aao.org/ppp.

Ansley, C. F., and R. Kohn (1982), A geometrical derivation of the fixed interval smoothing algorithm, *Biometrika*, *69*(2), 486–487.

Arora, V., and J. Johnson (2006), A model for building a standardized hand-off protocol, *The Joint Commission Journal on Quality and Patient Safety*, *32*(11), 646–655.

Arora, V., J. Johnson, D. Lovinger, H. Humphrey, and D. Meltzer (2005), Communication failures in patient sign-out and suggestions for improvement: a critical incident analysis, *Quality and Safety in Health Care*, *14*(6), 401–407.

Arora, V., J. Kao, D. Lovinger, S. C. Seiden, and D. Meltzer (2007), Medication discrepancies in resident sign-outs and their potential to harm, *Journal of General Internal Medicine*, *22*(12), 1751–1755.

Arora, V., J. Johnson, D. Meltzer, and H. Humphrey (2008), A theoretical framework and competency-based approach to improving handoffs, *Quality and Safety in Health Care*, *17*(1), 11–14.

Athans, M. (1972), On the determination of optimal costly measurement strategies for linear stochastic systems, *Automatica*, *8*(4), 397 – 412.

Ayer, T., O. Alagoz, and N. K. Stout (2012), A pomdp approach to personalize mammography screening decisions, *Operations research*, *60*(5), 1019–1034.

Azadeh, M. A., B. M. Shoja, P. Kazemian, and Z. T. Hojati (2013), A hybrid ant colony-computer simulation approach for optimum planning and control of maritime traffic, *International Journal of Industrial and Systems Engineering*, *15*(1), 69–89.

Balas, E., and C. H. Martin (1980), Pivot and complement–a heuristic for 0-1 programming, *Management Science*, *26*(1), 86–96.

Bansal, R., and T. Başar (1989), Simultaneous design of measurement and control strategies for stochastic systems with feedback, *Automatica*, *25*(5), 679–694.

Bard, J. F., and H. W. Purnomo (2005), Short-term nurse scheduling in response to daily fluctuations in supply and demand, *Health Care Management Science*, *8*(4), 315–324.

Baron, S., and D. L. Kleinman (1969), The human as an optimal controller and information processor, *Man-Machine Systems, IEEE Transactions on*, *10*(1), 9–17.

Beaulieu, H., J. A. Ferland, B. Gendron, and P. Michelon (2000), A mathematical programming approach for scheduling physicians in the emergency room, *Health care management science*, *3*(3), 193–200.

Berg, B. P., B. T. Denton, S. A. Erdogan, T. Rohleder, and T. Huschka (2014), Optimal booking and scheduling in outpatient procedure centers, *Computers & Operations Research*, *50*, 24–37.

Bertsekas, D. P., D. P. Bertsekas, D. P. Bertsekas, and D. P. Bertsekas (1995), *Dynamic programming and optimal control*, vol. 1, Athena Scientific Belmont, MA.

Borman, K. R., A. T. Jones, and J. A. Shea (2012), Duty hours, quality of care, and patient safety: general surgery resident perceptions, *Journal of the American College of Surgeons*, *215*(1), 70–77.

Carter, M. W., and S. D. Lapierre (2001), Scheduling emergency room physicians, *Health Care Management Science*, *4*(4), 347–360.

Centers for Medicare & Medicaid Services (2014), National health expenditures 2014 highlights, *Online verfügbar unter https://www.cms.gov/research-statistics-data-and-systems/statistics-trends-and-reports/nationalhealthexpenddata/downloads/highlights.pdf*.

Cheung, D. S., et al. (2010), Improving handoffs in the emergency department, *Annals of emergency medicine*, *55*(2), 171–180.

Chhatwal, J., O. Alagoz, and E. S. Burnside (2010), Optimal breast biopsy decision-making based on mammographic features and demographic factors, *Operations research*, *58*(6), 1577–1591.

Choplin, N. T., and R. P. Edwards (1995), *Visual field testing with the Humphrey Field Analyzer*, Slack Inc.

Clark, C. J., S. L. Sindell, and R. P. Koehler (2011), Template for success: using a resident-designed sign-out template in the handover of patient care, *Journal of surgical education*, *68*(1), 52–57.

Cohn, A., S. Root, C. Kymissis, J. Esses, and N. Westmoreland (2009), Scheduling medical residents at boston university school of medicine, *Interfaces*, *39*(3), 186–195.

Davis, K., K. Stremikis, D. Squires, and C. Schoen (2014), Mirror, mirror on the wall, *How the performance of the US Health care system compares internationally. New York: CommonWealth Fund*.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977), Maximum likelihood from incomplete data via the em algorithm, *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38.

Digalakis, V., J. R. Rohlicek, and M. Ostendorf (1993), Ml estimation of a stochastic linear system with the em algorithm and its application to speech recognition, *Speech and Audio Processing, IEEE Transactions on*, *1*(4), 431–442.

Donchin, Y., D. Gopher, M. Olin, Y. Badihi, M. Biesky, C. Sprung, R. Pizov, and S. Cotev (2003), A look into the nature and causes of human errors in the intensive care unit, *Quality and Safety in Health Care*, *12*(2), 143–147.

Ederer, F., D. Gaasterland, E. Sulivan, and A. I. A. Investigators (1994), The advanced glaucoma intervention study (agis): 1. study design and methods and baseline characteristics of study patients, *Controlled clinical trials*, *15*(4), 299–325.

Erdogan, S. A., and B. T. Denton (2011), Surgery planning and scheduling, *Wiley Encyclopedia of Operations Research and Management Science*.

Ernst, A. T., H. Jiang, M. Krishnamoorthy, and D. Sier (2004), Staff scheduling and rostering: A review of applications, methods and models, *European journal of operational research*, *153*(1), 3–27.

Everett, J. E. (2002), A decision support simulation model for the management of an elective surgery waiting system, *Health Care Management Science*, *5*(2), 89–95.

Fone, D., et al. (2003), Systematic review of the use and value of computer simulation modelling in population health and health care delivery, *Journal of Public Health*, *25*(4), 325–335.

Frankel, H. L., A. Foley, C. Norway, and L. Kaplan (2006), Amelioration of increased intensive care unit service readmission rate after implementation of work-hour restrictions, *Journal of Trauma and Acute Care Surgery*, *61*(1), 116–121.

Gandhi, T. K. (2005), Fumbled handoffs: one dropped ball after another, *Annals of internal medicine*, *142*(5), 352–358.

Gardiner, S. K., and D. P. Crabb (2002), Examination of different pointwise linear regression methods for determining visual field progression, *Investigative ophthalmology & visual science*, *43*(5), 1400–1407.

Gascon, V., S. Villeneuve, P. Michelon, and J. A. Ferland (2000), Scheduling the flying squad nurses of a hospital using a multi-objective programming model, *Annals of Operations Research*, *96*(1-4), 149–166.

Gendreau, M., J. Ferland, B. Gendron, N. Hail, B. Jaumard, S. Lapierre, G. Pesant, and P. Soriano (2006), Physician scheduling in emergency rooms, in *Practice and Theory of Automated Timetabling VI*, pp. 53–66, Springer.

Ghahramani, Z., and G. E. Hinton (1996), Parameter estimation for linear dynamical systems, *Tech. rep.*, Technical Report CRG-TR-96-2, University of Totronto, Dept. of Computer Science.

Greenberg, C. C., S. E. Regenbogen, D. M. Studdert, S. R. Lipsitz, S. O. Rogers, M. J. Zinner, and A. A. Gawande (2007), Patterns of communication breakdowns resulting in injury to surgical patients, *Journal of the American College of Surgeons*, *204*(4), 533–540.

Gul, S., B. T. Denton, J. W. Fowler, and T. Huschka (2011), Bi-criteria scheduling of surgical services for an outpatient procedure center, *Production and Operations management*, *20*(3), 406–417.

Gupta, D., and B. Denton (2008), Appointment scheduling in health care: Challenges and opportunities, *IIE transactions*, *40*(9), 800–819.

Gupta, D., and L. Wang (2008), Revenue management for a primary-care clinic in the presence of patient choice, *Operations Research*, *56*(3), 576–592.

Gupta, V., T. H. Chung, B. Hassibi, and R. M. Murray (2006), On a stochastic sensor selection algorithm with applications in sensor scheduling and sensor coverage, *Automatica*, *42*(2), 251–260.

Gutjahr, W. J., and M. S. Rauner (2007), An aco algorithm for a dynamic regional nurse-scheduling problem in austria, *Computers & Operations Research*, *34*(3), 642–666.

Harvey, A. C. (1990), *Forecasting, structural time series models and the Kalman filter*, Cambridge university press.

Heijl, A., P. Buchholz, G. Norrgren, and B. Bengtsson (2013), Rates of visual field progression in clinical glaucoma care, *Acta ophthalmologica*, *91*(5), 406–412.

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

Henriksen, K., et al. (2008), Using six sigma methodology to improve handoff communication in high-risk patients, *Agency for Healthcare Research and Quality (US)*.

Hilkhuysen, G., J. Oudhoff, M. Rietberg, G. Van der Wal, and D. Timmermans (2005), Waiting for elective surgery: a qualitative analysis and conceptual framework of the consequences of delay, *Public health*, *119*(4), 290–293.

Ho, C.-J., and H.-S. Lau (1992), Minimizing total cost in scheduling outpatient appointments, *Management science*, *38*(12), 1750–1764.

Horwitz, L. I., H. M. Krumholz, M. L. Green, and S. J. Huot (2006), Transfers of patient care between house staff on internal medicine wards: a national survey, *Archives of internal medicine*, *166*(11), 1173–1177.

Horwitz, L. I., T. Moin, and M. L. Green (2007), Development and implementation of an oral sign-out skills curriculum, *Journal of general internal medicine*, *22*(10), 1470–1474.

Hutter, M. M., K. C. Kellogg, C. M. Ferguson, W. M. Abbott, and A. L. Warshaw (2006), The impact of the 80-hour resident workweek on surgical residents and attending surgeons, *Annals of surgery*, *243*(6), 864.

Institute of Medicine, Committee on Quality of Health Care in America (2001), *Crossing the quality chasm: A new health system for the 21st century*, National Academy Press.

Jagsi, R., B. T. Kitch, D. F. Weinstein, E. G. Campbell, M. Hutter, and J. S. Weissman (2005), Residents report on adverse events and their causes, *Archives of Internal Medicine*, *165*(22), 2607–2613.

Jampel, H. D. (1997), Glaucoma care update target pressure in glaucoma therapy., *Journal of glaucoma*, *6*(2), 133–138.

Jonnalagadda, R., E. Walrond, S. Hariharan, M. Walrond, and C. Prasad (2005), Evaluation of the reasons for cancellations and delays of surgical procedures in a developing country, *International journal of clinical practice*, *59*(6), 716–720.

Kalman, R. E. (1960), A new approach to linear filtering and prediction problems, *Journal of Fluids Engineering*, *82*(1), 35–45.

Kazemian, P., Y. Dong, T. R. Rohleder, J. E. Helm, and M. P. Van Oyen (2014), An ip-based healthcare provider shift design approach to minimize patient handoffs, *Health care management science*, *17*(1), 1–14.

Kazemian, P., J. E. Helm, M. S. Lavieri, J. Stein, and M. P. Van Oyen (2016a), Dynamic monitoring and control of irreversible chronic diseases with application to glaucoma, *Available at SSRN: http://ssrn.com/abstract=2733399*.

Kazemian, P., M. Y. Sir, K. S. Pasupathy, and M. P. Van Oyen (2016b), Coordinating clinic and surgery appointments to meet access delay service level for elective surgery, *Available at SSRN*.

Kemp, C. D., et al. (2008), The top 10 list for a safe and effective sign-out, *Archives of Surgery*, *143*(10), 1008–1010.

Kitch, B. T., J. B. Cooper, W. M. Zapol, J. E. Marder, A. Karson, M. Hutter, and E. G. Campbell (2008), Handoffs causing patient harm: a survey of medical and surgical house staff, *The Joint Commission Journal on Quality and Patient Safety*, *34*(10), 563–570D.

Kohn, L. T., J. M. Corrigan, M. S. Donaldson, et al. (2000), To err is human: building a safer health system. a report of the committee on quality of health care in america, institute of medicine.

Koopmanschap, M., W. Brouwer, L. Hakkaart-van Roijen, and N. van Exel (2005), Influence of waiting time on cost-effectiveness, *Social Science & Medicine*, *60*(11), 2501–2504.

LaGanga, L. R., and S. R. Lawrence (2007), Clinic overbooking to improve patient access and increase provider productivity*, *Decision Sciences*, *38*(2), 251–276.

Landrigan, C. P., et al. (2004), Effect of reducing interns' work hours on serious medical errors in intensive care units, *New England Journal of Medicine*, *351*(18), 1838–1848.

Lavieri, M. S., M. L. Puterman, S. Tyldesley, and W. J. Morris (2012), When to treat prostate cancer patients based on their psa dynamics, *IIE Transactions on Healthcare Systems Engineering*, *2*(1), 62–77.

Leape, L. L., D. S. Swankin, and M. R. Yessian (1999), A conversation on medical injury., *Public Health Reports*, *114*(4), 302.

Lee, P. P., et al. (2006), A multicenter, retrospective pilot study of resource use and costs associated with severity of disease in glaucoma, *Archives of ophthalmology*, *124*(1), 12–19.

Li, P., H. T. Stelfox, and W. A. Ghali (2011), A prospective observational study of physician handoff for intensive-care-unit-to-ward patient transfers, *The American journal of medicine*, *124*(9), 860–867.

Liang, F., Y. Guo, and R. Y. Fung (2015), Simulation-based optimization for surgery scheduling in operation theatre management using response surface method, *Journal of medical systems*, *39*(11), 1–11.

Liu, N., S. Ziya, and V. G. Kulkarni (2010), Dynamic scheduling of outpatient appointments under patient no-shows and cancellations, *Manufacturing & Service Operations Management*, *12*(2), 347–364.

Lockley, S. W., C. P. Landrigan, L. K. Barger, and C. A. Czeisler (2006), When policy meets physiology: the challenge of reducing resident work hours., *Clinical orthopaedics and related research*, *449*, 116–127.

Lokketangen, A., and F. Glover (1998), Solving zero-one mixed integer programming problems using tabu search, *European Journal of Operational Research*, *106*(2), 624–658.

MacCormick, A. D., W. G. Collecutt, and B. R. Parry (2003), Prioritizing patients for elective surgery: a systematic review, *ANZ Journal of Surgery*, *73*(8), 633–642.

Mangasarian, O. L., W. N. Street, and W. H. Wolberg (1995), Breast cancer diagnosis and prognosis via linear programming, *Operations Research*, *43*(4), 570–577.

Mason, J. E., B. T. Denton, N. D. Shah, and S. Smith (2014), Optimizing the simultaneous management of blood pressure and cholesterol for type 2 diabetes patients, *European Journal of Operational Research*, *233*(3), 727–738.

Mehra, R. K. (1976), Optimization of measurement schedules and sensor designs for linear dynamic systems, *Automatic Control, IEEE Transactions on*, *21*(1), 55–64.

Meier, L., J. Peschon, and R. M. Dressler (1967), Optimal control of measurement subsystems, *Automatic Control, IEEE Transactions on*, *12*(5), 528–536.

Mullen, P. M. (2003), Prioritising waiting lists: how and why?, *European Journal of Operational Research*, *150*(1), 32–45.

Murray, M., and D. M. Berwick (2003), Advanced access: reducing waiting and delays in primary care, *Jama*, *289*(8), 1035–1040.

Musch, D. C., P. R. Lichter, K. E. Guire, C. L. Standardi, C. S. Group, et al. (1999), The collaborative initial glaucoma treatment study: study design, methods, and baseline characteristics of enrolled patients, *Ophthalmology*, *106*(4), 653–662.

Nasca, T. J., S. H. Day, and E. S. Amis Jr (2010), The new recommendations on duty hours from the acgme task force, *New England Journal of Medicine*, *363*(2), e3.

Naylor, C. D. (1991), A different view of queues in ontario, *Health Affairs*, *10*(3), 110–128.

Nemeth, C. P. (2012), *Improving healthcare team communication: building on lessons from aviation and aerospace*, Ashgate Publishing, Ltd.

Patrick, J., M. L. Puterman, and M. Queyranne (2008), Dynamic multipriority patient scheduling for a diagnostic resource, *Operations research*, *56*(6), 1507–1525.

Patterson, E. S., and R. L. Wears (2010), Patient handoffs: standardized and reliable measurement tools remain elusive, *The joint commission journal on quality and patient safety*, *36*(2), 52–61.

Petersen, L. A., T. A. Brennan, A. C. O'Neil, E. F. Cook, and T. H. Lee (1994), Does housestaff discontinuity of care increase the risk for preventable adverse events?, *Annals of internal medicine*, *121*(11), 866–872.

Philibert, I., and D. Leach (2005), Re-framing continuity of care for this century, *Quality and Safety in Health Care*, *14*(6), 394–396.

Rein, D. B., et al. (2006), The economic burden of major adult visual disorders in the united states, *Archives of Ophthalmology*, *124*(12), 1754–1760.

Risser, D. T., M. M. Rice, M. L. Salisbury, R. Simon, G. D. Jay, S. D. Berns, M. R. Consortium, et al. (1999), The potential for improved teamwork to reduce medical errors in the emergency department, *Annals of emergency medicine*, *34*(3), 373–383.

Saaty, T. L., and L. G. Vargas (1998), Diagnosis with dependent symptoms: Bayes theorem and the analytic hierarchy process, *Operations Research*, *46*(4), 491–502.

Saure, A., J. Patrick, S. Tyldesley, and M. L. Puterman (2012), Dynamic multi-appointment patient scheduling for radiation therapy, *European Journal of Operational Research*, *223*(2), 573–584.

Sayed, A. H. (2011), *Adaptive filters*, John Wiley & Sons.

Schell, G. J., M. S. Lavieri, J. E. Helm, X. Liu, D. C. Musch, M. P. Van Oyen, and J. D. Stein (2014), Using filtered forecasting techniques to determine personalized monitoring schedules for patients with open-angle glaucoma, *Ophthalmology*, *121*(8), 1539–1546.

Schuman, J. S., C. A. Puliafito, J. G. Fujimoto, and J. S. Duker (2004), *Optical coherence tomography of ocular diseases*, Slack New Jersey:.

Sexton, A., C. Chan, M. Elliott, J. Stuart, R. Jayasuriya, and P. Crookes (2004), Nursing handovers: do we really need them?, *Journal of nursing management*, *12*(1), 37–42.

Shechter, S. M., M. D. Bailey, A. J. Schaefer, and M. S. Roberts (2008), The optimal time to initiate hiv therapy under ordered health states, *Operations Research*, *56*(1), 20–33.

Sherali, H. D., M. H. Ramahi, and Q. J. Saifee (2002), Hospital resident scheduling problem, *Production Planning & Control*, *13*(2), 220–233.

Solet, D., J. Norvell, G. Rutan, and R. Frankel (2004), Physician-to-physician communication: Methods, practice and misgivings with patient handoffs., in *Journal of General Internal Medicine*, vol. 19, pp. 108–108, BLACKWELL PUBLISHING INC 350 MAIN ST, MALDEN, MA 02148 USA.

Solet, D. J., J. M. Norvell, G. H. Rutan, and R. M. Frankel (2005), Lost in translation: challenges and opportunities in physician-to-physician communication during patient handoffs, *Academic Medicine*, *80*(12), 1094–1099.

Somasekar, K., P. Shankar, M. Foster, and M. Lewis (2002), Costs of waiting for gall bladder surgery, *Postgraduate medical journal*, *78*(925), 668–669.

Sommer, A., J. M. Tielsch, J. Katz, H. A. Quigley, J. D. Gottsch, J. Javitt, and K. Singh (1991), Relationship between intraocular pressure and primary open angle glaucoma among white and black americans: the baltimore eye survey, *Archives of ophthalmology*, *109*(8), 1090–1095.

Stein, J. D., D. S. Kim, L. M. Niziol, N. Talwar, B. Nan, D. C. Musch, and J. E. Richards (2011), Differences in rates of glaucoma among asian americans and other racial groups, and among various asian ethnic groups, *Ophthalmology*, *118*(6), 1031–1037.

Sutcliffe, K. M., E. Lewton, M. M. Rosenthal, et al. (2004), Communication failures: an insidious contributor to medical mishaps, *ACADEMIC MEDICINE-PHILADELPHIA-*, *79*(2), 186–194.

Tham, Y.-C., X. Li, T. Y. Wong, H. A. Quigley, T. Aung, and C.-Y. Cheng (2014), Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis, *Ophthalmology*, *121*(11), 2081–2090.

Thrall, J. H. (2005), Prevalence and costs of chronic disease in a health care system structured for treatment of acute illness 1, *Radiology*, *235*(1), 9–12.

Tielsch, J. M., A. Sommer, K. Witt, J. Katz, and R. M. Royall (1990), Blindness and visual impairment in an american urban population: the baltimore eye survey, *Archives of Ophthalmology*, *108*(2), 286–290.

Ulmer, C., D. Wolman, M. Johns, et al. (2008), Committee on optimizing graduate medical trainee (resident) hours and work schedules to improve patient safety, institute of medicine. resident duty hours: Enhancing sleep, supervision, and safety.

Vajaranant, T. S., S. Wu, M. Torres, and R. Varma (2012), The changing face of primary open-angle glaucoma in the united states: demographic and geographic changes from 2011 to 2050, *American journal of ophthalmology*, *154*(2), 303–314.

Vasilakis, C., B. Sobolev, L. Kuramoto, and A. Levy (2007), A simulation study of scheduling clinic appointments in surgical care: individual surgeon versus pooled lists, *Journal of the Operational Research Society*, pp. 202–211.

Vidyarthi, A. R., V. Arora, J. L. Schnipper, S. D. Wall, and R. M. Wachter (2006), Managing discontinuity in academic medical centers: Strategies for a safe and effective resident sign-out, *Journal of Hospital Medicine*, *1*(4), 257–266.

Vitus, M. P., W. Zhang, A. Abate, J. Hu, and C. J. Tomlin (2012), On efficient sensor scheduling for linear dynamical systems, *Automatica*, *48*(10), 2482–2493.

Volpp, K. G., and D. Grande (2003), Residents suggestions for reducing errors in teaching hospitals, *N Engl J Med*, *348*(9), 851–855.

Wang, W.-Y., and D. Gupta (2011), Adaptive appointment systems with patient preferences, *Manufacturing & Service Operations Management*, *13*(3), 373–389.

Ward, B. W., J. S. Schiller, and R. A. Goodman (2014), Multiple chronic conditions among us adults: A 2012 update, *Preventing chronic disease*, *11*.

Weinbroum, A. A., P. Ekstein, and T. Ezri (2003), Efficiency of the operating room suite, *The American Journal of Surgery*, *185*(3), 244–250.

Wilensky, J. T., D. K. Gieser, M. L. Dietsche, M. T. Mori, and R. Zeimer (1993), Individual variability in the diurnal intraocular pressure curve, *Ophthalmology*, *100*(6), 940–944.

Witsenhausen, H. S. (1971), Separation of estimation and control for discrete time systems, *Proceedings of the IEEE*, *59*(11), 1557–1566.

Yang, Y., J. D. Goldhaber-Fiebert, and L. M. Wein (2013), Analyzing screening policies for childhood obesity, *Management science*, *59*(4), 782–795.

Ye, K., D. McD Taylor, J. C. Knott, A. Dent, and C. E. MacBean (2007), Handover in the emergency department: deficiencies and adverse effects, *Emergency Medicine Australasia*, *19*(5), 433–441.

Zhang, Z., and X. Xie (2015), Simulation-based optimization for surgery appointment scheduling of multiple operating rooms, *IIE Transactions*, *47*(9), 998–1012.