Operations Research & Statistical Learning Methods to Monitor the Progression of Glaucoma and Chronic Diseases

by

Isaac A. Jones

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
2021

Doctoral Committee:

Professor Mark Van Oyen, Chair
Professor Amy E. Cohn
Associate Professor Mariel Lavieri
Professor Clayton D. Scott

Isaac A. Jones
isaacaj@umich.edu
ORCID iD: 0000-0001-5009-471X

DEDICATION

To my mother,
my father,
my sisters,
my family,
my close friends

ACKNOWLEDGMENTS

In pursing my PhD in Industrial and Operations Engineering, I received a lot of support from the University of Michigan and more specifically its' department of Industrial and Operations Engineering. I would like to thank my PhD advisor, Dr. Mark Van Oyen. Mark, if you are reading this, you have played a significant role in me reaching this accomplishment. Your help and guidance did not go unnoticed. Additionally, I would like to thank Dr. Mariel Laveri, my PhD committee, and my research group for their help throughout this entire process.

I would like to give a special thanks to my academic advisor at Wayne State University, Ms. Gail Evans. She encouraged me to pursue my Doctorate in Industrial Engineering and played a major role in my collegiate academic achievements thus far.

Finally, I wish to give a wholeheartedly thanks to my mother, father, sisters, family, friends, and God. While my father is not here to see me finish this accomplishment, I know he is extremely proud of the man I have become. And mom, thank you for everything. I really cannot say it enough. Without you there would be no me. I know I give you headaches at times, but you truly inspire me. I understand how blessed I am to have been put in a position to reach this accomplishment and am therefore internally grateful for any and everyone that played a role, however small, in me getting here.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE

LIST OF TABLES

TABLE

LIST OF APPENDICES

ABSTRACT

This thesis focuses on advancing operations research and statistical learning methods for medical decision making to improve the care of patients diagnosed with chronic conditions. Because the National Center for Disease Prevention (2020) estimates chronic conditions affect approximately 60% of the US adult population, improving the care of patients with chronic conditions will improve the lives of most Americans. Patients diagnosed with chronic conditions face lifestyle changes, rising treatment costs, and frequently reductions in quality of life. To improve the way in which clinicians treat patients with chronic conditions, treatment decisions can be supplemented by evidenced-based, data driven algorithmic decision-making methods.

This thesis provides data-driven methodologies of a general nature that are instantiated for several medical decision-making problems. In chapter two we proactively identify the time of a patient's primary open angle glaucoma (POAG) progression under high measurement error conditions using a soft voting ensemble classification model. When medical tests have low residual variability (e.g., empirical difference between the patient's true and recorded value is small) they can effectively, without the use of sophisticated methods, identify the patient's current disease phase; however, when medical tests have moderate to high residual variability this may not be the case. We present a solution to the latter case. We find rapid progression disease phases can be proactively identified with the combination of denoising and supervised classification methods.

In chapter three, we determine the optimal time to next follow-up appointment for patients with the chronic condition of ocular hypertension (OHTN). Patients with OHTN are at increased risk of developing glaucoma and should be observed over their lifetime. Follow-up appointment schedules that are chosen poorly can result in, at minimum, delay in the detection of a patient's progression to glaucoma, and at worse, yield poor patient outcomes. To this end, we present a personalized decision support algorithm that uses the fitted Q-iteration reinforcement learning algorithm to recommend personalized time-to-next follow-up schedules that are based on a patient's medical state. We find personalized follow-up appointments schedules produced by reinforcement learning methods are superior to both 1-year and 2-year fixed interval follow-up appointment schedules.

In chapters four and five, we examine and compare several criteria for determining progression from OHTN to POAG and evaluate the use of a collective POAG conversion rule in predicting future occurrences of patients' POAG conversion. We find age, race, and sex are statistically significant determinants in progression for all compared criteria. However, there exists broad conversion discordance between the criteria, as demonstrated by statistically different survival curves and the limited overlap in eyes that progressed by multiple criteria. Ultimately, to permit machine learning models to predict conversion from OHTN to POAG, it is essential to have quantitative reference standards for POAG conversion for researchers to use. Additionally, using the collective POAG conversion rule, we find machine learning models can successfully predict future OHTN conversion events to POAG.

This research was conducted in collaboration with clinical disease/domain experts. All the medical decision-making research herein addresses real world healthcare issues, that, if solved, have the potential to improve vision care if implemented. While these methodologies primarily focus on chronic conditions affecting the eyes (e.g., OHTN and POAG), it is important to note that much of the work produced offers methods applicable to other chronic diseases.

CHAPTER 1

Introduction

1.1 Motivation

The National Center for Disease Prevention [11] estimates chronic conditions affect 60% of the adult population. [16] estimates the direct costs of treating chronic conditions (e.g., cancers, hypertension, mental disorders, diabetes, heart disease, pulmonary conditions, and stroke) exceeds 278 billion, and the indirect costs (e.g., lost workdays, caregiver, etc.) exceeds 1 trillion. In 2019 [20] estimated that approximately 18% (or 3.6 trillion) of the United States gross domestic product (GDP) is attributed to healthcare expenditures. This is an alarming figure as the healthcare expenditures in 2008 was estimated at approximately 16% (or 3 trillion) of GDP, indicating a percentage increase of 17% from 2008 to 2019. In addition to the economic burden imposed by chronic conditions, there is also a quality-of-life burden; patients face treatment concerns, life-style changes, and the possibility of physical constraints. Taken holistically, this information indicates chronic conditions place a significant burden on society. It is thereby in society's best interest, from a personal and financial perspective, to identify more effective and efficient ways to provide treatment for people affected by chronic conditions. This is our aim.

In this thesis we address open challenges faced by healthcare professionals in treating patients with chronic conditions, in particular ocular diseases, and provide potential solutions to them. In this effort, we make use of multiple key glaucoma clinical trials data sets and leverage the clinical expertise of a glaucoma expert in all research presented. Ocular diseases were selected as the application focus area because I am a member of a research team that includes an ophthalmologist specializing in glaucoma; a bio-statistician specializing in ocular diseases; and two engineering faculty with expertise related to glaucoma as well as engineering methods. Additionally I had access to 3 of the most important landmark clinical trials in glaucoma, and access to some of the leading researchers in glaucoma at other institutions. The sections to follow provide a

chapter-by-chapter summary of each thesis chapter.

1.2 Chapter 2: Predicting Rapid Progression Phases in Glaucoma Using a Soft Voting Ensemble Classifier Exploiting Kalman Filtering

Our first problem addresses the problem of identifying patients who have glaucoma who are most likely to be experience rapid worsening/progression of the disease. We develop a supervised machine learning model tailored to the needs of this healthcare problem. Data from two randomized clinical trials (the Advanced Glaucoma Intervention Study (AGIS) and the Collaborative Initial Glaucoma Treatment Study (CIGTS)) are used for this purpose.

A patient, diagnosed with a chronic condition, may transition between phases of remission, stability, intermediate progression, and rapid progression. In cases important to primary open angle glaucoma (POAG), the application focus, the patient may fluctuate between phases of rapid progression (RP) and non-rapid progression (Non-RP). Predicting the timing of rapid progression before it occurs has significant value as it can inform disease management decisions and subsequently prevent adverse outcomes. ***As such, the main research contribution of this chapter is to develop a method for proactively predicting future instances of rapid progression.***

Existing methods fail to dynamically adapt to a constantly changing disease phases and/or assume the patients' medical tests (e.g., optical coherence tomography, standard automated perimetry etc.) are accurate estimates of the patient's condition. When this is not the case, as in the case of POAG, it is often hard to accurately predict the future classification of the patient's disease phase.

When the measurement noise is small, even a single measurement can be reliable for decision-making. On the other hand, when measurement noise is moderate to large, a clinician may interpret noise as a sign of progression. For mild, moderate, and even advanced glaucoma [30] shows the standard deviation of MD of patients of European or African descent ranges between 1.2 – 2dB for patients with MD values in the range of -5 to -25 db. This implies that if a patient's true MD is -8, using a 2 standard deviations confidence interval, their observed value could range between -4, indicating slightly abnormal eyesight, and -12, indicating moderately severe visual impairment. Implying competing conclusions can be drawn by a clinician based on a MD reading. We focus on this setting

using the term “moderate to severe residual error” to describe applications in which noise is a key challenge in predicting the current and future values of measurements of interest as well as the patient’s current disease phase.

To solve this problem, we propose the joint use of statistical learning and stochastic systems theory methods. More precisely, we seek to integrate a soft voting ensemble classifier with an Interacting Multiple Model (IMM) filter. The purpose of the IMM Kalman Filter is two-fold. First, it is used to adapt dynamically to the patient’s disease phase. And second, to reduce the uncertainty associated with highly variable patient medical tests; the second point is a critical step in the context of glaucoma, as various patient test measurements (e.g., standard automated perimetry and functions thereof) are characterized by moderate to severe residual variability [56]. In general terms, we outline our framework as follows: (1) identify the important longitudinal data with moderate to severe residual variability. The identification of such features is typically done so by a domain expert (e.g., clinician, research staff, engineer, etc.). All other patient data that does not fit this criterion is held out. (2) For the longitudinal data identified in step 1, process it using a measurement error reduction method such as a Kalman filter. (3) Combine the processed data with the data that is initially held out in step 1, and (4) using the combined data build a supervised prediction model to identify future patient disease phases.

We applied our framework to predicting whether a patient, diagnosed with POAG, will be in a phase of RP or Non-RP within the next two (or three) years from their current visit. We found the two-year disease phase prediction performance (AUC) of our models increased by approximately 7% (0.752 to 0.819) when the filtered results of the IMM Filter were incorporated as additional covariates. These results suggest the combination of filters and statistical learning methods in clinical health domains have significant benefits.

1.3 Chapter 3: Reinforcement Learning Methods for Constructing Personalized Monitoring Schedules for Patients with Chronic Conditions: Application to Glaucoma

Our second problem addresses the concept of monitoring patient’s overtime; determining monitoring intervals between patient appointments that are neither too long nor too short. From the prior chapter, one’s concern for fast progressors requires a method that can predict when and how much a patient may deteriorate over a given time interval to give a

clinician an idea of when the patient should be seen for their next follow-up appointment. The method presented in this chapter provides a solution to this issue.

Determining the time to next (TNT) follow-up schedule is a challenging task as it should strike a balance between detecting the disease progression in a timely manner while avoiding unnecessary treatment for heterogeneous patients. If the time to next schedule is too short, the patient may receive unnecessary treatment; if the time to next schedule is too long, the patient may be at risk of undetected disease progression. There are clear tradeoffs between these two extremes. This supports the need for personalized patient follow-up scheduling, which depends on the medical need and state of the patient. In this chapter, we present a reinforcement learning (RL) methodology for determining personalized monitoring schedules for patients with ocular hypertension. ***The main research contribution of this chapter is to develop a framework for constructing follow-up scheduling policies using the available patient-specific data from electronic health record and providing personalized schedules that are superior to fixed interval scheduling policies (e.g., 1-year and 2-year fixed intervals). Our proposed scheduling policies have three underlying goals: (i) maximizing time between follow-up visits, (ii) maximizing the scheduling efficiency, which is the percentage of scheduled visits indicating disease progression may be near, and (iii) minimizing time to detect disease progression.***

We ultimately employ RL to build a TNT recommendation model that is personalized for each patient, can dynamically adapt to a patient changing medical state, can be used starting from the patient's initial visit, and provides an integrated approach. To the best of our knowledge, we are the first to tackle this problem in a manner that meets these four conditions.

While the application focus of this work is on ocular hypertension (OHTN), our conceptual framework can be used to build TNT models for other chronic diseases as well. We formalize our RL framework using a Markov Decision Process (MDP). Where the MDP is formalized as follows:

- The state space S is a continuous state space such that at time t, the patient is in state $s_t \in S$. The state s_t contains the variables (e.g., test measurements) essential for TNT decision making.

- The action space A is a set that the clinician may only choose an action $a_t \in A$ at time t. For the general TNT problem, the action, a_t, indicates the length of time the patient must wait until the next follow-up appointment.

- The transition probability function defines the dynamics of the system. That is, function $P : S \times A \times S \to [0, 1]$ gives the probability of the next states at time $t + 1$, given the action a_t and the patient's state s_t.

- The reward function, $r : S \times A \to \mathbb{R}$ is a function of the current state, chosen action, and next state outcome that evaluates the immediate effect of the chosen action. The reward function $r(s_t, a_t, s_{t+1})$ is used to incentivize the best action at time t but does not provide information on the long-term effects of the action. For instance, a reward function can incentivize timing of a follow-up visit either shortly before or when disease progression events are likely to occur.

To find an optimal policy, π we use fitted q-iteration (FQ) algorithm which is an off-policy batch mode reinforcement learning. The goal of FQ is to provide an estimate of the Q-function which can be directly used to find the optimal policy. Using the TNT policy proposed we developed two polices, namely, RL policy 1 and RL policy 2. The key difference between the two is how their visit delay reward was weighted. RL policy 2 puts less weight on the visit delay reward, thereby increasing the number of patient follow-up visits. This significantly reduced the diagnostic delay and the average time between follow-up visits compared to RL policy 1.

Comparisons between the 1- and 2-year fixed interval testing scheduling policies and RL scheduling policies indicated the RL policies outperformed the fixed interval policies on all but one evaluation criteria, average time between follow-up visits. The 2-year fixed interval follow-up policy had the largest average time between follow-up visits (2 years), followed by RL policy 1 (1.55 years). RL policy 1 had the highest scheduling efficiency (34%) followed by RL policy 2 (32%). RL policy 2 had the smallest diagnostic delay (2.63 months) followed by RL policy 1 (3.89 months).

The experimental results suggest the TNT model provides better follow-up recommendations than fixed interval scheduling. Comparing the TNT visit recommendations for RL-1 and RL-2 with 1-, and 2-year fixed interval scheduling policies, showed the algorithm can detect POAG progression more efficiently (RL policies' scheduling efficiency at least 33% larger than the best fixed interval policy's scheduling efficiency) and sooner (RL policies' diagnostic delay at least 48% smaller than the best fixed interval policy's diagnostic delay). For patients who do not progress, the algorithm schedules less follow-up visits compared to those who did progress.

1.4 Chapter 4: Comparison of Alternative Criteria for the Identification of Conversion from Ocular Hypertension to Primary Open-Angle Glaucoma

Unlike the prior work in glaucoma, we turn now to patients with ocular hypertension (OHTN). It is estimated approximately 10% of patients with OHTN will at some time progress to glaucoma, a much more serious condition. Early identification and initiation of treatment for patients with OHTN can reduce vision related morbidity and the possibility of progression to glaucoma. However, determining progression from ocular hypertension to POAG (Primary Open Angle Glacoma) can be challenging due to the inherent variability of visual field tests and the need for multiple measurements over time. Machine learning approaches to automate the detection of conversion from ocular hypertension to POAG could be useful as decision-support systems as well as in several other settings, including in tele-ophthalmology and resource-limited areas with limited access to ophthalmologists.

The development of such machine learning algorithms requires sizable numbers of eyes which do and do not develop primary open angle glaucoma POAG. The Ocular Hypertension Treatment Study (OHTS) represents the largest, longest-followed inception cohort of patients with and without POAG. Yet even in the OHTS, only a fraction (3.9%, n = 127) of patients developed POAG at ten years. Furthermore, among the patients which did develop POAG by the OHTS endpoint criteria, only approximately one third of these occurred based on abnormal visual fields.

Several alternative strategies to identify glaucomatous visual field progression have been proposed since the start of the OHTS (e.g., [43, 3, 39, 49]. Some criteria may classify patients with greater confidence (greater quality) whereas others may provide a larger pool of greater conversions (greater quantity). The optimal approach, however, remains unclear. ***The main research contribution of this chapter is to compare four alternative criteria to identify conversion from ocular hypertension to POAG based on visual fields changes, paying particular attention to those which identify conversion more rapidly or identify a larger cohort of eyes with POAG, as these may be useful for algorithm development.***

A key motivating factor for this chapter is the absence of a reference standard (e.g., a gold standard criterion for accurately assessing whether a patient converted to POAG). This imposes a necessary limitation, as there is no way to assess relevant characteristics including sensitivity and specificity of each approach. As such, several POAG progression criteria were compared including two global, event-based approaches which assess dif-

ferent visual field summary statistics (the OHTS endpoint criterion and the MD criterion), a pointwise event-based approach which assesses change across visual field test points (cluster deterioration), and a pointwise trend-based approach which identifies worsening at individual test points (TD trend).

For each criteria time to conversion was compared. Cumulative incidence curves for each POAG conversion criteria were drawn (with 95% confidence intervals) based on Kaplan-Meier estimates and compared using pairwise Paired Prentice-Williams Tests. The Bonferroni correction was applied to p-values. Additionally, a multivariate Cox model was used to estimate the hazard ratios associated with age, race, and sex. Confidence intervals were built using robust standard errors as each patient had two eyes in the dataset. Whether the associations between time to conversion and age, race, and sex differed by progression type were investigated with Cox Regression models, stratified by conversion criteria and with frailty terms enabling correlations between times to conversion within the same patient. Likelihood ratio tests of the interaction terms between type of conversion and the other fixed effects were used to determine whether covariate effects differed by conversion criteria.

Results indicated race, sex, and age were statistically significant determinants in progression for all four criteria. However, there was broad discordance between the four criteria, as demonstrated by the statistically different survival curves and the limited overlap eyes that progressed by multiple criteria. This suggests that these criteria may be tailored to the type of damage under investigation, particularly in the absence of a visual fields-based reference standard. Notably, all criteria demonstrated the least concordance with the OHTS endpoint criteria, which may, at least in part, reflect changes in the means of evaluating visual fields since the commencement of the OHTS in the mid-1990s.

1.5 Chapter 5: Machine Learning Prediction of Conversion from Ocular Hypertension to Open Angle Glaucoma

As per chapter 4, one of the key strengths of machine learning approaches is their ability to automate, the otherwise manual task of detecting a patient's conversion from ocular hypertension to POAG (Primary Open-Angle Glaucoma). At present, there exist several alternative criteria to identify POAG conversion [43, 3, 39, 49], but notably no gold standard. Trade-offs amongst criteria are often made to balance a criterion's sensitivity

(measures the proportion of converting patients that are correctly identified as converting) and specificity (measures the proportion of non-converting patients that are correctly identified as non-converting). Since correctly identifying converting patients often takes precedence over incorrectly labeling a non-converting patient as converting, a criterion that has a high sensitivity is often preferred.

As such, we elucidate and study several conversion criteria, and develop an ML method to proactively identify POAG converting patients. The conversion criteria were as follows: (A) the OHTS Endpoint – Committee - this was the approach used in OHTS for progression to POAG by visual fields; (B) Decline in Mean Deviation - conversion from OHTN to POAG defined as 2 consecutive MD values at least 3 DB below baseline. Baseline MD was defined as the average of the patient's two initial MD values prior to enrollment in the clinical trial [49]; (C) Decline in Pointwise Total Deviation- this is a trend-based assessment, first described by Kummet and colleagues [43], involving performing pointwise linear regression on each of the 52 visual field test locations in the total deviation plot; and (D) Deterioration of Points on Total Deviation Clusters – this is a clustered based approach developed in [39].

The main research contribution of this chapter is to determine the predictive performance of a collective conversion criterion (e.g., conversion criteria combining all 4 criterion). In particular, using predictive classification algorithms to determine if a patient will progress to PAOG, as defined by one of the four criteria, anytime within two years from their sixth visit, signaling POAG conversion. Conversions events were not identified sooner (i.e., before the patient's 6th visit) because the Kalman Filter, a key data preprocessing step, required the first six patient visits for model training and calibration.

Several supervised learning classification models were assessed: Logistic Regression, Random Forest, Gradient Boosted Decision Tree, and Neural Network. Results indicated the random forest classifier performed best. The classification models that included the Kalman filtered data had slightly better performance than the models that did not. For the Random Forest that did not include KF data the group 5-fold cross validation ROC AUC performance was mean ± SD of 0.81 ± 0.01. The testing performance across the performance measures were balance accuracy of 0.79, sensitivity of 0.75, specificity of 0.82, accuracy of 0.81, positive predicted value of 0.36, negative predicted value 0.96, and ROC AUC of 0.86. For the Random Forest that included the KF data the group 5-fold cross validation ROC AUC performance was mean ± SD of 0.81 ± 0.01. The testing performance across the performance measures were balance accuracy of 0.80, sensitivity of 0.77, specificity of 0.83, accuracy of 0.82, positive predicted value of 0.38, negative predicted value 0.96, and ROC AUC of 0.86.

1.6 Chapter 6: Summary and Conclusions

In this chapter, we summarize the key contributions of the thesis and discuss several notable areas of future research.

We note, the research presented in this thesis can be categorized into two groups: the optimization of monitoring frequency by which patients are seen, and the optimization of disease identification methods. We illustrate in Figure 1.1 how the thesis chapters fit into this categorization.

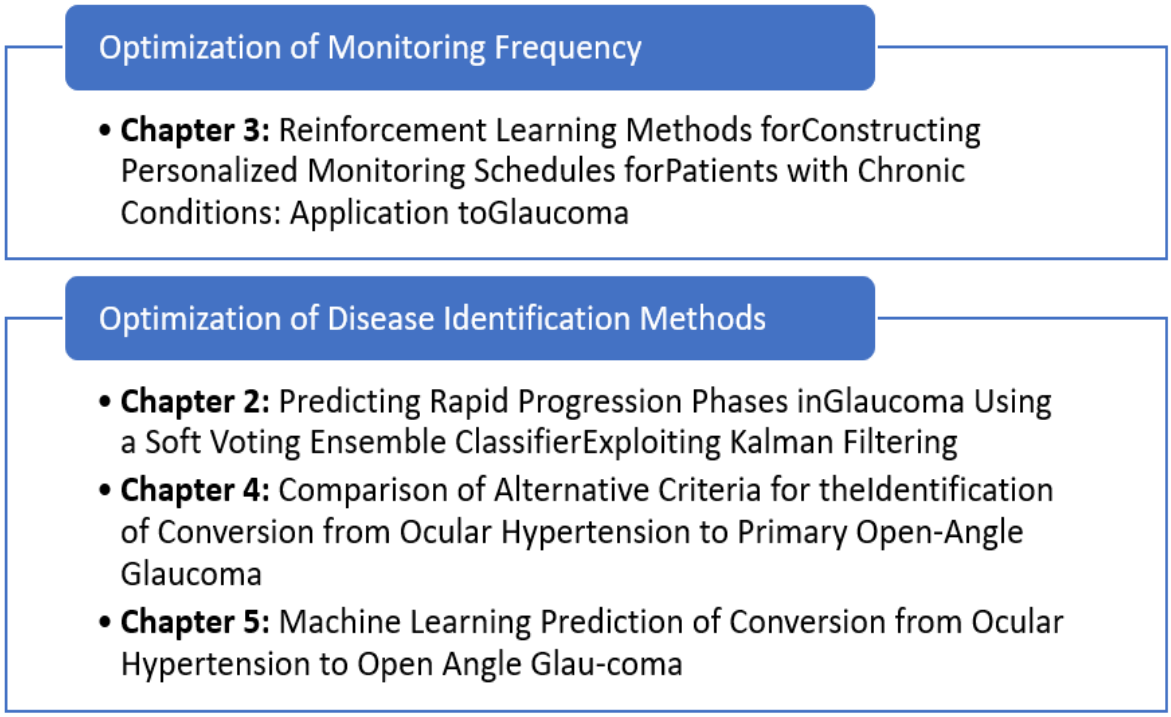


Figure 1.1: Categorization of Thesis Chapters

The works presented in chapters 2-5 make contributions in several key areas relevant to the care of patients with chronic conditions. They can be outlined as follows: (1) the proactive identification of periods where a patients' disease will progress at a rapid/fast rate and the denoising of clinical visual field measurements with high measurement error; (2) the dynamic and personalized recommendation of patients' follow-up appointment schedules; (3) the evaluation of criteria for determining patient conversion from ocular hypertension to primary open angle glaucoma; and (4) the development of a new criterion for identifying POAG progression and the evaluation of its ability to predict future progression/conversion events. Each of the four contributions advance the care of for patients diagnosed with ocular illnesses and more broadly chronic conditions.

CHAPTER 2

Predicting Rapid Progression Phases in Glaucoma Using a Soft Voting Ensemble Classifier Exploiting Kalman Filtering

2.1 Introduction

Chronic conditions affect 60% of the adult population [1]. Typically, they require monitoring and treatment over time. A patient, diagnosed with a chronic condition, may transition between phases of remission, stability, intermediate progression, and rapid progression. In cases important to primary open angle glaucoma (POAG), the application focus of this work, the patient may fluctuate between phases of rapid progression (RP) and non-rapid progression (Non-RP). Predicting the timing of rapid progression before it occurs can inform disease management decisions and subsequently prevent adverse outcomes.

Precision medicine encourages us to use all available population and patient information to generate personalized classifications of patients' disease phases and identify how these disease phases may change over time. Hence, the development of statistical models that can predict the future disease phase of the patient is key. There has been limited research in this area in the context of POAG. Existing methods fail to dynamically adapt to a constantly changing disease phases and/or assume the patients' medical test measures (e.g. optical coherence tomography, standard automated perimetry etc.) are accurate estimates of the patient's condition. When this is not the case, as in the case of POAG, it is often hard to accurately predict the future classification of the patient's disease phase.

When the measurement noise is small, even a single measurement is fairly reliable for decision-making. On the other hand, when measurement noise is moderate to large, a clinician may interpret noise as a sign of progression. For mild, moderate, and even advanced glaucoma, [30] shows the standard deviation of MD of patients of European or

African descent ranges between 1.2 – 2dB for patients with MD values in the range of -5 to -25 db. This implies that if a patient's true MD is -8, using a 2 standard deviations confidence interval, their observed value could range between -4, indicating slightly abnormal eyesight, and -12, indicating moderately severe visual impairment.

In this chapter, we focus on this setting using the term "moderate to severe residual error" to describe applications in which noise is a key challenge in predicting the current and future values of measurements of interest as well as the patient's current disease phase. One potential solution to this problem is augmenting predictive methods with filtering models that not only estimate the true values of the patients' tests results under moderate to severe residual variance/error conditions, but also adapt to changing chronic disease conditions.

We propose the joint use of statistical learning and stochastic systems theory methods. More precisely, we seek to integrate a soft voting ensemble classifier with an Interacting Multiple Model (IMM) filter. The purpose of the IMM Kalman Filter is two-fold. First, it is used to adapt dynamically to the patient's disease phase. And second, to reduce the uncertainty associated with highly variable patient medical tests; the second point is a critical step in the context of glaucoma, as various patient test measurements (e.g., standard automated perimetry and functions thereof) are characterized by moderate to severe residual variability [56].

We applied our framework to predicting whether a patient, diagnosed with POAG, will be in a phase of RP or Non-RP within the next two (or three) years. We found the two-year disease phase prediction performance (AUC) of our models increased by approximately 7% (0.752 to 0.819) when the filtered results of the IMM Filter were incorporated as additional covariates. These results suggest the combination of filters and statistical learning methods in clinical health domains have significant benefits.

The organization of our chapter is as follows. Section 2.2 reviews related literature. Section 2.3 describes the modeling framework. Section 2.4 presents the application of this framework to POAG. Lastly, section 2.5 concludes the paper and discusses future research.

2.2 Literature Review

Patient disease trajectories are uncertain. It is not uncommon for two patients with the same disease prognosis and treatment protocol to transition into two different disease phases, with one patient transitioning into a favorable phase (Non-RP in our application) and the other transitioning into a non-favorable phase (RP in our application). For the

Table 2.1: Definition of Phases (Corbin, 1991)

Phase	Definition
Pre-trajectory	Before the illness course begins, the preventive phase, no signs or symptoms present
Trajectory Onset	Signs and symptoms are present, includes diagnostic period
Crisis	Life-threatening situation requiring emergency/ critical care
Acute	Active illness or complications that require hospitalization for management
Stable	Illness course/symptoms controlled by regimen
Unstable	Illness course/symptoms not controlled by regimen but not requiring hospitalization
Dying	Immediate weeks, days, hours preceding death

sake of generality of perspective, we discuss a more general framework. [13] defines phases as the different changes in status that a chronic condition can undergo over the course of the disease. The scope outlined in [13] is much broader than what we focus on in our case study, but the methods we develop can be extended to a more elaborate framework incorporating multiple disease phases. The author lists eight phases a patient's diseases trajectory can take: pre-trajectory, trajectory onset, crisis, acute, stable, unstable, downward, and dying. Table 1 provides an overview of the 8 phases.

A patient's health seldom deteriorates or improves in a perfectly predictable manner. The disease phases, excluding pre-trajectory, can be viewed as levels of disease progression. For the case study, discussed later in this chapter, we focus on two of these diseases' phases: RP (downward) and Non-RP (stable).

The work presented in this chapter can be described as classification, identification, or detection of disease progression. Past research in progression identification [63] presents a comparison study on the use of various machine learning models to predict the severity of chronic kidney disease progression. [42] examines the ability of a random forest classification model to predict calciphylaxis risk successfully for patients with chronic kidney disease. And [15] develops a convolution neural network to identify and determine age-related macular degeneration disease severity.

In the domain of early identification of chronic diseases [2] develops a clinical decision support tool to help with the automatic identification of chronic obstructive pulmonary disease diagnosis. More recently, [68] develops an adaptive k-means clustering algorithm to predict patient health trends and [64] develops a risk prediction model, XGboost classifier, of incident essential hypertension within the following year. The model developed in [64] served two purposes – feature selection and risk prediction. We expect chronic disease

detection and progression identification to continue to be a fundamental research aim in these areas.

The research conducted in these two areas have similarities to this chapter; each use patient demographics and disease indicators for prediction purposes. However, the disease indicators outlined in their work (e.g., weight, medical laboratory tests) have low residual variability, making their true values easier to infer. They do not address situations when disease indicators contain high residual variability. This chapter relatively addresses this research gap.

Additionally, while it does not attempt to perform classification, identification, or detection of disease progression, there is existing work within the ophthalmology research domain in the use of Kalman filtering for forecasting the future outcomes of patients, e.g., for normal tension glaucoma [23], and forecasting visual field and intraocular pressure trajectory in patients with ocular hypertension [22]. While these works highlight the measurement prediction error of filtering methods in disease modeling, they do not identify periods where patients' disease progresses at a rapid rate as we do here.

2.3 Modeling Framework

To assess the condition of patients with chronic diseases, the longitudinal series of information obtained on successive patient visits must be modeled and analyzed. The focus here being longitudinal information that possesses moderate to severe residual variability, defined as the variability of the difference between a variable's true value and measured value. This variability can be attributed to a multitude of factors (e.g. variation in the measurement instrument, variation in the patient using the instrument, test-retest variability, etc.).

We outline our framework in general terms as follows: (1) identify the important longitudinal data with moderate to severe residual variability. The identification of such features is typically done so by a domain expert (e.g. clinician, research staff, engineer, etc.). All other patient data that does not fit this criteria is held out. (2) For the longitudinal data identified in step 1, process it using a measurement error reduction method such as a Kalman filter. (3) Combine the processed data with the data that is initially held out in step 1, and (4) using the combined data build a supervised prediction model to identify future patient disease phases. In the following two sections we discuss steps (2) and (4) in detail. The presentation will aim for generality where possible; however, as this chapter is application focused,for ease in reading, we will discuss them in the context of denoising longitudinal healthcare data and classifying patient diseases phases.

2.3.1 IMM Filter Formulation

For clarity, we will begin with a motivating example. Consider a patient with a chronic disease who, as evident in their past longitudinal test results, has experienced periods of no or slow progression, and rapid progression (Non-RP and RP respectively). We have reason to believe their test results contain measurement noise, as is well known to be the case for POAG measurements, and seek to filter the readings. Kalman filters have been shown to reduce measurement noise, but because the patient's disease trajectory is non-stationary (e.g. the patient moves in and out of several disease phases) a single Kalman filter (KF) will have trouble modeling both modes. For example a zero-order KF (assumes the patient's disease is not progressing) will likely under-estimate the change when the patient's disease is RP. A first-order filter (that assumes the patient's disease is progressing at a stable rate) or second-order filter (that assumes the patient's disease is progressing at a constantly increasing rate) will likely over-estimate changes when the patient's disease is Non-RP. We would intuitively prefer a meta-filter that dynamically chooses the most appropriate filter(s) to use based on the current disease phase of the patient; if the patient is not progressing we want the denoised estimate to be based mainly on the zero-order filter and similarly if the patient is progressing we want the denoised estimate to be based on the a first and/or second order filter. We employ the IMM Filter to accomplish this task.

The IMM Filter is described in [7, 6, 44]. The idea behind the IMM filter, as briefly discussed in the motivating example, is to have one Kalman filter for each mode of the system. As the system changes from one mode to another, in discrete time, the IMM filter dynamically adapts by primarily using the state estimates associated with the most probable Kalman filters; the more likely filters modify the state estimates of the less likely filters, and these collection of estimated states are blended to form a more accurate state estimate [44].

To build an IMM filter, mode probabilities ($\boldsymbol{\mu}$) and mode transition probabilities ($\boldsymbol{M}$) need to be estimated. Mode probabilities describe the belief in each type of system behavior, (e.g., probability the patient's disease is stable or rapidly progressing). Mode transition probabilities describe how the system will transition from one mode to the next (e.g. the probability the patient, who is currently stable (Non-RP), transitions to RP). For example, with a two disease phase model (Non-RP phase and an RP phase), the mode transition probability matrix gives the probability of transitioning from a non-rapid progressing phase to a rapidly progressing phase and vice-versa. Both $\boldsymbol{\mu}$ and $\boldsymbol{M}$ can either be estimated from data or obtained using domain knowledge.

For generality, we present the mathematical framework for an IMM filter composed of n KFs. However, in practice the number of KFs used is dictated by the number of unique behaviors of the system. In the two disease phase example above and in the IMM filter developed in our case study, we employ one KF to model the Non-RP phase disease dynamics and another to model the the RP phase disease dynamics. We leave to the reader the mathematical underpinnings of the KF (see [67] for a treatment).

The IMM filter performs a prediction and subsequent update in the usual manner of a KF. However, because the IMM filter uses a bank of KFs, the underlying math is different. Below we outline the initialization, update, and prediction steps used by the IMM filter.

Initialization: We denote the initial mode probabilities by the vector $\boldsymbol{\mu_0}$ and the mode transition probability matrix by $\boldsymbol{M}$.

$$\boldsymbol{\mu_0} = \begin{pmatrix} \mu_1 & \mu_2 & \dots & \mu_n \end{pmatrix} \tag{2.1}$$

$$\boldsymbol{M} = \begin{pmatrix} \mu_{11} & \dots & \mu_{1n} \\ \vdots & \ddots & \vdots \\ \mu_{n1} & \dots & \mu_{nn} \end{pmatrix} \tag{2.2}$$

The IMM filter and each of the n KFs in the filter bank are initialized using their initial state, $\boldsymbol{x_i}$, initial covariance, $\boldsymbol{P_i}$, initial mode probability, $\boldsymbol{\mu}$, and mode transition probability matrix, $\boldsymbol{M}$. The index, i, refers to the i^{th} KF in the IMM filter bank. Since we have n filters, the dimensions of $\boldsymbol{\mu}$ and $\boldsymbol{M}$ are $n \times 1$ and $n \times n$ respectively. The dimensions of $\boldsymbol{x_i}$ and $\boldsymbol{P_i}$ depend on the number of state variables. In our POAG model presented in the case study, the state is 3-dimensional; composed of the patient's current measurement, the rate of change of the patient's current measurement (velocity), and the rate at which the patient's current measurement is changing (acceleration). In this 3-dimension case the dimension of $\boldsymbol{x_i}$ and $\boldsymbol{P_i}$ are 3×1 and 3×3 respectively. The initial state is taken as the starting position (e.g. a patient's measurement, measurement velocity, and measurement acceleration at the start of treatment). The initial covariance matrix captures the joint variability of the state variables. In our example, the initial covariance matrix captures the variance of patient's current measurement, the variance of the measurement's velocity, the variance of the measurement's acceleration, and the covariance of the three state variables.

The initial state, $\boldsymbol{x_i}$ and covariance, $\boldsymbol{P_i}$ can be more precisely denoted as $\boldsymbol{x_{i,0|0}}$ and $\boldsymbol{P_{i,0|0}}$; where the index $i, 0|0$ denotes the state estimate at time 0 given information up to

and including time 0 for the i^{th} Kalman filter. The following initialization relationships are defined as follows:

$$\bar{c}_0 = \mu_0 M \tag{2.3}$$

$$\omega_{0,ij} := \mu_{(i,0)} \times M_{ij} \quad \forall\, i, j = 1, \ldots, n \tag{2.4}$$

$$\omega_0 = \|\omega_0\|. \tag{2.5}$$

The row vector $\bar{\boldsymbol{c}}_{\mathbf{0}}$ denotes the initial mode probabilities after accounting for the initial probabilities of system transitions. For example, in the two disease phase example, $\bar{\boldsymbol{c}}_{(1\times 2)} = \mu_{(1\times 2)} M_{(2\times 2)}$, is a 2-dimensional row vector capturing the probability a patient is non-rapid progressing in the next time period given they could be non-rapid progressing or rapid progressing in the current time period, and the probability a patient is rapid progressing in the next time period given they could be non-rapid progressing or rapid progressing in the current time period. ω_0 denotes the initial unnormalized mixing probabilities, and $\|\boldsymbol{\omega_0}\|$ denotes the normalized (e.g. rows sum to one) initial mixing probabilities. The mixing probabilities represent how filter estimated values (e.g. states and covariance matrices) should be weighted so as to incorporate estimates from the probable and improbable filters. The intention is for all filters to obtain improved state and covariance estimates by appropriately mixing them [44]. Next, the prediction and update steps are performed in a recursive manner. Each step is described below.

Predict: Each of the n filters' states and covariance matrices are weighted using the mixing probabilities to form the mixed state and mixed covariance as follows:

$$x^m_{i,t-1|t-1} = \sum_{i=1}^{n} \omega_{ij,t-1} \cdot x_{i,t-1|t-1} \quad \forall\, j = 1, \ldots, n \tag{2.6}$$

$$\begin{aligned} P^m_{i,t-1|t-1} = & \sum_{j=1}^{n} \omega_{ij,t-1} \cdot [(x_{j,t-1|t-1} - x^m_{i,t-1|t-1}) \\ & (x_{j,t-1|t-1} - x^m_{i,t-1|t-1})^T + P_{i,t-1|t-1}] \\ & \forall\, j = 1, \ldots, n. \end{aligned} \tag{2.7}$$

Using the mixed states, $\boldsymbol{x^m_{i,t-1|t-1}}$, and mixed covariance matrices, $\boldsymbol{P^m_{i,t-1|t-1}}$, a prediction step is performed by each KF. This results in n predicted mixed states, $\boldsymbol{x^m_{i,t|t-1}}$, and covariance matrices, $\boldsymbol{P^m_{i,t|t-1}}$, where the index, $t|t-1$, indicates the prior state and

prior covariance estimates at time t given observations up to and including time $t-1$. The IMM *predicted* state and *predicted* covariance matrix at time t are calculated as the expected value (using the mode probabilities) of the n KFs' mixed predicted states and mixed predicted covariances as follows:

$$x_{t|t-1} = \sum_{i=1}^{n} \mu_{i,t-1} \cdot x^m_{i,t|t-1} \quad \forall\, i = 1, \ldots, n \tag{2.8}$$

$$\begin{aligned} P_{t|t-1} = \sum_{i=1}^{n} \mu_{i,t-1} \cdot [(x_{t|t-1} - x^m_{i,t|t-1})(x_{t|t-1} - x^m_{i,t|t-1})^T \\ + P^m_{t|t-1}] \quad \forall\, i = 1, \ldots, n. \end{aligned} \tag{2.9}$$

In the case of the two disease phase example presented in the case study, the prediction step consists of both the non-rapid progressing KF and rapid progressing KF performing a prediction step. This results in two predicted states and two predicted covariances matrices. Using the mixing probabilities the predicted states and predicted covariances matrices are mixed with one another to form more probable state and covariance estimates. Next, using the mode probabilities (the probabilities the patient disease is non–rapid progressing or rapid progressing), the expected value of two mixed states and covariances are computed. This results in the IMM filter's predicted estimate of the patient's state (e.g. patient's predicted measurement, patient's predicted measurement velocity, and patient's predicted measurement acceleration) at the next time index.

Update: Each of the n Kalman filters in the IMM filter bank computes the likelihood, $L_{i,t}$ for $i \in [1, \ldots, n]$, for $t \in [1, \ldots, T]$, of observing measurement $\boldsymbol{z}$ (e.g. a patient's observed measurement) at discrete time t, and subsequently performs an update step using measurement $\boldsymbol{z}$. The likelihood of measurement $\boldsymbol{z}$ is calculated as the probability density of measurement $\boldsymbol{z}$ under a multivariate normal with mean $z - Hx$ and covariance $HPH^T + R$. Where H is the measurement matrix of the Kalman filters and R is the covariance matrix of measurement noise.

Using the vector of likelihoods, $\boldsymbol{L_t}$, the mode probabilities, $\bar{c}$, and the mixing probabilities, ω, are updated as follows:

$$\mu_t = \|L_t \odot \bar{c}_{t-1}\| \tag{2.10}$$

$$\bar{c}_t = \mu_t M \tag{2.11}$$

$$\omega_{ij,t} := \mu_{i,t} \times M_{ij} \quad \forall\, i,j = 1,\ldots,n \tag{2.12}$$

$$\omega_t = \|\omega_t\|, \tag{2.13}$$

where $\boldsymbol{\mu_t} = \|\boldsymbol{L_t} \odot \bar{\boldsymbol{c}}_{\boldsymbol{t}}\|$, denotes a normalized element-wise product between the vectors $\boldsymbol{L_t}$ and $\bar{\boldsymbol{c}}_{\boldsymbol{t}}$. The IMM *updated* state and *updated* covariance matrix at time t is calculated as the expected value (using the updated mode probabilities) of the n KFs' mixed updated states and mixed updated covariances as follows:

$$x_{t|t} = \sum_{i=1}^{n} \mu_{i,t} \cdot x^m_{i,t|t} \quad \forall\, i = 1,\ldots,n \tag{2.14}$$

$$\begin{aligned} P_{t|t} = \sum_{i=1}^{n} \mu_{i,t} \cdot [&(x^m_{i,t|t} - x_{t|t})(x^m_{i,t|t} - x_{t|t})^T \\ &+ P^m_{i,t|t}] \quad \forall\, i = 1,\ldots,n. \end{aligned} \tag{2.15}$$

The index, $t|t$, indicates the posterior state and covariance estimates at time t given all observations up to and including t have been used. Being recursive in nature, as time progresses and measurements are taken, the IMM recursively predicts and updates.

At each time instant, the denoised state estimates resulting from the update steps are used to approximate the true state of the system. In the case study, this implies the IMM filter's denoised measurements from the patient's initial tests to their current tests will be used to approximate the true course of their disease. Not their observed measurements during this period.

2.3.2 Disease Phase Identification

Over time patients can transition from one disease phase to another; certain disease phases are more favorable than others (e.g., a phase of stability or slow progression versus a phase of rapid progression). It is the physician's goal to maximize the patient's quality of life by controlling and/or maintaining a favorable disease phase. Therefore, the ability to successfully predict a patient's future disease phase has value, as it enables the physician to take preventive actions before adverse events occur, the worst of which is

blindness. For example, if the patient's predicted future disease phase is a less favorable phase, then it is prudent for the clinician to alter the patient's treatment in a way that minimizes the chance of the patient transitioning into the predicted disease phase. In order to accomplish this, we propose the development of a disease phase prediction model.

Formally, the model is a supervised learning prediction model. Define the set of disease phases a patient can be in as $d_i \in \boldsymbol{D}$ for $i = 1, \ldots, m$ where m denotes the total number of patient disease phases. The purpose of the supervised learning prediction model is to create a mapping function, $f(x)$, from the n dimensional feature vector, $\boldsymbol{x} \in \mathbb{R}^n$, to the probability mass function, $p_{i,t+\Delta_t} \in \boldsymbol{P}$, where $p_{i,t+\Delta_t}$ denotes the probability of being in disease phase, d_i at time $t + \Delta_t$. The n dimensional feature vector, $\boldsymbol{x}$, is composed of the IMM filter state estimates, and relevant patient information (e.g. demographic information, patient test results with low residual variability, engineered features, etc.); the feature vector should give a clear indication of the current status of the patient's chronic disease. The most likely disease phase, d_i, at time period $t + \Delta_t$, is computed as,

$$\underset{p_{i,t} \in P}{\arg\max} \{p_{i,t+\Delta_t}\}. \tag{2.16}$$

Time period $t + \Delta_t$, is taken as a future time period, where Δ_t denotes the prediction window length. The prediction window length, Δ_t, should determined based on clinical use case. For RP phase identification described in the case study we use a Δ_t equal to 2 and 3 years. Hence, the prediction d_i represents the disease phase the patient is most likely to be in at time $t + \Delta_t$, not the disease phase the patient is currently in. We leave the description of the classification model abstract, as any supervised learning model that outputs class probabilities is sufficient. For the glaucoma RP identification problem addressed in the case study, a soft-voting ensemble model is developed. The soft-voting ensemble model is discussed in section 2.4.3.6. In the next section we apply our methodology to the case of POAG rapid progression phase identification.

2.4 Open Angle Glaucoma Case Study

Glaucoma is a chronic disease characterized by gradual vision loss. This makes it difficult for clinicians to provide treatment for patients early on, as there may be a delay between the initial onset of glaucoma progression and patients experiencing symptoms of glaucoma progression. There are several types of glaucoma: open-angle glaucoma, angle-closure glaucoma, normal-tension glaucoma, pigmentary glaucoma, and glaucoma

in children. This proposed case study focuses on primary open-angle glaucoma (POAG), the most common form of the disease, estimated to affect more than 3.3 million people in the US in 2020 [31].

One of the main methods for detecting functional visual field loss (e.g., vision deterioration) is standard automated perimetry (SAP). SAP is a subjective psycho-physiological test for measuring retinal sensitivity at different locations in the visual field. The test in conducted by presenting patients with localized light stimuli of varying intensities and instructing them to press a button when each stimulus is perceived [65]. The result of SAP is a map of local retinal sensitives. The local retinal sensitives can be summarized into a single number by computing their mean deviation (MD). MD is the average, age and race adjusted, deviation of a patients retinal sensitivities from a patient with normal vision. MD values typically range between 2 dB to -30 dB. Subjects, who are able to see dimmer stimuli than others of similar age and race will have positive values for their MD, while subjects who require brighter stimuli will have negative MD values. The lower a subject's MD value, the worse their vision.

Using a patient's SAP measurements, their MD over time can be obtained. If their MD over time has an ordinary least squares slope of ≤ -1 DB/year we characterize this patient as progressing at rapid rate [35, 41]. This characterization is important because although POAG is often a slow, progressive chronic condition, at any point in time a patient with POAG may experience RP. If RP is not detected early and treated promptly, it can lead to irreversible vision loss and blindness.

As such, the purpose of this case study is to develop a model to *proactively identify instances in time were patients will experience RP, before observing their actual MD measurements*. We examine predictions made 2 years in advance as our base case. Then we revisit the problem for the case where the time window is the next 3 years from the current visit. This enables the clinical care team to act before more aggressive (likely riskier and costlier) treatment options are warranted.

This case study focuses on the use of what is termed in the ophthalmology community as trend analysis. In trend analysis, a patient's longitudinal measurements, MD in this case, are evaluated with linear or other forms of regression analysis to estimate rates of change and statistical significance [62]. While this approach is open for debate, as a patient's disease trajectory may not be linear, we use it because it is a well accepted approach in both clinical and research communities. There have been numerous works that either describe the behavior of MD measurements over time and/or predict future values [41, 30, 59], however this proposed case study differs from these works in two ways. First, we focus on proactively *classifying* patients with respect to their predicted

behavior over a future interval of time. Second, our method allows us to predict if patients will experience RP at each time a test is taken. That is to say, it is a dynamic, recursive method that does not require rebuilding the model at each time instance. This is more clinically relevant than identifying instances of disease progression after they have already occurred [62]. Our focus is an important one, because patients that will exhibit RP need intervention to avoid irreversible vision loss.

We underscore that SAP measurements or functions thereof (e.g., MD, pattern standard deviation, etc.) are only estimates of their true value [56, 57]. Clinicians consider MD values to be one of the most important sources of information when managing glaucoma care. We describe the MD tests as having high residual (true value - measured value) variability because there is sufficient measurement noise that one must consider in order to avoid misinterpreting the longitudinal test results of a patient. We emphasize that the variability does not render the data useless; rather, mechanisms to estimate the true values should be employed. Clinicians are used to making important decisions using this data, which they know has uncertainty. Therefore, the focus of this section is to present a mechanism to accomplish this.

2.4.1 Data

The data used in this work comes from two landmark randomized, longitudinal clinical trials: 1) Advanced Glaucoma Intervention Study (AGIS), and 2) the Collaborative Initial Glaucoma Treatment Study (CIGTS). AGIS has 591 patients with advanced glaucoma followed for up to 11 years. CIGTS has 607 patients with mild to moderate glaucoma followed for up to 10 years. Both provide demographic information (e.g. age, sex, and race) and longitudinal test results with three key continuous-valued measurements: Intraocular Pressure (IOP) - measures the pressure in the eye; MD - measures mean age and race adjusted deviation of a patient's retinal sensitivities; and pattern standard deviation (PSD) - measures the uniformity of a patient's retinal sensitivities. All values were recorded on 6 month intervals. A detailed description of the AGIS and CIGTS clinical trials can be found in [41, 59].

Both trials required participants to have a diagnosis of glaucoma in at least 1 eye, with elevated IOP at trial entry. For this study, we included only patients who were randomized to receive medical therapy or argon laser trabeculoplasty. Patients who had been randomized to trabeculectomy are excluded because incisional surgery can dramatically affect IOP and disease progression dynamics, making it beyond our scope. Furthermore, during follow-up, trial participants who later required incisional surgery are censored at

the time they underwent trabeculectomy. Additionally, because our methods used the first eight years of each patient's data, a patient's eye was excluded if it had fewer than eight years of data (i.e., eligibility required atleast 16 IOP measurements and 16 visual field tests using a Humphrey Field Analyzer). If both eyes of a participant met the eligibility criteria, we randomly selected 1 of the 2 eyes for inclusion in our analyses. These resulted in 295 patients meeting our eligibility requirements. Table 2.2 provides a summary of the study population.

We place emphasis on the fact the physiological mechanisms driving glaucoma are only partially understood. A foundational understanding of disease phases is not established for glaucoma in the way we wish it were. The only longitudinal information on health state comes through the MD, IOP, and PSD measurements.

Table 2.2: Summary of Study Population

	Black	White	Asian	**Total**
Total Number of Observations	1569	1347	97	**3013**
Rapid Progressed (RP) Instances	368	285	28	**681**
Patients w/ 1+ Instances of RP	117	96	7	**220**
Mean (SD) number of RP instances per patient for patients w/ 1+ Instances of RP	3.1 (1.9)	2.9 (2.1)	4 (3.0)	**3.0 (2.0)**
Number of Patients	149	137	9	**295**
Sex, n (%)				
Male	5 7(40.4)	81(57.4)	3(4.2)	**141 (47.7)**
Female	92 (59.7)	56 (36.3)	6 (3.8)	**154 (52.2)**
Baseline statistics				
Mean Deviation (MD), mean ± SD	-8.4 $\pm$ 5.4	-6.5 $\pm$ 5.3	-4.4 $\pm$ 5.7	**-7.4** $\pm$ **5.5**
Pattern Standard Deviation (PSD), mean ± SD	6.9 $\pm$ 3.8	6.0 $\pm$ 3.8	3.4 $\pm$ 3.2	**6.4** $\pm$ **3.9**
Intraocular Pressure (IOP), mean ± SD	17.3 $\pm$ 3.7	17.6 $\pm$ 3.4	20.4 $\pm$ 5.4	**17.6** $\pm$ **3.7**

2.4.2 Identifying Periods of Rapid Progression

Following a common convention, we previously defined RP as an ordinary least squares MD slope of ≤ -1 DB/year. For the purpose of proactively identifying patients who will exhibit RP in the future we define *RP instances* as 2 year windows (and later we repeat

the analysis again for 3 years) over which a patient's MD slope is less than or equal to -1 DB/year. Clinicians have a good sense of what it means for a patient to be rapid progressing or not over the next 2 years. It's a summary of patient behavior over the entire interval; therefore, an ordinary least squares summary measure of slope is consistent with their sense of characterization of RP over a 2 year time interval.

We underscore that RP classification (or Non-RP) is not intended to signify that the glaucoma's phase is fixed over the 2-year time interval, it merely serves as a measure of the patient's rate of change over the 2 years. This summary measure partially obscures the full picture of what may be happening over the 2-year period, because patients are transitioning between phases of RP and Non-RP and are not fixed in either phase. However, when making a prediction we focus on which phase best represents patient behaviour over this period.

Figure 2.1 provides a visual depiction of how 2-year future RP is calculated for a patient when assessed post-hoc from their actual readings. From time T_n to T_{n+4} the target label (MD slope) is calculated from the 5 MD values using Ordinary Least Squares Regression (OLSR), where n represents the starting period of the 2-year prediction. A period consists of a 6-month time interval. Hence, the 5 periods span a time interval of 2 years and one day, nominally.

This implies each patient has multiple opportunities to exhibit 2 year rapid progression. If the OLSR slope is ≤ -1 for periods T_n to T_{n+4}, the target for period T_n is 1, indicating RP is present 2 years in the future from time n (0 elsewhere). To predict whether a patient will experience RP from periods T_n to T_{n+4}, only the data from periods T_0 to T_n is used. Hence, no data from future periods is used to determine if the patient will experience future 2-year rapid progression. This process continues for all periods $n \in \{3, \ldots, N\}$, where N denotes the last period of available data. Similarly, for the 3-year case, the OLSR slope from periods T_n to T_{n+6} is used to determine the RP target (e.g. 0 or 1). A minimum of 4 periods of data is needed for both 2 and 3 year predictions. This is because some of the covariates used as input to the Soft Voting Ensemble Classifier require at least 4 periods of data. Appendix A.1 provides a table of all model inputs.

2.4.3 Model

In the following subsections, 2.4.2.1 – 2.4.2.5, we discuss the IMM Filter and disease phase prediction model for proactive rapid progression identification.

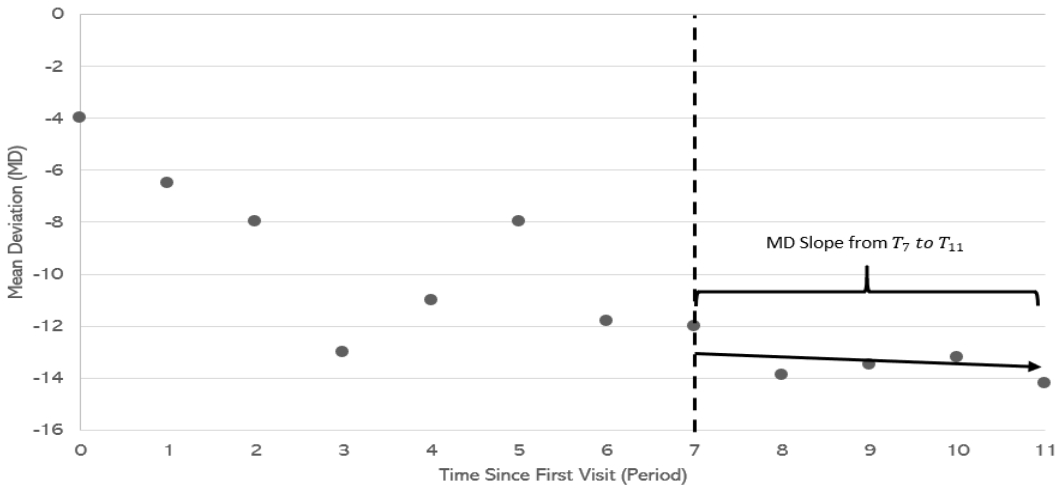


Figure 2.1: Illustration of the calculation of MD slope over a window of 2 years. MD = mean deviation.

2.4.3.1 IMM Filter

In section 2.3.1 we briefly discussed our motivation for using an IMM filter. This choice is supported by the fact we find glaucoma frequently progresses in spurts. There is a controversy in the field concerning whether the progressions are slowly changing or whether it's periods of inactivity followed by periods of rapid progression. However, due to the flexibility of the IMM filter, either stance can be supported given the right IMM filter setup. In the following subsections we discuss our developed IMM filter in detail. We start with the composition of the IMM filter bank.

2.4.3.2 IMM Filter Bank

The IMM filter uses two second order (constant acceleration) KFs, each with a varying degree of process noise. We denote the varying degrees of process noise as KF process noise covariance matrices denoted as Q_1 and Q_2. We use two KFs with different process noise matrices to model two assumed disease phase modes: non-rapid progression, and rapid progression. The state of IMM filter is 3-dimensional. It is composed of the MD value at time t; MD velocity, the rate of change of the MD at time t; and MD acceleration, the rate of change of MD velocity at time t, where time t, refers to the time index of the state variables outlined in Section 2.3.1.

2.4.3.3 IMM Filter Measurements

The filter uses a three-dimensional *measurement vector*. It is composed of two measurements of MD and one measurement of MD velocity. The first MD measurement, z_1, is from the patient's clinical trial results (i.e. AGIS or CIGTS). The standard deviation, σ_1, of this measurement is estimated as the mean standard deviation of all MD residuals obtained by the method outlined in [30].

The second MD measurement, z_2, is a pseudo MD measurement computed using a Random Forest regression (RF) model. The purpose of the RF is to estimate the MD at the current time t, using the patient's IOP, PSD, age, and race at time t. Since IOP, PSD, age, sex, and race are not used directly as state variables in the IMM filter, the pseudo MD measurement, a non-linear function of IOP, PSD, age, and race, leverages the information present in these variables in the IMM filter's update step. The standard deviation, σ_2, of this measurement is computed as the standard deviation of the RF MD prediction. The prediction of the RF model, z_2, and standard deviation of the prediction, σ_2, are estimated as follows:

$$z_2 = \frac{1}{B}\sum_{b=1}^{B}\mu_b, \tag{2.17}$$

where B is the number of estimators in the random forest ensemble, and μ_b is the mean value of the training leaf nodes the data instance falls in. Similarly, the standard deviation is estimated as,

$$\sigma_2 = \sqrt{(\frac{1}{B}\sum_{b=1}^{B}\sigma_b^2 + \mu_b^2) - z_2^2}, \tag{2.18}$$

where σ_b^2 is the variance of leaf nodes that the data instance falls in. For a more detailed treatment of this procedure see [37].

The third measurement, MD velocity, z_3, is obtained from the slope coefficient of OLSR over a moving window of four MD observations. The choice of four MD observations was determined using a grid search procedure. Past work confirmed OLSR is a valuable means of providing less noisy velocity estimates, resulting in better filter behavior [41]. The standard deviation of the MD velocity, σ_3, is calculated as the standard error of the slope coefficient from OLSR. We assumed the measurements z_1,z_2, and z_3 were uncorrelated.

2.4.3.4 IMM Filter Parameter Initialization

The initial values for the mode probabilities (μ) and mode transition probability matrix (M) were estimated from the data; the IMM filter is not sensitive to initial values of μ [44], as it is updated at each prediction/update step using the equations previously outlined (2.3-2.4) and (2.10-2.12). The state initialization for both KF models was taken to be the patients' baseline MD (e.g. their initial MD at enrollment in the clinical trial) and values of zero for both MD velocity and MD acceleration. A complete overview of the IMM filter parameter initializations are in Appendix B. All KFs parameters were determined using a grid search procedure.

2.4.3.5 IMM Filter Evaluation

As an established method in the area of glaucoma [30], the performance of our filter is evaluated by calculating the RMSE of the filtered MD values compared to the fitted values from OLSR through *all* the patient's MD values. The OLSR estimated MD values are treated as the true MD state of the patient. Although OLSR estimates of patient's state have noise, the high level of noise in MD measurements compared to the level of noise in OLSR estimates is a favorable compromise.

The IMM filtered MD estimates had a RMSE of 1.181. For comparison, a single KF had a RMSE of 1.738. This indicates the IMM Filter performed better than using a single KF. IMM filter performance could have potentially been improved by using additional KFs to model an expanded set of disease phases however, we *underscore point-wise prediction of MD values was not the goal of this work. Instead, the goal was future Rapid Progression identification*. The outputs of the IMM filter (e.g. filtered MD, filtered MD velocity values, and filtered MD acceleration) were used as additional features for the supervised learning model discussed in Section 2.4.3.6. The addition of these three features increased the performance of our supervised learning model by approximately 7%; providing us with evidence the IMM model produces outputs that provide additional discriminatory information relevant for the task of proactive rapid progression identification. An illustration of the IMM filter outputs for a selected patient is shown in Appendix C.1. In the next section we detail the supervised classification model we developed to proactively identify patients' instances as rapidly progressing.

2.4.3.6 Soft Voting Ensemble Classifier (SVE)

Ensemble methods are learning algorithms that construct a set of classifiers and then classify new data points using a (weighted) vote of their predictions [17]. In particular, a soft voting ensemble is a ensemble method that classifies new data points using a (weighted) vote of *predicted probabilities*, instead of *predicted classes*.

The goal of the Soft Voting Ensemble Classifier (SVE) is to identify patient instances as rapidly progressing (RP). This identification comes before observing the patients future MD values. Hence, the purpose of the SVE is to provide the clinicians with analytical insights in order to help them make proactive patient treatment decisions. Our SVE is an ensemble classifier composed of four supervised learning models. These models are: Support Vector Machine - Gaussian Kernel (SVM), Random Forest (RF), K-Nearest Neighbors (KNN), and Gradient Boosted Decision Trees (GBDT).All four supervised learning models are appropriate for multi-class classification. Each can be used to estimate the probability the patient instance will experience RP within the next two (or three) years. Other methods including a neural network, linear discriminant analysis, and quadratic discriminant analysis were explored, but they did not improve the performance of the classification model. For the interested reader, a through overview of these four models is found in [21].

The four models are individually trained, and their predicted probability of progression is fed into a Logistic Regression meta-classifier for final output. This model is commonly referred to as a soft voting ensemble classifier (SVE) [10, 24]. A diagram of the supervised learning model is in Figure 2.2.

Each of the four models, given a training/testing instance, computes the probability of the instance belonging to one of the two classes (1 for RP or 0 for Non-RP). The four probabilities are subsequently fed into the Logistic Regression classifier to obtain the probability the patient's eye will exhibit rapid progression behavior within next two (three) years. We confirmed feeding the probability of RP, instead of a binary indicator indicating whether RP is present or not (e.g. 0/1), into the Logistic Regression meta classifier resulted in the best performance.

2.4.3.7 Model Validation

Since the data contains a temporal component (observations are independent between patients, but not independent within patients), traditional approaches such as k-fold cross validation are prone to overestimating model generalizability [5]. We resolve this issue using walk-forward validation [61]: A general description of walk-forward validation is below

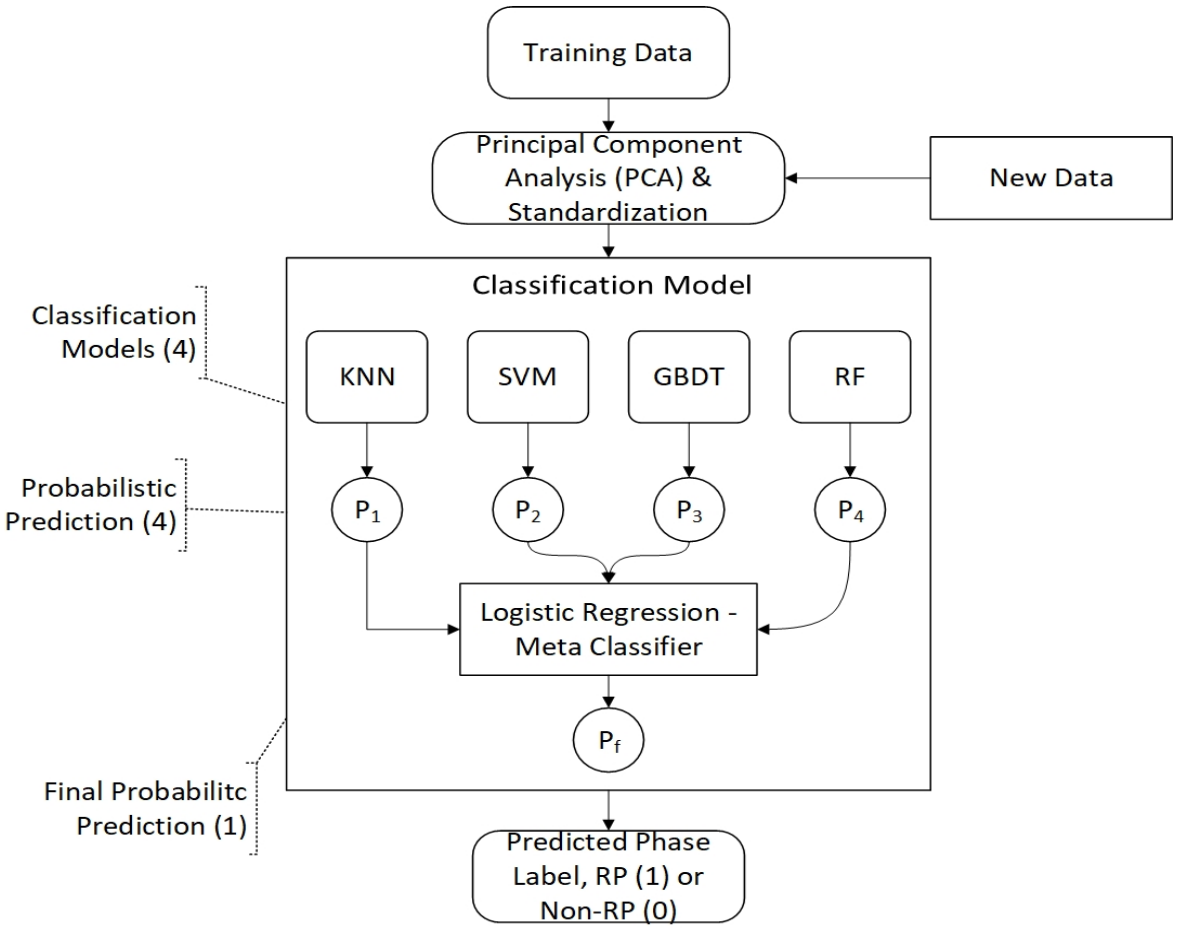


Figure 2.2: Soft voting ensemble classifier diagram

and illustrated in Figure 2.3:

1. Starting at the beginning of the time series, the minimum number of samples (i.e., starting from year 2) are used to train a model.
2. The model makes a prediction for the next time step (i.e., 6 months)
3. The prediction is stored and evaluated against the known value/label, as obtained from the method discussed.
4. The training window is expanded to include the next period of data and the process is repeated for the next iteration (go to step 1.)
5. After all longitudinal iterations are complete, the model's overall validation performance is obtained by averaging the model's performance across all held-out validation sets.

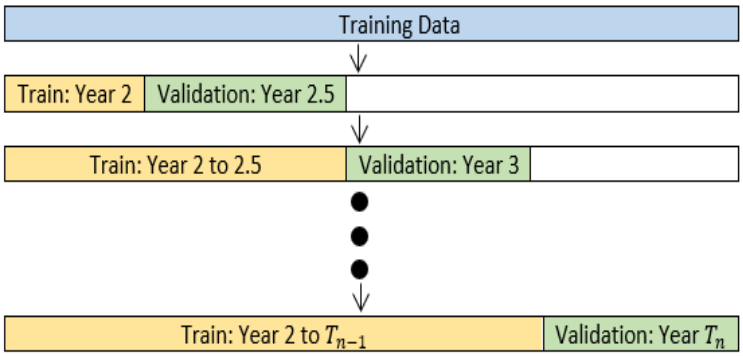


Figure 2.3: Walk forward validation illustration, not to scale

We use walk-forward validation not only to validate our model, but also to optimize the models' hyperparameters (i.e. the hyperparameters for all 5 supervised learning models). This was accomplished using grid search. We choose the hyperparameters that maximized our performance metrics: balanced accuracy (e.g. average recall for both targets), and receiver operating area under the curve (AUC). For a review of balanced accuracy and AUC see [9] and [24] respectively.

The use of a soft voting ensemble, instead of one of the four supervised learning models, is an important addition. Ensemble models have been shown to increase prediction performance by reducing variance of each sub-model's predictions at the expense of a small increase in bias [17]. Accordingly, we found the soft voting ensemble had a validation receiver operating area under curve (AUC), of more than 15% higher, in the relative sense, than any of the aforementioned classifiers. AUC is computed as the mean AUC across all walk-forward validation folds. For example, Table 2.3 and 2.4 provides an overview of the individual AUC validation performance using testing procedure 1 (to be discussed in section 2.4.4). As can be seen, the soft voting ensemble performance is superior to all individual classifiers. AUC and walk-forward validation are each discussed in the next section.

The log-odds coefficients of the logistic regression meta classifier for 2-year and 3-year RP prediction for K-Nearest Neighbor, Support Vector Machine, Gradient Boosted Decision Tree, and Random Forest classifiers under testing procedure 1 were respectively 14.11, 3.04, 1.56, 0.88 for 2-year prediction; and 11.80, 3.90, 1.67, 1.03 for 3-year prediction. Both models had intercept terms of approximately -9. The relatively large log-odds coefficients for K-Nearest Neighbors suggests the K-Nearest Neighbors models had a high true-positive rate (sensitivity). We recognize the classifiers are trained on the same data and therefore their predictions are likely correlated. Hence, the log-odds co-

efficients (classifier weights) are effected by multicollinearity. However, because our goal is prediction accuracy and not interpretation of log-odds coefficients it does not affect our modeling approach.

Table 2.3: 2 Year Testing Procedure 1 - Validation AUC Performance of Individual Models Vs. Soft Voting Ensemble

	KNN	GBDT	RF	SVM	**SVE**
AUC	0.68 ± 0.02	0.70 ± 0.02	0.69 ± 0.03	0.69 ± 0.03	$\mathbf{0.748 \pm 0.03}$

Table 2.4: 3 Year Testing Procedure 1 - Validation AUC Performance of Individual Models Vs. Soft Voting Ensemble

	KNN	GBDT	RF	SVM	**SVE**
AUC	0.70 ± 0.02	0.72 ± 0.03	0.69 ± 0.05	0.70 ± 0.04	$\mathbf{0.825 \pm 0.04}$

2.4.4 Results

The following two testing procedures are were used to evaluate our model performance:

Testing Procedure 1 (TP1): The model is trained and validated using all patient data except the last period of data. The last period of available data is used as the held-out (test) set. This is done in order to estimate how well our model predicts patients' eyes at the next visit, given it has been trained on data from the previous visits. Hence, TP1 estimates model performance for predicting future RP for patients whose past data it has already been trained on.

Testing Procedure 2 (TP2): The model is trained and validated using a subset of available patients (90%). The subset is obtained by stratifying on age, race, and sex, as to keep the patients in the training and held-out (testing) sets as similar as possible. For the remaining 10% of patients, the first available visit, year 2, is used as our held-out (test) set. Hence, TP2 estimates the performance of the model at predicting RP eyes for patients it has not been previously trained on, i.e., new patients. Tables 2.5 and 2.6 provide an overview of our results.

The TP1 results in Table 2.5 show a 2-year prediction BA and AUC of 0.717 and 0.819 respectively, and a 3-year prediction BA and AUC of 0.737 and 0.758 respectively. We believe our 3-year predictions for BA to be slightly better due to the inherent noise in the

Table 2.5: TP1 Balanced Accuracy (BA) and ROC AUC performance for 2- and 3-year prediction models

	Walk-Forward Validation			Testing Procedure 1 (TP1)
2 Year Predictions	Metric	Mean	Standard Deviation	-
	BA	0.702	0.023	0.717
	AUC	0.748	0.038	0.819
3 Year Predictions	Metric	Mean	Standard Deviation	-
	BA	0.759	0.012	0.737
	AUC	0.825	0.045	0.758

Table 2.6: TP2 Balanced Accuracy (BA) and ROC AUC performance for 2- and 3-year prediction models

	Walk-Forward Validation			Testing Procedure 2 (TP2)
2 Year Predictions	Metric	Mean	Standard Deviation	-
	BA	0.702	0.023	0.672
	AUC	0.768	0.036	0.772
3 Year Predictions	Metric	Mean	Standard Deviation	-
	BA	0.759	0.024	0.703
	AUC	0.817	0.050	0.758

Table 2.7: TP1 testing performance categorized by race. Note there were only 9 Asian participants for this study.

	Race	BA	AUC
2 Year Prediction	Black	0.762	0.859
	White	0.678	0.784
	Asian	0.654	0.739
3 Year Prediction	Black	0.741	0.747
	White	0.676	0.754
	Asian	1	1

MD readings. Given we define RP as an OLSR slope of ≤ -1 MD/year, as the window of time increases (e.g., 2 to 3 years), the noise in the slope estimate is reduced, and as a result the measured slope is more likely to be more representative of the true rate of progression. Additionally we note the probability threshold used for determining whether a patient's eye will exhibit RP in the next two (or three) years was 0.50 (e.g. if the predicted probability of RP was ≥ 0.50 classify the patient's eye as RP, else classify as non-RP).

The results for TP2, which are shown in Table 2.6, follow the results seen in TP1. The key difference between these two testing procedures is the results for TP2 are slightly inferior. We believe this is due to the fact for TP2 we are testing our model on patients that have no prior data in the training datasets. These patients, while stratified to be as

similar as possible, are not the same patients. Hence, their disease trajectory is likely different.

In addition, for TP1, we examined the testing BA and AUC categorized by race. We choose not to do this for TP2 because after categorizing by race the the number of patients in each category were too small to draw meaningful conclusions. The results are located in Table 2.7. We note, from Table 2.2, there were only 9 patients of Asian descent in our study population. Hence, while the results may suggest for Asian patients we predict future TP1 3 year rapid progression perfectly, it needs investigated further using a larger population of Asian patients. The results from Table 2.7 are largely inline with the results from Table 2.5. There does appear to be slightly better TP1 performance for black patients, however this could be attributed to black patients making up the majority of our study population. It would be worthwhile in future works to expand our study population and see if these results remain the same.

2.5 Conclusion

This chapter has provided models and a framework that can be adapted to a variety of chronic diseases in which medical tests over time have moderate to severe residual variability. We utilize the notion of disease phase, which may change over time, and apply it to the ocular condition glaucoma to identify whether a patient at time t is a rapid progressor or not. We found that to achieve high accuracy for the difficult problem of glaucoma phase classification, the method needed to adapt dynamically to changing disease phases. This was accomplished by integrating the outputs of an interacting multiple model (IMM) KF with supervised learning classification methods. Incorporating the IMM filter estimates increased TP1's 2-year AUC from 0.752 to 0.819. A similar effect holds for TP2 and 3-year predictions as well.

The results indicate the viable use of our supervised learning algorithms to inform clinicians on instances of RP between the current visit and the next 2 or 3 years for AGIS/CIGTS. This is of significant consequence, as identifying patients who will experience RP provides the clinician with key information for proactive treatment. We have shown our model is able to achieve acceptable levels of performance (e.g., TP1: 2-year (3-year) AUC of .82 (.76) in the walk-forward cross-validation). When tested on a held-out test set, we obtained a 2-year (3-year) AUC of .77 (.76). We conjecture, as the data used to train these models is increased, their performance will continue to improve, increasing their value to clinical practice. Future work is warranted to expand the study population

from patients with moderate to advanced glaucoma, to patients with early onset/moderate glaucoma.

More generally, these results suggest that the framework presented in this chapter is successful in the proactive identification of disease phases. In cases where medical tests have high residual variance, estimating the true reading with high accuracy is likely to be very difficult. The combination of filtering and statistical learning serves as a useful option when faced with this issue. The methodology presented in this chapter has far reaching implications, as it enables the clinician to make more informed decisions regarding the treatment of patients and increases the likelihood the clinician is able to maintain a satisfactory quality of life for the patient. Future works may be able improve our methodology by incorporating mechanisms for the dynamic control of the disease. For example, examining how clinical treatments affect the disease transition from one phase to another, and more insightfully, how this can be controlled so the patient time in a "favorable" disease phase is maximized. Additionally we may consider predicting the disease phase at smaller time intervals (e.g. 6 month periods). In this way, the clinician will have greater detail of patients' future disease trajectories, instead of a summary measures over 2 (or 3-year) time windows.

Ethics Statement

Ethics approval for this work was obtained via University of Michigan IRB HUM00079342.

Acknowlegements

Supported by the National Eye Institute, NIH R01 EY026641

CHAPTER 3

Reinforcement Learning Methods for Constructing Personalized Monitoring Schedules for Patients with Chronic Conditions: Application to Glaucoma

3.1 Introduction

Glaucoma is one of the leading causes of visual impairment in the United States [12, 31, 54]. It is estimated that over 3 million Americans have glaucoma [31], and up to 6 million have ocular hypertension (OHTN). Patients diagnosed with OHTN are at increased risk of developing glaucoma [55], and it is vital for these patients to seek treatment in a timely and recurrent manner. Nevertheless, there exists no consensus on the optimal monitoring frequency by which patients with OHTN should be scheduled for follow-up appointments [38]. Determining the time to next follow-up schedule is a challenging task, as it should strike a balance between detecting glaucoma progression in a timely manner while avoiding unnecessary treatment for heterogeneous patients. For example, if the time to next schedule is too short, the patient may receive unnecessary treatment; if the time to next schedule is too long, the patient may be at risk of undetected glaucoma progression. There are clear tradeoffs between these two extremes. This supports the need for personalized patient follow-up scheduling, which depends on the medical need and state of the patient.

The aim of this chapter is to develop a framework for constructing follow-up scheduling policies using the available patient-specific data from electronic health records. We present a reinforcement learning (RL) methodology for determining personalized monitoring schedules for patients with ocular hypertension. We chose to employ an RL framework because it is easily implementable, does not require warm-up periods (e.g., can be used

starting from the patient's first appointment), is model-free, and can dynamically adapt to patients' changing medical state. To the best of our knowledge, we are the first to apply RL to address this research problem.

Our proposed scheduling policies have three underlying goals: (i) maximizing time between follow-up visits, (ii) maximizing the scheduling efficiency, which is the percentage of scheduled visits indicating glaucoma progression may be near, and (iii) minimizing time to detect glaucoma progression. The organization of our chapter is as follows: Section 3.2 reviews background and literature. Section 3.3 outlines the proposed methods. Section 3.4 presents and discusses the numerical results of the OHTN case study. Finally, Section 3.5 concludes the chapter and discusses the future work.

3.2 Background and Literature

There are two primary areas of research relevant to our approach: (1) RL applications within healthcare; and (2) personalized patient time-to-next test (TNT) models (i.e., data-driven models for determining personalized patient follow-up visit schedules).

Reinforcement Learning Applications in Healthcare. In healthcare, there has been numerous data-driven approaches to advance the way patients are scheduled and treated. RL has been employed successfully in a variety of these settings. According to [66], the relevant RL studies can be classified into three categories: dynamic treatment regimes, automated medical diagnosis, and general domain areas.

In dynamic treatment regimens, [52] used RL to identify a treatment protocol for weaning patients off mechanical ventilation. Their model focused on improving three aspects of patient ventilation weaning: (1) time into ventilation, (2) physiological stability (i.e., whether vitals are steady and within expected ranges), and (3) failed reintubation or breathing trials. They showed that their model outperformed, on average, the currently used ventilation weaning policy by clinicians in regulating patients' vitals reintubations. [19] developed an RL algorithm to optimize anemia treatment amongst hemodialysis patients by determining the most appropriate erythropoiesis-simulating agent (ESA) dosages for patients. Although perspective validation was required, the authors demonstrated their algorithm performed superior to a currently used ESA dosage policy. And more recently, [14] developed an RL system that dynamically assists with communications to patients. In the case of patients with physical disabilities or cognitive disabilities, the system automatically searches for the most effective way to communicate and remind the daily treatment plan to the patient.

In automated medical diagnosis, [50] provided an illustration of RL for effectively classifying lung nodules as benign or malignant. The methods employed are based directly on applying RL to lung lesions CT images thereby automating the analysis and detection of the disease. Similarly, but more recently, [66] employed RL for lung cancer detection by combining deep RL methods with medical big data generated by IoT. [47] used Trust Region Policy Optimization (TRPO), a type of reinforcement learning, for automated joint surgical gesture segmentation and classification. The motivation of their work is to provide an effective means of surgical skill assessment and efficient surgery training for surgeons.

Last, in general domain areas, [25] used RL for primary care appointment slot scheduling. In their work they develop an RL algorithm to schedule appointments slots in a set of increasingly challenging primary care environments. [60] used approximate dynamic programming, a type of RL, to solve capacity allocation problems. Last, in the drug discovery and development research domain [51] used RL to fine tune a recurrent neural network to generate molecules with certain desirable properties through augmented episodic likelihood. In all cases examined, RL was successfully employed as a methodology for improving healthcare decisions.

Personalized Time-to-Next Test Patient Models. There has been limited works in this area. Prior works commonly model a low-dimensional (often discretized) state space (e.g., [45, 48] or generally focus on performing screenings to detect the first incidence of a disease [4, 34, 46], rather than monitoring an ongoing one. Recently, with the advances of computing, there has been advances in personalized monitoring through the application of medical wearable devices. For example, [32] surveys numerous wearable devices, providing an overview of their clinical application, and discusses opening challenges in this area. However, the use of wearable devices commonly tackles the disease monitoring problem from a disease surveillance perspective; focusing on the large volume of real-time clinical data the devices produce, instead of how this information can be used to inform patient monitoring schedules for clinical interventions or situations with costly measurements.

The studies of [59] and [36] are the closest works to ours. In these works, they not only consider the frequency of monitoring decisions but also incorporate the dynamic updating of information enabling the personalized scheduling of patients. They developed a Kalman filtered based TNT algorithm parameterized using a clinical trial dataset from the Advanced Glaucoma Intervention Study (AGIS) and the Collaborative Initial Glaucoma Treatment Study (CIGTS). Their aim, analogous to ours, was to develop personalized patient TNT monitoring schedules. Our approach differs from these works in several ways. First, we propose an *integrated model* that directly provides a control policy (follow-

up scheduling policies) which is close to the optimal policy. This provides an easy-to-implement method that can be used in real-world practice. Unlike the RL method we devise, they address the TNT problem in a multi-step manner which includes, estimating the next state using a Kalman Filter model, estimating the probability of disease progression using a regression model, and implementing an optimization step to determine TNT using a confidence interval constructed based on the estimated next states with respect to different time intervals. Second, as it is important in practice *we do not require warm up periods* (i.e., periods for parameterizing the model for each patient) before our algorithm can recommend time until the next follow-up appointment. Our method can generate TNT recommendations starting from a patient's first appointment. The TNT algorithms developed in [59] and [36] required 4 periods (spanning 18 months) to generate TNT recommendations. This marks a notable difference, as delays to recommend a patient an appropriate TNT can lead to over-monitoring if it is shown the patient did not require a follow-up visit; or may lead to under-monitoring, if it is shown the patient should have been scheduled a follow-up visit sooner.

Third, we employ a *model-free algorithm*, which assumes no knowledge of the dynamics of the system being modeled. This obviates the need to directly model how a patient's state transitions between appointments. Their methods required the disease dynamics to be effectively modeled using linear Gaussian systems and are not well suited for discrete patient covariates (e.g., sex or race of patient). In particular, their Kalman Filter required model components (e.g., state transition matrix, observation model, process noise matrix, etc.) to accurately reflect the dynamics of the system (e.g., the patients' disease), which has both benefits (e.g., interpretability) and weaknesses (e.g., possibility of poor model specification). Lastly, our RL-based methodology is *flexible* in a sense that different criteria can be considered to find a near-optimal scheduling policy rather than a conservative estimate of the patient's probability of progression used in prior studies. We ultimately employ the use of RL to build a TNT recommendation model that is personalized for each patient, can dynamically adapt to a patient changing medical state, can be used starting from the patient's initial visit, and provides an integrated approach. To the best of our knowledge, we are the first to tackle this problem in a manner that meets these four conditions.

3.3 Methods

3.3.1 Conceptual Framework

While the application focus of this work is on OHTN, our conceptual framework can be used to build TNT models for other chronic conditions as well. We formalize our RL framework using a Markov Decision Process (MDP). An MDP is a modeling framework for optimally solving the sequential decision-making problems. An MDP is defined by a state space, action space, transition probability function, and reward function. These components are formally defined as follows:

- The state space S is a continuous state space such that at time t, the patient is in state $s_t \in S$. The state s_t contains the variables (e.g., test measurements) essential for TNT decision making.
- The action space A is a set that the clinician may only choose an action $a_t \in A$ at time t. For the general TNT problem, the action, a_t, indicates the length of time the patient must wait until the next follow-up appointment.
- The transition probability function defines the dynamics of the system. That is, the function $P : S \times A \times S \to [0, 1]$ gives the probability of the next states at time $t + 1$, given the action a_t and the patient's state $s_t \in S$.
- The reward function, $r : S \times A \to \mathbb{R}$ is a function of the current state, chosen action, and next state outcome that evaluates the immediate effect of the chosen action. The reward function $r(s_t, a_t, s_{t+1})$ is used to incentivize the best action at time t, but it does not provide information on the long-term effects of the action. For instance, a reward function can incentivize timing of a follow-up visit either shortly before or when disease progression events are likely to occur.

The goal of the MDP is to learn a scheduling policy that maximizes the cumulative reward over time, called an optimal policy. Let $\pi : S \to A$ denote an optimal policy which is a function that maps every state $s_t \in S$ at time t to a decision choice $a = \pi(s)$. The current action chosen by following the policy, π, maximizes the total expected sum of rewards, $R^\pi(s_0)$ given an initial state s_0. That is, we aim to maximize the time-discounted total reward over a planning horizon of infinite length rather than just the immediate reward. This is defined as:

$$R^\pi(s_0) = \lim_{T\to\infty} E_{s_{t+1}|s_t,\pi(s_t)}(\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, s_{t+1})), \tag{3.1}$$

where $\gamma \in [0, 1)$ is the discount factor to balance the effect of the immediate reward and future rewards. We leave to the reader the mathematical underpinnings of the MDP, see [53] for a detailed overview. The presentation of the TNT model will aim for generality where possible; however, for ease in reading and to explain the application, we will discuss the MDP formulation in the context of building a TNT model specifically for patients diagnosed with OHTN.

3.3.2 Data

We use the Ocular Hypertension Treatment Study (OHTS), a randomized clinical trial involving 1636 patients with OHTN recruited from 22 centers throughout the US between February 1994 and March 2009 [29, 28, 26]. Participants were randomized to either treatment with IOP-lowering medications or followed without treatment. Both groups were followed for up to 15 years using standardized measurements of tonometry and perimetry starting at baseline and every 6 months thereafter. To be eligible for participation in OHTS, individuals had baseline intraocular pressures (IOPs) of 21 to 32 mm Hg in both eyes, reliable and normal visual fields (Carl Zeiss Meditec, Dublin, CA), and no detectable structural evidence of glaucoma based on optic nerve evaluation. In our analyses, trial participants were censored when they experienced non-glaucomatous visual field loss or underwent any incisional intraocular surgery other than uncomplicated cataract surgery. Patients' eyes were excluded if they had fewer than 2 sets of tonometry or perimetry measurements. If either one or both eyes were eligible from a patient, we included them in our study. In total 1619 patients entered our study.

OHTS provides demographic information (e.g., age, sex, and race) and longitudinal medical information of three key continuous valued measurements: intraocular pressure (IOP) - measures the pressure in the eye, mean deviation (MD) - measures mean age and race adjusted deviation of a patient's retinal sensitivities, and pattern standard deviation (PSD) - measures the uniformity of a patient's retinal sensitivities. All values were recorded on approximately 6-month intervals. A summary of the data is given in Table 3.1.

Table 3.1: Description of OHTS Study Sample

Characteristic	**Mean/Count**	**SD/Percent**
No. of Patients	1619	-
No. of Eyes	3231	-
No. Eye Progressions	423	-

Table 3.1 continued from previous page

Characteristic	Mean/Count	SD/Percent
Sex, No (%)		
Male	697	43%
Female	922	57%
Race/ethnicity, No (%)		
White	1133	70%
Black	399	25%
Other	87	5%
Age at baseline, years	57	10
No. of visual field tests per eye	22	9
Follow-ups visit length, years	10.5	4
Baseline Reading		
MD, dB	0.17	1.15
PSD, dB	1.95	0.29
IOP, mm Hg	25.01	2.98

A participant's eye was labelled as progressing from OHTN to primary open-angle glaucoma (POAG) if it exhibited a drop in MD of 3 or more decibels (dB) from their baseline MD (e.g., initial MD measurement) on 2 consecutive MD tests [49]. If the progression criterion was met, the date of progression was taken to be the date of the second MD test. We use two consecutive MD tests, instead of one, to limit misclassifying fluctuating MD performance as progression.

All quantitative data elements (e.g., MD, IOP, PSD, etc.) were recorded in OHTS at intervals of approximately every 6 months. While most patients strictly adhered to the visit schedule for the trial, infrequently patients may have missed visits. In these cases, linear interpolation was employed. B-splines interpolation of various orders was tested; however, it was inferior to linear interpolation. To avoid creating progression artifacts (e.g., patient eyes that did not progress according to the original MD measurements, but progressed after linear interpolation was introduced), glaucoma progression was determined using the original MD measurements in the clinical trial dataset, and not their interpolated values.

3.3.3 TNT MDP Formulation for Patients with Ocular Hypertension

The MDP TNT formulation for patients with ocular hypertension is described below. Notation for the chapter is provided in Appendix D.1. The state, s_t, is a 13-dimensional vector composed of a patient's left or right eye features recorded at time t, where t is defined on intervals of 6-months (e.g., $t = 3 \rightarrow 18$ months). The features are as follows: patient's age, race, and sex; intraocular eye pressure (IOP), mean deviation (MD), and pattern standard deviation (PSD); baseline (or initial follow-up appointment) MD, IOP, and PSD; the change from baseline for MD, IOP, and PSD eye measurements; and the number of months from the patient's baseline visit to the current visit. Besides age, race, and sex, all features are recorded at the eye level.

The action space, A, represents the integer number of 6-month periods between a patient's current visit and their next visit. The maximum number of months between two consecutive follow-up visits is 24 months or 2 years; this ensures a patient has at least one follow-up visit every 2 years. Hence, the action space comprises 4 actions, $A \in 1, 2, 3, 4$. For example, given $t = 0$, an action $a_0 = 2$, indicates the patient's first follow-up visit will be scheduled 12 months (or two 6-month periods) after their initial (e.g., baseline) follow-up appointment. The choice of 6-month periods was used to match the same follow-up appointment resolution (patient appointments approximately every 6-months) used in the OHTS dataset.

The transition probability function is inferred from the OHTS patient data. This is because our RL model employs the fitted Q-iteration algorithm [18]. The fitted Q-iteration algorithm is model-free, indicating it does not assume any knowledge of the dynamics (e.g., state transitions) of the system being modeled. Hence, while it is common to explicitly define the transition function when describing an MDP, it is not required for this work.

The reward captures how the clinician should respond (i.e., what action should he/she take) given the patient's updated current state during a visit. Rewards are associated with each state-transition pair using the tuple $r_{t+a_t}(s_t, a_t, s_{t+a_t})$. The subscript $t + a_t$ refers to the time of next follow up appointment (e.g., number of 6-month periods from their current follow-up visit) after choosing action a_t. All normalized rewards and sub-rewards are bounded between 0 and 1. The reward at time, t, is composed of a linear combination of four sub-rewards, and four weights that are normalized to sum to one. We considered the following sub-rewards:

I. Visit delay $(r^{(VD)}_{t+a_t})$: To prevent unnecessary visits, we incentivize longer duration between visits by giving a linearly increasing reward to actions suggesting a longer

duration between two visits that ranges from $\frac{1}{4}$ to 1:

$$(r_{t+a_t}^{(VD)}) = (\frac{1}{4}) \cdot a_t, \quad a_t \in A \tag{3.2}$$

II. MD drop $(r_{t+a_t}^{(MD)})$: To promote higher scheduling efficiency, we reward suggesting a time to next visit for which the MD shows suspected progression or near suspected progression (e.g., single occurrence of MD drop from baseline ¡= -3):

$$(r_{t+a_t}^{(MD)}) = \frac{1}{1 + \exp^{-3(-(MD_t - MD_{Baseline}) - 2.5)}} \tag{3.3}$$

We incentivize recommending visits at periods most likely to have a drop in MD of about -2 dB below baseline. This allows our scheduling policy to potentially detect a patient's progression early, enabling a clinician to take proactive action.

III. Progression identification $(r_{t+a_t}^{(PI)})$: The reward is a value of 1 if the action at the current state is to schedule the next appointment in 6 months, and if both the current state's (e.g., current follow-up visit) and the MD and transitioning state (e.g., follow-up visit 6 months after current visit) indicates a drop of MD of 3 or more decibels (dB) from the state's baseline MD value; otherwise, it is 0. This incentivizes having a visit at the first period that satisfies the definition of progression.

$$(r_{t+a_t}^{(PI)}) = \begin{cases} 1, & (a_t = 1) \wedge (MD_t - MD_{Baseline} \leq -3) \wedge (MD_{t+a_t} - MD_{Baseline} \leq -3) \\ 0, & \text{Otherwise} \end{cases} \tag{3.4}$$

IV. MD stability $(r_{t+a_t}^{(MS)})$: Because clinicians and patients are uncomfortable with large changes between visits, we penalize follow-up visits with a drop in MD of 0.05dB or more:

$$(r_{t+a_t}^{(MS)}) = \begin{cases} 1, & MD_{t+a_t} - MD_t \geq -0.05 \\ 0, & \text{Otherwise} \end{cases} \tag{3.5}$$

Finally, the reward is formulated as follows:

$$r_{t+a_t} = \lambda^{(VD)} r_{t+a_t}^{(VD)} + \lambda^{(MD)} r_{t+a_t}^{(MD)} + \lambda^{(PI)} r_{t+a_t}^{(PI)} + \lambda^{(MS)} r_{t+a_t}^{(MS)}, \tag{3.6}$$

where the scalars $\lambda^{(VD)}, \lambda^{(MD)}, \lambda^{(PI)}$, and $\lambda^{(MS)}$, are weights applied to each sub-reward to determine their contribution to the total reward (see section 3.3.5 for their values). For example, if, in the extreme case, all but the visit delay reward is given a non-zero

positive weight, the visit recommendations would always recommend visits every 2-years to maximize the time between visits. In a clinical application, weights should be chosen under the supervision of a clinician. Taken holistically, the rewards encourage the MDP to recommend follow-up visits to maintain patients' MD measurements appropriate to the condition of the patient. When the patient is suspected of progressing to OAG; or when the patient is progressing to POAG, the TNT should be short. When neither of these adverse events are expected, the TNT should be as far out into the future as the model is confident in recommending. Plots of the sub-rewards are shown in Figure 3.1.

A conceptual illustration of the RL framework for TNT is presented in Figure 3.2. Figure 3.2 can be summarized as follows: (1) a RL scheduling model receives, as input, a patient's state, s_t, and estimated reward, r_t. (2) The model determines which action (e.g., 6, 12, 18, or 24 months until the patient's next follow-up appointment) maximizes the summation of current and future rewards. (3) The maximizing action, a_t, is selected, and the patient next follow-up appointment is scheduled. This 3-step process transitions the patient from their current state s_t to their transitioning state s_{t+1} through the selection of action a_t. Steps 1-3 proceed in a recursive manner.

3.3.4 Fitted Q-iteration

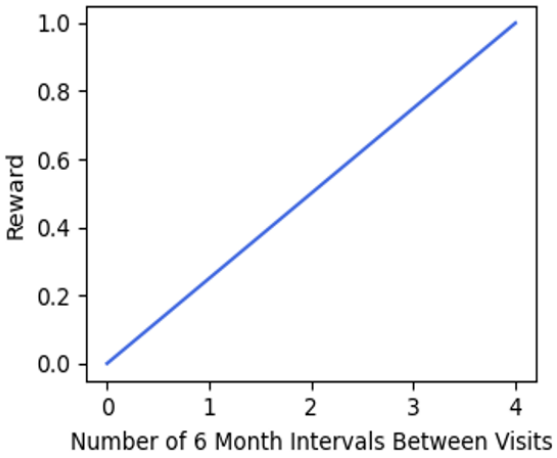
To find an approximately optimal policy, π, we use fitted q-iteration (FQ) algorithm, which is an off-policy batch mode RL algorithm. The goal of FQ is to provide an estimate of the Q-function, which can be directly used to find the optimal policy. The Q-function, defined as $Q^{\pi}(s,a) : S \times A \rightarrow R$, computes the expected reward starting at state, s, choosing action a, and thereafter following policy, $\pi(s)$. Given a set of one-step transitions of the form $F = ((s_t^n, a_t^n, s_{t+1}^n), r_{t+1}^n), n = 1, \ldots, |F|$, the TNT policy is estimated as follows:

Step 1 (Initialization): In the first step, we initialize Q-function $\hat{Q}_0$, where $\hat{Q}_0(s,a) = 0$ for all - $s \in S$ and $a \in A$.

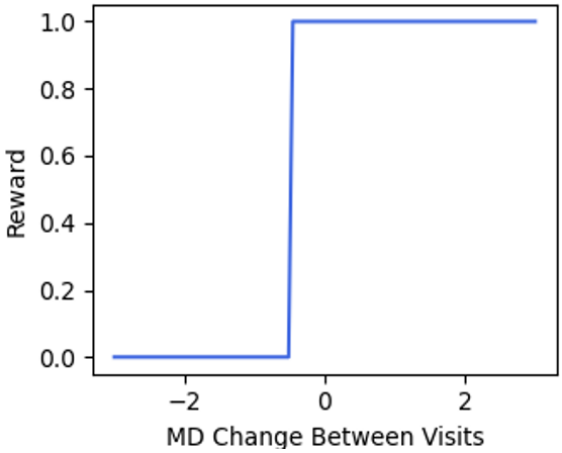
Step 2 (Approximate Q-function): In the second step, we approximate the Q-function over K iterations. In the k^{th} iteration, we have two sub-steps:

Forming the training set: We calculate $\hat{Q}_k(s_t^n, a_t^n)$ for all $(s_t^n, a_t^n) \in F$ according to the Bellman equation:

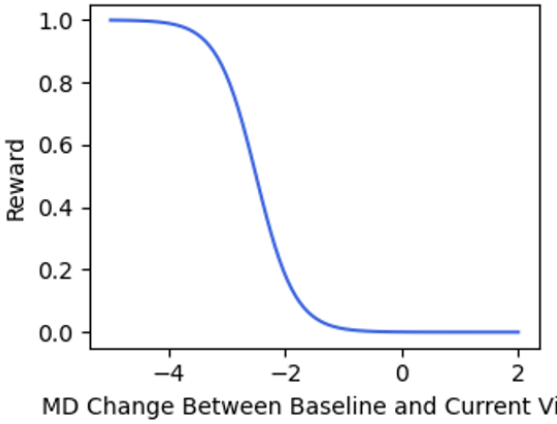
$$\hat{Q}_k(s_t, a) \leftarrow r_{t+1} + \gamma \max_{a \in A} \hat{Q}_{k-1}(s_{t+1}, a), \tag{3.7}$$



(a) Visit Delay Reward



(b) MD stability reward



(c) MD Drop Reward

Figure 3.1: Illustrations of the reward and cost structures for visit delay reward, MD stability cost, and MD drop reward functions. Visit delay and MD drop rewards range between 0 and 1, while the MD stability ranges between 0 and -1. The progression identification reward is not present because it is based on a set progression conditions being met. MD = mean deviation.

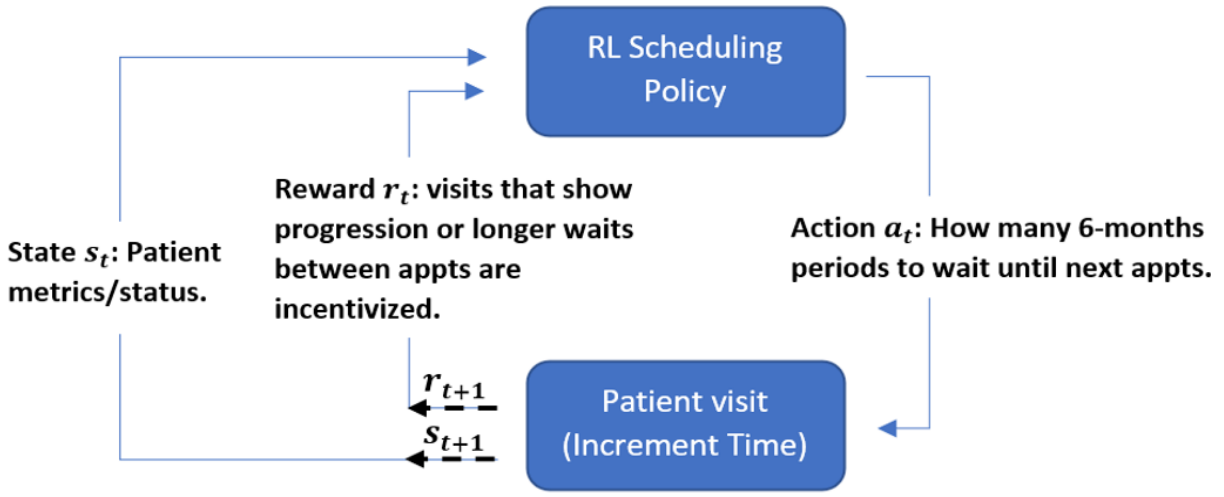


Figure 3.2: Conceptual Illustration of the RL framework for TNT. RL = reinforcement learning; TNT = time to next test.

where $\hat{Q}_{k-1}(.)$ is the function approximated in the prior iteration using a supervised learning regressor. Accordingly, we form the k^{th} training set, $T_k = \{((s_t^n, a_t^n), \hat{Q}_k(s_t^n, a_t^n)), n = 1, \ldots, |F|\}$.

Improving the estimated Q-function: We approximate the function $\hat{Q}_{k-1}(.)$ using the training set T_k and a supervised learning regressor. The supervised regressor can be viewed as predicting the maximum expected reward after transitioning to state,$s_{t+1} \in S$, and choosing action, $a \in A$. The supervised learning regressor takes state-action pairs, $(s_t, a_t) \in T_k$ as the input variables and $\hat{Q}_k(s_t, a_t)$ as the response variables. We use the Extremely Randomized Decision Trees (ERDT) proposed by [52, 18] as a supervised regressor, which is a tree-based supervised classification algorithm. The algorithm has three parameters that need to be specified: (1) the number of features to consider when computing a decision true node split, (2) the number of trees to include in the tree ensemble, and (3) the minimum number of samples in each leaf node. For the models presented, the number of features considered were: all 13 state variables, 50 trees included in the tree ensemble, and a minimum number of samples of 200 in each leaf node. The parameter values were chosen using the direction in [18].

Step 3 (Approximate Optimal Policy): In the last step, the policy, π, is computed as:

$$\hat{\pi}(s_t) = \arg\max_{a \in A} \hat{Q}_K(s_t, a)., \tag{3.8}$$

where $\hat{Q}_K(.)$ is the Q-function approximated in the prior step over K iterations.

It is worth noting that the convergence of FQ is guaranteed for commonly used supervised regressors if the structure of the regressor does not change from one iteration to the next [18]. The convergence of FQ can be measured as the average distance between subsequent FQ iterations. Since, FQ starts with an arbitrary approximation of the Q-function, (e.g., $Q_0(s_t, a_t) = 0$), that generally improves after each iteration, when there is no longer an improvement in the average distance between subsequent Q-functions, FQ, is said to have converged to an optimal Q-function. The distance between subsequent Q-iterations can be calculated as,

$$\Delta(\hat{Q}_k, \hat{Q}_{k-1}) = \frac{\sum_{(s,a)\in S\times A}(\hat{Q}_k(s,a) - \hat{Q}_{k-1}(s,a))^2}{|S \times A|} \tag{3.9}$$

where $|S \times A|$, is the cardinality of the cartesian product of the state and action space. For the model developed, we stopped after a fixed $K = 100$ iterations (i.e., our analysis of the Q-function over time indicated 100 iterations was sufficient for an accurate solution). However, we confirmed convergence using $\Delta(\hat{Q}_k, \hat{Q}_{k-1})$. In cases where early stopping is preferred, $\Delta(\hat{Q}_k, \hat{Q}_{k-1}) \leq \epsilon$, can be used. Where, ϵ, is treated as a threshold signaling convergence.

3.3.5 Evaluation

Using fitted Q-iteration proposed in section 3.3.4, we developed two polices, namely, RL policy 1 and RL policy 2. The key difference between the two is how their visit delay reward was weighted. RL policy 2 puts less weight on the visit delay reward, thereby increasing the number of patient follow-up visits. This significantly reduced the diagnostic delay and the average time between follow-up visits compared to RL policy 1. The un-normalized reward weights for RL policy 1 and RL policy 2 were respectively, $\lambda^{(VD)} = 0.05$ and 0.01. For $\lambda^{(MD)}, \lambda^{(PI)}$ and $\lambda^{(MS)}$ both polices used $\lambda^{(MD)} = 1, \lambda^{(PI)} = 1$, and $\lambda^{(MS)} = 1$.

To assess how well the 2 RL policies performed relative to fixed interval testing the following three metrics, suggested in [59], were used to assess the performance across the held-out test set:

I. Average time-to-next test: The average time, in years, between subsequent patient's eye follow-up visits (the higher the better).

II. The efficiency in scheduling follow-up visits: The percentage of scheduled visits that show a patient's eye MD loss of 3 decibels from baseline (i.e., suspected POAG

progression); It is computed at the patient eye level (the higher the better).

III. The diagnostic delay: The average number of months that a patient's eye glaucoma progression went undetected before the next scheduled visit); it is 0 if a visit occurs on the date of the confirmation visit (the lower the better).

To validate and test our proposed methodology, the OHTS data was divided into a training set (80%; a total 2587 OHTS patients' eyes) and a testing set (20%; 644 OHTS patients' eyes). There were 340 (13%) progressed eyes in the training set and 79 (12%) progressed eyes in the testing set. An 80/20 split across patients was used to ensure the model had enough training and testing instances to learn a scheduling policy and appropriately test it. Trial participants were randomly assigned to training or testing set. For the 1612 patients who contributed both eyes, their two eyes were assigned to either the training set or testing set. For the 7 patients who contributed one eye, their eye was either assigned to the training or testing set. The training set was used to create the time-to-next test scheduling policies (i.e., train the ERDT supervised machine learning model) using fitted Q-iteration proposed in section 3.3.4. The testing set was used to evaluate the scheduling policy's performance using the evaluation metrics.

3.4 Results and discussion

3.4.1 Policy Evaluation

We evaluate the performance of the proposed scheduling policies using the OHTS randomized clinical trial data. A common scheduling policy for TNT used by clinicians is fixed interval scheduling. Fixed interval scheduling refers to scheduling a patient's follow-up every x years, where x typically ranges between 0.5-years and 2-years. The larger the interval, x, the less aggressive the clinician is in monitoring the patient. For the application presented in this chapter, we use x equal to 1-year and 2-years. We use the fixed interval scheduling policies as our benchmarks.

Our proposed scheduling policy for TNT reflects the real-world practice in which a patient's TNT is dictated by their medical risk of deterioration. In cases where their medical status suggests their illness is under control, they can be seen at longer durations between follow-up appointments; conversely, if their illness is not under control, the patient' follow-up visit is scheduled with a shorter TNT. This approach can be seen as balancing the patient's quality of life (i.e., less inconvenience and possible discomfort) with their risk of poor health outcomes (e.g., glacoma progression).

The TNT RL policies were evaluated on the testing set. The evaluation consisted of patients' eyes, beginning at baseline, being recursively recommended their next follow-up visits by the TNT RL policies until reaching the end of their follow-up history. Average TNT was recorded for all testing eyes. Efficiency and diagnostic delay were recorded for each of the 79 progressed eyes in the testing set. The results, sorted by average TNT, are shown in Table 3.2.

Table 3.2: Two RL and Two Fixed-interval Policies Performance (bold indicates the max and min values achieved). RL = reinforcement learning.

Scheduling Policies	**Average TNT (years)**			**Scheduling Efficiency (Max)**	**Diagnostic Delay in Months (Min)**
	Overall	Non-Progressed	Progressed		
Two-year	2	2	2	22%	12.46
RL Policy 1	1.55	1.61	1.08	**34%**	3.89
One-year	1	1	1	24%	7.45
RL Policy 2	0.94	0.96	0.80	32%	**2.63**

Comparisons between the 1- and 2-year fixed interval testing scheduling policies and RL scheduling policies indicated the RL policies outperformed the fixed interval policies. The 2-year fixed interval follow-up policy, by design, had the largest average time between follow-up visits (2 years), followed by RL policy 1 (1.55 years). RL policy 1 had the highest scheduling efficiency (34%) followed by RL policy 2 (32%). RL policy 2 had the smallest diagnostic delay (2.63 months) followed by RL policy 1 (3.89 months). This suggests the RL policies can achieve higher scheduling efficiency and lower diagnostic delay than fixed one-year policy while requiring less visits from the patient.

To gain more insights, we investigated the 4 follow-up scheduling policies with respect to the diagnosis delay criterion in more detail. Figure 3.3 presents a boxplot comparison of the diagnostic delay of the 4 follow-up scheduling policies across the testing eyes. The figure visually confirms both RL policy 1 and 2 perform better than both fixed interval policies in minimizing diagnostic delay. Not only do the RL policies perform better on average, but also, have less variability in their performance than the fixed interval policies. The box-and-whisker plot illustrates the distribution of diagnostic delay between different scheduling policies. The diagnostic delay indicates the number of months a patient's true glaucoma progression went undetected before the next scheduled visit (the lower the better). The vertical line in each box plot represents the 50% percentile diagnostic delay. The cross, "+", represents the mean diagnostic delay. The four policies had mean diagnostic delays of 3.89 for RL policy 1, 2.63 for RL policy 2, 7.45 for 1-year fixed policy and 12.46 for 2-year fixed policy. The red "×" represents the mean after removing the

outliers (5 for RL policy 1 and 3 for RL policy 2) for each respective RL policy. After the removal, the means are 2.99 (a decrease from 3.89 for RL policy 1) and 2.11 (a decrease from 2.63 for RL policy 2). We note the removal of outliers is done only to illustrate the sizable impact a few patients' eyes have on the mean diagnostic delay. In evaluating the performance of our policies in the sections to follow, no patients' eyes were excluded.

ANOVA results confirmed the TNT policies had statistically significant differences in mean diagnostic delay ($p < 0.001$). We also conducted pairwise comparisons using paired t-tests. Family wise error rate (FWER) amongst the 5 pairwise comparisons at the $\alpha = 0.05$ confidence level was controlled using the Holm-Bonferroni method (see [33] for detail). Paired T-tests confirmed the average diagnostic delay for RL policy 2 is statistically smallest amongst the four policies. Paired t-test p-values comparisons amongst the policies were respectively $p_{(RL\ policy\ 2<RL\ policy\ 1)} = 0.016$, $p_{(RL\ policy\ 1<1-year\ fixed)} \leq 0.001$, and $p_{(RL\ policy\ 1<2-year\ fixed)} \leq 0.001$. Additionally, diagnostic delay of RL policy 1 is statistically smaller compared to both fixed interval policies ($p \leq 0.001$).

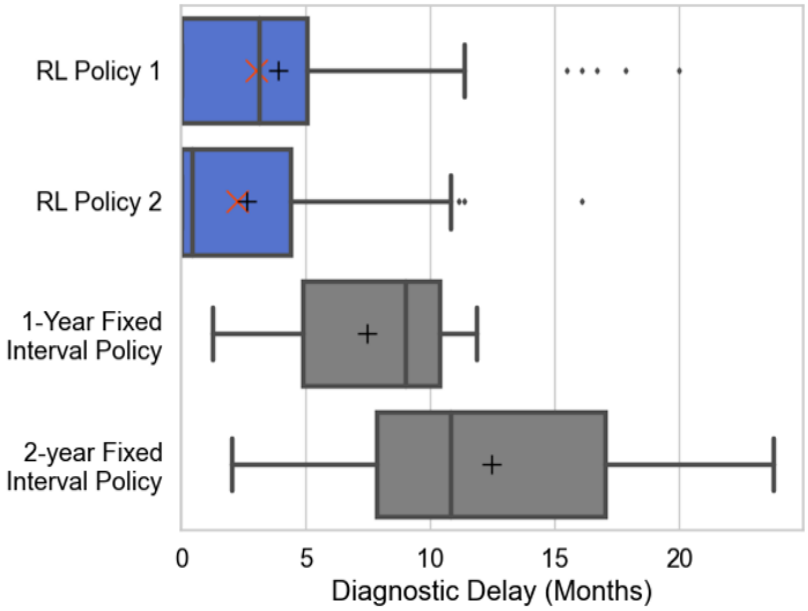


Figure 3.3: Diagnostic delay comparison amongst RL and fixed interval scheduling policies (the lower the better; only considers progressed patients). The boxes represent the twenty-fifth to seventy-fifth percentiles and the whiskers extend to the most extreme points within 1.5 the interquartile range. RL = reinforcement learning.

We also investigated the 4 follow-up scheduling policies with respect to the scheduling efficiently criterion in more detail. Figure 3.4 presents a boxplot comparison of the 4 follow-up scheduling policies' scheduling efficiency across the testing eyes. The figure visually confirms both RL policy 1 and 2 perform better than both fixed interval policies in

maximizing scheduling efficiency. The box-and-whisker plot illustrates the distribution of efficiency between different scheduling policies. The efficiency indicates the percentage of scheduled visits that show a MD loss of 3 decibels from baseline (i.e., suspected POAG progression); (the higher the better). The vertical line in each box plot represents the 5% percentile scheduling efficiency. For the four policies mean scheduling efficiency was 34% for RL policy 1, 32% for RL policy 2, 24% for 1-year fixed policy, and 22% for 2-year fixed policy.

ANOVA results confirmed the TNT policies had statistically significant differences in average scheduling efficiencies ($p \leq 0.001$). Paired t-test confirmed the average scheduling efficiency for RL policy 1 was statistically larger than policy 2 (i.e., $p_{(RL\ policy\ 1>RL\ policy\ 2)} \leq 0.040$). P-value comparisons for RL policy 2 and the 2 fixed interval policies were respectively $p_{(RL\ policy\ 2>1-year\ fixed)} \leq 0.001$, and $p_{(RL\ policy\ 2>2-year\ fixed)} \leq 0.001$. RL policy 1 scheduling efficiency was statistically larger than both fixed interval policies ($p \leq 0.001$). The two fixed interval policies had statistically the same scheduling efficiency, $p_{(1-year\ fixed>2-year\ fixed)} \leq 0.102$.

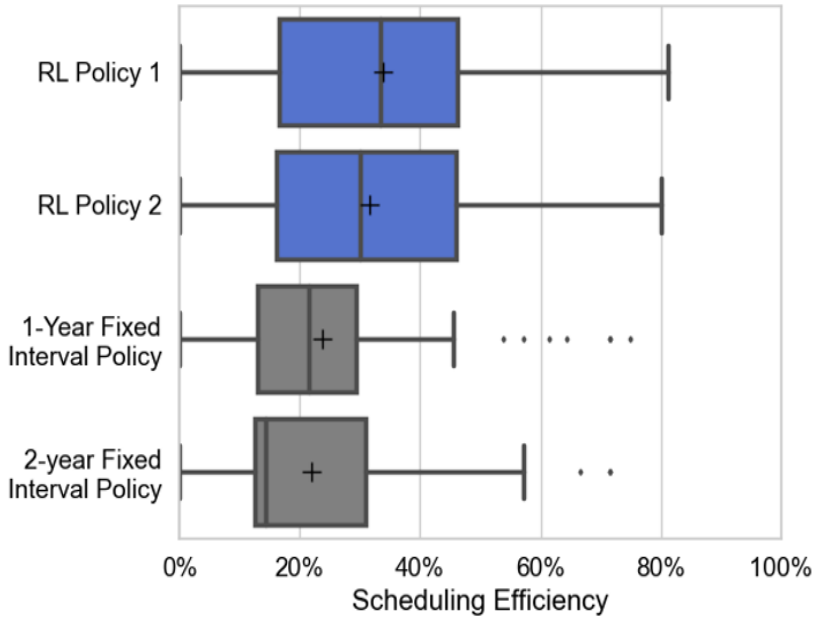


Figure 3.4: Scheduling efficiency comparison amongst RL policies and fixed interval scheduling (the higher the better; only considers progressed patients). The boxes represent the twenty-fifth to seventy-fifth percentiles and the whiskers extend to the most extreme points within 1.5 the interquartile range. RL = reinforcement learning.

3.4.2 State Feature Importance Scores

We investigated the mean feature importance scores for the Extra Trees regression models trained for RL policies 1 and 2. The importance scores are shown in Figure 3.5. The higher the feature importance score, the more important the feature was in predicting the expected reward from a chosen action (e.g., 6, 12, 18, or 24-month follow-up visit). Surprisingly while both RL policies share the same state features, they differ in how they value them. As it can be seen, Age and MD baseline differ the most. The differences are due to changes in the relative weight on the visit delay reward. It is worth noting that, while the policies differ in the way they value each of the state variables, they give the most weight to the same 7 state variables (age, MD baseline difference, MD, months form baseline, PSD, baseline MD, and PSD baseline difference). This implies these top features are the most important state variables in predicting the expected future reward from a chosen action (e.g., 6, 12, 18, or 24-month follow-up visit).

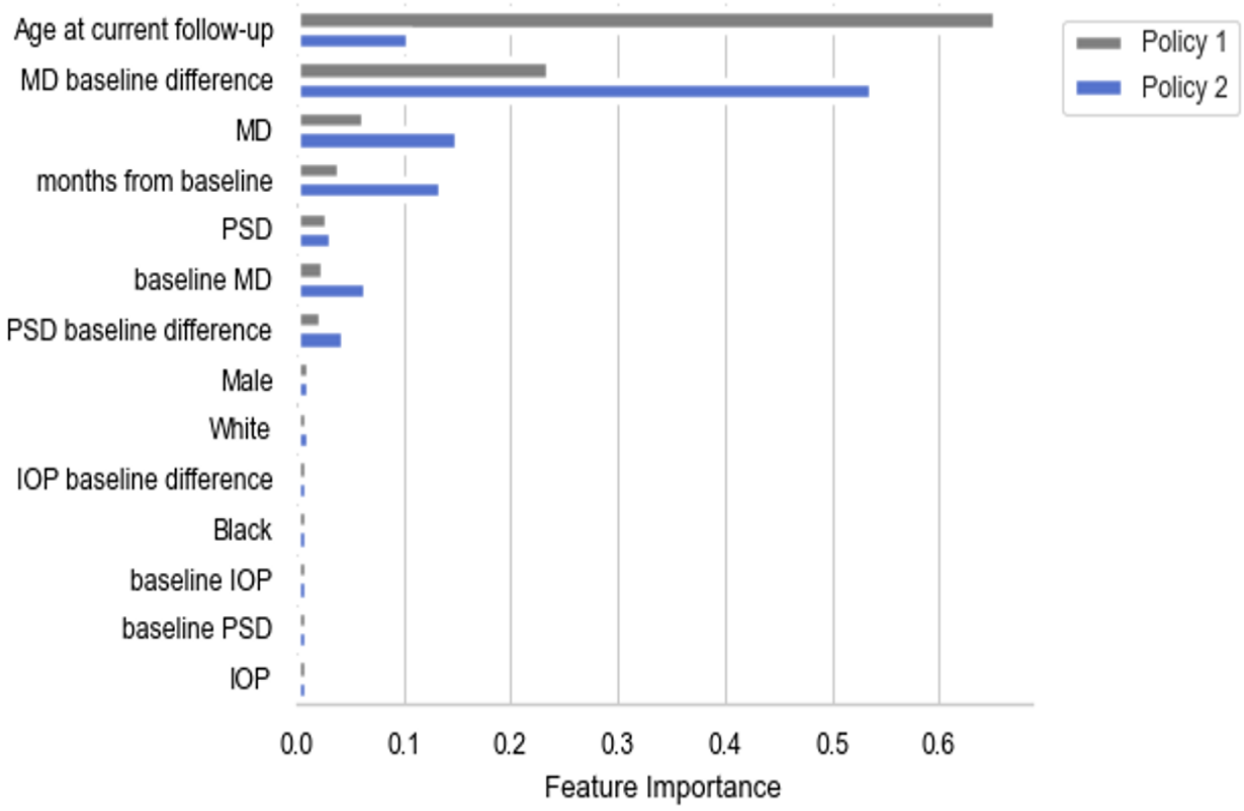


Figure 3.5: State feature importance scores for RL Policy 1 and RL Policy 2. RL = reinforcement learning.

3.4.3 RL Policies Action Evaluation

Table 3.3 summarizes the follow-up appointment recommendations, partitioned by follow-up type (e.g., action type in the MDP formulation), for RL policy 1 and RL policy 2. According to Table 3.3, progressed eyes are more likely to have at least one 6, 12, or 18-month interval than non-progressed eyes irrespective of the follow-up type. The table confirms RL policy 2 recommended a significantly greater number of follow-up visits than RL policy 1; for RL policy 1 the number of eyes that had at least one 6-month follow-up appointment recommendation was 194 (30%) compared to 545 (85%) for RL policy 2 (a 180% increase).

Table 3.3: Number of Eyes with at Least One Follow-up Appointment of Each Type

Follow-up policy	**Follow-Up type**	**Non-Progressed (n=565)**	**Progressed (n=79)**	**Overall (n=644)**
RL Policy 1	6-month	21% (n=118)	96% (n=76)	30% (n=194)
	12-month	5% (30)	20% (16)	7% (46)
	18-month	11% (61)	18% (14)	12% (75)
	24-month	100% (565)	100% (79)	100% (644)
RL Policy 2	6-month	82% (n=466)	100% (n=79)	85% (n=545)
	12-month	27% (150)	28% (22)	27% (172)
	18-month	17% (98)	20% (16)	18% (114)
	24-month	100% (565)	100% (79)	100% (644)

Additionally, to gain further insight into the inter-visit intervals prior to the date at which patient eye progression occurred, we looked at the distribution of follow-up appointment types scheduled by RL policy 1 and RL policy 2 at 6, 12, 18, and 24 months prior to the date of eye progression. The distribution of follow-up appointment types corresponding to RL policy 1 and RL policy 2 are shown in Figure 3.6. A policy recommending a follow-up appointment in 24 months, when a patient's eye is 24 months from its date of progression, implies that the next follow-up appointment will occur at the time of progression; however, we assume that shorter follow-up appointment recommendations (e.g., 6, 12, or 18-month) do not imply the patient's eye missed its progression date. For RL policy 1, the total number of follow-up appointment recommendations were 37, 38, 28, and 36 respectively for 6, 12, 18, and 24 months prior to the data of eye progression. For RL policy 2, the total number of follow-up appointments recommendations were 49, 51, 47, and 42 respectively for 6, 12, 18, 24 months prior to date of eye progression.

As it can be seen in Figure 3.6, when progressed eyes approach their progression date, for both policies, the number of 6-month follow-up appointments monotonically increases. For RL policy 1, the percentage of progressed patients assigned 6-month follow-up appointments increases from 39% at 24 months prior to progression to 95% at 6 months prior to progression. Similarly, for RL policy 2, the percentage of progressed patients with

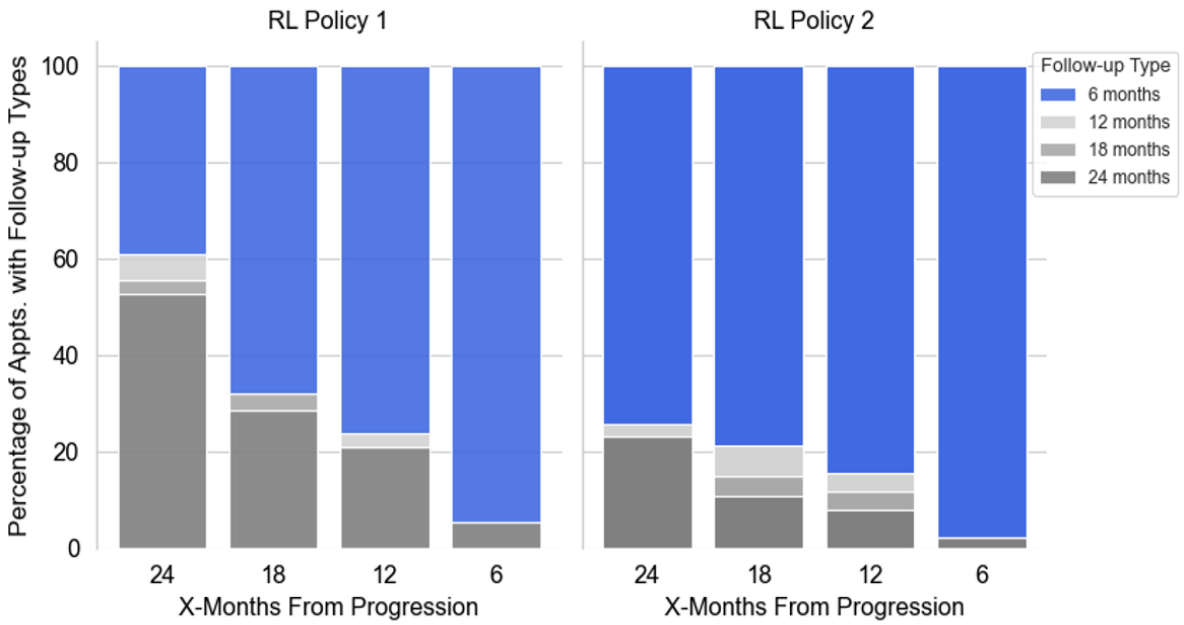


Figure 3.6: Follow-up appointment distribution for appointments made 6-24 months prior to POAG progression for follow-up types: 6,12,18, and 24-month (only progressed patients are plotted in this Figure).

6-month follow-up appointments increases from 74% at 24 months to 98% at 6-month follow-up appointments prior to progression. This implies that, as patients' eyes are near their progression date, both policies correctly schedule more 6-month follow-up visits. The likely cause of this behavior is that, as a patient's eye approach the progression date, the medical state (i.e., MD, PSD, IOP, age, etc.) is more likely to deviate from baseline, suggesting to the policy that the patient progression is near; therefore, a follow-up visit should be scheduled as soon as possible.

The policies do have a small tendency (21% for RL policy 1 and 8% for RL policy 2) to recommend 24-month follow-up visits within 12-months from progression. This is likely due to the inherent noise in ocular test measurements [30]; even as a patient nears progression, their observed test measurements can appear normal, signaling to the policies that there is no cause for concern. Hence, scheduling a follow-up appointment 24 months from their current visit is reasonable, when in hindsight a 6-month or 12-month follow-up would have been preferred.

3.5 Conclusion

As patients with chronic illnesses such as ocular hypertension require varied levels of care, personalized follow-up schedules are required. For convenience, physicians may choose to use fixed interval policies, which are clearly simpler. However, we cannot ignore the fact that fixed intervals of time are suboptimal and will often come at the cost of increased testing and hence cost; and a decrease in the patient's quality of life. It is thus in the best interest of the patient to only perform follow-up appointment visits when they are necessary. Motivated by this fact, we presented a new decision-aid tool using an offline reinforcement learning approach to schedule personalized follow-up visits for patients with chronic conditions.

We evaluated the performance of the proposed scheduling policies to schedule time to next visit (TNT) for patients with Ocular Hypertension using the Ocular Hypertension Treatment Study (OHTS) randomized clinical trial data. The experimental results indicated the two proposed policies provide better follow-up recommendations than fixed interval scheduling policies commonly used in practice. Comparing the TNT visit recommendations for our proposed RL policies with 1-, and 2-year fixed interval scheduling policies, we showed that our RL policies can detect POAG progression more efficiently (RL policies' scheduling efficiency was at least 33% larger than the best fixed interval policy's scheduling efficiency) and sooner (RL policies' diagnostic delay was at least 48% smaller than the best fixed interval policy's diagnostic delay). For patients who do not progress, our policies schedule less follow-up visits compared to those who did progress. This was as one would expect; for example, for RL policy 1, progressed patients were scheduled follow-up visits approximately every 1.0 years, while non-progressed patients were scheduled approximately every 1.6 years, consistent with the need for more intensity of monitoring for patients likely to progress. This contrasts with fixed interval policies where a constant intensity of monitoring is undertaken, regardless of whether there is a medical need to do so. Ideally, patients should only be seen by a clinician when a medical follow-up visit is warranted; however, we have required a maximum interval of 24 months as a clinically justified safeguard.

We note that our method works at the eye level, instead of the patient level. This was done intentionally. While POAG (and OHTN) generally affects both eyes, it does not always occur at the same time. Hence, patients' follow-up appointment decisions are commonly based on the visual impairment/deterioration of one eye. This does not hinder our model's ability to work with a patient that has two affected eyes. In the case of two affected eyes, the model can recommend appointment follow-ups for both eyes, and the

minimum follow-up recommendation can be used. In future work, targeting the implementation of this work, RL TNT follow-up schedules could be further personalized by altering the reward weights in the RL reward function through a user interface. Depending on the type of care a patient requires, a clinician can, in principle, modify weights to produce follow-up schedules that best adhere to their patients' needs. For instance, increasing the visit delay reward to encourage less visits, or increasing the 3 MD-drop reward to increase the likelihood of visits when suspected progression (e.g., first occurrence of 3MD drop from baseline) has occurred. In addition, it is valuable to investigate whether our results generalize to other chronic conditions and to expand our model's state by incorporating not only the current state of the patient, but also the state of the patient at prior visits. We believe this will further improve our model's performance, as historical patient information provides the model with additional patient insights.

CHAPTER 4

A Comparison of Different Approaches for Detecting Conversion from Ocular Hypertension to Primary Open-Angle Glaucoma Using Standard Automated Perimetry

4.1 Introduction

It is estimated approximately 10% of patients with OHTN will at some time progress to glaucoma, a much more serious condition. Early identification and initiation of treatment for patients with OHTN can reduce vision related morbidity and the possibility of progression to glaucoma. However, determining progression from ocular hypertension to POAG (Primary Open Angle Glacoma) can be challenging due to the inherent variability of visual field tests and the need for multiple measurements over time. Machine learning approaches to automate the detection of conversion from ocular hypertension to POAG could be useful as decision-support systems as well as in several other settings, including in tele-ophthalmology and resource-limited areas with limited access to ophthalmologists. A machine learning algorithm which predicts progression from ocular hypertension (OHTN) to primary open-angle glaucoma would have significant clinical utility. The development of such a model requires quantitative reference standards. Standard automated perimetry is integral to the diagnosis and management of POAG. However, there is currently no consensus reference standard for conversion from OHTN to POAG by perimetry.

Various approaches to define conversion have been used. Broadly, they are either trend based (using all information to generate a rate of change over time, which may then be classified as fast or slow) or event based (determining when a defined parameter surpasses a particular threshold). These approaches may use a global measure which summarizes the entire visual field (e.g., mean deviation) or regional measures which fo-

cus on subsets of the visual field (e.g., total deviation). Trend-based approaches capture longitudinal information but may be slow to identify conversion, whereas event-based approaches focus on a limited amount of available data but may identify change more rapidly. Global measures are more robust, but may miss regional change, whereas local approaches are more likely to capture true glaucomatous nerve fiber bundle defects but are also noisier. There are clear trade-offs one most consider when selecting a POAG conversion criterion to use.

Several alternative strategies to identify glaucomatous visual field progression have been proposed since the start of the OHTS (e.g., [43, 3, 39, 49]. Some criteria may classify patients with greater confidence (greater quality) whereas others may provide a larger pool of greater conversions (greater quantity). The optimal approach, however, remains unclear. The purpose of this chapter is to compare four alternative criteria to identify conversion from ocular hypertension to POAG based on visual fields changes, paying particular attention to those which identify conversion more rapidly or identify a larger cohort of eyes with POAG, as these may be useful for algorithm development. To facilitate the prediction of conversion from OHTN to POAG using machine learning methods it is essential to have qualitative reference standards. A desirable reference standard would be quick, accurate, sensitive, specific, and reproducible.

The organization of this chapter is as follows: Section 4.2 outlines the proposed methods. Section 4.3-4.4 presents and discusses discusses the numerical results. Finally, Section 4.5 concludes the chapter and discusses the future work.

4.2 Methods

This study used data from the Ocular Hypertension Treatment Study (OHTS) [2], which included 3,265 eyes from 1,636 patients with OHTN from 22 US centers who were followed between February 1994 and March 2009. In OHTS phase 1 (February 1994 to June 2002), patients were randomized to treatment with ocular hypotensive medication or observation and were followed for a mean of 7.5 years. In OHTS phase 2 (June 2002 to March 2009), all patients were treated with ocular hypotensive medication and followed for a mean of 5.5 additional years. Beginning with the patients initial baseline visit SITA Standard (SAP) Humphrey 30-2 visual fields (Carl Zeiss Meditec; Dublin, CA) were obtained twice per year. At baseline all patients had normal peripheral vision on SAP. Nine patients were excluded from this analysis because complete data for both eyes were not unavailable. All study participants provided informed consent. This study was approved by the Washington University and University of Michigan institutional review boards and

faithfully adhered to the tenets of the Declaration of Helsinki.

4.2.1 Conversion Criteria Studied

We compared four criteria researchers have used previously to identify conversion of an eye from OHTN to POAG based on perimetry: (A) the OHTS Endpoint Committee, (B) 3-MD drop from baseline, (C) TD trend analysis, and (D) Deterioration of Points on Total Deviation Clusters (cluster deterioration). The definitions are as follows:

1. *Endpoint*: The approach used in OHTS for progression to POAG by visual fields. If three consecutive sets of visual fields demonstrated change (corrected pattern standard deviation decline at the $p<0.05$ level or glaucoma hemifield test outside normal limits) these fields were evaluated by the Visual Field Reading Center (VFRC). If an abnormality was found to be consistent in character and location, all information including medical history, visual fields, and optic disc photographs for both eyes were reviewed by the OHTS Endpoint Committee. The OHTS Endpoint committee was composed of 3 glaucoma specialists. The committee was tasked with reviewing the report from the VFRC along with other relevant clinical information to determine if the abnormality was attributable to POAG. Attribution to POAG was made by determining if a consistent depression in pattern deviation compatible with a nerve fiber bundle defect, paracentral scotoma, or nasal step was found and consistent with the patient's other clinical information.

2. *3-MD Drop from Baseline*: Conversion from OHTN to POAG defined as 2 consecutive MD test measurements at least 3 DB below baseline. Baseline MD was defined as the average of the patient's two initial MD values prior to enrollment in the clinical trial [49]. If progression was confirmed on 2 consecutive MD test measurements, the timing of conversion was noted as the time of the second MD measurement. Note: This was the POAG conversion definition used in chapter 3.

3. *TD Trend Analysis*: This is a trend-based assessment, first described by Kummet and colleagues [43], involving performing pointwise linear regression on each of the 52 visual field test locations in the total deviation plot. For each time point, t, starting at the third measurement and ending at the penultimate measurement, a linear regression model of Total Deviation versus time was fit. If the slope was less than -1.2 dB/year and the p-value comparing this slope to the null hypothesis of a slope of 0 was <0.04, this procedure was repeated by fitting the linear regression model using the previous data in addition to data from the following time point (e.g.,

the first four observations) as well as data from the following two time points (in this case, the first five observations) for confirmation. If progression was confirmed, the timing of first evidence of progression was determined to be the initial time point in the series of time, t, $t+1$, and $t+2$. Otherwise, this procedure was continued until the last time point was reached. Total Deviation was assessed rather than Pattern Deviation because it has been shown to provide slightly better performance in determining progression for patients with ocular hypertension [3].

4. *Clusters of points that deteriorate based (Cluster Deterioration)*: The total deviation (TD) probabilities were extracted for each visual field location. An eye was flagged as "abnormal" at the 5% level on the first date at which at least five locations had TD probability < 5% (i.e., any of "< 5%", "< 1%", or "< 0.5%"). An eye was flagged as "confirmed abnormal" on the first date at which there were at least five locations with total deviation probability < 5%, and those same five locations had probability < 5% on the next test date. An eye was flagged as "persistent abnormal" on the first date at which at least five locations had TD probability < 5% on three consecutive test dates (the same five locations at each date). The exercise was repeated to find the first dates at which abnormality, confirmed abnormality, and persistent abnormality, occurred using the 1% and 0.5% confidence limits [39]. TD was assessed rather than Pattern deviation because it has been shown to provide slightly better performance in determining progression for patients with ocular hypertension [3]. An eye was labeled as converted if it was flagged as "persistent abnormality" at the 0.005 (0.5%) level. The timing of conversion was noted as the date at which the persistent abnormality was observed.

4.2.2 Analysis

Participant characteristics were summarized for the entire sample using means and SDs for continuous variables and frequencies and percentages for categorical variables. Ordinary least square regression (OLSR) slope p-values were calculated using the student's t-statistic and a null hypothesis of $\beta = 0$. For each eye and each conversion criterion, we recorded whether conversion occurred and either the first date of conversion (event) or the date of the last perimetric test (right censored). Time was measured as days since enrollment in OHTS. An patient's eye could convert according to 0 to 4 of the criteria during its' follow-up history.

4.2.2.1 Survival Analysis

Univariable, and multivariable survival analysis was conducted. For the univariable case, cumulative incidence curves for each POAG conversion definition were drawn (with 95% confidence intervals to model the uncertainty in the point estimates of the univariable survival functions [58]) based on Kaplan-Meier estimates and compared using pairwise Paired Prentice-Williams Tests. The Bonferroni correction was applied to p-values. The univariable survival curves can be seen in Figure 4.2.

For the multivariate case, for each progression criterion, a multivariate Cox proportional hazard model [8] was used to estimate the survival functions (or more importantly the log-hazard functions) associated with age, race, and sex. Confidence intervals were built using robust standard errors as each patient had two eyes in the dataset (i.e. we control for the correlations introduced by having both eyes from same patient in our data set). Whether the associations between time to conversion and age, race, and sex differed by progression type were investigated using multivariate Cox models, stratified by conversion criteria and with frailty terms allowing correlations between times to conversion within the same patient. Likelihood ratio tests of the interaction terms between type of conversion and the other fixed effects were used to determine whether covariate effects differed by conversion criteria. The multivariable survival coefficients for each of the four POAG conversion criteria models is in Table 4.3.

All statistical analyses were conducted using Python, version 3.6, and R version 3.6.2. Survival analysis models (univariable and multivariable) were developed using the python lifelines library and R survival package.

4.2.2.2 Ties for Earliest Conversion

The various test modalities were not all taken on the same day and visit in every case, so two or more different criteria could signal conversion at times that were close yet different. Multiple earliest conversions according to different criteria were considered to occur at the same time if a patient met multiple conversion criteria within 30 days of their earliest conversion. For example, if a patient's earliest conversion was 3-MD drop from baseline at time t, and within $t + 30$ days also met the criterion for TD Trend analysis, the patient was listed as meeting two conversion criteria for earliest conversion. A 30-day resolution was deemed an acceptable level of aggregation, because the tests were taken on roughly 6-month intervals. Table 4.1 outlines the mean time until earliest conversion, including ties, for eyes that were observed to had converted. Table 4.2 outlines the mean time until conversion analyzing each conversion criteria separately. We note because the mean

time until conversion is not adjusted for the eyes that were not observed to had POAG converted the times listed may be underestimated.

Table 4.1: Mean time until conversion to POAG under the criterion the earliest triggered conversion (only considers eyes that progressed)

Overall – 3256 eyes [1]	N eyes that earliest converted this way	Mean/SD time until first conversion (days)
3 MD drop	265	2469 (1387)
TD Trend Analysis	600	1518 (544)
Cluster Deterioration	117	2567 (1485)
Endpoint (Visual Field)	79	2187 (1382)

Table 4.2: Mean time until conversion to POAG, analyzing each definition separately (only considers eyes that progressed)

Overall – 3256 eyes	N eyes converted this way	Mean/SD time until conversion (days)
3 MD drop	423	2692 (1333)
TD Trend Analysis	731	1664 (757)
Cluster Deterioration	341	2816 (1321)
Endpoint (Visual Field)	127	2557 (1393)

4.3 Results

The mean ± SD age at baseline was 57 ± 10 years. 701 patients (43%) were male, and 927 (57%) were female. 1134 (70%) were white, 403 (24.5%) black, and 91 (5.5%) were another race. Patients underwent a mean ± SD of visual 22 ± 9 visual fields over 10.5 ± 4.1 years of follow-up. Their 3256 eyes had mean ± SD baseline MD of 0.17 ± 1.15 dB.

The number of eyes with evidence of progression by at least 1, 2, 3, and all 4 criteria were 1002 (30.8%), 421 (12.9%), 187 (5.7%), and 12 (0.3%), respectively. Of the 1002 eyes which progressed by at least one criterion, 731 (73%), 423 (42%), 341 (34%), and 127 (13%) converted by TD trend, 3 MD drop, cluster deterioration, and endpoint criteria, respectively. Of the 421 eyes which progressed by at least two criteria, 353 (84%), 283 (67%), 303 (72%), and 102 (24%) converted by TD trend, 3 MD drop, cluster deterioration, and endpoint criteria, respectively. And, of the 187 eyes which progressed by at least three criteria, 182 (97%), 143 (76%), 182 (97%), and 66 (35%) converted by TD trend,

[1] There was a total of 59 ties for earliest conversion (i.e. earliest conversion occurred within 30 days of another conversion): 3 MD drop (25), TD trend analysis (25), Endpoint (5), and Cluster Deterioration (4).

3 MD drop, cluster deterioration, and endpoint criteria, respectively. A total of 293 (18%) patients had both eyes exhibit conversion by 1 or more criteria.

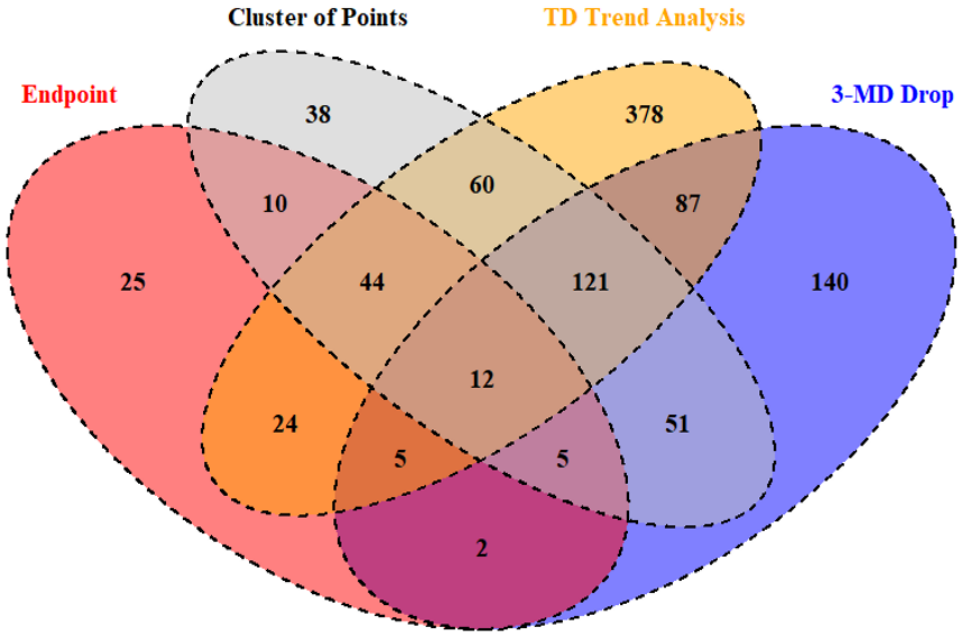


Figure 4.1: Agreement Among the Four Conversion Criteria. MD = mean deviation; TD = total deviation. Venn diagram comparison of the four criteria. The numbers represent the number of eyes that conversion due to each conversion criteria. The criteria are defined as (1) 3MD Drop – OAG conversion due to two successive MD measurements were found to be at least 3 dB less than baseline. (2) Endpoint – OAG conversion as noted by endpoint committee. (3) Cluster Deterioration - Conversion to OAG by identifying five TD locations where total deviation probability $<$5% on three consecutive test dates (the same five locations at each date). (4) TD Trend - Conversion from OHTN to POAG determined by total TD Ordinary Least Squares Regression point slopes of ≤ -1.2dB per year and p-value ≤ 0.04 for 3 consecutive periods (i.e., $\beta(t) \leq -1.2$, $\beta(t+1) \leq -1.2$, and $\beta(t+2) \leq -1.2$) at two or more of the 52 TD point locations.

4.3.1 Criteria Conversion Comparisons to Endpoint Criterion

We choose to only discuss the conversion agreement (and disagreement) of TD trend, 3 MD drop, and cluster of points had with the Endpoint criterion because the Endpoint criterion was the most involved criteria amongst the four; for the endpoint criterion before an eye could be labeled as POAG converting an endpoint committee, composed of a panel of medical experts, had to come a majority consensus. Hence, the Endpoint criteria can be seen as the most conservative criterion, and potentially the closest, amongst

the four criteria, towards conversion reference standard for identifying true POAG patient conversions.

The TD trend criterion had the greatest agreement with the Endpoint conversion criterion amongst the three criteria (e.g., TD trend, 3 MD drop, and cluster of points) with identifying a total of 85, out of a possible 127, endpoint conversions. This indicates 76% of all visual field endpoint conversions could have been identified using the TD trend conversion criterion. This is important because the TD trend criterion has advantages over the endpoint criterion in that it is a simpler approach to use and can identify conversion sooner. Endpoint, while possibly being the criterion with the highest specificity (i.e. high true positive rate) in identifying true POAG conversions, is in large part impractical to use in a clinical setting because of its reliance on a panel of experts to make a clinical decision of whether progression occurred. However, it should be noted that while TD trend has the highest agreement with endpoint, it also indicated the highest number of POAG conversions (731) out of all criteria. Hypothetically, this could mean that while TD trend may have a high sensitivity , it may suffer from low specificity. There is no data or research to confirm this, but if true, it would suggest that a mixed approach may be warrented.

Cluster of points had the second highest agreement with endpoint criteria with a total of 71 out of possible 127 endpoint conversions, or 56% conversion agreement. A key difference between the cluster of points criteria and the TD trend criterion is that the former indicated 47% less conversions. Hypothetically, this could mean that while the conversion agreement between cluster of points and endpoint is lower than the agreement between TD trend, cluster of points may not suffer from low specificity as much as TD trend.

3-MD drop had the lowest agreement with endpoint with a total of 24 out of 121 endpoint conversions, roughly a 20% agreement. For a detailed breakdown of conversion agreement amongst the criteria see Figure 4.1.

4.3.2 Survival Analysis

All four criteria's time to conversion were statistically different from one another in pairwise comparisons of the unadjusted survival curves (Bonferroni-corrected paired Prentice-Williams tests p-value < 0.001; Figure 4.2). TD trend exhibited the highest estimated proportion of conversion events over 160 months (731). In contrast, the OHTS Endpoint Committee criterion had the lowest (127).

The proportion of eyes converting within 10 years (120 months) of OHTS enrollment compared to total number of conversion for endpoint, 3-MD drop, TD trend, and cluster deterioration were respectively 74% (n=94), 71% (299), 96% (702), and 69% (234).

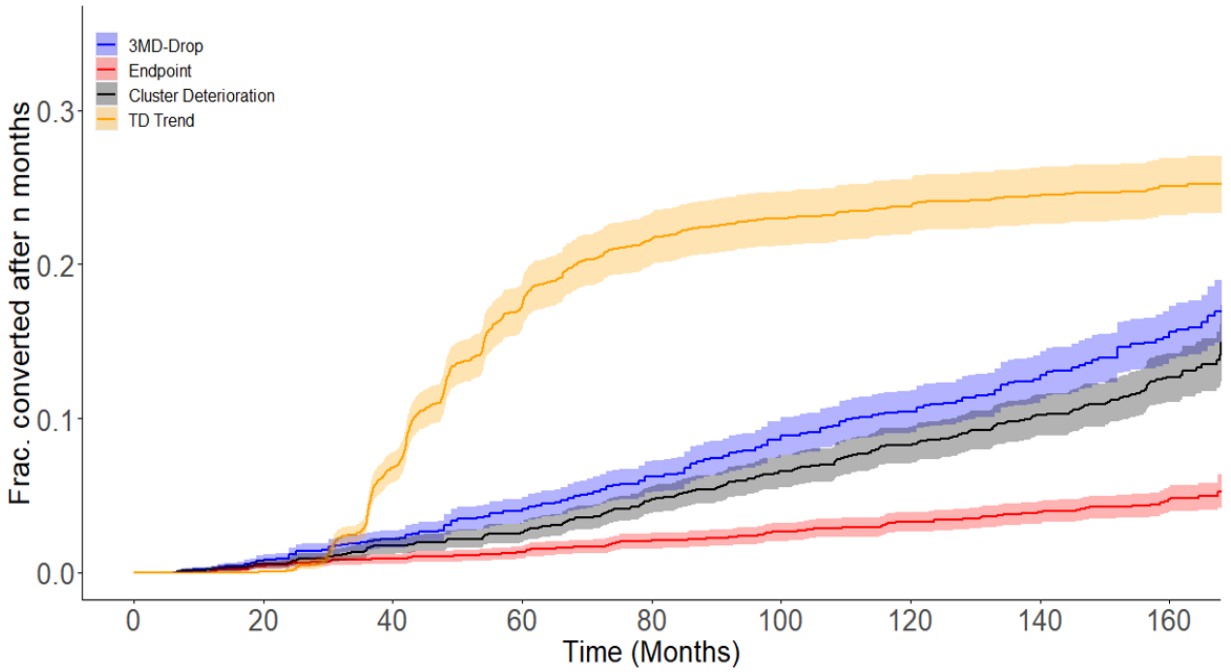


Figure 4.2: Kaplan-Meier Survival Plot for Time to First Conversion for Each Criterion. MD = mean deviation; TD = total deviation. Survival curves of the four criteria. The shaded regions represent confidence regions of the univariable survival models. The criteria are defined as (1) 3MD Drop – OAG conversion due to two successive MD measurements were found to be at least 3 dB less than baseline. (2) Endpoint – OAG conversion as noted by endpoint committee. (3) Cluster Deterioration - Conversion to OAG by identifying five TD locations where total deviation probability $< 5\%$ on three consecutive test dates (the same five locations at each date). (4) TD Trend - Conversion from OHTN to OAG determined by total TD Ordinary Least Squares Regression point slopes of ≤ -1.2 dB per year and p-value ≤ 0.04 for 3 consecutive periods (i.e., $\beta(t) \leq -1.2, \beta(t+1) \leq -1.2,$ and $\beta(t+2) \leq -1.2)$ at two or more of the 52 TD point locations.

In multivariable Cox regression the effects of race, sex, and age on each conversion type depended statistical significantly on the conversion criterion (likelihood ratio tests p-value < 0.05 each covariate). For example, while all conversion types show black patients have a higher relative risk of conversion compared to white patients, for endpoint conversion the hazard ratio greatest. This indicates race is a stronger indicator of time-to-conversion under the endpoint committee criterion compared to the other conversion criteria. Effects, therefore, are reported separately for each conversion criterion (Table 4.3).

For all four criteria, older (vs younger) age was associated statistical significantly with considerably higher hazard of converting. The magnitude of this association was much greater for 3 MD and Cluster deterioration compared to the other two criteria. Blacks

had 22%, 39%, 124%, and 87% statistical significant higher hazard of conversion compared with whites for the TD trend, 3-MD drop, Endpoint, and cluster deterioration criteria, respectively. Other race had 18%, 37%, 17%, and 70% higher hazard of conversion compared with whites for the TD trend, 3-MD drop, Endpoint, and cluster deterioration criteria, respectively. Only the results for TD trend and cluster deterioration were statistically significant. The hazard of conversion for females relative to males was either higher (MD Decline), lower (endpoint, cluster deterioration), or about the same (TD trend).

4.4 Discussion

This chapter compares several criteria for determining progression from OHTN to POAG on perimetry. Identification of conversion to POAG by changes on perimetry is an attractive approach as perimetry is integral to the management of POAG. Large volumes of these data are available in clinical trial data and data repositories and are amenable to automated analysis. Early detection of progression to POAG can enrich research samples, which may be useful for the development of automated algorithms. Alternatively, more conservative criteria may result in more reliable classification.

Several POAG conversion criteria were compared. The criteria were two global, event-based approaches which assess different visual field summary statistics (the OHTS endpoint criterion and the 3-MD drop criterion), a pointwise event-based approach which assesses change across visual field test points (cluster deterioration), and a pointwise trend-based approach which identifies worsening at individual test points (TD trend).

Race, sex, and age were statistically significant determinants of progression for all four criteria. However, there was broad discordance between the four criteria, as demonstrated by the statistically different survival curves and the limited overlap demonstrated in Figure 4.1. This suggests that these criteria may be tailored to the particular type of damage, which is under investigation, particularly in the absence of a visual fields-based reference standard. In addition, the multivariable survival analysis (multivariable box cox regression) indicated: (i) for all criteria blacks have the highest risk of progressing from OHTN to POAG; (ii) the older the patient, the higher the risk of glaucoma progression; (iii) the risk of conversion to POAG is statistically the same for both males and females for all criteria except endpoint; and (iv) the hazard coefficients for race, sex, and age differ significantly from one another depending on the conversion criterion used (likelihood ratio tests p-value less than or equal to 0.05 for each covariate). The last point is particularly interesting because it implies the rate at which a patient will be successfully identified as POAG converting is in part, related to their demographics. Suggesting the type of POAG

Table 4.3: The Association of Demographic Factors with the Hazard of Progression from Ocular Hypertension to Glaucoma Using Each Progression Criterion. MD = mean deviation; TD = total deviation; UCI = Upper 95% confidence interval; LCI = lower 95% confidence interval. The reference categories were race is white, sex is male, and age is between 40 and 50. The criteria are defined as (1) 3MD Drop – OAG conversion due to two successive MD measurements were found to be at least 3 dB less than baseline. (2) Endpoint – OAG conversion as noted by OHTS Endpoint Committee. (3) Cluster Deterioration - Conversion to OAG by identifying five TD locations where total deviation probability $< 5\%$ on three consecutive test dates (the same five locations at each date). (4) TD Trend - Conversion from OHTN to OAG determined by total TD Ordinary Least Squares Regression point slopes of ≤ -1.2 dB per year and p-value ≤ 0.04 for 3 consecutive periods (i.e. $\beta(t) \leq -1.2$, $\beta(t+1) \leq -1.2$, and $\beta(t+2) \leq -1.2$) at two or more of the 52 TD point locations.

Criteria	Comparison	Hazard Ratio	LCI	UCI	P-Value
TD Trend Analysis	Black vs White	1.22	1.00	1.47	0.04
	Other vs White	1.18	0.78	1.78	0.04
	Female vs Male	0.98	0.82	1.15	0.77
	50s vs 40s	1.51	1.17	1.94	<0.01
	60s vs 40s	2.48	1.93	3.16	<0.0001
	70s/80s vs 40s	3.17	2.38	4.20	<0.0001
3-MD Drop from Baseline x2	Black vs White	1.39	1.06	1.80	0.01
	Other vs White	1.37	0.82	2.29	0.22
	Female vs Male	1.23	0.97	1.55	0.07
	50s vs 40s	2.55	1.66	3.89	<0.0001
	60s vs 40s	4.99	3.30	7.52	<0.0001
	70s/80s vs 40s	9.28	6.01	14.34	<0.0001
Endpoint	Black vs White	2.24	1.57	3.17	<0.0001
	Other vs White	1.17	0.47	2.88	0.73
	Female vs Male	0.50	0.35	0.70	<0.0001
	50s vs 40s	1.86	1.06	3.24	0.028
	60s vs 40s	2.58	1.49	4.47	<0.0001
	70s/80s vs 40s	4.87	2.70	8.78	<0.0001
Cluster Deterioration	Black vs White	1.87	1.42	2.45	<0.0001
	Other vs White	1.70	1.00	2.89	<0.01
	Female vs Male	0.92	0.71	1.18	0.52
	50s vs 40s	2.51	1.56	4.04	<0.0001
	60s vs 40s	5.07	3.21	8.00	<0.0001
	70s/80s vs 40s	9.79	6.01	15.92	<0.0001

conversion criterion used by a clinician should hypothetically differ based on the patients' demographics.

Notably, all criteria demonstrated a degree of dis-concordance with the OHTS endpoint criteria, which may, at least in part, reflect changes in the means of evaluating visual fields since the commencement of the OHTS in the mid-1990s. The OHTS endpoint criterion was the most conservative approach, identifying only 3% of eyes as POAG converting. It should be noted though, evidence of progression by the OHTS endpoint criterion should always be treated as a true POAG conversion, however as discussed previously the endpoint criterion only detects a very small number of POAG conversion, requires lengthy time to identify progression, and requires skilled graders. Other approaches detect a greater percentage of POAG conversions and detect conversion a lot sooner, though some of these conversions may not be clinically meaningful (i.e. false positive POAG conversions). Researchers should consider these trade-offs when deciding which criteria to employ into POAG conversion machine learning prediction algorithms.

4.4.1 Limitations

A key motivating factor for this analysis, the absence of a reference standard, also imposes a necessary limitation, as there is no way to assess relevant characteristics including sensitivity and specificity of each approach. The OHTS endpoint criterion is unique among the four criteria in that additional information including optic disc photographs and clinical information were used by the Endpoint Committee in their determination of progression to POAG. In fact, approximately half of the eyes with suspicious visual fields were not confirmed to have developed POAG [27], which may explain at least part of the discordance between this criterion and the remaining criteria.

Additionally for this work we focused primarily on functional measurements of visual field acuity (e.g. MD, TD, etc.); we only consider functional damage on perimetry for identifying POAG conversion. We do not consider measurements that measure structural damage (optical coherence tomograpy (OCT) measurements, retinal nerve fiber layer (RNFL) measurements, etc.). We acknowledge this research gap.

4.5 Conclusion

Different criteria capture different characteristics of progression to POAG. Machine learning models may benefit from using multiple definitions of conversion, as there was clear discordance among the compared criteria. The choice of criteria may be tailored based

on the desired outcome measure. More liberal criteria (e.g., TD trend) may be used to determine conversion from ocular hypertension to POAG earlier or in a greater number of eyes, whereas more conservative criteria (e.g., endpoint) may be used to develop models which can identify conversion with greater confidence. One question our results, which cannot be answered at this time, of whether the Endpoint, which the study team intended to be quite conservative, is closer or further from the truth relative to any of the other criteria. Given there is currently no reference gold standard (e.g., a conversion criterion that has been shown to accurately identify all converting patients) a hybrid approach combing multiple conversion criteria may be the next best option. Nonetheless, to permit machine learning models to predict conversion from ocular hypertension (OHTN) to primary open-angle glaucoma (POAG) it is essential to have quantitative reference standards for conversion for researchers to use. More research is needed to further examine the multitude of conversion criteria used in clinical settings for determining conversion to POAG.

In future work, we plan to use the outlined criteria to determine if they can be used to proactively identify whether a patient will convert to POAG at some point in the future. Correctly identifying converting patients is important as it signals to the clinician treatment may be needed. However, being able to proactively predict converting patients enables the clinician to take preventative measurements before the patient converts to POAG, at which point unreversible vision loss may have already occurred. One question, which could not be answered, of whether the Endpoint criterion, which the study team intended to be quite conservative, is closer or further from the truth relative to any of the other criteria.

CHAPTER 5

Machine Learning Prediction of Conversion from Ocular Hypertension to Primary Open Angle Glaucoma

5.1 Introduction

One of the key strengths of machine learning approaches is their ability to automate, the otherwise manual task of detecting a patient's conversion from ocular hypertension to POAG (Primary Open-Angle Glaucoma). At present, there exist several alternative criteria to identify POAG conversion (e.g., [43, 3, 39, 49]), but notably no gold standard. Tradeoffs amongst criteria are often made to balance a criterion's sensitivity (measures the proportion of converting patients that are correctly identified as converting) and specificity (measures the proportion of non-converting patients that are correctly identified as non-converting). Since correctly identifying converting patients often takes precedence over incorrectly labeling a non-converting patient as converting, a criterion that has a high sensitivity is often preferred.

As such, we elucidate and study several conversion criteria, and we also develop a machine learning (ML) method to maximize the number of patients identified as converting. The criteria are as follows: (A) the OHTS Endpoint – Committee - this was the approach used in OHTS for progression to POAG by visual fields; (B) Decline in Mean Deviation - conversion from OHTN to POAG defined as 2 consecutive MD values at least 3 DB below baseline. Baseline MD was defined as the average of the patient's two initial MD values prior to enrollment in the clinical trial [49]; (C) Decline in Pointwise Total Deviation- this is a trend-based assessment, first described by Kummet and colleagues [43], involving performing pointwise linear regression on each of the 52 visual field test locations in the total deviation plot; and (D) Deterioration of Points on Total Deviation Clusters – this is a clustered based approach developed by [39].

The main purpose of this chapter is to determine the predictive performance of a collective conversion criterion (i.e., conversion criteria combining all 4 criteria). We use predictive classification algorithms to determine if a patient will convert to POAG, as defined by at least one of the four criteria indicating POAG conversion anytime within two years from their sixth visit. Several supervised learning classification models were assessed: Logistic Regression, Random Forest, Gradient Boosted Decision Tree, and Neural Network. Results indicated the random forest classifier performed best. The classification models that included the Kalman filtered data had slightly better performance than the models that did not (about 1% improvement across all testing metrics). For the Random Forest that included the Kalman filtered covariates the group 5-fold cross validation ROC AUC performance was mean $\pm$ SD of 0.81 ± 0.01. The testing performance across all evaluation metrics was a balanced accuracy of 0.80, sensitivity of 0.77, specificity of 0.83, accuracy of 0.82, positive predicted value of 0.38, negative predicted value 0.96, and ROC AUC of 0.86. To the best of our knowledge, we are the first to combine the four conversion criteria and assess their collective prediction performance.

The organization of this chapter is as follows: Section 5.2 outlines the proposed methods. Section 5.3-5.4 presents and discusses the numerical results. Section 5.5 provides a discussion of results, and last section 5 concludes chapter.

5.2 Methods

5.2.1 Data

The data used for this work comes from the Ocular Hypertension Treatment Study (OHTS), conducted between February 1994 and March 2009 ([29, 28]. In OHTS, 1636 patients with ocular hypertension (OHTN) from 22 US centers were randomized to either treatment with IOP-lowering medications or followed without treatment. Patients were followed for up to 15 years based on the judgement of the endpoint committee using standardized measurements of tonometry and perimetry starting at baseline and every 6 months thereafter. The study was approved by the University of Michigan and Washington University Institutional Review Boards and patients provided written informed consent.

5.2.2 Sample Selection

To be eligible, OHTS participants had baseline IOPs of 21 to 32 mm Hg in both eyes, reliable and normal 30-2 Humphrey visual fields (Carl Zeiss Meditec), and no detectable

evidence of glaucoma based on optic nerve evaluation. In our study, patients were censored when they experienced non glaucomatous visual field loss or underwent any incisional intraocular surgery other than uncomplicated cataract surgery. Patients required at least 6 follow-up visits from baseline. The 6 follow-up measurements ensured the Kalman Filters, developed according to the procedure in [22], could forecast using at least 3 warm-up measurements and validate at least 12 months into the future. Patients with only one eligible eye under the above inclusions were excluded.

5.2.3 Definition of POAG Conversion Types

Patients were labeled as OAG progressed if they experienced one of the following progression criteria two years from their sixth visit:

1. *Endpoint*: The approach used in OHTS for progression to POAG by visual fields. If three consecutive sets of visual fields demonstrated change (corrected pattern standard deviation decline at the $p<0.05$ level or glaucoma hemifield test outside normal limits) these fields were evaluated by the Visual Field Reading Center (VFRC). If an abnormality was found to be consistent in character and location, all information including medical history, visual fields, and optic disc photographs for both eyes were reviewed by the OHTS Endpoint Committee. The OHTS Endpoint committee was composed of 3 glaucoma specialists. The committee was tasked with reviewing the report from the VFRC along with other relevant clinical information to determine if the abnormality was attributable to POAG. Attribution to POAG was made by determining if a consistent depression in pattern deviation compatible with a nerve fiber bundle defect, paracentral scotoma, or nasal step was found and consistent with the patient's other clinical information.

2. *3-MD Drop from Baseline*: Conversion from OHTN to POAG defined as 2 consecutive MD test measurements at least 3 DB below baseline. Baseline MD was defined as the average of the patient's two initial MD values prior to enrollment in the clinical trial [49]. If progression was confirmed on 2 consecutive MD test measurements, the timing of conversion was noted as the time of the second MD measurement. Note: This was the POAG conversion definition used in chapter 3.

3. *TD Trend Analysis*: This is a trend-based assessment, first described by Kummet and colleagues [43], involving performing pointwise linear regression on each of the 52 visual field test locations in the total deviation plot. For each time point, t, starting at the third measurement and ending at the penultimate measurement, a

linear regression model of Total Deviation versus time was fit. If the slope was less than -1.2 dB/year and the p-value comparing this slope to the null hypothesis of a slope of 0 was $<$0.04, this procedure was repeated by fitting the linear regression model using the previous data in addition to data from the following time point (e.g., the first four observations) as well as data from the following two time points (in this case, the first five observations) for confirmation. If progression was confirmed, the timing of first evidence of progression was determined to be the initial time point in the series of time, t, $t+1$, and $t+2$. Otherwise, this procedure was continued until the last time point was reached. Total Deviation was assessed rather than Pattern Deviation because it has been shown to provide slightly better performance in determining progression for patients with ocular hypertension [3].

4. *Clusters of points that deteriorate based (Cluster Deterioration)*: The total deviation (TD) probabilities were extracted for each visual field location. An eye was flagged as "abnormal" at the 5% level on the first date at which at least five locations had TD probability < 5% (i.e., any of "< 5%", "< 1%", or "< 0.5%"). An eye was flagged as "confirmed abnormal" on the first date at which there were at least five locations with total deviation probability < 5%, and those same five locations had probability < 5% on the next test date. An eye was flagged as "persistent abnormal" on the first date at which at least five locations had TD probability < 5% on three consecutive test dates (the same five locations at each date). The exercise was repeated to find the first dates at which abnormality, confirmed abnormality, and persistent abnormality, occurred using the 1% and 0.5% confidence limits [39]. TD was assessed rather than Pattern deviation because it has been shown to provide slightly better performance in determining progression for patients with ocular hypertension [3]. An eye was labeled as converted if it was flagged as "persistent abnormality" at the 0.005 (0.5%) level. The timing of conversion was noted as the date at which the persistent abnormality was observed.

A comparison study between the conversion criteria listed can be found in chapter 4.

5.2.4 Variables of Interest

5.2.4.1 OHTS Variables

OHTS variables used in the classification models were a combination of visual field variables (e.g., mean deviation (MD), pattern standard deviation (PSD), total deviation pointwise plots (TD), intraocular eye pressure (IOP), demographic information (e.g., age, sex,

ace), medication information (e.g., ocular, and non-ocular medications), family medical history (i.e., does the patient have a parent or sibling with glaucoma), and feature engineered variables. In total 107 variables.

5.2.4.2 Kalman Filtered Variables

A Kalman filter (KF) is a statistical model that primarily functions as a measurement denoiser and measurement forecaster. The KF accomplishes these tasks by leveraging population level and patient-level disease dynamics. Unlike traditional forecasting techniques, KF uses past prediction errors to correct future estimates, dynamically updating forecasts as new measurements become available [44]. KFs were used to denoise and forecast mean deviation (MD), pattern standard deviation (PSD), and intraocular eye pressure (IOP) measurements. For the sake of brevity, we denote the KF denoised measurements as filtered MD, filtered PSD, and filtered IOP. The KFs were developed according to the procedure outlined in [22]. The filtered measurements (filtered MD, filtered PSD, filtered IOP, and functions thereof) were used as additional variables for the prediction model. A complete list of all variables, including the OHTS variables, can be found in Appendix E.1.

Additionally, because the KF required equally spaced measurements (e.g., measurements equally spaced 6-months apart), MD, IOP, and PSD were numerically interpolated on 6-month intervals using linear interpolation. The implication of this is every patient's 6th visit occurs at the same time relative to their OHTS clinical trial start date.

5.2.5 Predicting conversion to Primary Open Angle Glaucoma (POAG)

Supervised classification models were developed to use data from both eyes over the first 6 follow-up visits to predict whether patients will convert from OHTN to POAG within 2 years from their 6th follow-up visit. Figure 5.1 gives a visual illustration of this. As previously mentioned, conversions events were not identified sooner (i.e., before the patient's 6th visit) because the KFs require the first six patient visits for model calibration. This process is succinctly described in [22] as ensuring the KFs could forecast using 3 warm-up measurements and validate at least 12-months into the future. While the objective of [22] was to assess how well a KF can predict MD, IOP, and PSD values 5 years into the future for patients with OHTN, we extend their work by predicting patients' probability of conversion from OHTN to POAG. A 2-year prediction widow was used because it is commonly used in Glaucoma literature as a useful early warning [40].

5.2.5.1 Prediction Models

Supervised classification models are a class of statistical models whereby the training data labels (e.g., a binary indicator variable indicating if POAG conversion occurred (1) or did not occur (0)) are known in advance. The goal of supervised learning classification is to create a classification model using training data with known data labels, to apply it to data with unknown data labels. We evaluate the use of four supervised classification models: Logistic Regression, Random Forest, Gradient Boosted Decision Tree, and Neural Network. All four classification models are appropriate for binary-class classification. Binary-class classification refers to a two-class classification problem. In the case of this study, the two classes are either 1 (POAG conversion will occur within 2 years from the patient's 6th visit), or 0 (POAG conversion will not occur within 2 years from the patient's 6th visit). Each classification model can be used to estimate the probability the patient will experience one of the two classes. If the probability is greater than a user-defined threshold the patient is labeled as POAG converting; otherwise, the patient is labeled as non-POAG converting. For the interested reader, a through overview of each of the four models can be found in [21].

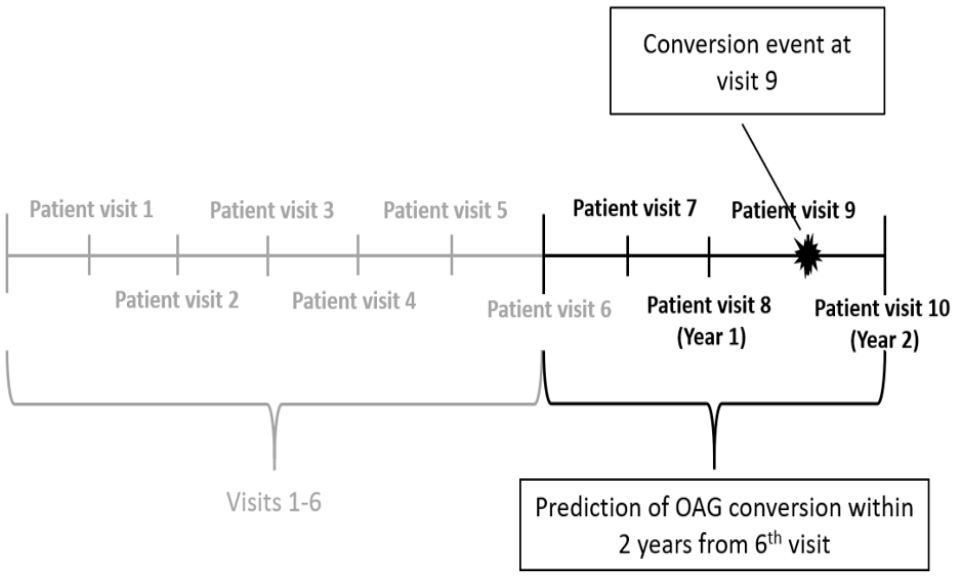


Figure 5.1: Illustration of the POAG conversion prediction timeline. An example of a patient's POAG conversion event. The conversion events occur at their 9th follow-up visit. The binary class target for this patient will be (1), indicating the patient converts to POAG within 2-years following their 6th visit. To predict this conversion event, the classification models will use only information known at the conclusion of the patient's 6th visit.

5.2.6 Performance Measures

To evaluate the performance of the models developed the following performance measures were evaluated. (1) the sensitivity - proportion of eye conversions that were correctly classified as converting; (2) the specificity - proportion of non-converting eyes that were correctly classified as non-converting; (3) the positive predicted value (PPV) - the ratio of correctly classified eye conversions to all classified conversions; (4) the negative predicted value (NPV) - the ratio of correctly classified non-eye conversions to all classified non-conversions; (5) the accuracy - proportion of correctly classified patient eyes to all patient eyes; (6) the balanced accuracy - the average sensitivity and specificity; and (7) the receiver operating characteristic area under the curve. All performance measures were computed for each of the four classification models.

5.2.6.1 Training and Validation Procedure

Prior to training and validation all numeric inputs where centered and scaled to have zero mean and unit variance. Models were trained using 70% (1,838 eyes) of the patients and tested on the remaining 30% (786 eyes) of patients. Training and testing datasets were divided so both the patients' left and right eyes were in the same training or testing dataset. The total number of patients converting in the training set were 252. Of the 252 conversion, 45 (18%) were attributed to 3MD-Drop, 18 (7%) to cluster deterioration, 216 (85%) to TD trend analysis, and 12 (5%) to endpoint. The total does not add up to 252 because a patient's POAG conversion event can be attributed to multiple conversion criteria (e.g., a patient's eye converting by both TD trend and endpoint criteria) within the 2-year period following their 6th visit. For the test set, there were a total of 96 conversions. Of the 96 conversions, 19 (20%) were attributed to 3MD-Drop, 9 (9%) to cluster deterioration, 84 (88%) to TD trend analysis, and 4 (4%) to endpoint. Group 5-fold cross validation was used to train and optimize all models' hyperparameters. The purpose of group 5-fold cross validation was to ensure both the patient's left and right eyes were contained in the same cross validation fold.

All statistical analysis was conducted between January 2020 and June 2020 using Python, version 3.8. Patient characteristics were summarized for the entire sample using means and SDs for continuous variables and frequencies and percentages for categorical variables. ANOVA tests were used to determine if validation performance was statistically different across classification models. All Kalman Filters, classification models, statistical tests, and associated code were developed using the R, MATLAB, and Python programming languages.

5.3 Results

5.3.1 Study Sample

The study sample consisted of 1312 patients (1 patient was removed for having an unknown sex and race). The mean ± SD age at baseline was 56 ± 10 years. Among these patients 1312 patients, 543 (41%) were male, and 769 (59%) were female; the racial breakdown (by self-reporting) included 924 (70%) white participants, 318 (24%) black participants, and 70 (6%) classified as other. There was a total of 348[1] POAG patients' conversions within 2-years from their 6th visit (e.g., between patients' 6th and 10th follow-up visits); 64 (18%) were due to 3MD-Drop; 27 (8%) to cluster deterioration; 300 (86%) to TD trend; and 16 (5%) to endpoint. Patients underwent a mean ± SD of visual 26 ± 7 visual fields over 11.5 ± 3 years of follow-up. The mean ± SD baseline MD was 0.21 ± 1.16 dB. For additional study sample statistics see Table 5.1.

Table 5.1: Description of study sample. MD = mean deviation; PSD = pattern standard deviation; IOP = intraocular eye pressure. Description of the OHTS study sample.

Characteristics	Mean/Count	SD/Percent
No. of Patients	1312	-
No. of Eyes	2624	-
No. of OATS Eye Conversions	348	-
3MD-Drop	64	-
Cluster Deterioration	27	-
TD Trend Analysis	16	-
Endpoint	300	-
Sex, No (%)		
Male	543	41 %
Female	924	59 %
Race/ethnicity, No (%)		
White	924	70 %
Black	318	24 %
Other	70	6 %
Age at baseline, years	56	10
No. of visual field tests per eye	26	7

[1]The number of eye conversions does not add up to 348 because a patient's eye could have converted due to multiple conversion criteria within the 2-years following their 6th follow-up visit.

Follow-ups visit length, years	11.5	3.0
Baseline Readings		
MD, dB	0.21	1.16
PSD, dB	1.95	0.28
IOP, mm Hg	25.12	2.95
Change [2]		
MD, dB	-1.32	2.33
PSD, dB	0.40	0.94
IOP, mm Hg	-3.90	4.64

5.3.2 Predicting performance of conversion from OHTN to POAG

5.3.2.1 Classification Performance

The probability threshold used to classify whether a patient will POAG convert or not was calculated as the value that resulted in the largest average geometric-mean (G-mean) statistic from group 5-fold cross validation on the training dataset. The G-mean statistic was used to determine the optimal probability threshold, because it captures a balance between sensitivity and specificity (i.e., G-mean $= \sqrt{(Sensitivity \times Specificity)}$; maximizing the G-mean is equivalent to choosing the point on receiver operating curve that maximizes the difference between true positive rate and false positive rate.

The classification models that included the Kalman filtered covariates had slightly better performance than the models that did not. The Random Forest classification model had the best performance with or without the KF covariates. For the Random Forest that did not include KF covariates the group 5-fold cross validation ROC AUC performance was mean ± SD of 0.81 ± 0.01. Its testing performance, across the performance measures, was a balance accuracy of 0.79, a sensitivity of 0.75, a specificity of 0.82, an accuracy of 0.81, a positive predicted value of 0.36, a negative predicted value 0.96, and a ROC AUC of 0.86. For the Random Forest that included the KF covariates the group 5-fold cross validation ROC AUC performance was mean ± SD of 0.81 ±0.01. Its' testing performance, across the performance measures, was a balance accuracy of 0.80, a sensitivity of 0.77, a specificity of 0.83, an accuracy of 0.82, a positive predicted value of 0.38, a negative predicted value 0.96, and a ROC AUC of 0.86. Table 5.2 provides the classification per-

[2]Change was computed, for eyes that converted to POAG, as the reading at a patient's 6th visit subtracted from their baseline reading.

formance of all the classification models partitioned by whether they were trained using data that included KF features or not.

Table 5.2: Classification Performance Overview. ROC AUC = Receiver operating characteristic area under the curve. Validation and testing performance of the four classification models. To test whether the validation AUC performance was statistically different between the four models an ANOVA test was computed. The ANOVA test indicated the four models did not (p-value ≤ 0.05) share, statistically, the same validation performance.

Data Set	Performance	Evaluation Criteria	Logistic Regression	Random Forest	Neural Network	Gradient Boosted Decision Tree	P Value
Without Kalman Filtered Data	Validation	ROC AUC – Mean (Std)	0.83 (0.01)	0.86 (0.01)	0.82 (0.03)	0.84 (0.01)	0.01
	Testing	Balanced Accuracy	0.77	0.79	0.76	0.77	-
		Sensitivity	0.74	0.75	0.74	0.71	-
		Specificity	0.80	0.82	0.76	0.83	-
		Accuracy	0.80	0.81	0.75	0.82	-
		Positive Predicted Value	0.35	0.36	0.30	0.36	-
		Negative predicted Value	0.95	0.96	0.95	0.95	-
		ROC AUC	0.84	0.86	0.82	0.84	-
With Kalman Filtered Data	Validation	ROC AUC - Mean (Std)	0.83 (0.01)	0.86 (0.01)	0.82 (0.02)	0.85(0.01)	$<$0.01

Table 5.2 continued from previous page

	Testing	Balanced Accuracy	0.77	0.80	0.75	0.79	-
		Sensitivity	0.74	0.77	0.72	0.77	-
		Specificity	0.80	0.83	0.77	0.80	-
		Accuracy	0.80	0.82	0.76	0.80	-
		Positive Predicted Value	0.35	0.38	0.30	0.35	-
		Negative predicted Value	0.95	0.96	0.95	0.96	-
		ROC AUC	0.83	0.86	0.82	0.86	-

5.3.3 Comparison and Analysis of Model Performance

The Random Forest classification model with the inclusion of the Kalman filtered features performed best. The following two subsections go into detail of this model's testing performance.

5.3.3.1 Top 25 Most Important Model Features

Figure 5.2 illustrates the top 25 most important features for determining progression to POAG for the random forest classification model that included the Kalman filtered data. Most of the top covariates are either associated with TD trend analysis (e.g., Points TD Criteria 1, Beta 1, etc.) or provide summary information of the patient's visual field and how it is changed overtime (e.g., MD, MD filtered, MD velocity, etc.). This is likely because the TD trend analysis conversion criterion made up a significant proportion of all conversion events (86% of all POAG conversion were attributed to the TD trend conversion criterion). Additionally, several filtered covariates are included in the top 25 (e.g., filtered PSD, filtered MD T-4, etc.), implying the Kalman filtered features helped in identifying future occurrences of POAG conversions.

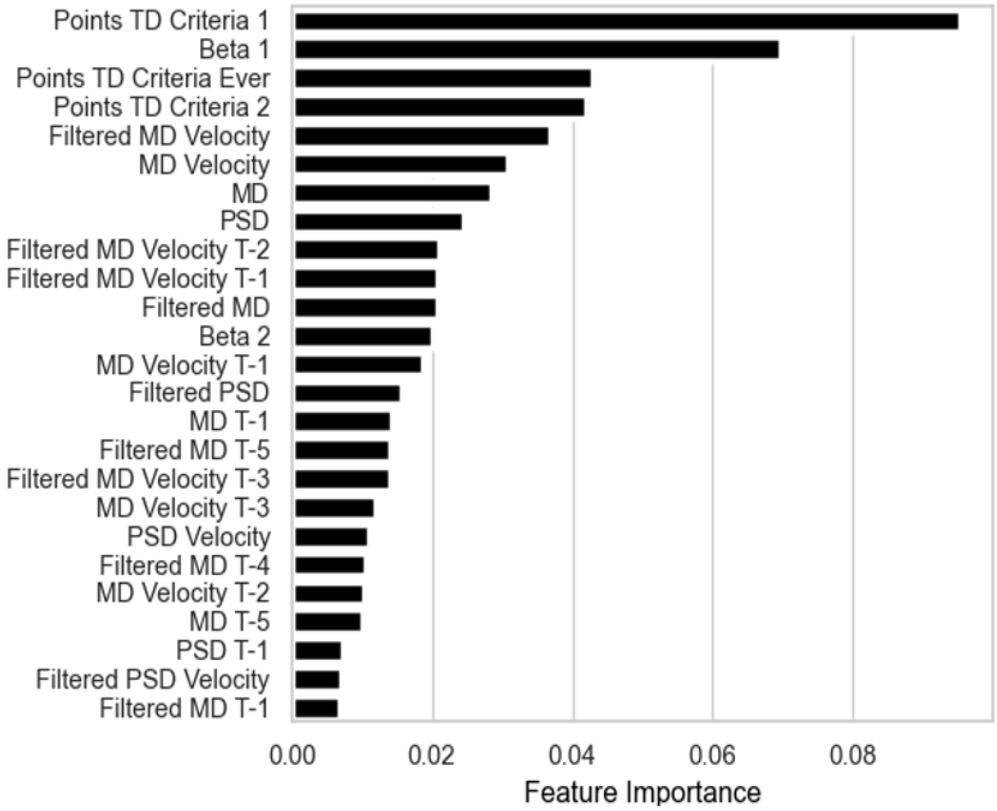


Figure 5.2: Top 25 covariates from Random Forest model with Kalman filtered data. MD = mean deviation; PSD = pattern standard deviation; IOP = intraocular eye pressure. The figure illustrates the mean feature importance scores for the Random Forest regression models. The higher the feature importance score, the greater the feature's importance in determining whether the patient will convert to OAG 2 years from there 6th visit by one of the four conversion criteria.

5.3.3.2 Breakdown of Correctly Classified Conversions

Figure 5.3 displays the confusion matrix of the RF model, with Kalman filtered covariates, for a classification probability threshold of 0.44 (the probability threshold that had the largest average G-mean statistic from group 5-fold cross validation on the training data set). Of the 76 correctly predicted POAG conversions, the conversion criteria breakdown was: 3MD Drop (12), Endpoint (2), Cluster Deterioration (6), and Trend Analysis (68). The total does not add up to 76 because patient conversion events were not mutually exclusive. For the 22 conversions that were misclassified 7 were 3MD Drop, 2 were endpoint, 3 were cluster deterioration, and 16 were trend analysis. Again, the total does not add up to 22 because patient conversion events were not mutually exclusive.

Of the 96 eyes that converted from OHTN to POAG, for those that converted due to one of the four criteria (80 eyes) the RF model predicted 64 (80%) of them correctly. For

the eyes that converted due to two criteria (12) the RF model predicted them 6 (50%) correctly, and for eyes that progressed due to three criteria (4) the RF model predicted 100% of them correctly. No eyes in the test set progressed by all four criteria.

	Predicted Non-Progression	Predicted Progression
True Non-Progression	570 (0.72)	118 (0.14)
True Progression	22 (0.03)	76 (0.10)

Figure 5.3: Random Forest Testing Confusion Matrix. A testing classification confusion matrix of the Random Forest model that included the KF features. The results were obtained using a classification probability threshold of 0.44; if the probability of POAG conversion was greater than 0.44 the patient's eye was labeled as POAG converting within 2-years from their 6th follow-up visit. If it was less than 0.44 the patient's eye was labeled as non-POAG converting within 2-years from their 6th follow-up visit.

5.4 Discussion

Predicting whether a patient will convert 2 years (4 visits) from their 6th visit has been shown to be feasible with reasonable accuracy. While the random forest classification model performed best, all machine learning models demonstrated success in predicting future POAG conversion (e.g., balanced accuracy of all models greater than or equal to 0.75). In clinical practice, depending on whether a clinician values sensitivity (e.g., recall) or specificity (e.g., true negative rate) more the classification threshold (the probability threshold used to determine at which a prediction probability signifies the target class) can be adjusted to meet their respective goal. It is this flexibility that enables supervised classification models to be uniquely qualified for classifying future instances of POAG conversion. In the absence of a POAG conversion reference standard, supervised learning models, or more broadly machine learning models, provide a means to quantify POAG conversion uncertainty into the form of a probability of POAG conversion, in turn, aiding clinicians in making more informed clinical decisions for their patients.

The implication of the results suggest clinicians can use supervised machine learning methods to inform proactive treatment decisions for patients. While the discussed supervised learning models have low positive predicted values (true positive rate) ranging

between 0.30 to 0.40, they also have high negative predicted values (true negative rate) ranging between 0.95 to 0.97. This implies the models can predict, with a high level of certainty the patients that will not progress, but struggles predicting the patients that will progress. While this is not ideal, given the progressive loss of vision POAG causes and its' adverse impact on patients' quality of life, it is better to proactively treat patients at risk of developing POAG, even if in the end they do not, than it is wait for progressive vision loss to have occurred before treating the patient. While not all POAG treatments are appropriate for patients not diagnosed with POAG, at the very least, the patients the model predicts will develop POAG can be monitored by a clinician at a higher follow-up frequency than if they had not; striking a balance between overtreatment and undertreatment.

Additionally, as discussed, we use a collective conversion criterion (conversion criteria composed of 4 seperate definitions of POAG conversion). This was done for several reasons. First, the endpoint and cluster deteriorations POAG conversion criteria are highly conservative, indicating only a small portion of possible POAG conversions (e.g., high sensitivity, but low recall). By only using endpoint or cluster deteriorations criterion a large portion of patients who progressed, would fail to be identified. Second, as the endpoint and cluster deteriorations criteria likely have high precision, but low recall, and the TD trend and 3-MD drop criteria have high recall, but low precision, combining all four criteria into a single criterion leverages the strengths of each criterion. Third, by combining the four criteria into a single criterion combines the strengths and weaknesses of trend-based conversion analysis (evaluating a series of visual field measurements using linear or other form of regression analysis to estimate rates of change and potentially statistical significance) and event-based conversion analysis (visual field measurements are compared with a baseline value and considered significant if the difference between the visual field measurement and baseline value exceed a predetermined threshold) into a single criterion. The strength of trend-based analysis is it can not only identify if a patient converts, but also the rate of conversion, however it requires multiple measurements to do so; the strength of event-based analysis is it can identify conversion using a single measurement but lacks information on the rate of conversion. Our hope is the developed criterion is closer to a POAG conversion reference standard than existing methods. Thereby, enabling clinicians to more effectively identify POAG converting patients.

5.5 Conclusion

This chapter investigated the use of supervised classification methods for predicting whether a patient will convert 2 years from their 6th visit using only information known

at the conclusion of their 6th visit. The conversion criterion used was based on a joint conversion definition composed of 4 separate conversion criteria: (A) the OHTS Endpoint Committee, (B) Decline in Mean Deviation, (C) Decline in Pointwise Total Deviation, and (D) Deterioration of Points on Total Deviation Clusters; if any one of the four criteria indicated POAG conversion the patients was labeled as POAG converting. The results indicated, supervised classification methods, in particular the random forest classifier, are capable of proactively identifying future conversions events and the Kalman filtered preprocessed data, all be it a small improvement, improved classification performance. In particular, the Random Forest model that included the Kalman filtered data the group 5-fold cross validation ROC AUC performance was mean ± SD of 0.81 ± 0.01. The testing performance across the evaluated performance measures were balance accuracy of 0.80, sensitivity of 0.77, specificity of 0.83, accuracy of 0.82, positive predicted value of 0.38, negative predicted value 0.96, and ROC AUC of 0.86.

More work is needed to understand the individual prediction performance of the conversion criteria and to investigate how well the model generalizes to other patient populations outside of the OHTS clinical trial candidates. In future work we plan to investigate these issues, in addition to expanding our conversion criteria to incorporate additional POAG conversion criteria. Given, there is currently no gold standard for identifying POAG conversion (e.g., a conversion criterion that has been shown to accurately identify all converting patients), more research is needed to determine which, if any, conversion criteria are best suited for identifying and classification POAG conversion. If a suitable model is identified patients POAG conversion can be identified sooner, thereby reducing the chance the patient experiences progressive or irreversible vision loss.

CHAPTER 6

Summary and Conclusions

As mentioned, chronic conditions affect more than half the United States' adult population [11]. The direct and indirect cost of treating patients with chronic conditions exceeds 1.2 trillion [16]. Given the complexities involved with treating patients with chronic conditions; the medical resources required (e.g., clinical personnel, medications, physical infrastructure, etc.); the quality-of-life impact on patients; and the costs, it is essential for more effective methods for treating patients with chronic conditions to be developed. Currently there has been a rise in the use of healthcare data to inform clinical decisions. We see this as an opportunity to advance the state of medical decision-making research.

This thesis focuses on advancing operations research and statistical learning methods for improving the care of patients diagnosed with ocular chronic conditions. We focused on four research problems: (i) proactive identification of patients' disease progression under high measurement error conditions; (ii) identification of the optimal time to next follow-up appointment for patients with OHTN; (iii) examining and comparing several criteria for determining progression from OHTN to primary open angle glaucoma (POAG); and (iv) the evaluation of the use of a joint POAG conversion criterion in predicting future occurrence of POAG conversion.

This research was conducted in collaboration with clinical disease/domain experts. All the medical decision-making research herein addresses real world healthcare issues, that if solved, have the potential to improve peoples' lives. While these methodologies primarily focus on chronic conditions affecting the eyes (e.g., OHTN and POAG), it is important to note that much of the work produced offers methods applicable to other chronic diseases.

6.1 Chapter Summaries

In chapter two we addressed the problem of identifying whether a patient, who has been diagnosed with glaucoma, will experience rapid worsening/progression of their disease. We developed a supervised meta classification model tailored to solving this problem. Data from two randomized clinical trials (the Advanced Glaucoma Intervention Study (AGIS) and the Collaborative Initial Glaucoma Treatment Study (CIGTS)) were used for this purpose. As patients diagnosed with a chronic condition may transition between different phases of progression, we found it important to predict the timing of rapid progression before it occurs to better inform disease management decisions.

The chapter provided a framework that can be adapted to a variety of chronic diseases settings in which medical tests have moderate to severe residual variability. When medical tests have low residual variability (e.g., empirical difference between the patient's true and recorded value is small) they can effectively, without the use of sophisticated methods, identify the patient's current disease phase; however, when medical tests have moderate to high residual variability this may not be the case. The framework presented solves the latter case. The framework accomplished this by using the filtered outputs of an interacting multiple model Kalman filter as additional inputs for a supervised learning meta classifier. The purpose of this integration was to estimate the true values of patients' disease metrics by directly modeling for rapid and non-rapid disease phases, and dynamically adapting to changes in these values over time. We applied our framework to classifying whether a patient with POAG will experience rapid progression over the next two or three years from the time of their current visit. The performance (AUC) of our model increased by approximately 7% (increased from 0.752 to 0.819) when the Kalman filtering results were incorporated as additional features in the supervised learning model. These results suggest the combination of filters and statistical learning methods in clinical health has significant benefits.

We showed that the presented model is able to achieve acceptable levels of performance (e.g., TP1: 2-year (3-year) AUC of .82 (.76) for the walk-forward cross-validation performance). We conjecture that, as the data used to train these models is increased, their performance will continue to improve, increasing their value to clinical practice. Future work is warranted to expand the study population from patients with moderate to advanced glaucoma, to patients with early onset, mild, and moderate glaucoma. The methodology presented has far reaching implications, as it enables the clinician to make more informed decisions regarding the treatment of patients with glaucoma and increases the likelihood the clinician is able to maintain a satisfactory quality of life for these patients.

Although this chapter applied its' methodology to POAG, the methodology developed is applicable to other chronic conditions.

Future works may be able improve the presented methodology by incorporating mechanisms for the dynamic control of the disease. For example, examining how clinical treatments affect the disease transition from one phase to another, and more insightfully, how this can be controlled so patients' time in a "favorable" disease phase is maximized. Additionally, we may consider predicting disease phases at smaller time intervals (e.g., 6-month periods). In this way, the clinician will have greater detail of patients' future disease trajectories, instead of a summary measures over 2 or 3-year time windows.

In chapter three we used reinforcement learning to develop scheduling policies for patients with ocular hypertension (OHTN). Patients with OHTN are at increased risk of developing glaucoma and should be observed over their lifetime. Follow-up appointment schedules that are chosen poorly can result in, at minimum, delay in the detection of a patient's progression to glaucoma, and at worst, poor patient outcomes. To this end, we presented a personalized decision support tool that uses an off-policy reinforcement learning algorithm to recommend personalized patient time-to-next follow-up schedules.

We proposed two scheduling policies, each of which considered three underlying criteria: (i) average time between follow-up visits, (ii) scheduling efficiency (measured by the percentage of eyes with scheduled visits at the earliest indication of the glaucoma progression), and (iii) delay to detect the glaucoma progression. We evaluated the performance of the proposed scheduling policies using the Ocular Hypertension Treatment Study (OHTS) randomized clinical trial data. The numerical results indicated that our proposed scheduling policies are superior to both 1-year and 2-year fixed interval scheduling policies with respect to scheduling efficiency (at least 33% higher than the best fixed interval policy) and time to detect glaucoma progression (at least 48% lower than the best fixed interval policy), and comparable in average time between scheduled follow-up visits. For patients who do not progress, the policies schedule less follow-up visits compared to those who did progress. This was as one would expect; for example, for RL policy 1, progressed patients were scheduled follow-up visits approximately every 1.0 years, while non-progressed patients were scheduled approximately every 1.6 years, consistent with the need for more intensity of monitoring for patients likely to progress. This contrasts with fixed interval policies where a constant intensity of monitoring is undertaken, regardless of whether there is a medical need to do so. Ideally, patients should only be seen by a clinician when a medical follow-up visit is warranted; however, we have required a maximum interval of 24 months as a clinically justified safeguard. While our primary focus in this chapter was determining time to next test for patients with OHTN, the methodology

developed can be applied to other chronic conditions.

In future work, targeting the implementation of this work, RL TNT follow-up schedules could be further personalized by altering the reward weights in the RL reward function through a user interface. Depending on the type of care a patient requires, a clinician can, in principle, modify weights to produce follow-up schedules that best adhere to their patients' needs. For instance, increasing the visit delay reward to encourage less visits, or increasing the 3 MD-drop reward to increase the likelihood of visits when suspected progression (e.g., first occurrence of 3MD drop from baseline) has occurred. In addition, it is valuable to investigate whether our results generalize to other chronic conditions and to expand our model's state by incorporating not only the current state of the patient, but also the state of the patient at prior visits. We believe this will further improve our model's performance, as historical patient information provides the model with additional patient insights.

In chapter four we examined and compared several criteria for determining progression from ocular hypertension to primary open angle glaucoma (POAG). We did this for several reasons: to increase the number of detected POAG conversions (e.g., an increase in POAG recall), to identify conversions earlier, to establish conversions with more confidence, or to understand how patients' demographics affect the time to POAG conversion.

The compared criteria were: (A) Endpoint, (B) Decline in Mean Deviation, (C) Decline in Point-wise Total Deviation (e.g., TD trend), and (D) Deterioration of Points on Total Deviation Clusters (e.g., cluster of points). We indicated the criteria can be divided into either trend-based or event-based conversion definitions. The categorization is as follows: decline in point-wise total deviation is trend-based; decline in mean deviation and deterioration of points on total deviation clusters is event based; and the OHTS Endpoint Committee is likely a mixture of trend-based methods, event-based methods, and glaucoma domain expertise. Trend based methods look at changes over a period of time by estimating the slope of the time series. Event-based methods compare each new measurement and test if the change from the baseline method exceeds a threshold or trigger level.

We found there were a total of 1,002 POAG patient eye conversions between all compared criteria. Endpoint had 127, cluster of points had 341, TD trend had 731, and decline in MD had 423. The total did not add up to 1,002 because a patient's eye could progress due to multiple criteria over the observed timeframe. Using univariable survival analysis we confirmed the rates of conversion for all investigated criteria were statistically different from one another. We noted TD trend had the highest rate of conversion with roughly 30% of all patient eyes being identified as POAG converted by the end of the observed

time frame. In addition, an analysis of the multivariable survival analysis' (i.e., multivariable cox regression) coefficients indicated: (i) under each of the four criteria blacks have the highest risk of progressing from OHTN to POAG (i.e., blacks had the highest risk of conversion across all criteria and p-value ≤ 0.05); (ii) the older the patient, the higher the risk of glaucoma progression; (iii) the risk of conversion to POAG is statistically the same, for all all criteria, excluding endpoint, for both males and females; and (iv) the hazard coefficients for race, sex, and age differ significantly from one another depending on the conversion criterion used (likelihood ratio tests p-value less than or equal to 0.05 for each covariate). The last point is of interest because it implies the rate at which a patient will be successfully identified as POAG converting is in part, related to their demographics. Hypothetically suggesting the type of POAG conversion criterion used by a clinician should differ based on patients' race.

In future work, we plan to investigate whether the described criteria can be used to proactively identify whether a patient will convert to POAG. Correctly identifying converting patients is important as it signals to the clinician treatment is required. Likewise, being able to proactively predict converting patients enables the clinician to take preventative action before the patient converts to POAG, at which point unreversible vision loss may have already occurred. One question, which could not be answered, of whether the Endpoint criterion, which the OHTS study team intended to be quite conservative, is closer or further from the truth relative to any of the other criteria. We hope to answer this additionally answer this question in future work.

In chapter five we evaluated the use of a collective POAG conversion criterion in predicting future occurrence of a patient converting from ocular hypertension to POAG. The collective criterion labeled a patient's eye as progressing if they progressed according to any one of the following criteria two years from their sixth visit: (A) the OHTS Endpoint Committee, (B) Decline in Mean Deviation, (C) Decline in Point-wise Total Deviation (e.g., TD trend), and (D) Deterioration of Points on Total Deviation Clusters (e.g., cluster of points). A secondary focus of this chapter was to also evaluate the inclusion of Kalman filtered covariates in improving prediction performance.

We found the random forest classification model that included the Kalman filtered data performed best. The group 5-fold cross validation ROC AUC performance was mean ± SD of 0.81 ± 0.01. The testing performance of the model across the evaluated performance measures were balance accuracy of 0.80, sensitivity of 0.77, specificity of 0.83, accuracy of 0.82, positive predicted value of 0.38, negative predicted value 0.96, and ROC AUC of 0.86.

In future work we plan to study the individual prediction performance of each POAG

conversion criterion the collective conversion criterion consists of (e.g., endpoint, decline in MD, TD trend, and cluster deterioration) and investigate how well the presented model generalizes to other patient populations outside of patients with ocular hypertension. Additionally, we plan to expand our conversion criteria to incorporate other POAG conversion criteria. While we primarily focused on event-based conversion criteria in the inclusion of the collective conversion criterion we plan to include additional trend-based conversion criteria. Given, there is currently no gold standard for identifying POAG conversion (e.g., a conversion criterion that has been shown to most accurately identify converting patients), more research is needed to determine which, if any, conversion criteria are best suited for identifying and classifying POAG conversion. If a suitable criterion is identified, a patient's POAG conversion can be identified sooner, thereby reducing the chance the patient's vision experiences irreversible vision loss.

APPENDIX A

Detailed Description of Case Study Model Inputs

Table A.1: Overview of data variables used in the rapid progression classification model

Featured Measured	Feature Engineering Function	Description	Data Used
Progression	Expanding window	Using an expanding window measure progression from periods 3 to T, engineer the features OLSR slope, OLSR slope p-value, and a binary indicator indicating if the slope is statistically significant at the level of 0.05	MD, KF MD

Table A.1 continued from previous page

Feature Measured	Feature Engineering Function	Description	Data Used
		Calculate summary statistics of mean, standard deviation, quantiles at various levels, min, and max for each covariate described in the above row.	
	Moving window	Calculate a moving window of 3 most recent periods from periods 3 to T for OLSR slope, and OLSR p-value.	
Patient Test Results	Expanding window	Using an expanding window calculate summary statistics mean, standard deviation, various quantiles, min, and max from periods 0 to T.	MD, PSD, IOP, KF MD

Table A.1 continued from previous page

Feature Measured	Feature Engineering Function	Description	Data Used
	Moving window	Using a moving window of 3 & 4 periods, calculate summary statistics (e.g., mean, standard deviation, quantiles, min, and max).	
	Shifted/ Time-lag features	Time lag features in addition to time T: T-3, T-2, T-1.	
Demographics	N/A	Unchanging covariates based on patient's demographic information.	Age, Sex, Race
Patient Appointment Information	N/A	Follow-up period.	Visit Period (6-month interval)

APPENDIX B

IMM Initial Filter Model Parameters

The bank of Kalman Filters was iteratively tuned using a grid search procedure. The objective of this procedure was to determine the best parameters for the Kalman filter models. The initial estimates for M and μ_0 were obtained using the data. M was estimated using the number of transitions between RP and RP; RP and Non-RP; Non-RP and RP; and Non-RP and Non-RP. μ_0 was estimated using the normalized frequency of RP instances and Non-RP instances. The estimated values for M and μ_0 are found in Tables B.1 and B.2 respectively.

Table B.1: IMM initial transition probability matrix, M

	Stable Phase	Rapid Phase
Stable Phase	0.89	0.11
Rapid Phase	0.12	0.88

Table B.2: IMM Initial mode probability matrix, μ

Stable Phase	Rapid Phase
0.78	0.22

The elements of the F matrix are shown in Table B.3. The F matrix captures the system transition of Kalman Filter. The F matrix was built using the linear vector difference equation model for the following physics equations:

$$MD_{predicted} = MD + (MD_{velocity} \cdot \Delta t) + (MD_{acceleration} \cdot \frac{\Delta t^2}{2})$$
$$MD_{predictedvelocity} = MD_{velocity} + MD_{acceleration} \cdot \Delta t$$
$$MD_{predictedacceleration} = MD_{acceleration},$$

where $\Delta t = 6$ months, and state variables $x = \langle MD \;\; MD_{velocity} \;\; MD_{acceleration} \rangle$.

Table B.3: Transition Matrix, F

	MD	MD Velocity	MD Acceleration
MD	1	6	18
MD Velocity	0	1	6
MD Acceleration	0	0	1

The Q (process) noise covariance matrices are shown in Table B.4 and B.5. The Q matrix captures the noise introduced into our system due to external factors we do not directly model for. The Q matrix captures this noise using a random process centered at 0. In our model we assumed the Q matrices have a piecewise white noise model, where the noise follows a discrete time Wiener process. This noise model assumes the noise for the highest order term (e.g. acceleration) is constant for the duration of each time period, but differs for each time period, and each of these is uncorrelated between time periods [44].

Table B.4: Non-RP Process Noise Covariance Matrix, Q_1

	MD	MD Velocity	MD Acceleration
MD	0.20	0.06	0.01
MD Velocity	0.06	0.02	0.00
MD Acceleration	0.01	0.00	0.00

Table B.5: RP Process Noise Covariance Matrix, Q_2

	MD	MD Velocity	MD Acceleration
MD	3.24	1.08	0.18
MD Velocity	1.08	0.36	0.06
MD Acceleration	0.18	0.06	0.01

The initial estimates for the MD, MD velocity, and MD acceleration variances are high, so as to model the uncertainty associated with these three measurements when you see a new patient.

Table B.6: Initial Covariance Matrix, P

	MD	MD Velocity	MD Acceleration
MD	1	0	0
MD Velocity	0	0.5	0
MD Acceleration	0	0	0.25

As a clinician with a new patient there is no current information to suggest an appropriate MD rate of change. There is only a single MD baseline value. As such the initial state

estimate for each patient is

$$x = \begin{pmatrix} MD_{baseline} \\ 0 \\ 0 \end{pmatrix}$$

Last, the measurement function, H, and measurement noise, R are defined respectively as,

$$H = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

$$R = \begin{pmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \sigma_3 \end{pmatrix}.$$

APPENDIX C

Illustration of the IMM Kalman filtered MD Results

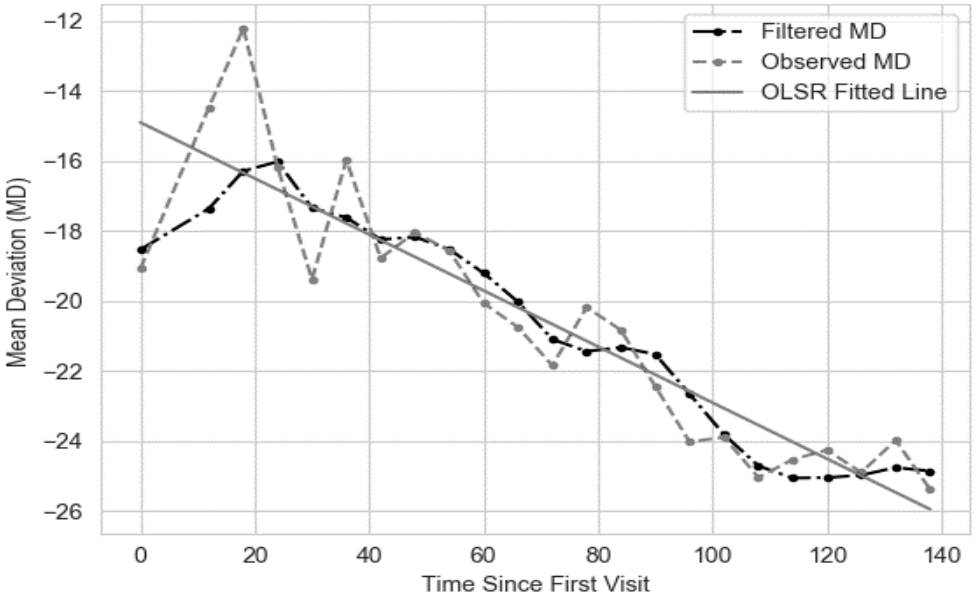


Figure C.1: IMM Kalman filtered MD results compared to a patient's observed MD results, and the MD "true state" estimated by ordinary least squares regression (OLSR)

APPENDIX D

Chapter 4 Table of Notation

Symbol	Description
t	Time in periods of 6-months.
a_t	Action taken at time t
s_t	The state of the patient at period t. The state describes the patient's demographics and test measurements at period t.
$r_{t+a_t}(s_t, a_t, s_{t+a_t})$	Reward received from transitioning from state s_t to state s_{t+a_t} after taking action a_t.
r_{t+a_t}	Shorthand for $r_{t+a_t}(s_t, a_t, s_{t+a_t})$.
$\pi(s)$	Learned policy from Q-iteration. Determines the action to take given state s.
$Q^{\pi}(s, a)$	Q-function. Computes the expected sum of rewards received from starting in state s taking action a, and thereafter following policy $\pi(s)$.
F	One-step transitions. A set comprised of all possible patient eye state/action transitions.
K	The integer number of iterations used to approximate the Q-functions.
$\hat{Q}_k(s, a)$	Approximation of the Q-function at iteration $k \in [1, K]$.
T_k	The training set at iteration k. The training set is composed of state action pairs and associated estimates of the Q-function value that results from taking action a in state s.
$\hat{Q}_k$	Q-function estimated at iteration k of fitted Q-iteration.
$\Delta(\hat{Q}_k, \hat{Q}_{k-1})$	Distance between subsequent Q-iterations.
λ	Reward weights
p	P-value

Table D.1 continued from previous page

Symbol	Description
α	Type-1 (false-positive) error rate.
$R^{\pi}(s_0)$	Total expected sum of rewards starting from state s_0.

APPENDIX E

Detailed Description of Case Study Model Inputs for POAG Conversion Prediction

Table E.1: Overview of data variables at the 6th visit of the patient, denoted as time T.

Feature Measured	Feature Engineering Function	Description	Data Used
Patient Test Results	Shifted/ Time-lag of patient medical test	Time lag features in addition to time T: T, T-1, T-2, T-3, T-4, T-5	Observed MD, PSD, IOP, MD velocity, PSD velocity, IOP velocity, MD acceleration, PSD acceleration, and IOP acceleration
	Shifted/ Time-lag of patient medical test rates of change		
	Shifted/ Time-lag of patient medical test acceleration		
	Kalman filtered medical tests	KF filtered measurements at times T, T-1, T-2, T-3, T-4, T-5	Kalman filtered MD, PSD, IOP, MD velocity, PSD velocity, IOP velocity, MD acceleration, PSD acceleration, and IOP acceleration
		Forecasted KF measurements at times T+1, T+2, T+3, T+4	

Table E.1 continued from previous page

Feature Measured	Feature Engineering Function	Description	Data Used
Cup-to-disc ratio	Average, mean, minimum	Descriptive statistics of a patient's cup-to-disc ratio since time T	Cup-to-disc ratio
Demographics History	N.A.	Unchanging covariates based on patient's demographic information	Age, race, and sex, family history of glaucoma
Medication Use	N.A.	Binary indicator indicating a patient's medical use	Ocular and non-ocular meds
Total Deviation (TD) Trend Analysis	Points TD Criteria Ever: How many TD locations ever met TD trend criteria	TD plot engineered features	TD plots of a patient's visual field
	Points TD Criteria 1: How many points met TD criteria at 5th visit		
	Points TD Criteria 2: How many points met TD criteria at 5th and 6th visit		
	Beta 1 (Beta 2): Minimum TD OLSR slope of the points that met criteria at 5th visit (and 6th visit) respectively		

Table E.1 continued from previous page

Feature Measured	Feature Engineering Function	Description	Data Used
	P-value 1 (P-value 2): Minimum TD slope p-value of the points that met criteria at 5th visit (and 6th visit) respectively		
	Missing 1: Indicator variable if the patient does not have at least one TD reading before their last 2nd to last visit		
	Missing 2: Indicator variable if the patient does not have at least two TD readings before their last visit		

BIBLIOGRAPHY

[1] National center for chronic disease prevention and health promotion, May 2020.

[2] Jorge LM Amaral, Agnaldo J Lopes, José M Jansen, Alvaro CD Faria, and Pedro L Melo. Machine learning algorithms and forced oscillation measurements applied to the automatic identification of chronic obstructive pulmonary disease. *Computer Methods and Programs in Biomedicine*, 105(3):183–193, 2012.

[3] Paul H Artes, Balwantray C Chauhan, John L Keltner, Kim E Cello, Chris A Johnson, Douglas R Anderson, Mae O Gordon, Michael A Kass, Ocular Hypertension Treatment Study Group, et al. Longitudinal and cross-sectional analyses of visual field progression in participants of the ocular hypertension treatment study. *Archives of ophthalmology*, 128(12):1528–1532, 2010.

[4] Turgay Ayer, Oguzhan Alagoz, and Natasha K Stout. Or forum—a pomdp approach to personalize mammography screening decisions. *Operations Research*, 60(5):1019–1034, 2012.

[5] Christoph Bergmeir and José M Benítez. On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191:192–213, 2012.

[6] Henk AP Blom. An efficient filter for abruptly changing systems. In *The 23rd IEEE Conference on Decision and Control*, pages 656–658. IEEE, 1984.

[7] Henk AP Blom and Yaakov Bar-Shalom. The interacting multiple model algorithm for systems with markovian switching coefficients. *IEEE Transactions on Automatic Control*, 33(8):780–783, 1988.

[8] Mike J Bradburn, Taane G Clark, Sharon B Love, and Douglas G Altman. Survival analysis part ii: multivariate data analysis–an introduction to concepts and methods. *British journal of cancer*, 89(3):431–436, 2003.

[9] Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M Buhmann. The balanced accuracy and its posterior distribution. In *2010 20th International Conference on Pattern Recognition*, pages 3121–3124. IEEE, 2010.

[10] Cagatay Catal and Mehmet Nangir. A sentiment classification model based on multiple classifiers. *Applied Soft Computing*, 50:135–141, 2017.

[11] National center for chronic disease prevention and health promotion. Chartbook. In *Health, United States, 2020 [Internet]*. ational center for chronic disease prevention and health promotion (US), 2020.

[12] Nathan Congdon, Benita O'Colmain, CC Klaver, Ronald Klein, Beatriz Muñoz, David S Friedman, John Kempen, Hugh R Taylor, Paul Mitchell, et al. Causes and prevalence of visual impairment among adults in the united states. *Archives of Ophthalmology (Chicago, Ill.: 1960)*, 122(4):477–485, 2004.

[13] Juliet M Corbin. A nursing model for chronic illness management based upon the trajectory framework. *Research and Theory for Nursing Practice*, 5(3):155, 1991.

[14] Antonio Coronato and Muddasar Naeem. A reinforcement learning based intelligent system for the healthcare treatment assistance of patients with disabilities. In *International Symposium on Pervasive Systems, Algorithms and Networks*, pages 15–28. Springer, 2019.

[15] Arun Das, Paul Rad, Kim-Kwang Raymond Choo, Babak Nouhi, Jonathan Lish, and James Martel. Distributed machine learning cloud teleophthalmology iot for predicting amd disease progression. *Future Generation Computer Systems*, 93:486–498, 2019.

[16] Ross DeVol, Armen Bedroussian, Anita Charuworn, Anusuya Chatterjee, I Kim, Soojung Kim, and Kevin Klowden. An unhealthy america: The economic burden of chronic disease. *Santa Monica, CA: Milken Institute*, 326:2010–2060, 2007.

[17] Thomas G Dietterich. Ensemble methods in machine learning. In *International Workshop on Multiple Classifier Systems*, pages 1–15. Springer, 2000.

[18] Damien Ernst, Pierre Geurts, and Louis Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6:503–556, 2005.

[19] Pablo Escandell-Montero, Milena Chermisi, Jose M Martinez-Martinez, Juan Gomez-Sanchis, Carlo Barbieri, Emilio Soria-Olivas, Flavio Mari, Joan Vila-Francés, Andrea Stopper, Emanuele Gatti, et al. Optimization of anemia treatment in hemodialysis patients via reinforcement learning. *Artificial intelligence in medicine*, 62(1):47–60, 2014.

[20] National Center for Health Statistics (US et al. Chartbook. In *Health, United States, 2018 [Internet]*. National Center for Health Statistics (US), 2019.

[21] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The Elements of Statistical Learning*, volume 1. Springer Series in Statistics New York, 2001.

[22] Gian-Gabriel P Garcia, Mariel S Lavieri, Chris Andrews, Xiang Liu, Mark P Van Oyen, Michael A Kass, Mae O Gordon, and Joshua D Stein. Accuracy of Kalman filtering in forecasting visual field and intraocular pressure trajectory in patients with ocular hypertension. *JAMA Ophthalmology*, 137(12):1416–1423, 2019.

[23] Gian-Gabriel P Garcia, Koji Nitta, Mariel S Lavieri, Chris Andrews, Xiang Liu, Elizabeth Lobaza, Mark P Van Oyen, Kazuhisa Sugiyama, and Joshua D Stein. Using Kalman filtering to forecast disease trajectory for patients with normal tension glaucoma. *American Journal of Ophthalmology*, 199:111–119, 2019.

[24] Aurélien Géron. *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. " O'Reilly Media, Inc.", 2017.

[25] Tiago Salgado Magalhães Taveira Gomes. Reinforcement learning for primary care e appointment scheduling. 2017.

[26] Mae O Gordon, Julia A Beiser, James D Brandt, Dale K Heuer, Eve J Higginbotham, Chris A Johnson, John L Keltner, J Philip Miller, Richard K Parrish, M Roy Wilson, et al. The ocular hypertension treatment study: baseline factors that predict the onset of primary open-angle glaucoma. *Archives of ophthalmology*, 120(6):714–720, 2002.

[27] Mae O Gordon, Eve J Higginbotham, Dale K Heuer, Richard K Parrish II, Alan L Robin, Patricia A Morris, Deborah A Dunn, Bradley S Wilson, Michael A Kass, and Ocular Hypertension Treatment Study. Assessment of the impact of an endpoint committee in the ocular hypertension treatment study. *American journal of ophthalmology*, 199:193–199, 2019.

[28] Mae O Gordon and Michael A Kass. What we have learned from the ocular hypertension treatment study. *American journal of ophthalmology*, 189:xxiv–xxvii, 2018.

[29] Mae O Gordon, Michael A Kass, Ocular Hypertension Treatment Study Group, et al. The ocular hypertension treatment study: design and baseline description of the participants. *Archives of Ophthalmology*, 117(5):573–583, 1999.

[30] Carolina PB Gracitelli, Linda M Zangwill, Alberto Diniz-Filho, Ricardo Y Abe, Christopher A Girkin, Robert N Weinreb, Jeffrey M Liebmann, and Felipe A Medeiros. Detection of glaucoma progression in individuals of african descent compared with those of european descent. *JAMA Ophthalmology*, 136(4):329–335, 2018.

[31] Eye Diseases Prevalence Research Group et al. Prevalence of open-angle glaucoma among adults in the united states. *Archives of Ophthalmology*, 122(4):532, 2004.

[32] Kyeonghye Guk, Gaon Han, Jaewoo Lim, Keunwon Jeong, Taejoon Kang, Eun-Kyung Lim, and Juyeon Jung. Evolution of wearable devices with real-time disease monitoring for personalized healthcare. *Nanomaterials*, 9(6):813, 2019.

[33] W Haynes. Holm's method. *Encyclopedia of systems biology*, pages 902–902, 2013.

[34] Tianhua He, Kan Li, Mark S Roberts, Anne C Spaulding, Turgay Ayer, John J Grefenstette, and Jagpreet Chhatwal. Prevention of hepatitis c by screening and treatment in us prisons. *Annals of internal medicine*, 164(2):84–92, 2016.

[35] Anders Heijl, M Cristina Leske, Boel Bengtsson, Bo Bengtsson, Mohamed Hussein, and EMGT Group. Measuring visual field progression in the early manifest glaucoma trial. *Acta Ophthalmologica Scandinavica*, 81(3):286–293, 2003.

[36] Jonathan E Helm, Mariel S Lavieri, Mark P Van Oyen, Joshua D Stein, and David C Musch. Dynamic forecasting and control algorithms of glaucoma progression for clinician decision support. *Operations Research*, 63(5):979–999, 2015.

[37] Frank Hutter, Lin Xu, Holger H Hoos, and Kevin Leyton-Brown. Algorithm runtime prediction: Methods & evaluation. *Artificial Intelligence*, 206:79–111, 2014.

[38] Nomdo M Jansonius. Towards an optimal perimetric strategy for progression detection in glaucoma: from fixed-space to adaptive inter-test intervals. *Graefe's Archive for Clinical and Experimental Ophthalmology*, 244(3):390–393, 2006.

[39] Chris A Johnson, Pamela A Sample, George A Cioffi, Jeffrey R Liebmann, and Robert N Weinreb. Structure and function evaluation (safe): I. criteria for glaucomatous visual field loss using standard automated perimetry (sap) and short wavelength automated perimetry (swap). *American journal of ophthalmology*, 134(2):177–185, 2002.

[40] Isaac A Jones, Mark P Van Oyen, Mariel S Lavieri, Christopher A Andrews, and Joshua D Stein. Predicting rapid progression phases in glaucoma using a soft voting ensemble classifier exploiting kalman filtering. *Health Care Management Science*, pages 1–16, 2021.

[41] Pooyan Kazemian, Jonathan E Helm, Mariel S Lavieri, Joshua D Stein, and Mark P Van Oyen. Dynamic monitoring and control of irreversible chronic diseases with application to glaucoma. *Production and Operations Management*, 28(5):1082–1107, 2019.

[42] Ross S Kleiman, Eric R LaRose, Jonathan C Badger, David Page, Michael D Caldwell, James A Clay, and Peggy L Peissig. Using machine learning algorithms to predict risk for development of calciphylaxis in patients with chronic kidney disease. *AMIA Summits on Translational Science Proceedings*, 2018:139, 2018.

[43] Colleen M Kummet, KD Zamba, Carrie K Doyle, Chris A Johnson, and Michael Wall. Refinement of pointwise linear regression criteria for determining glaucoma progression. *Investigative ophthalmology & visual science*, 54(9):6234–6241, 2013.

[44] Roger Labbe. *Kalman and Bayesian Filters in Python*. book, 2015.

[45] Chris P Lee, Glenn M Chertow, and Stefanos A Zenios. Optimal initiation and management of dialysis therapy. *Operations Research*, 56(6):1428–1449, 2008.

[46] Elliot Lee, Mariel S Lavieri, and Michael Volk. Optimal screening for hepatocellular carcinoma: A restless bandit model. *Manufacturing & Service Operations Management*, 21(1):198–212, 2019.

[47] Daochang Liu and Tingting Jiang. Deep reinforcement learning for surgical gesture segmentation and classification. In *International conference on medical image computing and computer-assisted intervention*, pages 247–255. Springer, 2018.

[48] Wesley J Marrero, Mariel S Lavieri, and Jeremy B Sussman. Optimal cholesterol treatment plans and genetic testing strategies for cardiovascular diseases. *Health Care Management Science*, pages 1–25, 2021.

[49] David C Musch, Brenda W Gillespie, Paul R Lichter, Leslie M Niziol, Nancy K Janz, and CIGTS Study Investigators. Visual field progression in the collaborative initial glaucoma treatment study: the impact of treatment and other baseline factors. *Ophthalmology*, 116(2):200–207, 2009.

[50] Stelmo Magalhães Barros Netto, Vanessa Rodrigues Coelho Leite, Aristófanes Corrêa Silva, Anselmo Cardoso de Paiva, and Areolino de Almeida Neto. Application on reinforcement learning for diagnosis based on medical image. *Reinforcement learning*, 379, 2008.

[51] Marcus Olivecrona, Thomas Blaschke, Ola Engkvist, and Hongming Chen. Molecular de-novo design through deep reinforcement learning. *Journal of cheminformatics*, 9(1):1–14, 2017.

[52] Niranjani Prasad, Li-Fang Cheng, Corey Chivers, Michael Draugelis, and Barbara E Engelhardt. A reinforcement learning approach to weaning of mechanical ventilation in intensive care units. *arXiv preprint arXiv:1704.06300*, 2017.

[53] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[54] Harry A Quigley and Aimee T Broman. The number of people with glaucoma worldwide in 2010 and 2020. *British journal of ophthalmology*, 90(3):262–267, 2006.

[55] Jean-Francois Rouland, Gilles Berdeaux, and Antoine Lafuma. The economic burden of glaucoma and ocular hypertension. *Drugs & aging*, 22(4):315–321, 2005.

[56] Richard A Russell, David P Crabb, Rizwan Malik, and David F Garway-Heath. The relationship between variability and sensitivity in large-scale longitudinal visual field data. *Investigative Ophthalmology & Visual Science*, 53(10):5985–5990, 2012.

[57] Richard A Russell, David F Garway-Heath, and David P Crabb. New insights into measurement variability in glaucomatous visual fields from computer modelling. *PloS One*, 8(12):e83595, 2013.

[58] S Sawyer. The greenwood and exponential greenwood confidence intervals in survival analysis. *Applied survival analysis: regression modeling of time to event data*, pages 1–14, 2003.

[59] Greggory J Schell, Mariel S Lavieri, Jonathan E Helm, Xiang Liu, David C Musch, Mark P Van Oyen, and Joshua D Stein. Using filtered forecasting techniques to determine personalized monitoring schedules for patients with open-angle glaucoma. *Ophthalmology*, 121(8):1539–1546, 2014.

[60] Hans-Joerg Schuetz and Rainer Kolisch. Approximate dynamic programming for capacity allocation in the service industry. *European Journal of Operational Research*, 218(1):239–250, 2012.

[61] Roger M Stein. Benchmarking default prediction models: Pitfalls and remedies in model validation. *Moody's KMV, New York*, 20305, 2002.

[62] Eija Vesti, Chris A Johnson, and Balwantray C Chauhan. Comparison of different methods for detecting glaucomatous visual field progression. *Investigative Ophthalmology & Visual Science*, 44(9):3873–3879, 2003.

[63] Jing Xiao, Ruifeng Ding, Xiulin Xu, Haochen Guan, Xinhui Feng, Tao Sun, Sibo Zhu, and Zhibin Ye. Comparison and development of machine learning tools in the prediction of chronic kidney disease progression. *Journal of Translational Medicine*, 17(1):119, 2019.

[64] Chengyin Ye, Tianyun Fu, Shiying Hao, Yan Zhang, Oliver Wang, Bo Jin, Minjie Xia, Modi Liu, Xin Zhou, Qian Wu, et al. Prediction of incident hypertension within the next year: Prospective study using statewide electronic health records and machine learning. *Journal of Medical Internet Research*, 20(1):e22, 2018.

[65] Siamak Yousefi, Tobias Elze, Louis R Pasquale, and Michael Boland. Glaucoma monitoring using manifold learning and unsupervised clustering. In *2018 International Conference on Image and Vision Computing New Zealand (IVCNZ)*, pages 1–6. IEEE, 2018.

[66] Chao Yu, Jiming Liu, and Shamim Nemati. Reinforcement learning in healthcare: A survey. *arXiv preprint arXiv:1908.08796*, 2019.

[67] Paul Zarchan and Howard Musoff. *Fundamentals of Kalman Filtering: A Practical Approach*. American Institute of Aeronautics and Astronautics, Inc., 2013.

[68] Qingyuan Zhou, Zongming Zhang, and Yuancong Wang. Wit120 data mining technology based on internet of things. *Health Care Management Science*, pages 1–9, 2019.