INDIVIDUALIZED THERAPY FOR CYSTIC FIBROSIS USING ARTIFICIAL INTELLIGENCE

Yiyun Tang

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics, School of Public Health.

Chapel Hill
2010

Approved by:

Advisor: Professor Michael Kosorok

Reader: Professor Gary Koch

Reader: Professor Jason Fine

Reader: Professor Donglin Zeng

Reader: Professor Lisa LaVange

Reader: Professor George Retsch-Bogart

ABSTRACT

YIYUN TANG: Individualized Therapy for Cystic Fibrosis Using Artificial Intelligence

(Under the direction of Dr. Michael R. Kosorok)

Optimal clinical management of inherited chronic diseases, such as Cystic Fibrosis (CF), requires a dynamic approach which updates treatments to cope with the evolving course of illness and to tailor medicines and dosages for individual patients. The chronic progressive nature of CF and heterogeneity across patients lead to challenges of developing optimal regimens. An adaptive individualized therapy provides a solution and a means toward these goals. In this dissertation, we examine the problem of computing optimal adaptive individualized therapy for CF patients. A temporal difference reinforcement learning method called fitted Q-iteration is utilized to discover the optimal treatment regimen directly from clinical data. We propose multi-state discrete-time Markov process to model the disease dynamic for cystic fibrosis patients with *Pseudomonas aeruginosa* infection with the model parameters tuned and estimated from the published data in Wisconsin CF neonatal screening project. Our study results indicate that reinforcement learning and the clinical reinforcement trial framework can be an effective tool in discovering and developing personalized therapy which optimises the benefit-risk trade off in multi-stage decision making and improves long term outcomes in chronic diseases.

ACKNOWLEDGEMENTS

I am grateful to the many people who have taught, assisted and encouraged me throughout five years of graduate study at the University of North Carolina at Chapel Hill.

The first person whom I would like to thank is my advisor, Dr. Michael Kosorok, who led me to this exciting research field and patiently guided me through the dissertation process. I wish to express my deepest appreciation to him from the bottom of my heart for his enormous support, continuous encouragement and mentoring throughout my whole doctoral studies. This dissertation could not have been written without my advisor. From him, I learned how to conduct research, pursue my goal consistently, and organize thinking, both theoretically and practically.

I wish to express my sincere thanks my committee members. I owe many thanks to Dr. Gary Koch for his encouraging inspirations, kind guidance, support in many aspects and valuable advice all these years. I am deeply grateful to Dr. George Retsch-Bogart for his in-depth knowledge of cystic fibrosis and insightful discussions on many issues. I wish to express my special thanks to Dr. Jason Fine for his invaluable advice and kindness throughout this research. I also appreciate enormously helpful discussions with Dr. Donglin Zeng; working with him has been a wonderful learning experience for me. I would like to thank Dr. Lisa LaVange for constructive suggestions and comments on my research.

I was very fortunate to have the opportunity to work at the Biometric Consulting Laboratory under the direction of Dr. Gary Koch, the North Carolina Translational and Clinical Sciences Institute under the guidance of Dr. Michael Kosorok and Dr. Jianwen Cai, and the Center for Environmental Medicine, Asthma and Lung Biology

in University of North Carolina at Chapel Hill under the guidance of Dr. Haibo Zhou. A special thanks goes to them, for their financial support of my graduate studies and research.

Dr. Philip Farrell, Dr. Zhanhai Li and Wisconsin Cystic Fibrosis Neonatal Screening Study Group deserved a special thank for providing the Wisconsin neonatal screening data in this study.

I am greatly indebted for all my friends and colleagues in the Reinforcement Learning Group at the University of North Carolina at Chapel Hill, particularly Yufan Zhao, Kai Ding and Yingqi Zhao, whose friendship has made this journey more enjoyable and memorable.

Finally, it's impossible to have completed this journey without the love, support and encouragement from my dad, Wanan Tang, my mom, Huiling Wang, my husband, Chang Zeng, my sister, Yishan T. Zander, and brother-in-law, Michael R. Zander. This dissertation is dedicated to them.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

Cystic Fibrosis (CF) is the most common lethal hereditary disorder in Caucasians. It affects approximately 30,000 people in the United State and 70,000 people worldwide (Cystic Fibrosis Foundation, 2008). The most fundamental pathogenesis of CF is that the CF transmembrane conductance regulator (CFTR) protein is encoded by a defective gene on chromosome 7 which leads to life-threatening lung infections and obstruction of the pancreas (Rowe, et al., 2005). The prognosis of the disease is substantially dependent on chronic respiratory infection, a hallmark of CF.

In clinical practice, treatment of many inherited chronic diseases, such as CF, is a dynamic process involving a series of therapeutic decisions over time. For example, in treating CF patients with chronic lung infections by the most common and significant pathogen, *Pseudomonas aeruginosa* (*Pa*), clinicians routinely modify therapy in the face of infection severity, toxicity and antibiotics resistance, reducing the duration, dose, or switching medication (Döing, et al., 2000; Flume, et al., 2007) Essentially, these treatment decisions are made based on clinical judgement sequentially over time combined with accruing information on the patient. The quality of life, length of survival and cost of care are commonly determined by the success of the entire sequence of antibiotic treatment over many years.

The unique characteristics of the disease require personalized, time varying and multistage consideration in order to improve patient longterm outcome. There are three primary issues to consider. First, various defective CFTR mutations lead to different cellular consequences (Rowe, et al., 2005). Second, the frequent infection relapse and progression require timely treatment modification (Flume, et al., 2007). Third, the chronic nature of CF leads to repeated courses of potentially toxic drugs

for many years, increasing risk of cumulative side-effect, such as drug resistance, impairment of renal function and hearing (Döing, et al., 2000; Döing, et al., 2004; Flume, et al., 2007). These characteristics reflect in multidimensional heterogeneities, consisting in part of variation between patients due to genetic factors and within-patient heterogeneities over time.

These aspects of the disease pose increasingly difficult challenges for studying CF therapies, because standard, single-decision trials are unable to correct for individual differences and prior history in assessing treatments. The reviews of clinical trials in CF (Döing, et al., 2007; Langton Hewer, et al., 2009; Retsch-Bogart, 2009; Ryan, et al., 2000; Waters, et al., 2008) have found the common dilemma between limited number of CF patients and the need to control for confounding factors including mutation class, age, disease severity, and prior treatment, among other factors. The increasing evidence and growing recognition of the influence of prior and subsequent treatments has led to considerable interest in studying the prolonged treatment effect and to evaluate entire treatment sequences. For example, early aggressive Pa eradication therapy is of significant interest because it might be able to improve overall survival in the long term (Taccetti, et al., 2005; Treggiari, et al., 2009; Treggiari, et al., 2007); specifically, the strategy of intermittent administration of inhaled tobramycin may reduce the risk of resistance development (Ramsey, et al., 1999). Moreover, even if the value of a specific antibiotic therapy has been established, significant questions remain as to optimum dosage, duration of treatment and frequency of administration.

In this thesis, we present a "clinical reinforcement trial" procedure to discover optimal personalized therapy for CF which seeks to address the above questions and to leverage patient differences in order to improve the entire decision-making process. The clinical reinforcement trial approach based on Q-learning for discovering effective regimens was first introduced for potentially irreversible diseases such as cancer. This framework was further refined for clinical trials in non-small cell lung cancer after adaptation to handle right-censored survival data. This clinical reinforcement trial

framework is an extension and melding of earlier work on dynamic treatment regimens in counterfactual frameworks (Murphy, 2005; Murphy, et al.,2001; Robins, 2004) and sequential multiple assignment randomized trials (SMART) (Murphy, 2005; Thall, et al., 2002) which have been applied to behavioral and psychiatric disorders (Murphy, et al.,2007; Pineau, et al., 2007). There are, however, several fundamental differences between the challenge of identifying personalized therapy for CF and the tailored therapy settings for the other therapeutic areas studied in previous work. To begin with, CF patients are usually diagnosed by neonatal screening at birth as described in (Southern, et al., 2009), acquire *Pa* infection in early childhood, and experience frequent reinfection (Kosorok, et al., 2001; Li, et al., 2005). CF patients are usually monitored and treated at regular intervals, with three month intervals being typical, throughout a life time with median survival between 30 and 40 years of age (Cystic Fibrosis Foundation, 2008). A significant therapeutic goal is to delay acquisition of the mucoid variant of *Pa*, which usually occurs a median of 13 years after initial *Pa* infection, since mucoid *Pa* is associated with marked decline in lung function (Li, et al., 2005). As a consequence, the decision making process involves more stages over a much longer period of time in CF than in many other therapeutic areas. Thus the degree of adaptation and modification of previous methodologies required to meet the challenges of CF therapy is significant.

We propose a new clinical reinforcement trial design wherein patients are enrolled at various age ranges in order to capture the known age-specific feature of *Pa* infection. Aiming at exploring the possible treatment sequences, the proposed multiple courses trial involves a fair randomization of patients among different treatment options as well as collection of clinically relevant outcomes and biomarkers at each time point. We then propose to estimate from the resulting data a personalized therapeutic regimen which synthesizes all patient information available at each decision point as input and dictates treatments that result in the most desirable long term outcomes, with particular emphasis on delaying mucoid *Pa*.

In order to efficiently inform the therapy in a manner clinically useful for patients at all ages and decision times, we utilize fitted Q-iteration (Ernst, et al., 2005) in reinforcement learning (RL) (Kalbfleisch, 1985) to estimate the optimal therapy. In some applications of RL to inform multi-stage therapies, such as STI strategies for HIV (Ernst, et al., 2006), the procedures involve a mixture of learning and confirming, which is analogous to response adaptive randomization during trial conduct. This approach does not appear to be fruitful in the CF setting, due in part to the generally irreversible progression of lung disease in CF (Farrell, et al., 2003), and so we propose instead to conduct a second, confirmatory trial to validate the estimated optimal therapy by comparing to existing, standard-of-care alternatives.

Due to limited actual clinical data on treatment mechanism, in-silico modeling of disease dynamics is a cost effective tool for examining the feasibility of using the proposed procedure to identify optimal therapy. We utilize a simple, multistate disease model of *Pa* infection which has been tuned to approximately match published clinical outcome data from the Wisconsin CF neonatal screening project (Li, et al., 2005). The model expresses disease dynamics as a discrete time non-homogeneous Markov chain with stochastic transitions among three phenotypically distinguishable states, *Pa* free, non-mucoid *Pa* infection, and mucoid *Pa*.

The remainder of this dissertation is organized as follows. In Section 2.1, we provide a background introduction of Cystic Fibrosis and antibiotic therapy against *Pa* in CF. The review of clinical trial design with particular attention given to dynamic treatment regimes and personalized medicines are provided in Section 2.2. In Chapter 3, we formulate the problem within a reinforcement learning context in Section 3.1, specifically Q-learning in Section 3.2, followed by the fitted Q-iteration algorithm for estimating the required Q-functions without the time index in Section 3.3. We describe one of the extensions of support vector machine (SVM), support vector regression (SVR), which makes fitting Q-functions feasible for clinical data sets in Section 3.4.

In Chapter 4, we propose a discrete time non-homogeneous Markov model for the *Pa* infection disease dynamic in Cystic Fibrosis, with clinical and biological rationale provided in Section 4.1 and the probability model presented in Section 4.2. In Section 4.3, we tune the model parameter based on the literature research results as one approach to obtain the data generative model in our simulation studies in Chapter 5. Another approach to obtain the model parameters is based on the data analysis of the Wisconsin Nenotal Screening Project in Section 4.4.

In Chapter 5, we provide the reinforcement learning procedure to discover optimal adaptive personalized therapy within the "clinical reinforcement trial" framework. We specialize our overall approach to Cystic Fibrosis clinical trials in Section 5.1 and the details of estimating optimal therapy in Section 5.2. To demonstrate the reinforcement learning's potential in discovering optimal therapies, in Chapter 6, we apply our proposed method to virtual randomized sequential trials, which are based on the disease models and parameters obtained from Chapter 5. This study examines the performance of reinforcement learning via SVR and demonstrates that the therapy found using Q-learning is superior to any constant-dose regimen.

Finally, we summarize our proposed methods in Chapter 7 and discuss some challenges for future research.

2 BACKGROUND

2.1 Cystic Fibrosis and Antibiotic Therapy against Pseudomonas Aeruginosa Infection in CF

Cystic Fibrosis (CF) is the most common lethal hereditary disorder with autosomal recessive heredity in Caucasians (Ratjen, 2001) that affects approximately 30,000 people in the United State, and 70,000 people worldwide. It occurs in 1:3,500 live births among the Caucasian population (CFF Patient Registry, 2008). The CF transmembrane conductance regulator (CFTR) protein is encoded by a defective gene on chromosome 7 (Rowe, et al., 2005). The protein products cause the body to produce unusual thick and sticky mucus that clogs the lungs, leads to life-threatening lung infections, obstructs the pancreas and stops natural enzymes from helping the body break down and absorb food. The prognosis of the disease is substantially dependent on chronic respiratory infection, a hallmark of CF, which may start very early (Burns, et al., 2001) and has been recognized as having the greatest role in morbidity and mortality leading to premature death in 90% of patients (Gibson, et al., 2003a).

Pseudomonas aeruginosa (*Pa*), a ubiquitous environmental bacterium, is the most common and significant pathogen for patients with CF. After variable time periods, children with CF usually acquire nonmucoid *Pa*, which is transient and can possibly be eradicated by aggressive anti-*Pa* antibiotics (Ratjen, et al., 2001; Valerius, et al., 1991). Mucoid *Pa*, a mutant phenotype of *Pa*, develops at subsequent stage (Rosenfeld, et al., 2001), and lives in a defensive mode of growth called biofilm (Hentzer, et al., 2001). Hence it confers resistance to phagocytosis, antibiotics and is much more

difficult to treat and eradicate (Prince, 2002). Early acquisition of mucoid *Pa* was associated with 4-fold greater decrease in cumulative survival. Antibiotic-resistant, biofilm-forming mucoid *Pa* is believed to play a dominant role in the progression of lung disease in patients with CF (Li, et al., 2005). Therefore, therapy to prevent or delay the onset of chronic *Pa* infection is an essential component of CF clinical care. The quality of life, length of survival and cost of care are commonly determined by the success or failure of the antibiotic treatment of the initial *Pa* infection in early childhood, and by subsequent antibiotic treatments. The scope of our study is to improve the antibiotic treatment against *Pa* in patients with CF.

The approaches to management of *Pa* infection and the pulmonary sequelae of CF include early eradication, chronic suppression and acute exacerbation therapies (Waters and Ratjen, 2008; Retsch-Bogart, 2009; Smyth, et al., 2009; Ryan, et al., 2009; UK CF Trust, 2009).

For chronic stable CF patients, the well-studied inhaled antibiotic for chronic suppression of *Pa* is tobramycin solution for inhalation (TOBI). Another macrolide antibiotic is azithromycin. They were approved by the FDA for patients older than 6 years and persistent infection with *Pa* cultures. Other nebulizer antibiotics, such as aztreonam lysine for inhalation (AZLI), TOBI Inhaled Powder (TIP), inhaled colistin, liposomal formulation of amikacin inhaled (SLIT-amikacin), and inhaled fluoroquinolones ciprofloxacin, are still under investigation. In general, mucus treatments with Deoxyribonucleic (DNase) and Hypertonic Saline are strongly recommended in patients with and without *Pa* infection.

For early infection CF patients, there is increasing evidence that a window of opportunity exists to eradiate nonmucoid and antibiotic-sensitive *Pa* and to provide potential long term benefit by delaying or preventing chronic infection (Treggiari, et al., 2009, 2007). The benefits of eradication must be weighed against the potential harms of prolonged antibiotic therapy.

Due to the differences of acute pulmonary exacerbation in rate and volume of dis-

tribution of elimination for many antibiotics caused by *Pa* patients, high doses and shorter intervals may be required. Depending upon the severity of exacerbation, sensitivity to different antibiotics from sputum cultures of patients, combination therapy with an aminoglycoside and a beta-lactam (e.g. carbapenem) anti-*Pa* antibiotic will be chosen. The optimal frequency, mode of delivery (oral, intravenous, inhaled) and time to switch are currently debated and multiple parameters of clinical status are needed to help to determine.

Based on all approaches mentioned above, the repeated courses of potentially toxic drugs over multiple years impairs renal function, hearing and has other adverse effects, such as hypersensitivity, allergic reactions, and the risk of bacterial resistance. Also, the complex treatment history and different risk levels which are caused by the underlying various mutations of the defective gene lead to huge heterogeneity in an individual's response. Additionally, screening, diagnosis and ongoing monitoring of infection, antibiotic resistance, and toxicity are crucial. Hence increasing the time varying and personalized consideration in CF treatment is necessary and important.

Uncertainty continues as to the best antibiotic regimens for each individual with different prognostic factors including genetic mutation, treatment and response history, resistance, toxicity profile, age, disease severity and stage, etc. As a consequence, many of the following questions remain unsolved: best suited combination of antibiotic, dosage schedule, modes of administration, duration, and timing to initiate or alter treatment. Benefits and risks in each individual at different time points need to be assessed and taken into account in order to improve long term health outcome.

Based on prognostic factors, it is therefore timely to develop optimal sequential antibiotic therapy tailored to each individual. For instance, how and when to alter the intensity, type, dosage, or delivery of treatment at the critical decision time points? The goal of our study is to deliver the right drug to the right person at the right dose, mode and time in the management of this chronic disease to improve the patients' long term health outcomes. If this can be achieved in a well-structured study, it will

lead to a huge impact in clinical practice and in CF patient life.

2.2 Dynamic Treatment Regime and Individualized Therapy

Optimal clinical management of a chronic disease, such as CF, requires a dynamic approach which updates treatments to cope with the evolving course of illness. Because there exist heterogeneities of patients as well as delayed effects in an ongoing process, the optimal therapy should contain personalized care, time varying adaptation and long term benefit as key components. A one-size-fits-all or once-and-for-all may not be appropriate in this therapeutical area. Specifically, the heterogeneities include within patient temporal difference and between patients inherent difference in treatment responses. For example, in CF, due to its genetic pathogenesis and chronic nature, these heterogeneities reflect in the following ways: various response and side-effects linked to different mutant classes of CFTR gene pathologically, reoccurrence of infection over time, gradually progression to mucoid *Pa*, and increasing patient burden and risk of antibiotics resistance. The delayed effects lead to controversial issues on schedule of antibiotics and increased research activities to assess the long term benefit of the early aggressive eradication therapy of *Pa*. To summarize, there are three key desired features in the solution: tailoring to individual patient, dynamic adapting to time varying prognostic factors, and incorporating delayed effects to maximize long term benefit.

The closest related previous works have been referred to variously as "dynamic treatment regime", "adaptive treatment strategies". An adaptive treatment strategy is characterized by a sequence of individually tailored decision rules, each of which specifies how to treat a patient at the critical decision point based on observation of the patient up to that point in the medical care process. These strategies were utilized in a variety of health related areas, such as the treatment of alcoholism, smoking cessation, cocaine abuse, depression and hypertension (Murphy, et al., 2005a,

2007a, 2007b; Lavori, et al., 2000a; and Collins, et al., 2004), and acute HIV infection (Altfel and Walker, 2001; Albert and Yun, 2000). Adaptive treatment strategies operationalize the clinical practice of adapting treatment options based on patient's progress thereby facilitating systematic study and refinement.

In developing adaptive treatment strategies, one might ask if we can use the usual meta-analysis on multiple trials in separate courses. This simple "piece together" approach may not appropriate, because it ignores the potential delayed treatment effects stimulated by the synergy of early-stage and late-stage treatments and the potential cohort effects leading to study population shift. From a study type point of view, there are two approaches to identify and compare adaptive treatment regimes: experimental and observational approaches. The experimental approach is based on a Sequential Multiple Assignment Randomization Trial (SMART) (Murphy, 2005a) in which subjects are randomized to follow different regimes in multiple stages. Examples of SMARTs include: the CATIE 2001 trial involving the treatment of psychosis in Alzheimer's disease, schizophrenia patients (Schneider, et al., 2001); the STAR*D 2003 trial investigating the treatment of depression; the ongoing Pelham study involving the treatment of ADHD (Rush, et al., 2003; Lavori, et al., 2001); ongoing Oslin trial investigating the treatment of Alcohol Dependence (Oslin, et al., 2003). The observational approach is based on observational longitudinal data, where treatments actually received over time have been recorded along with other information for each subject. Implementation examples include constructing adaptive treatment strategies from cancer and leukemia group B (CALGB) protocol 8923 study; the Enhanced Suppression of the Platelet IIb/IIIa Receptor with Integrilin Therapy (ESPRIT) trial in nonurgent coronary stenting (O'Shea, et al., 2001). In our preliminary study, we apply the experimental approach in a simulation study, and discuss more details in Section 3.

Methodological challenges for developing adaptive treatment strategies primarily come from two considerations: methods must incorporate the effect of future treat-

ment decisions when evaluating the present treatment decision; the specified model associated with the methods needs to be chosen appropriately to the application scenarios and driven by practical consideration. Murphy and Robins have pioneered statistical methods for inferring the optimal regime from both experimental and observational data (Murphy, 2002; Robins, 2004). Murphy proposed semiparametric methods for estimating the optimal rules when the multivariate distribution of covariates and outcome is unknown. The parametric part was used to estimate those optimal rules by modeling the regret function. The iterative minimization procedure was utilized to identify the optimal rules. Robins proposed g-estimation methods in structural nested mean models (SNMM). Under some regularity conditions, these estimates were most efficient. Another novel Bayesian-frequentist compromise approach was also proposed by Robins. Moodie et al showed that Murphy's and Robin's approaches are closely related (Moodie, et al., 2006).

The other related work in the literatures involves personalized medicine discovery. The term "personalized medicine" usually refers to the application of genomic and molecular data to better target the delivery of health care. Essentially, personalized medicine is in many ways an extension of traditional clinical medicine taking advantage of the cutting edge of genetics research. For instance, there is cutting edge research on the impact of genetic variation on the efficacy and safety of medication in pharmacogenetics, the disease-causing mutation to inform "at risk" individuals, and targeted therapy designed to target aberrant molecular pathways in cancer management. Some of the first instances of personalized therapy at work include Herceptin in treating breast cancer with HER2 over expressed patients (Slamon, et al., 2001; Romond, et al., 2005), and Gleevec in treating chronic myelogenous leukemia (CML) patients with Bcr-Abl mutation (Druker, et al., 2001, 2006). Compared to personalized therapy that we described in earlier chapters, personalized medicine is contained within the individualized or tailoring therapy, which is a larger concept that encompasses the many different types of personalized approaches to medicine.

Adaptive clinical trials hold the promise of radically altering the clinical development process and boosting the biopharmaceutical industry's return on investment in drug development. The adaptation on study design aspects based on accruing information on the ongoing trial is the key component of adaptive design. The terminology "adaptive" in these examples include the adaptative nature of treatment itself and the adaptation in the design aspect, for example, adaptive randomization allocation to the next cohort in the trial. There are a lot of implementations and research activities. The adaptive designs in developing individualized therapy usually are biomarker-adaptive design. From the design prospective, biomarkers can be used in targeted design, biomarker-stratified design, and enrichment design (see, e.g., Wang, 2007; Simon, 2008). From the design adaptation prospective, biomarkers and early responses can serve as bases for adaptive randomization and sample size reassessment for enriching subgroups. From the modeling perspective, many researchers study the relationship between treatment, biomarkers and outcome by parametric models. For example, under the Bayesian framework (see, e.g. Thall, et al., 2000, 2002), some biomarker response adaptive designs are very efficient and have been successfully implemented in many trials. Sequentially randomized trials with adaptive randomization had been addressed in a Bayesian framework and analyzed for optimal regimes using a likelihood approach (Thall, et al., 2002). However, usually these trials did not take advantage of their multiple course sequential nature, but rather treated each phase as a separate trial. Moreover, frequent adaptation based on earlier biomarker or immediate treatment outcome indeed does not fit the goal of long term benefit in our chronic disease management setting.

Computer scientists have also developed parallel reinforcement learning techniques to tackle these problems (see, e.g., Sutton, 1998). They have been formulated in a framework under which a summarized scalar value outcome is obtained from sequential interactions with the environment and solved by backward and/or recursive algorithms. Reinforcement learning has been applied to treating behavioral disorders,

where each patient typically has multiple opportunities to try different treatments (Pineau et al, 2007; Ernst et al, 2006). Q-learning is one of the most important breakthroughs in reinforcement learning, for constructing decision rules for chronic psychiatric disorders, since such chronic conditions often require sequential decision making to achieve the best clinical outcomes. The extension and melding of adaptive treatment strategy, reinforcement learning and statistical learning have been introduced and investigated in solving dosage and timing issues on therapy for advanced non-small cell lung cancer (NSCLC) patients (Zhao et al, 2009). Their research results showed that some current statistical learning methods have the advantage of flexibility and facility to handle many variables and complex nonlinear relationships, and thus are capable of avoiding many of the problems caused by model misspecification. However, the time-indexed solutions in previous works are not suitable to handle the frequently repeated courses of age related treatment, which are common and challenging in CF and in other chronic diseases. In our study, we employ a similar combination approach but without a time-index in the CF therapeutic area, and discuss more details about fitted Q iteration reinforcement learning methods in Chapter 5.

3 REINFORCEMENT LEARNING IN MEDICAL DECISION MAKING

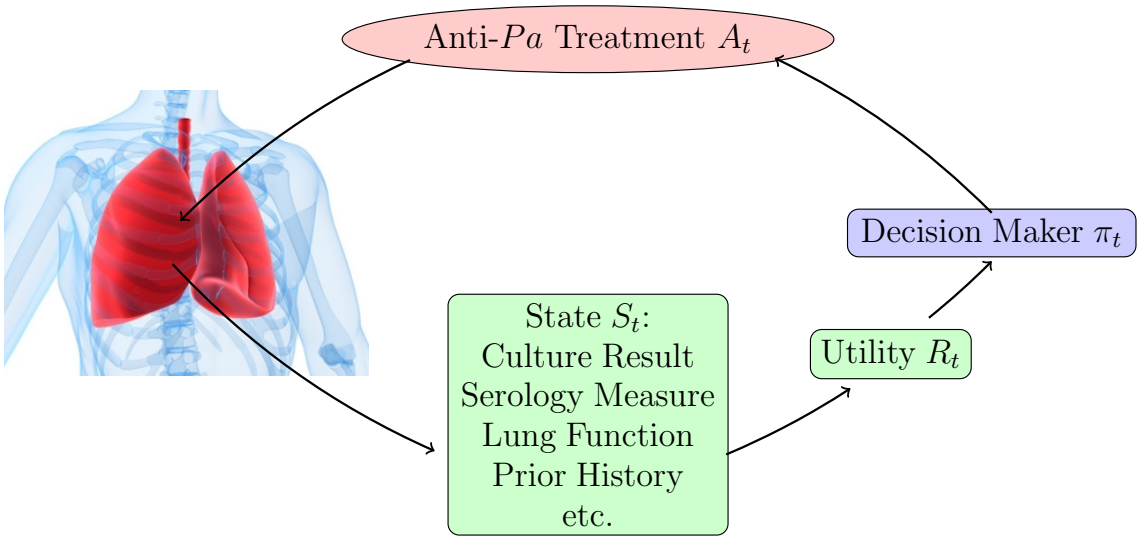
In this section, our main aim is to introduce the reinforcement learning framework, particularly, Q-learning and fitted Q iteration combined with SVR which will be used to discover optimal adaptive personalized therapy for a cystic fibrosis reinforcement clinical trial presented in Section 4. In section 3.1, we introduce the key elements of reinforcement learning in the setting related to our study and three fundamental classes of methods for solving the reinforcement learning problem along with a unified view. In section 3.2, we take a step closer to Q-learning which can be efficiently combined with many supervised learning methods.

3.1 Reinforcement Learning Framework

Reinforcement learning (RL) is an active sub-area of machine learning and artificial intelligence concerned with how an agent ought to take actions in an environment so as to maximize some notion of long-term reward. RL is a powerful artificial intelligence technique in which an agent learns to optimize sequences of actions in an evolving system by exploring possible action sequences, receiving both the long and short term consequences for those actions, and estimating the relationship between actions and consequences (Kaelbling, et al., 1996; Sutton, et al., 1998).

Reinforcement learning uses a formal framework defining the interaction between a learning agent and its environment in terms of states, actions, and rewards. The agent, for example, the physician, interacts with the dynamic environment, which represents the complex system consisting of the CF patient body and more sources of

Figure 1: Reinforcement learning in anti-Pa therapy treating lung infection for cystic fibrosis



error and restrictions on what can be measured in our study. Based on the patient's states, such as clinical status, the agent makes decisions, and gives an action that assigns some treatment to the patient. By measuring the patient's clinical status at the next visit, the agent gets feedback, the so called reward for the previous action. While these interactions continually happen, the agent chooses a sequence of actions applied to the patient, and gets feedback from the response to those actions from patients. If the feedback is positive, when facing the same situation in the future, the agent is more likely to apply that action, and vice versa: this is what meant by the term "reinforcement". The goal is to maximize the long term cumulative reward. A large class of real world problems can be formulated as such stochastic multistage decision processes. The focus is more on goal-directed learning from interaction than on other approaches. A detailed survey of the reinforcement learning literature can be found in Sutton and Barto (1998).

First, we introduce the key elements of reinforcement learning in the more specific setting related of the medical decision making. The key elements of reinforcement

learning include "state" S_t, "action" A_t and incremental "reward" R_t at the t-th decision time, $t = 0, ..., T$. Let S_t represents the set of environmental "states", corresponding to the vector of patient information at that time, such as time-varying sputum culture results in CF, serology measures, pulmonary function tests, prior response, treatment history and baseline characteristics including mutation class, etc.

The action A_t refers to the treatment given at that decision point. Specifically, we use upper case letters, such as S and A, to denote random variables, and use lower case letters, such as s and a, to denote the realized values of the random variables S and A, respectively. Let $\bar{S}_t = (S_0, ..., S_t)$ and $\bar{A}_t = (A_0, ..., A_t)$ represent histories of state and action.

At each time step, after a series of treatments, the agent receives a numerical reward R_t. The reward is defined as a function of action and state, which maps the previous states, actions series, and next state to a single number, i.e., $R_t = r_t(\bar{S}_{t+1}, \bar{A}_t)$. It is typically a clinically meaningful outcome, which reflects the benefit-risk assessment base on previous treatment history and responses at different time points at the individual level. It reflects the immediate utility that contributes to the ultimate patient outcome of interest. For example, the immediate status of Pa infection stage and lung function contribute to future transition to mucoid Pa status and overall survival of CF patients.

The discounted cumulative return cr_t is given by the following equation, where γ is the discount rate ($0 \leq \gamma \leq 1$), balancing the weights of a patient's immediate rewards and future rewards, i.e.

$$cr_t = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \ldots + \gamma^T r_{t+T} = \sum_{k=0}^{T} \gamma^k r_{t+k},$$

In this equation, γ is the discount rate ($0 \leq \gamma \leq 1$), which means, rewards that are received in the future are geometrically discounted according to γ. Additionally, we can interpret γ in another way. It can be seen as a control to the balance the agent's immediate rewards and future rewards. If $\gamma = 0$, we easily see that $R_t = r_t$, we only

need to learn how to choose a_t so as to maximize/minimize the immediate reward r_t. As γ approaches 1, we take future rewards into account more strongly. In the extreme case, when $\gamma = 1$, we fully maximize/minimize rewards over the long run.

Figure 1 gives a schematic of the fundamental components of reinforcement learning described above in the anti-Pa therapy context for CF. The available data from either clinical practice, observational studies or sequential randomized trials, are realizations of the time-order random variables

$$(S_0, A_0, R_0, ..., S_T, A_T, R_T, S_{T+1}).$$

The "policy" $\pi_t(\bar{s}_t, \bar{a}_{t-1}) = a_t$ maps from the state-action history $(\bar{s}_t, \bar{a}_{t-1})$ to the probability that action a_t is taken. In the deterministic setting, the policy is the decision rule about which treatment to assign to patients given the history, i.e. $\pi_t(\bar{s}_t, \bar{a}_{t-1}) = a_t$. We denote the distribution of a patient's longitudinal trajectories as P_π, and expectations with respect to P_π as E_π, when the policy π is applied to generate actions. The goal and principle of reinforcement learning are learning what to do, how to map situations from state space S to action space A, how to choose a_t, to find the optimal policy resulting in the maximum expected discounted return $\sum_{t=0}^{T} \gamma^t r_t$. By seeking action sequences that maximize the cumulative return, we optimize benefit to achieve a favorable outcome. This corresponds to our aim to discover the optimal personalized therapy which achieves the most beneficial ultimate outcome in the long run.

To accomplish this goal, a "state-value function" $V_t(\bar{s}_t)$, which is formulated as a function of the history of state $\bar{s}_t$, represents the total amount of rewards expected to accumulated in the future for the agent to start from some state and following some policy π afterward. The optimal value function $V_t^*(\bar{s}_t)$ is defined by maximizing over $\pi \in \Pi$:

$$V_t(\bar{s}_t) = E_\pi\Big[\sum_{k=0}^{T}\gamma^k R_{t+k}\Big|\bar{S}_t = \bar{s}_t\Big],$$

$$V_t^*(\bar{s}_t) = \max_{\pi\in\Pi} V_t(\bar{s}_t) = \max_{\pi\in\Pi} E_\pi\Big[\sum_{k=0}^{T}\gamma^k R_{t+k}\Big|\bar{S}_t = \bar{s}_t\Big].$$

Using the dynamic programming concept and the Bellman equation (Bellman, 1957), the optimal policy is the decision rule that we expect to yield the highest long term reward:

$$\pi_t^*(\bar{s}_t) \in arg\max_{a_t} V_t^*(\bar{s}_t),$$

$$\pi_t^*(\bar{s}_t) \in arg\max_{a_t} E\Big[R_t + \gamma V_{t+1}^*(\bar{S}_{t+1})\Big|\bar{S}_t = \bar{s}_t, \bar{A}_t = \bar{a}_t\Big].$$

A fundamental property of value functions used throughout reinforcement learning is that they satisfy particular recursive relationships. To see this, first let $T = \infty$, then we extend equation (2.2) as follows,

$$\begin{aligned}
V_t(\bar{s}_t, \bar{a}_{t-1}) &= E_\pi\Big[\sum_{k=0}^{\infty}\gamma^k r_{t+k}\Big|\bar{S}_t = \bar{s}_t, \bar{A}_{t-1} = \bar{a}_{t-1}\Big] \\
&= E_\pi\Big[r_t + \gamma\sum_{k=0}^{\infty}\gamma^k r_{t+k+1}\Big|\bar{S}_t = \bar{s}_t, \bar{A}_{t-1} = \bar{a}_{t-1}\Big] \\
&= E_\pi\Big[r_t + \gamma V_{t+1}(\bar{S}_{t+1}, \bar{A}_t)\Big|\bar{S}_t = \bar{s}_t, \bar{A}_t = \bar{a}_t\Big] \\
&= \sum_{a_t}\pi_t(\bar{s}_{t+1}, \bar{a}_t)\sum_{s'}\mathcal{P}_{ss'}^a\Big[\mathcal{R}_{ss'}^a + \gamma V_{t+1}(s')\Big],
\end{aligned}$$

where $\mathcal{P}_{ss'}^a = Pr\{s_{t+1} = s'|\bar{s}_t = s, \bar{a}_t = a\}$ and $\mathcal{R}_{ss'}^a = E\Big[r_t|\bar{s}_t = s, \bar{a}_t = a, s_{t+1} = s'\Big]$. The last two equations are two forms of the *Bellman equations* for $V_t(\bar{s}_t, \bar{a}_{t-1})$. The Bellman equation was first introduced by Bellman (1957). The Bellman equation expresses the relationship between the value of a state and the values of its successor states: the value of the start state is equivalent to the value of the expected next state plus the expectation of the reward along the way. It's worth noting that the value function $V_t(\bar{s}_t, \bar{a}_{t-1})$ is the unique solution to its Bellman equation.

To summarize, the key elements in reinforcement learning include agent-environment interface (state, action, reward), a goal and return (immediate and discounted cumulative rewards), a policy (maps from state-action history to next action), and a value

function (performance, in terms of expected future return). Additionally, the Markov property is assumed in many reinforcement learning problems and serves as a foundation case for extension to more complex and non-Markovian cases.

Modern techniques and methods for estimating optimal value functions or optimal policies can be categorized into one of the following three classes: dynamic programming, Monte Carlo method and temporal difference learning (Sutton, 1998; Kaelbling, et al., 1996; Bertsekas, et al., 2000). Assuming the environment is a finite Markov decision process (MDP), its dynamics are given by a set of transition probabilities, and the expected immediate rewards. If the Markov assumption holds for the environment, then the environment's response at $t+1$ depends only on the state and action representations at t, and we can replace $\bar{s}_t$ with s_t and $\bar{a}_t$ with a_t. Overall, at each time step t, (s_t, a_t, s_{t+1}, r_t) provides the knowledge of full information for the agent. The key idea is to use a step-by-step based generalized policy iteration (GPI), which is the generalized idea of two interacting processes revolving around an approximate policy and an approximate value function. The valuation process takes the policy as given and performs some form of policy valuation. The improvement process takes the value function as given and performs some form of policy improvement in order to make it better. Each process changes the bases for the other, overall they work together to find a joint solution, which is optimal when both the policy and value function are unchanged by either process. The reason that this iteration can be realized in a step-by-step manner is that we assume the perfect model of the environment as an MDP exists. The backup updates the value of one state based on the values of all possible one step successor states and their transition probabilities from the known model.

In the other hand, the Monte Carlo method requires the sample episodes instead of a model of the environment's dynamic. In the interacting GPI processes, Monte Carlo methods provide an alternative policy evaluation process. Rather than use a model to compute the values of each state, they average many returns that start in

the state and don't update the value estimates based on other value estimates. Thus, the GPI process is incrementally implemented on an episode-by-episode basis in a model-free manner.

The third method, temporal difference (TD) learning, is a combination of DP and Monte Carlo ideas. Like Monte Carol methods, TD methods can learn directly from raw experience in a model-free manner and avoid the harm of the violation of the Markov property. Like DP methods, TD methods update estimates based in part on other learned estimates, without waiting to the end of the sample episode for the final outcome. For control to find an optimal policy, TD, DP and Monte Carlo methods all employ some variation of GPI, for example, a greedy search or some searching with exploitation and exploration balance or softmax selection, or learning in an off-line or on-line fashion.

One fundamental expression of TD-learning is the incremental implementation, which requires less memory and computation. The general form is

$$\text{new estimate} \leftarrow \text{old estimate} + \text{stepsize}\Big[\text{target} - \text{old estimate}\Big].$$

Specifically, if we replace *estimate* with value function, *target* with reward function, and denote *stepsize* as α, then in this case TD learning becomes

$$V_t(\bar{S}_t, \bar{A}_{t-1}) \leftarrow V_t(\bar{S}_t, \bar{A}_{t-1}) + \alpha\Big[r_t + \gamma V_{t+1}(\bar{S}_{t+1}, \bar{A}_t) - V_t(\bar{S}_t, \bar{A}_{t-1})\Big].$$

Roughly speaking, the TD method bases its incremental implementation in part on an existing estimate. Recalling the Bellman equation, we know that

$$\begin{aligned} V_t(\bar{s}_t, \bar{a}_{t-1}) &= E_\pi\Big[\sum_{k=0}^{\infty} \gamma^k r_{t+k} \Big| \bar{S}_t = \bar{s}_t, \bar{A}_{t-1} = \bar{a}_{t-1}\Big] \\ &= E_\pi\Big[r_t + \gamma \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \Big| \bar{S}_t = \bar{s}_t, \bar{A}_{t-1} = \bar{a}_{t-1}\Big] \\ &= E_\pi\Big[r_t + \gamma V_{t+1}(\bar{S}_{t+1}, \bar{A}_t) \Big| \bar{S}_t = \bar{s}_t, \bar{A}_t = \bar{a}_t\Big]. \end{aligned}$$

In these equations, under a policy π, each V represents the true value of a state-action pair, but this is not known. Thus, the TD target uses the current estimate V

instead of the true V. TD learning as discussed above is also known as TD(0) learning, which is a special case of TD(λ) learning. Almost any TD(λ) learning belongs to the "eligibility traces" problem. The one step TD and Monte Carlo methods are actually corresponding to two cases in TD(λ) as 0 or 1. The choice of λ depends on the uncertainty on a state's value estimates. By adjusting λ, we can place eligibility trace methods anywhere between a continuum from the Monte Carlo to one-step TD methods. It protects against both long-delayed rewards and non-Markovian tasks with less memory requirement and computational expense.

3.2 Q-learning

A key break through and the most widely used algorithm in reinforcement learning off-policy TD learning is Watkins' Q-learning (Watkins, 1989; Watkins and Dayan, 1992). Q-learning no longer requires estimating the value function, it estimates a Q-function instead. From a statistical perspective, the optimal time-dependent Q-function is

$$Q_t^*(\bar{s}_t, \bar{a}_t) = E\Big[r_t + V_{t+1}^*(\bar{S}_{t+1}, \bar{A}_t)\Big|\bar{S}_t = \bar{s}_t, \bar{A}_t = \bar{a}_t\Big].$$

Note that since $V_t^*(\bar{s}_t, \bar{a}_{t-1}) = \max_{a_t} Q_t^*(\bar{s}_t, \bar{a}_{t-1}, a_t)$, we have

$$\pi_t^*(\bar{s}_t, \bar{a}_{t-1}) = arg\max_{a_t} Q_t^*(\bar{s}_t, \bar{a}_{t-1}, a_t)$$

as an optimal policy. So once one has Q^*, it is relatively easy to determine an optimal policy.

Under some appropriate and rigorous assumptions, Q_t has been shown to converge to Q^* with probability 1 (Watkins and Dayan, 1992). More general convergence results were proved by Jaakkola, et al. (1994) and Tsitsiklis, et al. (1994).

This approach has advantages of the simple recursive form, minimal computational expense and the capacity for effectively combining with many statistical learning methods. Instead of maximizing the value function, Q-learning optimizes the Q-

function which has the direct relationship with value function, and is easy to use to determine the optimal policy.

First, we describe the fitted Q procedure to learn the optimal policy under both non-Markovian and Markovian settings in batch mode.

The estimator of the non-Markovian policy based on a training data set will be denoted by the series of time-indexed Q-functions, Q_t, where $t = 0, 1, \ldots, T$. Its fitted procedure is based on the one-step Q-learning recursive form:

$$Q_t(\bar{s}_t, \bar{a}_t) = E\Big[R_t + \gamma \max_{a_{t+1}} Q_{t+1}(\bar{S}_{t+1}, \bar{A}_{t+1})\Big|\bar{S}_t = \bar{s}_t, \bar{A}_t = \bar{a}_t\Big]. \tag{3.1}$$

Backward induction is the key point of the optimization, which starts at the end with the effects of the last treatment at $t = T$ and works backward through time $t = T-1, \ldots, 1$, until to Q_0 at the beginning of the trajectories. The input/output pairs at the previous time point are obtained by merely refreshing the output values by the one-step recursive form, and are used to approximate Q_t by any (parametric or non-parametric) regression architecture. After this backwards approximation is done, we can easily estimate the optimal policies $\widehat{\pi}_t$ from the sequence $\{\widehat{Q}_0, \widehat{Q}_1, \ldots, \widehat{Q}_T\}$ by $\widehat{\pi}_t(\bar{s}_t) = arg\max_{a_t} \widehat{Q}_t(\bar{s}_t, \bar{a}_t; \theta_t)$, for $t = 0, 1, \ldots, T$.

In learning a non-stationary non-Markovian policy with one set of finite horizon trajectories (training data set)

$$\{S_0, A_0, R_0, S_1, A_1, R_1, \ldots, A_T, R_T, S_{T+1}\},$$

we denote the estimator of the optimal Q-functions based on this training data by $\hat{Q}_t$, where $t = 0, 1, \ldots, T$. According to the recursive form of Q-learning in (2.8), we must estimate Q_t backwards through time $t = T, T-1, \ldots, 1, 0$, that is, estimate Q_T from the last time point back to Q_0 at the beginning of the trajectories. And we set Q_{T+1} equal to 0 in the first equation. One-step Q-learning has a simple recursive form

$$Q_t(\bar{S}_t, \bar{A}_t) \leftarrow r_t + \gamma \max_{a_{t+1}} Q_{t+1}(\bar{S}_{t+1}, \bar{A}_t, a_{t+1}).$$

In order to estimate each Q_t, we denote $Q_t(\bar{s}_t, \bar{a}_t; \theta)$ as a function of a set of parameters θ, and we allow the estimator to have different parameter sets for different

time points t. Once this backwards estimation process is done, we save the sequence of $\{\widehat{Q}_0, \widehat{Q}_1, \ldots, \widehat{Q}_T\}$ for estimating optimal policies,

$$\widehat{\pi}_t = arg \max_{a_t} \widehat{Q}_t(\bar{s}_t, \bar{a}_t; \theta_t), \quad t = 0, 1, \ldots, T.$$

and thereafter use these optimal policies to test or predict for a new data set.

The most suitable type of Q-learning for our setting is model-free Q-learning in batch mode with an approximation algorithm (Watkins, 1989). This is because in complicated diseases the relationship between disease dynamics and the unknown treatment effects are impossible to know in advance and should thus be nonparametrically modeled. The batch offline learning mode is more ethical in some medical settings because it protects against potential risks to patients due to inadequately trained solutions in the early stages of online learning. Because algorithms with a tabular representation are infeasible in many real-life medical applications which typically have a continuous state space, continuous and nonparametrically modeled Q-functions are needed.

Additionally, the properties of other promising learning methods based on modification or extension of Q-learning, for example, A-learning (Blatt, et al., 2004) and the Sarsa algorithm (Rummery and Niranjan, 1994), have not been carefully investigated. Thus, we will restrict our attention to Q-learning for our future study's methodology and application.

3.3 Fitted Q-Iteration Algorithm

Under the Markovian assumption, the dynamic environment's response depends only on the current state and action. These one-step dynamics enable us to predict the next state and expected next reward, and serve as the base for choosing actions. In contrast to the non-Markovian case, where the fitted Q-functions have a time-index, we fit the stationary Q-functions through the backward recursive iteration.

The fitted Q-iteration algorithm (Ernst, et al., 2005) makes use of a set of *one-step* dynamic system transition samples in a Markov decision process (MDP) $\mathcal{F} = \{(s_t^l, a_t^l, r_t^l, s_{t+1}^l)\}_{l=1}^{\#\mathcal{F}}$. The recurrence relation of (3.1) in the discrete MDP problem becomes:

$$Q_N^\star(S_t, A_t) = E\Big[R(S_t, A_t) + \gamma \max_a Q_{N-1}^\star(S_{t+1}, a)|S_t, A_t\Big], \forall N > 1 \tag{3.2}$$

with $Q_1^\star(S_t, A_t) = R(S_t, A_t)$. As N increases, this sequence converges in infinity norm to the optimal stationary Q-function. The resulting optimal policy is $\pi_N^*(s) = arg\max_a Q_N^\star(s, a)$.

At each iteration, using the empirical r_t, the approximation (3.2) can be formulated as a sequence of standard supervised learning steps on the kth training sample, taking the form

$$\mathcal{TS}_k = \{(s_t^l, a_t^l, r_t^l + \gamma max_a \hat{Q}_{k-1}(s_{t+1}^l, a))\}_{l=1}^{\#\mathcal{F}}, \forall k > 1$$

with $\mathcal{TS}_0 = \{(s_t^l, a_t^l, r_t^l, s_{t+1}^l)\}_{l=1}^{\#\mathcal{F}}$, $\hat{Q}_0(s, a) = 0, \forall s, a$. The estimated stationary policy is

$$\hat{\pi}_N(s) = arg\max_a \hat{Q}_N(s, a). \tag{3.3}$$

Hence, fitted Q-iteration can be combined with any regression algorithm to fit the Q-function. The extensive testing of fitted Q-iteration in standard RL simulation (Ernst, et al., 2005) and in clinical applications in HIV (Ernst, et al., 2006) and epilepsy (Guez, et al., 2008) demonstrate encouraging performance, even with high-dimensional state spaces, and efficient use of training data. In chronic disease treatment, the frequent regular monitoring provides relatively complete transition samples, and a stationary Q-function is often the most useful policy in practice. Age-specific characteristics can be accommodated by adding age as a covariate.

3.4 Approximation Methods

There are many similarities between reinforcement learning and supervised learning, in both of which a general function is learned from from samples. Especially, the temporal difference family of algorithms can be viewed as supervised learning algorithms in which the training examples consist of the approximation of the true state value $(s, \hat{V}(s))$, or approximation of true state action value $(s, a, \hat{Q}(s, a))$ pairs.

Ormoneit and Sen (2002) formulated the Q-function problem as a sequence of kernel-based regression problems by applying fitted value iteration ideas (Gordon, 1999) to kernel based reinforcement learning. This framework takes full advantage in the context of reinforcement learning of the generalization capabilities of any regression algorithm, and this is not limited to parametric function approximators. Ernst, et al. (2005) and Geurts, et al. (2006) utilized a nonparametric tree based method such as pruned CART, KD-tree, random forests and extremely randomized tree in a batch mode in a simulation of HIV infection (Ernst, et al., 2006) and adaptive treatment of Epilepsy (Guez, et al., 2008).

This fitted-Q iteration has consistency properties when combined with kernel-based, tree-based regressors (Ernst, et al., 2003). In chronic disease treatment, the frequent regular monitoring and repeated treatment provide relatively complete states, which leads to an appropriate Markovian working assumption and the need for the stationary Q-function as the simplified policy in practice. In the Wisconsin CF neonatal screening project (Farrell, et al., 1997), where the patients are diagnosed at birth and continuously monitored afterward, we add age as one of the covariates in the Q-function to preserve the time-varying age-specific characteristic in the personalized therapy.

Due to challenges that may arise from the complexity of the true Q-function, including the non-smooth maximization operation and the potential high-dimension of the state and action variables, we apply support vector regression (SVR) (Watkins,

1989) as the main approximation method for fitting the Q-functions.

As one of the most popular extensions of the support vector machine, SVR is derived within the reproducing kernel Hilbert space (RKHS) context to minimize the ϵ-insensitive loss function, which is defined as $L(f(\bar{x}_i), y_i) = (|f(\bar{x}_i) - y_i| - \epsilon)_+$, $\epsilon > 0$ (Vapnik, 1995; Vapnik, et al., 1997). Given training data $\{(\bar{x}_i, \bar{y}_i) \in X \times Y\}_{i=1}^n$, SVR solves the following optimization problem:

$$\min_{\bar{w},b,\xi,\xi'} \frac{1}{2}\|\bar{w}\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i'),$$

$$\text{subject to } \ (\bar{w}^T\Phi(\bar{x}_i) + b) - y_i \leq \epsilon + \xi_i, \ \ y_i - (\bar{w}^T\Phi(\bar{x}_i) + b) \leq \epsilon + \xi_i',$$

where $\xi_i, \xi_i' \geq 0, i = 1, \ldots, n$. Since errors within deviation ϵ are considered acceptable and the data is mapped through the nonlinear transformation into a feature space, SVR is a more general and flexible approach compared to competing methods to handle the potentially complex nonlinear relationship between rewards and state-action pairs. Also, by minimizing the regularization term $\frac{1}{2}\|\bar{w}\|^2$ and the training error $C\sum_{i=1}^{n}(\xi_i + \xi_i')$, SVR can avoid overfitting the training data and yield both fast and high quality performance. SVR performs with similar or better reproducibility in clinical research settings (Zhao, et al., 2009) as extremely randomized trees (Geurts, et al., 2006), a popular, more computationally intense alternative also used for fitted Q-iteration.

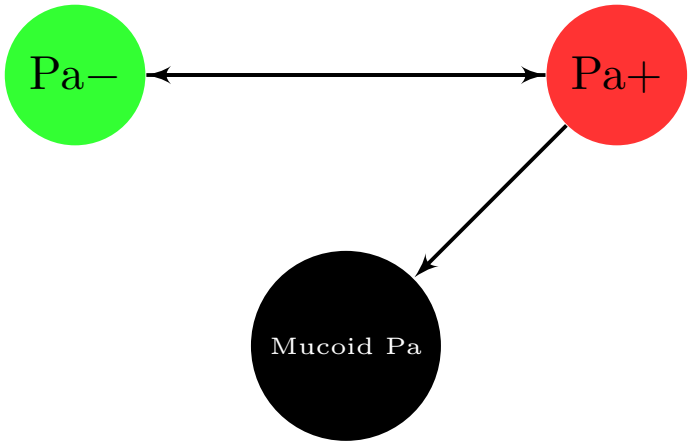
4 DISEASE DYNAMICS OF CYSTIC FIBROSIS

In this section, we propose a discrete time nonhomogeneous Markov model for the underlying disease dynamic in *Pa* infection. Dynamic analysis in longitudinal medical studies, where patients may experience several clinical states, is naturally carried out by use of multi-state models, which is a model for the stochastic process allowing individuals to move through a series of states. Based on the biological and clinical rationale in Section 4.1, we present the model structure in Section 4.2. We then present two approaches to obtain the model parameters. The first set of parameters in Section 4.3 is obtained by tuning the parameters to match the results in literature. The second set of parameters is obtained through maximum likelihood estimation procedure in Section 4.4. The model with two sets of parameters will be used in the virtual clinical trial studies and in the evaluation of long term outcome in Chapter 6.

4.1 Model Rationale

To obtain data which mimics real life clinical data for CF patients with *Pa* infection, we briefly review prior knowledge of this disease process. After being diagnosed at birth, children with CF usually acquire nonmucoid *Pa*, which is transient and can possibly be eradicated by aggressive anti-*Pa* antibiotics (Kosorok, et al., 2001; Meira-Machado, et al., 2009; Taccetti, et al., 2005; Treggiari, et al., 2007). Mucoid *Pa*, a mutant phenotype of *Pa*, develops at later stages, and lives in a defensive mode of growth called a biofilm (Prince, 2002). Hence it confers resistance to phagocytosis and antibiotics and is much more difficult to treat and eradicate (Gibson, et al., 2003). Therefore, there are three phenotypically distinguishable states: free of *Pa* (state 1),

Figure 2: *Pa* infection progression in 3-state Markov Model for cystic fibrosis



nonmucoid *Pa* (state 2), and irreversible mucoid *Pa* (state 3), as illustrated in Figure 2.

In CF clinical trials, there are three major classes of endpoints, clinical efficacy measures, surrogate endpoints and biomarkers. First, common clinical efficacy measure for definitive clinical trials in lung infection is pulmonary exacerbations. Secondly, FDA defines both FEV1 at a given time point and rate of decline in FEV1 as surrogate endpoints because they are a well established predictor of survival. Thirdly, one of the most established sets of biomarkers in CF is microbiological parameters relating to Pa (Murphy, et al. 2001). The progression to mucoid state is associated with irreversible damage of lung function (Li, et al., 2002), and many studies have demonstrated the reduction of Pa bacterial density or eradication of Pa lead to significant improvement in FEV1 and reduction in pulmonary exacerbations (Flume, et al., 2009).

Hence, the transitions between three states in *Pa* infection are closely related to both biological patheogensis and clinical outcomes. In this section, we propose a discrete time nonhomogeneous Markov model for this underlying disease dynamic in Pa infection.

Table 1: Patient outcomes and biomarkers collected in regular study visits

Patient Information	**Definition**
$\Delta F508H$	$\Delta F508/\Delta F508$ at CFTR residue 508 indicator
$Cul(t)$	Pa phenotype nonmucoid + isolated from respiratory culture
$Ser(t)$	Pa serology tests + indicator
$Muc(t)$	Pa phenotype mucoid + isolated from respiratory culture
$FEV_1(t)$	Pulmonary function test predicted FEV_1
$D_2(t)$	Cumulative duration in nonmucoid infection
$Sev(t)$	Severity: 50% Pa + in past year to divide as chronic or intermittent
$Cum_D(t)$	Cumulative intensity of drug D exposure
$Sus_D(t)$	Susceptibility tests result of drug D

We aim at optimizing the maintance therapy of Pa infection, while pulmonary exacerbation usually happens after the mucoid infection development. The patient outcomes that we simulated include the observed Pa infection state and severity and FEV_1 based on the underlying true state. In Chapter 5, we incorporate these outcomes into a benefit-risk assessment for guidance of the optimal therapy that we are seeking for. Table 1 shows the content of patient information and outcomes typically collected.

The transitions between three states in Pa infection are closely related to both biological pathogenesis and clinical outcome. Specifically, progression to the mucoid state is associated with irreversible damage of lung function (Kosorok, et al., 2001; Langton and Smyth, 2009), and many studies have demonstrated that reduction of Pa bacterial density or eradication of Pa leads to significant improvement in FEV_1 and reduction in pulmonary exacerbations (Gibson, et al., 2003). Motivated by regularity of clinical patient observations and the progressive nature of CF, we propose a discrete time non homogeneous Markov model for Pa infection.

4.2 Probability Model

The proposed multi-state model is expressed as a continuous stochastic process with a finite state space and time-homogeneous assumption, and is partly motivated by competing risks survival analyses from earlier work (Kalbfleisch, et al., 1985; Meira-Machado, et al., 2009; Putter, et al., 2007). For non-homogeneous processes, the model is either reduced to the homogeneous case or fitted through piecewise constant transition intensities between different time points (Meira-Machado, et al., 2009).

We propose a multi-state model that expresses the underlying disease dynamics as a discrete-time stochastic process $Y(t)$, for $t = 0, 1, \ldots$, with transitions between three states having covariate-dependent transition probabilities $p_{ij}(s, t, \mathbf{Z}(s))$ dependending on time-dependent covariates Z(s), denoted

$$p_{ij}(s, t, \mathbf{Z}(s)) = pr\{Y(t) = j | Y(s) = i, \mathbf{Z}(s)\}, (s < t),$$

and with the one time unit step transition matrix $P(t, t+1, \mathbf{Z}(t))$ having the structure

$$\begin{bmatrix} 1 - p_{12}(t, t+1, \mathbf{Z}(t)) & p_{12}(t, t+1, \mathbf{Z}(t)) & 0 \\ p_{21}(t, t+1, \mathbf{Z}(t)) & 1 - p_{21}(t, t+1, \mathbf{Z}(t)) - p_{23}(t, t+1, \mathbf{Z}(t)) & p_{23}(t, t+1, \mathbf{Z}(t)) \\ 0 & 0 & 1 \end{bmatrix}.$$

Based on longitudinal studies of *Pa* development (Burns, et al., 2001; Gibson, et al., 2003; Kosorok, et al., 2001; Li, et al., 2002; Rosenfeld, et al., 2001; Starner, et al., 2005), $p_{ij}(t, \bar{Z}(t))$ is related to individual characteristics through the time-dependent covariates $\bar{Z}(t))$, consisting of age, $\Delta F508H$, $Trt(t)$, $Cul(t)$, $Ser(t)$, $D_2(t)$, with corresponding definitions given in Table 1. First, the probability of first acquisition of nonmucoid, $p_{12}(t)$, depends on age, and mutation class in CFTR at residue 508. $\Delta F508H$ indicates $\Delta F508$ homozygosity or not. Secondly, the probability of successful eradication of nonmucoid Pa infection, $p_{21}(t)$, relates to treatment effect, age and

$\Delta F508H$. Because of the relatively low sensitivity of throat sputum cultures issue in CF, the detection of nonmucoid *Pa* infection can be improved by combining with serology measurements as reflected in antibody titer levels, as the detection criteria. In our model, the observated nonmucoid infection is determined by the product of culture *Pa* + indicator, $Cul(t)$, and serology tests + indicator, $Ser(t)$. These are Bernoulli random variables which are linked to the true state 2 by the published sensitivities of these tests. Thirdly, the probability of progression to mucoid *Pa* infection, $p_{23}(t)$ depends on the true cumulative duration in nonmucoid infection $D_2(t)$, *age* and $\Delta F508H$.

The following generalized logistic model accounts for the time varying treatment structure, biomarkers, and prognostic covariates. We denote the linear components at time t by

$$\eta_{12}(t, \mathbf{Z}(t)) = \beta_{12_0} + \beta_{12_1} t + \beta_{12_2} \Delta F508H,$$

$$\eta_{21}(t, \mathbf{Z}(t)) = \beta_{21_0} + \beta_{21_1} t + \beta_{21_2}(t) Trt(t)^{Cul(t) \times Ser(t)} + \beta_{21_3} \Delta F508H, \tag{4.1}$$

$$\eta_{23}(t, \mathbf{Z}(t)) = \beta_{23_0} + \beta_{23_1} t + \beta_{23_2} D_2(t) + \beta_{23_3} \Delta F508H.$$

We characterize the regression of one time unit step transition at time t on time-dependent covariates $\bar{Z}(t)$ by probability functions

$$p_{12}(t, t+1, \mathbf{Z}(t)) = \frac{exp(\eta_{12}(t, \mathbf{Z}(t)))}{1 + exp(\eta_{12}(t, \mathbf{Z}(t)))}$$

$$p_{21}(t, t+1, \mathbf{Z}(t)) = \frac{exp(\eta_{21}(t, \mathbf{Z}(t)))}{1 + exp(\eta_{21}(t, \mathbf{Z}(t))) + exp(\eta_{23}(t, \mathbf{Z}(t)))} \tag{4.2}$$

$$p_{23}(t, t+1, \mathbf{Z}(t)) = \frac{exp(\eta_{23}(t, \mathbf{Z}(t)))}{1 + exp(\eta_{21}(t, \mathbf{Z}(t))) + exp(\eta_{23}(t, mathbf Z(t)))}$$

The formulation in (4.1) and (4.2) also accommodates an arbitrary number of treatment courses as well as options for either discrete or continuous time. Because

we aim at optimizing the maintenance therapy of Pa infection, the patient outcomes simulated include the observed Pa infection state, severity, and FEV_1 based on the underlying true state.

Because for the patients who have never been infected, the transition probability p_{12} is assumed to be different from the patients who have been infected before, we use the logistic regression model with age as covariates to model p_{12}:

$$p_{12}(t, t+1) = \frac{exp(\beta'_{12_0} + \beta'_{12_1} t)}{1 + exp(\beta'_{12_0} + \beta'_{12_1} t)}.$$

4.3 Parameter Tuning

We tune the model so that when under standard care or some anti-Pa antibiotic treatment, the patient outcomes roughly match those in prior clinical studies (Armstrong, et al., 1996; Douglas, et al., 2009; Geurts, et al., 2006; Gibson, et al., 2003; Kosorok, et al., 2001; Li, et al., 2005; Mayer-Hamblett, et al., 2007; Pedersen, et al., 1987; Rosenfeld, et al., 1999; Taccetti, et al., 2005). The tuning parameters are $(\beta_{12_0}, \beta_{12_1}, \beta_{12_2})$=$(-1, 0.01, 0.01)$; $(\beta_{21_0}, \beta_{21_1}, \beta_{21_2})$=$(-2.5, -0.05, -0.02)$; $(\beta_{23_0}, \beta_{23_1}, \beta_{23_2}, \beta_{23_3})$=$(-4.8, 0.02, 0.01, 0.32)$ for after the first Pa acquisition. Before the first Pa infection, the parameters are $(\beta'_{12_0}, \beta'_{12_1})$=$(-1, 0.001)$. The anti-$Pa$ treatment effect parameter $\beta_{21_2}(t)$=0 as mentioned before. We list the important clinical outcomes in Table 2, including time to first acquisition and progression to mucoidy, pulmonary function and sensitivity of culture and serology tests, etc. The age prevalences given in (Li, et al., 2005) and average prevalence in 1000 simulated datasets consisting of the same numbers of CF patients are shown in Figure 3. This model not only reflects the important issues in CF clinical care, but also mimics the disease progression in a relatively realistic way.

Table 2: Literature and model generating patient outcomes

Patient Outcomes	**Literature**	**Model Output**
Time to first acquisition of nonmucoid *Pa* (yr)	1.0 (0.5−1.5) [1]	1.0 (0.5−2.5)†
Time to mucoid *Pa* (yr)	13.0 (10.0−14.9) [1]	13.6 (9.84−17.5)†
Pa free duration after eradication (month)	8 (3−25) [3],18 (4−80) [2]	9 (1.54−39)*
$\Delta FEV_1 \cdot yr^{-1}$ standard care	-4.69±2.95% [2]	-4.42±12.2%‡
$\Delta FEV_1 \cdot yr^{-1}$ some anti-*Pa* treatment	-1.63±1.60% [2]	-1.5±8.1%‡
Sputum culture sensitivity	83% [4]	85%
Serology markers sensitivity	93% [5]	93%

† Median (95%CI), * Median (Range) , ‡ Mean ± SD.

[1]: Li, et al., 2005

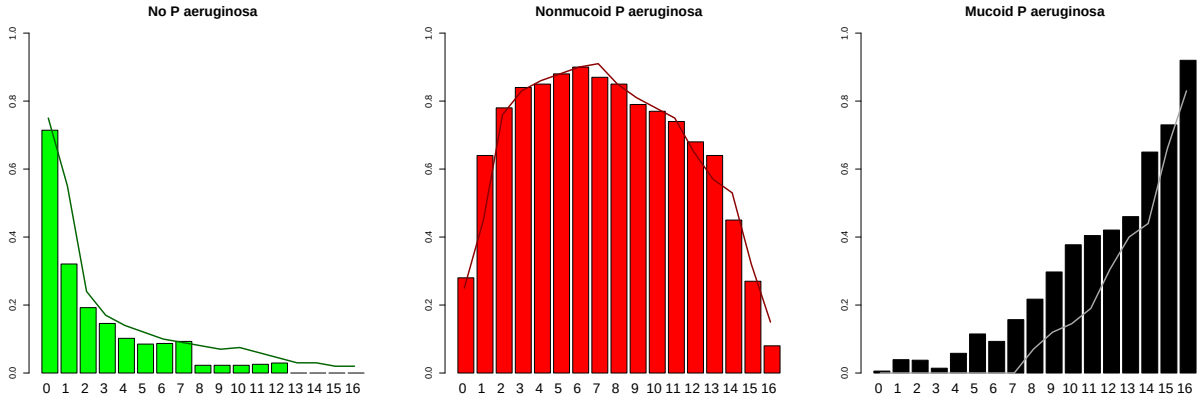
[2]: Starner, et al., 2005

[3]: Douglas, et al., 2009

[4]: Rosenfeld, et al., 1999

[5]: Pedersen, et al., 1987

Figure 3: Age-specific prevalence of no, nonmucoid, and mucoid *Pa* from birth to age 16 years (Li, et al., 2005) and prevalence in simulations. The barplot represents the prevalence in literature (Li, et al., 2005). The line represents the prevalence in simulations.



4.4 Wisconsin Neonatal Screening Data Analysis

The Wisconsin CF Neonatal Screening Project (Farrell, et al., 1997; Farrell, et al., 2003) is a randomized clinical trial conducted to assess neonatal screening for CF using a standardized evaluation and treatment protocol that prevented malnutrition. In this Section, we use a dataset for the study of longitudinal development of mucoid *Pa* infection (Li, et al., 2008), consisting of 69 CF patients at two CF centers in Madison and Milwaukee, Wisconsin, from birth up to age 16 years with 56 cases diagnoses made through the Wisconsin CF Neonatal Screening Project. In this dataset, the 69 cases were followed up every 6 weeks for their first year of life and every 3 months thereafter to age 16 years. The outcomes include culture result and antibiotic susceptibility testing based on respiratory secretions, which were obtained from patients every 6 months by protocol, either by sputum or vigorous oropharyngeal swabbing, with antibody responses to P aeruginosa from serum samples every 6 months coinciding with cultures. Lung function was examed by spirometry every 6 months after age 4 years, including forced expiratory volume in 1 second (FEV1), forced vital capacity (FVC), FEV1/FVC ratio, and forced expiratory flow between 25% and 75% of FVC (FEF25%-75%). The combined culture and serologic (cell lysate titer > 8) positive results were used to define the timing of the first appearance of nonmucoid *Pa* and mucoid *Pa*. All patients have genotype information of ΔF508 homozygosity or not.

The treatment effect of anti-*Pa* drug in this data is not obvious, because "at the discretion of physicians, patients could be treated with anti-*Pa* antibiotics if there were clinically significant infections, but patients did not routinely receive anti-*Pa* antibiotics after the first *Pa* detection, in accordance with the prevailing standard of care" (Li, et al., 2008). Additionally, the transitions with anti-*Pa* antibiotics are about 3% in the total transitions in the dataset. In the following analysis, we use the transitions without anti-*Pa* antibiotics to estimate the disease dynamics under the prevailing standard of care. We use the last observation carried forward to handle

the missing values.

The maximum likelihood estimation procedure is utilized to fit the model specified in (4.1) and (4.2). In the dataset, there are transitions from Pa free state to mucoid pa state, where nonmucoid Pa might happen between the regular 3 months clinical visits. We model the probability p_{13} as

$$p_{13}(t, t+1, \mathbf{Z}(t)) = p_{12}(t, t+1, \mathbf{Z}(t)) \times p_{23}(t, t+1, \mathbf{Z}(t)). \tag{4.3}$$

For reasons of simplicity in presentation, we first consider the situation of just one patient. The vectors for a transition as $Y = (y_{11}, y_{12}, y_{13}, y_{21}, y_{22}, y_{23}))^T$ for one patient, where y_{ij} represents the indicator of the ture underlying transition from state i to j with y_{ij}=1 and others are zero. The corresponding probability p_{ij} is specified in (4.1), (4.2) and (4.3). $\mathbf{z} = (z_1, ..., z_l)^T$ is the vector of covariates, and β_{ij} is the parameter vector corresponding to the ij-th transition category.

$$\begin{aligned} log\Pi_{ij} p_{ij}^{y_{ij}} &= y_{12}\beta_{12}^T\mathbf{z} + y_{13}(\beta_{12}^T\mathbf{z} + \beta_{23}^T\mathbf{z}) + y_{22}\beta_{22}^T\mathbf{z} + y_{23}\beta_{23}^T\mathbf{z} \\ &\quad -(y_{11} + y_{12} + y_{13})log[1 + exp(\beta_{12}^T\mathbf{z})] \\ &\quad -(y_{21} + y_{22} + y_{23} + y_{13})log[1 + exp(\beta_{21}^T\mathbf{z}) + exp(\beta_{23}^T\mathbf{z})] \end{aligned}$$

Let $\beta = (\beta_{12_0}, \beta_{12_1}, \beta_{12_2}, \beta_{21_0}, \beta_{21_1}, \beta_{21_2}\beta_{23_0}, \beta_{23_1}, \beta_{23_2}, \beta_{23_3})^T$ denote the 10 parameters. We have for the partial derivative as following, with $m, m^{'} \in (1, ..., l)$

$$\frac{\partial L}{\partial \beta_{12_m}} = (y_{12} + y_{13})z_m - (y_{11} + y_{12} + y_{13})\frac{exp(\beta_{12}^T\mathbf{z})}{1 + exp(\beta_{12}^T\mathbf{z})}z_m$$

$$\frac{\partial L}{\partial \beta_{21_m}} = y_{21}z_m + (y_{21} - y_{22} + y_{23} + y_{13})\frac{exp(\beta_{21}^T\mathbf{z})}{1 + exp(\beta_{21}^T\mathbf{z}) + exp(\beta_{23}^T\mathbf{z})}z_m$$

$$\frac{\partial L}{\partial \beta_{23_m}} = (y_{23} + y_{13})z_m - (y_{21} + y_{22} + y_{23} + y_{13})\frac{exp(\beta_{23}^T\mathbf{z})}{1 + exp(\beta_{21}^T\mathbf{z}) + exp(\beta_{23}^T\mathbf{z})}z_m$$

The observed information can be computed to be

$$-\frac{\partial^2 L}{\partial \beta_{12_m} \partial \beta_{12_{m'}}} = (y_{11} + y_{12} + y_{13})(\hat{p}_{12} - \hat{p}_{12}^2)z_m z_{m'}$$

$$-\frac{\partial^2 L}{\partial\beta_{21_m}\partial\beta_{21_{m'}}} = (y_{21}+y_{22}+y_{23}+y_{13})(\hat{p}_{21}-\hat{p}_{21}^2)z_m z_{m'}$$

$$-\frac{\partial^2 L}{\partial\beta_{23_m}\partial\beta_{23_{m'}}} = (y_{21}+y_{22}+y_{23}+y_{13})(\hat{p}_{23}-\hat{p}_{23}^2)z_m z_{m'}$$

$$-\frac{\partial^2 L}{\partial\beta_{21_m}\partial\beta_{23_{m'}}} = (y_{21}+y_{22}+y_{23}+y_{13})(-\hat{p}_{21}\hat{p}_{23})z_m z_m^{'}$$

Where

$$\hat{p}_{12} = \frac{exp(\beta_{12}^T\mathbf{z})}{1+exp(\beta_{12}^T\mathbf{z})}$$

$$\hat{p}_{21} = \frac{exp(\beta_{21}^T\mathbf{z})}{1+exp(\beta_{21}^T\mathbf{z})+exp(\beta_{23}^T\mathbf{z})}$$

$$\hat{p}_{23} = \frac{exp(\beta_{23}^T\mathbf{z})}{1+exp(\beta_{21}^T\mathbf{z})+exp(\beta_{23}^T\mathbf{z})}$$

The other elements in the observed information matrix are zeros. We use the Newton-Raphson optimation algorithm to estimate the parameter β iteratively, with zeros as initial values and 1e-06 as convergence stopping criteria. This is the estimation procedure when we assume that there is no measurement error in the data.

If we consider the sensitivity issue in CF, we need to incoporate the misclassification matrix E, whose entry $e(r,s)$ represents the probability of observing s state when ture state is r. The absorbing state mucoid infection is assumed to be observed without measurement error. According to the sensitivity and specificity of throat culture in literature (Rosenfeld, et al., 1999; Pedersen, et al., 1987), we have specificity $\lambda_1 = 93\%$ and sensitivity $\lambda_2 = 85\%$.

$$\begin{bmatrix} \lambda_1 & 1-\lambda_1 & 0 \\ 1-\lambda_2 & \lambda_2 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

The predict value matrix V, whose entry $v(s,r)$ represents the probability of predicting true state as r when s is observed. According to nd the predictive values of positive and negative throat culture in literature (Bonnie, et al., 1991), we have

predictive value of positive culture $\lambda_3 = 83\%$ and predictive value of negative culture $\lambda_4 = 70\%$.

$$\begin{bmatrix} \lambda_4 & 1-\lambda_4 & 0 \\ 1-\lambda_3 & \lambda_3 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Let $O = (o_{12}, o_{12}, o_{13}, o_{21}, o_{22}, o_{23})^T$ denote the observed transition corresponding to $Y = (y_{11}, y_{12}, y_{13}, y_{21}, y_{22}, y_{23}))^T$. The corresponding probability $p_{O_{ij}}$ can be computed by p_{ij} in (4.1), (4.2) and (4.3) and the misclassification matrix E and predictive value matrix V. $p_{O_{ij}}$ is the (i, j) entry of the product of matrices:

$$\begin{bmatrix} \lambda_4 & 1-\lambda_4 & 0 \\ 1-\lambda_3 & \lambda_3 & 0 \\ 0 & 0 & 1 \end{bmatrix} \times P \times \begin{bmatrix} \lambda_1 & 1-\lambda_1 & 0 \\ 1-\lambda_2 & \lambda_2 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

The final probability matrix for observed transition P_O is

$$\begin{bmatrix} p_{O_{11}} & p_{O_{12}} & 0 \\ p_{O_{21}} & p_{O_{22}} & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$p_{O_{11}} = \lambda_1\lambda_4 p_{11} + \lambda_1(1-\lambda_4)p_{21} + (1-\lambda_2)\lambda_4 p_{12} + (1-\lambda_2)(1-\lambda_4)p_{22}$$

$$p_{O_{12}} = (1-\lambda_1)\lambda_4 p_{11} + (1-\lambda_1)(1-\lambda_4)p_{21} + \lambda_2\lambda_4 p_{12} + \lambda_2(1-\lambda_4)p_{22}$$

$$p_{O_{21}} = \lambda_1(1-\lambda_3)p_{11} + \lambda_1\lambda_3 p_{21} + (1-\lambda_2)(1-\lambda_3)p_{12} + (1-\lambda_2)\lambda_3 p_{22}$$

$$p_{O_{22}} = (1-\lambda_1)(1-\lambda_3)p_{11} + (1-\lambda_1)\lambda_3 p_{21} + \lambda_2(1-\lambda_3)p_{12} + \lambda_2\lambda_3 p_{22}$$

The likelihood for one observation is $\Pi_{ij} p_{O_{ij}}^{o_{ij}}$. Numerical methods need to be used to maximize the likelihood , e.g. a quasi-Newton algorithm, which does not require the specification of the derivatives of the objective function, which are computed by finite differences and used to estimate the Hessian at the maximum (Dennes and Schnabel, 1983).

Because the probabilities of acquisition of nonmucoid *Pa* is quite different between the patients who were never infected and those who had been infected before, we separate the transitions in the dataset by having been infected by nonmucoid *Pa* or not. For the transitions from patients who have been infected before, the parameter β are presented in Table 3. The transition probabilities are significantly associated with age, which indicates the higher risk of developing nonmucoid and mucoid *Pa* infection and greater treatment difficulty as patients grow older. There are significant association between D_2 and p_{23}, which indicates that the longer the patients stay in nonmucoid *Pa* infection, the higher the probability of developing mucoid infecttion. The results of parameters related to ΔF508H reveal the difference between two subpopulation in terms of transition probabilities. The patients with ΔF508 homozygosity are more easily acquire nonmucoid *Pa* infection. As for the probability of p_{21} and p_{23}, although the estimates are not significant, the trends match the previous research results that the patients with ΔF508 homozygosity are harder to treat, with more severe disease and being more susceptible to mucoid *Pa* infection.

For the transitions from patients who have never been infected, we use the logistic regression model with age as covariates to model p_{12}:

$$p_{12}(t, t+1) = \frac{exp(\beta'_{12_0} + \beta'_{12_1} t)}{1 + exp(\beta'_{12_0} + \beta'_{12_1} t)}.$$

The parameter estimates are provided at the bottom of Table 3.

Analysis of the state transitions on this dataset by fitting the proposed model provide insights into the dynamics of *Pa* infection and refine the data generative model in simulation studies. We simulated 68 patients in order to compare the simulated longitudinal data based on the proposed model and estimated parameters to the original Wisconsin neonatal screening project dataset with 69 patients. Figure 4 shows the Kaplan-Meier plots of time to first nonmucoid *Pa* infection $T1$, time to mucoid *Pa* infection $T2$ and the time between first acquisition of nonmucoid and mucoid infection $T12$ respectively. The darker line represents the patients data from the

Table 3: Parameter estimates for Wisconsin neonatal screening project data

Parameter	**Covariate**	**Estimate**	**St erros**	**Est/SE**	**p-value**
β_{12_0}	Intercept	-0.876464	5.56839	-0.1574	0.8749273
β_{12_1}	Age	0.017504	0.004941	3.5429	0.0003958
β_{12_2}	ΔF508H	0.008112	0.003575	2.2692	0.0232556
β_{21_0}	Intercept	-0.155778	0.119471	-1.3039	0.1922808
β_{21_1}	Age	-0.005863	0.003148	-1.8626	0.0625213
β_{21_3}	ΔF508H	-0.029767	0.030761	-0.9677	0.3332058
β_{23_0}	Intercept	-3.722290	0.395525	-9.4110	$<$ 2.2e-16
β_{23_1}	Age	0.097886	0.019947	4.9074	9.23e-07
β_{23_2}	D_2	0.215752	0.060433	3.5701	0.0003569
β_{23_3}	ΔF508H	0.033001	0.0455311	0.7248	0.4685742
Patients who never been infected					
β'_{12_0}	Intercept	-1.575928	0.510372	-3.0878	0.0020163
β'_{12_1}	Age	0.004296	0.001321	3.2511	0.0011494

original Wisconsin neonatal screening project dataset, where there are some censored data. The lighter line corresponds to the simulated patients data by the proposed disease model and parameters estimated in this section. We can see the simulated data reflect the disease progression in the dataset in a realistic way.

The analysis of pulmonary function reveals little influence of the change from no *Pa* to nonmucoid *Pa*, as well as the infection severity level changes before the development of mucoid infection on the change of lung function.

In order to exam the capacity of the reinforcement learning procedure to discover optimal therapy in such disease dynamics, we simulate the treatment effect scenarios with time-varying efficacy and toxicity through a model parameter $\beta_{21_2}(t)$. The details will be provided in the presentation of the simulation study given in Chapter 6.

Figure 4: Kaplan-Meier plot of time to first acquisition of nonmucoid *Pa* infection $T1$.

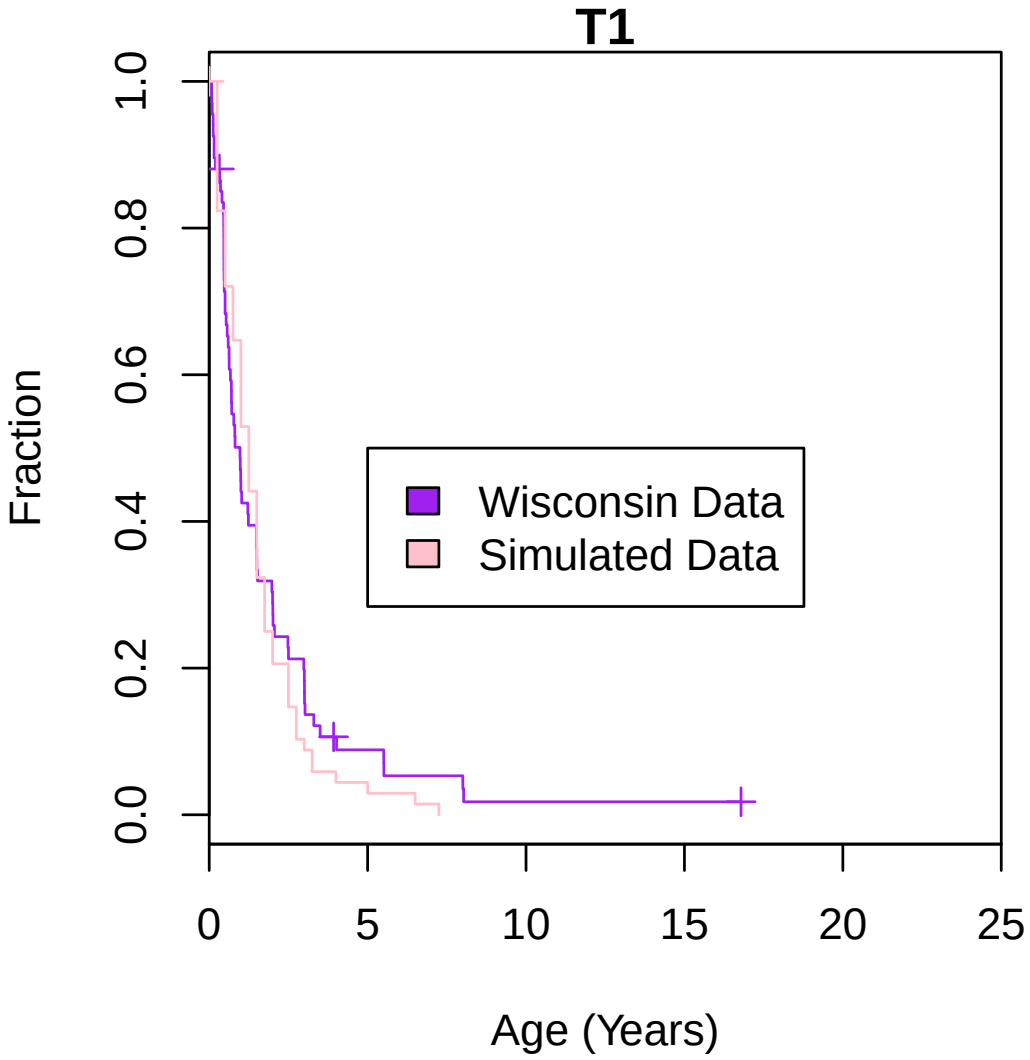


Figure 5: Kaplan-Meier plot of time to mucoid *Pa* infection $T2$.

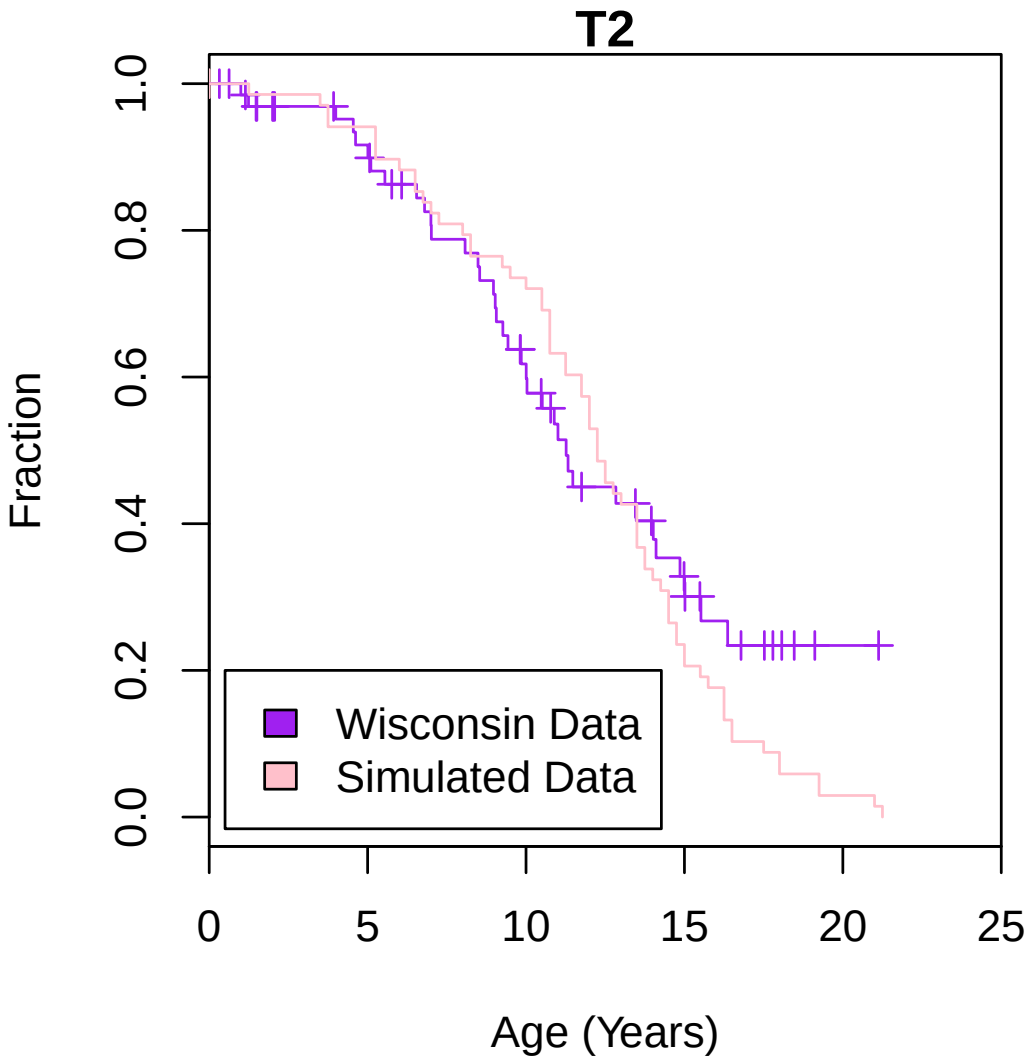
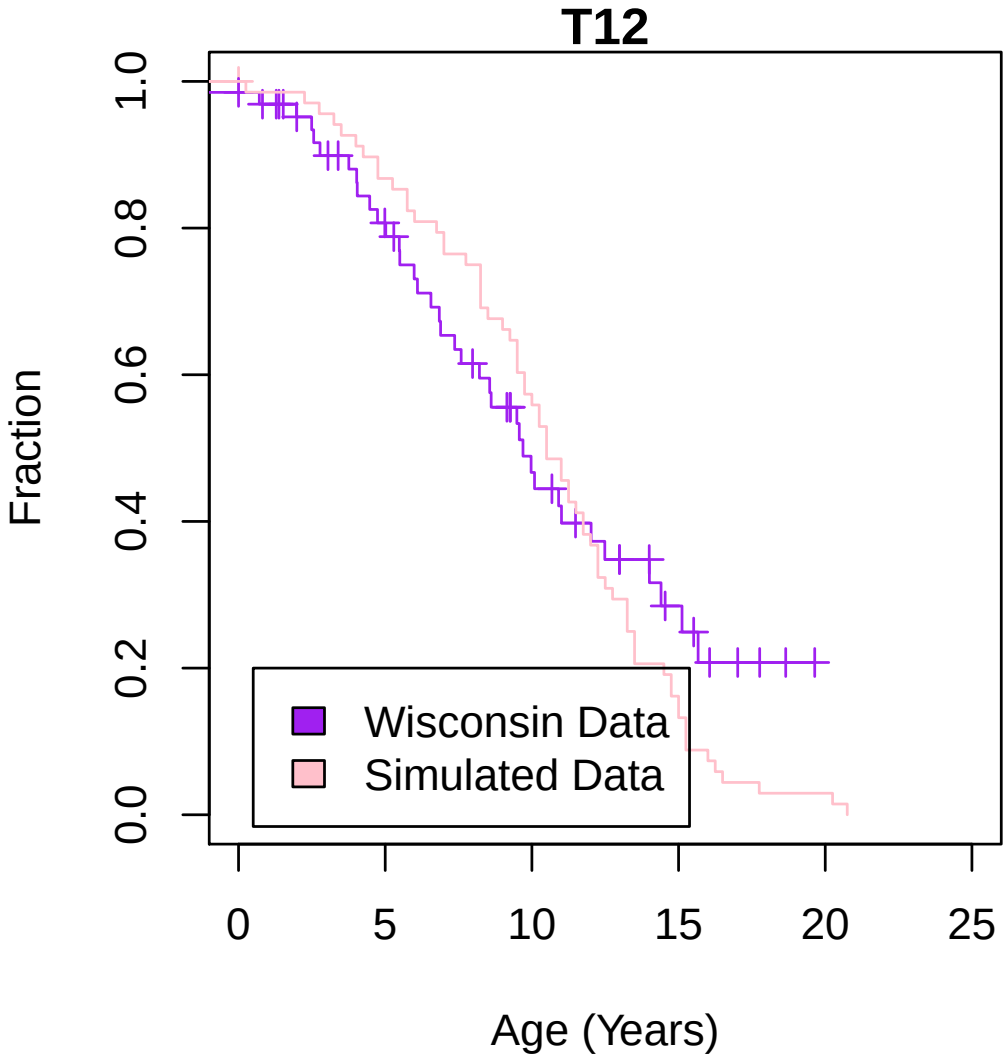


Figure 6: Kaplan-Meier plot of time between first acquisition of nonmucoid and mucoid infection $T12$.



5 CYSTIC FIBROSIS CLINICAL REINFORCEMENT TRIALS

In developing adaptive personalized therapy, the usual meta-analysis on multiple trials in separate disease stages or treatment courses is not appropriate because the potential delayed treatment effects and the potential cohort effects leading to the shift of the study population will be ignored. There are three advantages of using an experimental approach based on sequential randomization comparing to the observational approach based on observational longitudinal data. First, the sequential randomization at multiple decision times ensures the assigned treatment is independent of potential future responses to treatment, conditional on the history to date. The sequential ignorability and stable unit treatment value assumptions (SUTVA) always holds under these experimental studies, while it is not testable in observational studies. Secondly, one is forced to limit the options and minimize some confounding effects by randomization, while in the observational context, it will be impossible to control such variations. Thirdly, the complete information at each decision time on patients who experiened all possible beneficial treatment sequences by prototol facilitate the construction of optimal personalized therapy.

The proposed "clinical reinforcement trial" consists of both a learning stage (phase IIb) and a confirmatory stage (phase III) trial to optimize and validate the personalized therapy. As mentioned in Section 1, background for the general strategy and key aspects of clinical reinforcement trial designs can be found in (Zhao, et al., 2009; Zhao, et al., 2010) and for SMART designs can be found in (Murphy, 2005; Murphy, et al., 2007; Thall, et al., 2002). Based on the published results from previous CF trials (Döring, et al., 2007), the CF neonatal screening project (Gibson, et al., 2003;

Taccetti, et al., 2005), and a contemporary CF trial (Treggiari, et al., 2009), we develop a virtual clinical reinforcement trial that provides a realistic approximation to a potential real CF trial.

5.1 Clinical Reinforcement Trial Conduct

1. Learning stage trial design.

   In a randomized trial in CF for patients 1–25 years of age, N_1 trial participants are sequentially fairly randomized at enrollment and at each decision time based on detection of *Pa* from quarterly respiratory cultures (culture-based therapy) to one of the five treatment options *A-L*, *A-H*, *B-L*, *B-H* and *S-C* defined below with an equal allocation ratio for L_1 years. The randomization is stratified by patient indicator of mutation class $\Delta F508$ homozygosity. The primary endpoint is the time to presence of mucoid isolated from *Pa* respiratory culture. The secondary clinical endpoint is the decline in pulmonary function FEV_1. The secondary microbiological endpoint is the proportion of patients with new *Pa*-positive respiratory cultures during the study. Patient clinical outcomes and biomarker values are collected at each quarterly clinical visit.

   For simplicity and without loss of generality, we here consider four active anti-*Pa* treatments, consisting of two anti-*Pa* antibiotic drugs *A* and *B* having different intensity levels high (*H*) and low (*L*). The choice of drug could, for example, be based on FDA approved inhaled antibiotics tobramycin and consensus panel supported oral ciprofloxacin (Döring, et al., 2000; Treggiari, et al., 2009). The treatment *S-C* represents a the prevailing standard of care, a hypothetical placebo, or other treatment without targeted anti-*Pa* antibiotics.

2. Learning stage rationale and goal.

   The rationale of culture-based therapy is based on the clinical guidance for *Pa* infection in CF patients (Döring, et al., 2000; Döring, et al., 2004; Flume, et

al., 2007) Usually anti-Pa treatment is applied only when Pa is detected, since risk of nephrotoxicity due to long term preventive treatment may out-weigh benefit. For patients in the mucoid stage, a high intensity treatment such as IV anti-Pa treatments are required (Flume, et al., 2007). The scientific goal of this trial is to uncover the optimal strategy based on existing treatments to prolong the time to the mucoid stage for young CF patients whenever nonmucoid Pa is detected.

3. Learning stage utility.

   The relatively short study duration is one of the common characteristics in phase II trials. Due to its strong relationship to both time to mucoid Pa and nonmucoid Pa infection severity, FEV_1 serves as a surrogate endpoint or biomarkers in our phase IIb trial. A utility function, i.e., a reward in the reinforcement learning framework $r_t = R(s_t, a_t, s_{t+1})$, for $t = 0, 1, \ldots, 4L_1 - 1$, is prespecified and contains an appropriately weighted assessment of benefit and risk based on the outcomes available at each interval. We use a combination of three clinical meaningful components, lung function, infection status, and severity, as guidance for the optimal therapy we are seeking for (Langton Hewer, et al., 2009). Specifically, Table 4 and Table 7 are our reward functions for the simulation studies in Chapter 6. The utility/reward set-up in RL enables us to integrate benefits at the individual level and cumulated over time.

4. Estimating optimal therapy.

   (a) *Inputs*: State variables S_t consisting of age, $\Delta F508H$, $Cul(t)$, $Ser(t)$, $D_2(t)$, $Muc(t)$, $Sev(t)$, $FEV_1(t)$, $Cum_A(t)$, $Cum_B(t)$, $Sus_A(t)$, $Sus_B(t)$ and $Trt(t)$, as given in Table 1. The patients have longitudinal observations quaterly for L_1 years $\{s_{i0}, a_{i0}, r_{i0}, s_{i1}, a_{i1}, r_{i1}, ..., a_{i(4L_1-1)}, r_{i(4L_1-1)}, s_{i(4L_1)}\}_{i=1}^{N_1}$. The set of one-step system transitions is obtained after separation and standardization as a training set $\mathcal{TS}$ of 4-tuples of the form $\langle s, a, r, s' \rangle$. Hence,

$\mathcal{TS}_0 = \{(s_t^l, a_t^l, r_t^l, s_{t+1}^l)\}_{l=1}^{\#\mathcal{TS}_0}$ with $\#\mathcal{TS}_0 = 4 \times L_1 \times N_1$.

(b) Initialization: $\hat{Q}_0(s,a) = 0, \forall s, a$,

(c) Estimation: $Q_N^\star(s,a)$ sequence in (3.2) is fitted by Q-iteration:

- **repeat** at each iteration k,$k \geq 1$
  - **for all** $\langle s, a, r, s' \rangle$ on $\mathcal{TS}_{k-1}$ **do**
  - $r' \leftarrow r + \gamma \max_{a'} \hat{Q}_{k-1}(s', a')$
  - update $\langle s, a, r', s' \rangle$ as $\mathcal{TS}_k$
  - approximate $\hat{Q}_k(s,a)$ on $\mathcal{TS}_k$ by SVR
  - **end for**
- **until** stop criteria $\max_{\forall s,a} |\hat{Q}_k(s,a) - \hat{Q}_{k-1}(s,a)| \leq \epsilon$ is met.

We use the Gaussian kernel $K(\mathbf{x}, \mathbf{y}) = \exp(-\zeta \|\mathbf{x} - \mathbf{y}\|^2)$ in SVR approximation iterations. The tuning parameter pair (C, ζ) are selected by grid search over cost parameter $C = 2^{-5}, 2^{-3}, \ldots, 2^{15}$ and scale parameter $\zeta = 2^{-15}, 2^{-9}, \ldots, 2^3$ in 10-fold cross-validation.

(d) Output: Personalized therapy $\hat{\pi}^*(s) = arg\max_a \hat{Q}_N^\star(s,a)$

5. Confirmatory stage design.

In a separate, potentially longer duration L_2 years trial, N_2 trial participants are only randomized at enrollment to either one of four fixed therapies *A-L*, *A-H*, *B-L*, *B-H* or the new arm *R-L* with equal allocation in a conventional way. The therapy *R-L* represents the adaptive personalized therapy identified in Step 4. The randomization, stratification, endpoints and patients information are the same as those in the Step 1 trial. The objective is to investigate whether the adaptive personalized therapy prolongs time to mucoid infection and reduces the isolation of *Pa* from respiratory cultures, compared with the four fixed treatment options. The placebo arm is not included at this stage.

5.2 Estimating Optimal Therapy

In this section, we connect the estimating procedure described in previous section 5.1 step 4(c) to the optimal treatment strategy within counterfactual framework. We extend some of the previous work (Murphy, 2005; Zhao, et al., 2010) to the fitted Q iteration procedure. We denote a_t, a_{t+1} as the decision at time point t and $t+1$, respectively. We let $Q_N^\star(S_t, A_t)$ be the potential outcome as the N-th iteration Q-function value of the sequence by fitted Q-iteration in (3.2), after time t and before $t+1$. We also let $Q_{N-1}^\star(S_{t+1}, A_{t+1})$ be the potential outcomes from the N-1-th Q-function in fitted Q-iteration.

Moreover, we let S_t denote the states at time t. Under the Markovian working assumption, we let $S_{t+1}(a_t)$ denote the state at time $t+1$, after policy a_t and independent of any other previous action sequence. Within a counterfactual framework, we maximize the value state function in order to find an optimal treatment strategy:

$$E_{a_t}\left[Q_N^\star(a_t)\Big|S_t\right] = E_{a_t}\left[R(S_t, a_t) + \gamma \max_{a_{t+1}} Q_{N-1}^\star(S_{t+1}(a_t), a_{t+1})\Big|S_t\right]$$

$$= E_{a_t}\left[R(S_t, a_t) + \gamma \max_{a_{t+1}} E_{a_t, a_{t+1}}\left[Q_{N-1}^\star(a_{t+1})\Big|S_{t+1}(a_t)\right]\Big|S_t\right].$$

For the first iteration step, we have $E_{a_t}\left[Q_1^\star(a_t)\Big|S_t\right] = E_{a_t}\left[R(S_t, a_t)\right]$. Based on Q-iteration algorithm, the optimal policy in (3.3) can be obtained via the iterative step.

In the 1-th iteration step,

$$\pi_1^* = arg\max_{a_t} E_{a_t}\left[R(S_t, a_t)\right].$$

In the N-1-th iteration step,

$$\pi_{N-1}^*(\pi_N) = arg\max_{a_{t+1}} E_{\pi_N, a_{t+1}}\left[Q_{N-1}^\star(a_{t+1})\Big|S_{t+1}(\pi_N)\right]. \tag{5.1}$$

In the N-the iteration step,

$$\pi_N^* = arg\max_{a_t} E_{a_t}\left[R(S_t, a_t) + \gamma E_{a_t,\pi_{N-1}^*}\left[Q_{N-1}^\star(\pi_{N-1}^*)|S_{t+1}(a_t)\right]\Big|S_t\right]. \tag{5.2}$$

Assuming the stable unit treatment value assumptions (SUTVA) and no unmeasured confounders, which are guarantee under our sequential randomized designs. We have the relationship:

$$Q_1^\star(S_t, A_t) = \sum_{a_t} R(S_t, a_t) I(A_t = a_t),$$

$$Q_N^\star(S_t, A_t) = \sum_{a_t} Q_N^\star(S_t, a_t) I(A_t = a_t),$$

$$Q_{N-1}^\star(S_{t+1}, A_{t+1}) = \sum_{a_t, a_{t+1}} Q_{N-1}^\star(S_{t+1}, a_{t+1}) I(A_t = a_t, A_{t+1} = a_{t+1}).$$

Particularly, in the 1-th iteration, $E_{a_t}\left[Q_1^\star(a_t)\Big|S_t\right] = E_{a_t}\left[R(S_t, a_t)\right]$ can be estimated via estimating $E[R(S_t, A_t)|S_t, A_t]$.

Because in the fitted Q iteration, in all the iteration, four-tuples consist of the same S_t,A_t,S_{t+1}, these imply the quantity to be maximized in (5.1) as

$$\begin{aligned} E_{a_t,a_{t+1}}\left[Q_{N-1}^\star(a_{t+1})\Big|S_{t+1}\right] &= E_{a_{t+1}}\left[Q_{N-1}^\star(a_{t+1})\Big|S_{t+1}, A_t = a_t, A_{t+1} = a_{t+1}\right] \\ &= E\left[Q_{N-1}^\star(A_{t+1})\Big|S_{t+1}, A_{t+1} = a_{t+1}\right]. \end{aligned}$$

Hence, the function of the potential outcomes on the right hand side of (5.1) can be estimated via estimating $E[Q_{N-1}^\star|S_{t+1}, A_{t+1}]$.

Similarly, we can write the quantity to be maximized in the

$$E_{a_t}\left[R(S_t,a_t)+\gamma E_{a_t,\pi^*_{N-1}}\left[Q^\star_{N-1}(\pi^*_{N-1})|S_{t+1}(a_t)\right]\Big|S_t\right]$$

$$= E\left[R(S_t,a_t)+\gamma\max_{a_{t+1}}E\left[Q^\star_{N-1}|S_{t+1},A_{t+1}=a_{t+1}\right]\Big|S_t,A_t=a_t\right]$$

$$= E\left[R(S_t,A_t)+\gamma\max_{a_{t+1}}E\left[Q^\star_{N-1}|S_{t+1},A_{t+1}=a_{t+1}\right]\Big|S_t,A_t=a_t\right].$$

Therefore, the function regarding the potential outcome on the right hand side of (5.2) can be estimated via estimating $E[Q^\star_N|S_t,A_t]$.

6 REINFORCEMENT LEARNING TREATMENT STRATEGIES

In this section, we simulate virtual CF patients based on the underlying disease process using the proposed Markov model in Chapter 4, and implement the proposed sequentially randomized clinical reinforcement trial in Chapter 5. In order to examine the performance of the proposed design and methodology, we compare the long term outcome, time to mucoid *Pa* infection based on the same working model, with the other fixed regimens in extensive simulation studies. We conduct the simulation studies based on the Markov model in section 4.2, with the tuned parameters in section 4.3 and with the estimated parameters from section 4.4 respectively.

6.1 Clinical Scenarios

We evaluate the design under realistic clinical scenarios (Flume, et al., 2007; Ramsey, et al., 1999; Retsch-Bogart, 2009; Rosenfeld, et al., 2001; Sutton, et al., 1998; Treggiari, et al., 2009; Treggiari, et al., 2007; Valerius, et al., 1991) described in Table 4. There are differential treatment efficacy and side effects in terms of probability of successful eradication for a two mutation class population. Patients who are ΔF508 homozygous are a high risk population, generally more severe, more easily acquire mucoidy, and have greater difficulty eradicating *Pa* infection. Additionally, immediate efficacy is time varying and age specific. The optimal treatments are also different by age group as presented in Table 4.

For example, for the low risk population, antibiotic A is best in both high (H) and low (L) intensity regimens when the patient is under 8 years old; while antibiotic

Table 4: Efficacy and side-effects of treatments in simulations

	Age Range	Antibiotics	Intensity	NonΔF508H	ΔF508H
Immediate	Age ≤ 8	A	L	High	Low
Efficacy			H	High	Low
		B	L	Low	Medium
			H	Low	High
	8 $<$ Age	A	L	Low	Low
			H	Low	Medium
		B	L	Medium	Low
			H	High	Low

Delayed	No Off-drug Cycle	Susceptibility $\downarrow$
Side-effects	Life-time Exposure$>$ 20	Eradication $\downarrow$

B is best with high intensity (H) as the preferable regimen when the patient is older than 8 years old. For the high risk population, antibiotic B with high intensity (H) is the preferable regimen when the patient is under 8 years old; while antibiotic A is best when the patient is older than 8 years old; the higher intensity level regimens are more successful for bacteria eradication. However, in this more severe group of patients who are ΔF508 homozygous, the treatments have lower efficacy compared to the other group of patients.

In the bottom panel of Table 4, the delayed side effects are modeled when the cumulative drug use exceeds a threshold or repeated courses of the same anti-*Pa* drug without a "drug-off" period, antibiotic resistance will then develop, and consequently, the eradication probability will decrease. The "drug-off" or switching drug can lead to some degree of return of susceptibility, as has been observed with inhaled tobramycin (TOBI) (Flume, et al., 2007; Gibson, et al., 2003; Guez, et al., 2008; Ramsey, et al., 1999; Ratjen, et al., 2001).

6.2 Simulation Methods and Results

6.2.1 Study I with tuned parameters

We generate a virtual CF trial based on the disease model in section 4.3 with treatment effect scenarios described in Section 6.1. The conduct of the clinical reinforcement trial follows the procedure proposed in Section 5.1 with total sample sizes $N_1 = 1000$ and study durations $L_1 = 2$ years and $L_2 = 4$ years for the learning and confirmatory stages respectively. Without loss of generality, we assume equal numbers of patients in two subgroups defined by whether patients are $\Delta F508H$ in all studies. Besides the testing scenario in the confirmatory trial with $N_2 = 1000$ and 4 years of follow up, we exam the procedure with $N_2 = 1000$ in the scenario where we can apply the therapies from birth until a mucoid *Pa* event occurs. We use

Table 5: Reward/utility function setup I

State Variables	Change	Reward
Culture/Serology	Infected to free of Pa	1
Lung Function	$\uparrow > 10\%$	0.1
	no $\downarrow > 10\%$	0.1
Severity of Infection	progress to mucoid	0
	intermittent to chronic/stay	0.1
	chronic to intermittent	0.2

the threshold ϵ in fitted Q-iteration with stopping criteria 10^{-4} and discount factor $\gamma = 0.5$. The immediate reward function set up in Table 5.

6.2.2 Testing results of study I in virtual trial from birth till mucoid infection

In this testing scenario, we apply therapies and follow up all the virtual patients until the development of mucoid *Pa* infection. As shown in Figure 8, we provide the boxplot of the time to mucoid *Pa* infection corresponding to the fixed treatment regimens *S-C*, *A-L*, *A-H*, *B-L*, *B-H* and the adaptive personalized therapy denoted *R-L*. There are 200 patients in each arm with one half being Δ F508 homozygous. The empirical performances of these treatment regimens are illustrated in Figure 8 and Table 6. In terms of time to mucoid *Pa* infection, the fixed treatment regimens *S-C*, *A-L*, *A-H*, *B-L*, *B-H* have differential effects on the different risk groups classified by ΔF508 homozygosity. In Table 6, the other endpoints, nonmucoid *Pa* infection proportion, predicted $FEV_1\%$, and change per year, all demonstrate the same treatment effect patterns. This matches the treatment effect patterns in the clinical scenarios in

Section 5.1, Table 4, where the high risk population requires higher intensity level treatment to eradicate *Pa* infection and the right drug chosen in early childhood improves prognosis in both subpopulations.

Figure 7: Boxplot of distribution of time to mucoid *Pa* in study I. The gray and dark green represent patients with Δ F508 homozygosity, otherwise the colors are blue and light green.

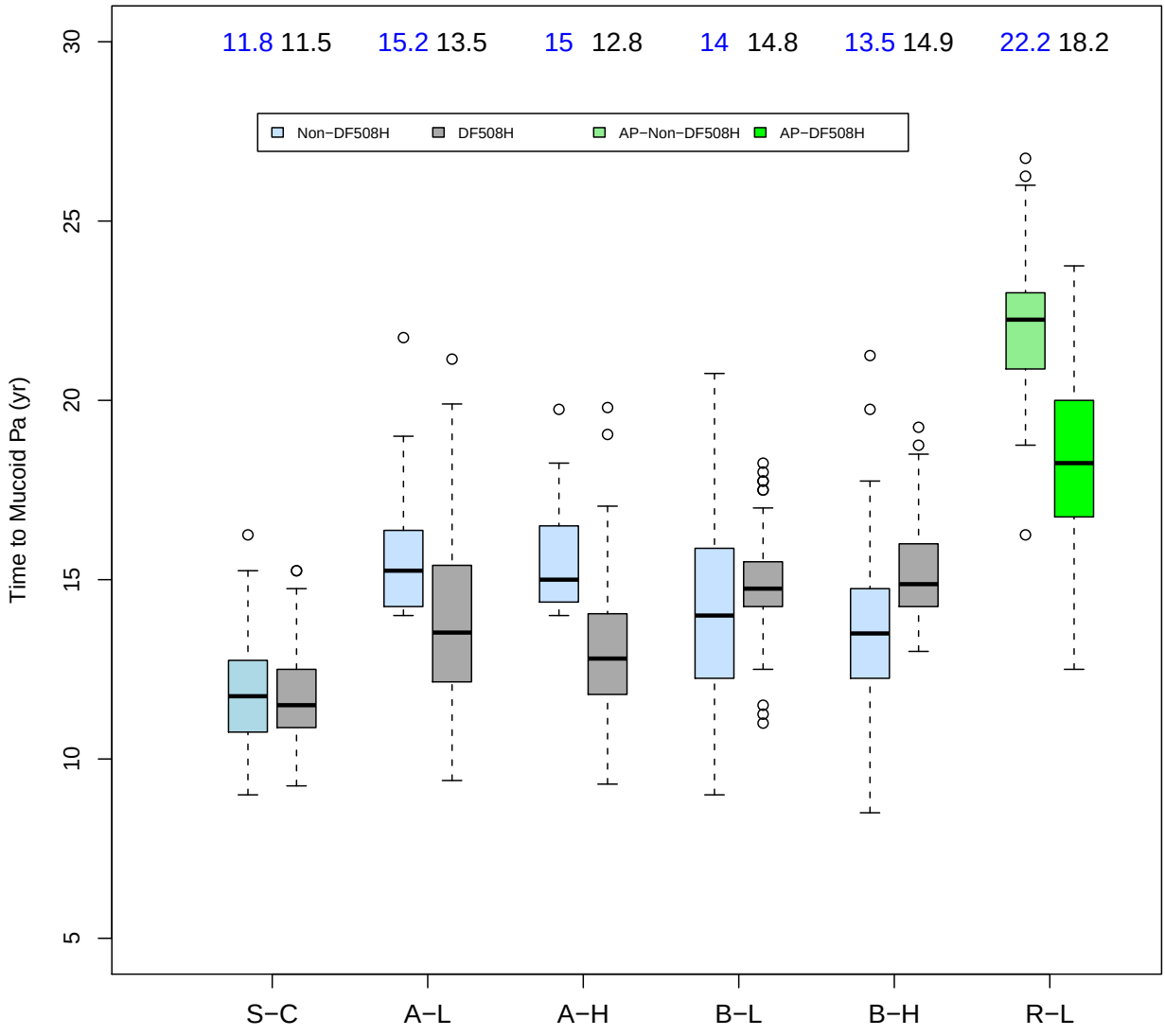


Figure 8: Boxplot of distribution of time to mucoid *Pa* grouped by two subpopulations in study I.

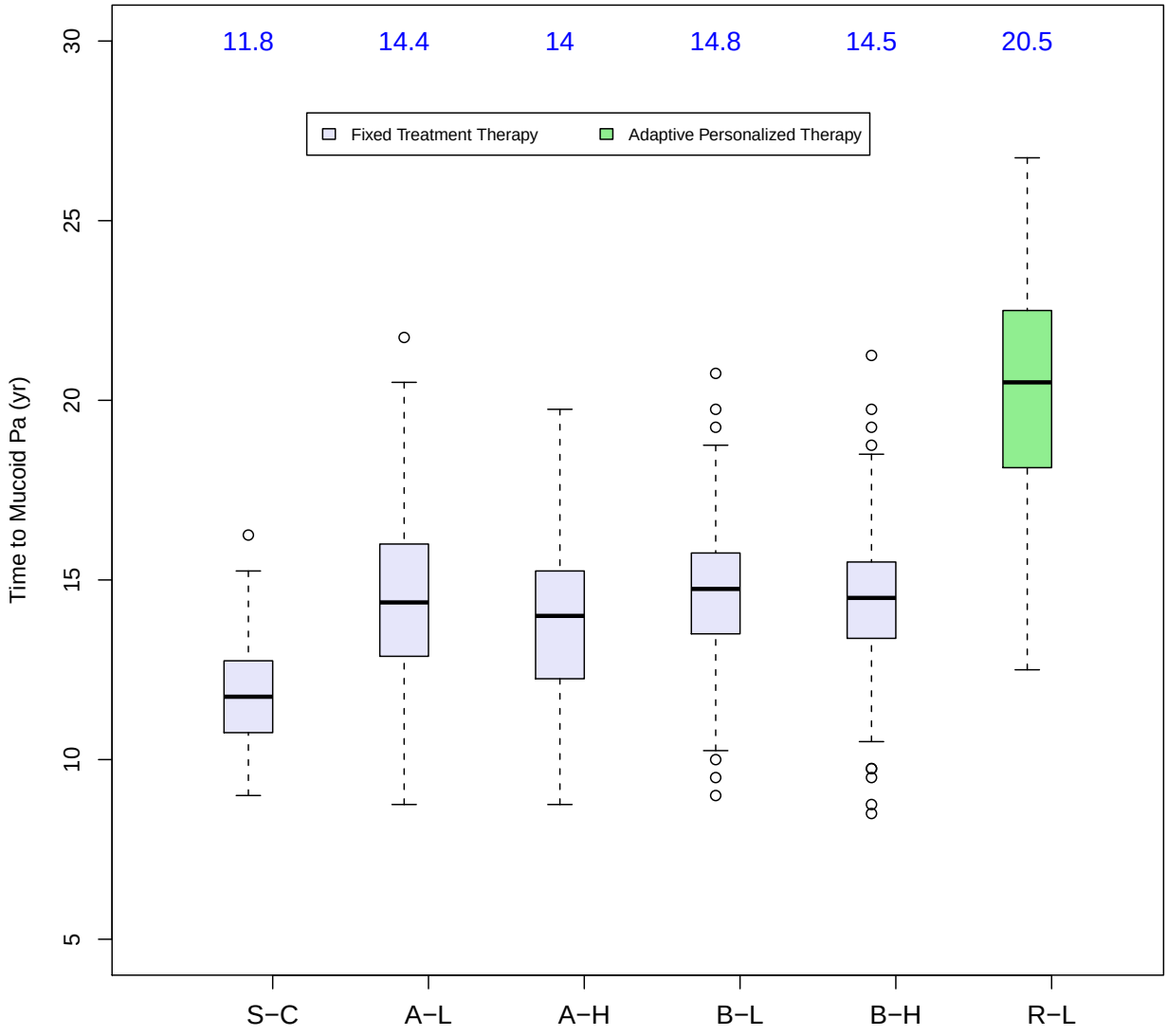


Table 6: Comparisons between fixed treatment regimens and estimated optimal therapy for time to mucoid *Pa* (year). Each training/testing dataset is of size 100/subgroup.

Group	Non ΔF508H | ΔF508H						All					
Therapy	SC	AL	AH	BL	BH	RL	SC	AL	AH	BL	BH	RL
						Time to Mucoid *Pa* (T_2) (Yr)						
Mean	11.8|11.6	15.6|13.9	15.5|13.0	14.1|15.0	13.5|15.3	21.1|18.1	11.8	14.7	14.3	14.5	14.3	19.6
SD	1.5|1.7	1.5|2.3	1.3|1.8	2.5|1.2	2.1|1.3	1.7|2.5	1.4	2.3	2.1	1.9	2.2	3.3
Min	9.0|8.5	14.0|9.4	14.0|9.3	9.0|11.0	8.5|13.0	16.3|12.5	9.0	8.8	8.8	10.0	9.8	14.3
Median	11.8|11.5	15.2|13.5	15.0|12.8	14.0|14.8	13.5|14.9	22.2|18.2	11.8	14.4	14.0	14.8	14.5	20.5
Max	16.3|15.3	21.8|21.2	19.8|19.8	20.8|18.3	21.3|19.3	26.8|23.8	16.0	19.5	19.5	19.8	20.8	26.3
						Nonmucoid *Pa* + over T_2 (%)						
Mean	62.2|61.4	40.8|47.5	43.2|49.3	45.1|45.1	44.5|39.7	37.9|38.3	61.8	41.7	46.7	44.5	42.1	39.1
						Predicted FEV_1 while non-mucoid (%)						
Mean	70.4|70.6	76.5|72.5	76.4|71.4	73.2|75.3	72.6|76.6	78.7|77.9	70.5	74.5	73.9	74.2	74.6	79.3
↓ Rate/Yr	5.78|6.45	3.54|4.62	3.60|4.91	4.03|3.58	4.11|3.96	0.54|2.32	6.10	3.90	4.25	3.82	4.03	1.43

In Figure 8, we illustrate results of the same simulated trial in a different way by grouping patients from the two subpopulations based on Δ F508 homozygosity together. If one ignores patient heterogeneity of mutation class, the benefits of treatment remain undetected among the fixed treatment therapies. For example, drug B's benefit to patients with $\Delta F508H$ is "averaged out" with outcomes for patients without this characteristics. Similarly, drug A's benefit to patients who are not $\Delta F508H$ is also masked. Figure 9 shows the observed frequency of the three states (Pa free (in green), nonmucoid Pa (in red) and mucoid Pa (in black)) among the 200 patients in each arm over time, demonstrating a similar pattern to time to mucoid Pa. By optimizing the usage of these existing drugs, the discovered personalized therapy achieves superior patient outcomes than any other fixed treatment therapies even in the mixture of the two subpopulations. The treatment benefits of these drugs may be missed by a traditional, single-decision point clinical trial.

The top two subplots of Figure 10 illustrate the discovered therapies for two individual patients who are not Δ F508 homozygous. When a patient is younger than 8 years old, the right antibiotic A is chosen at the effective and lower intensity level more frequently. When a patient is older than 8 years old, the right antibiotic B is chosen more frequently and with higher intensity level, with alternating patterns to avoid resistance or regain susceptibility.

The two subplots at the bottom of Figure 10 illustrate the discovered therapies for two individual patients who are Δ F508 homozygous. The discovered regimen chooses the right antibiotic B initially, and automatically switches to the more suitable antibiotic A at the correct age of 8 years old. In this more severe group, the higher intensity level is chosen more often than the lower intensity level. At the same time, switching the drug or lowering the intensity level, alternatively, is achieved and preferable in order to avoid resistance development and to lower the cumulative burden to the patient.

The discovered adaptive personalized therapy by the reinforcement learning proce-

Figure 9: Barplot of *Pa* infection states average proportions over time using different therapies in a simulated trial with follow up till development of mucoid *Pa* in Study I.

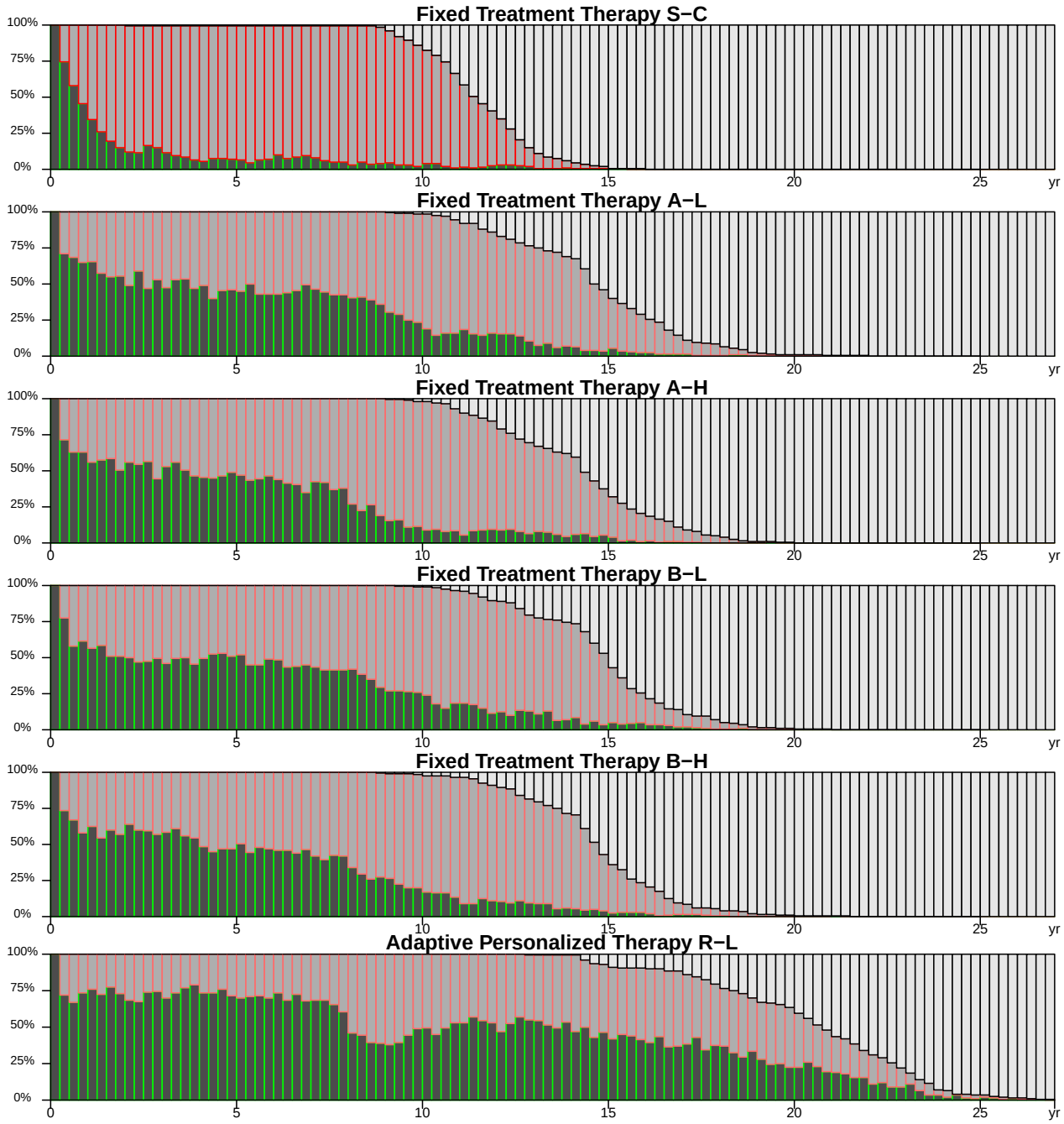


Figure 10: Representation of the optimal adaptive regimens for four individuals who are not Δ F508 homozygous on the top and Δ F508 homozygous at the bottom.

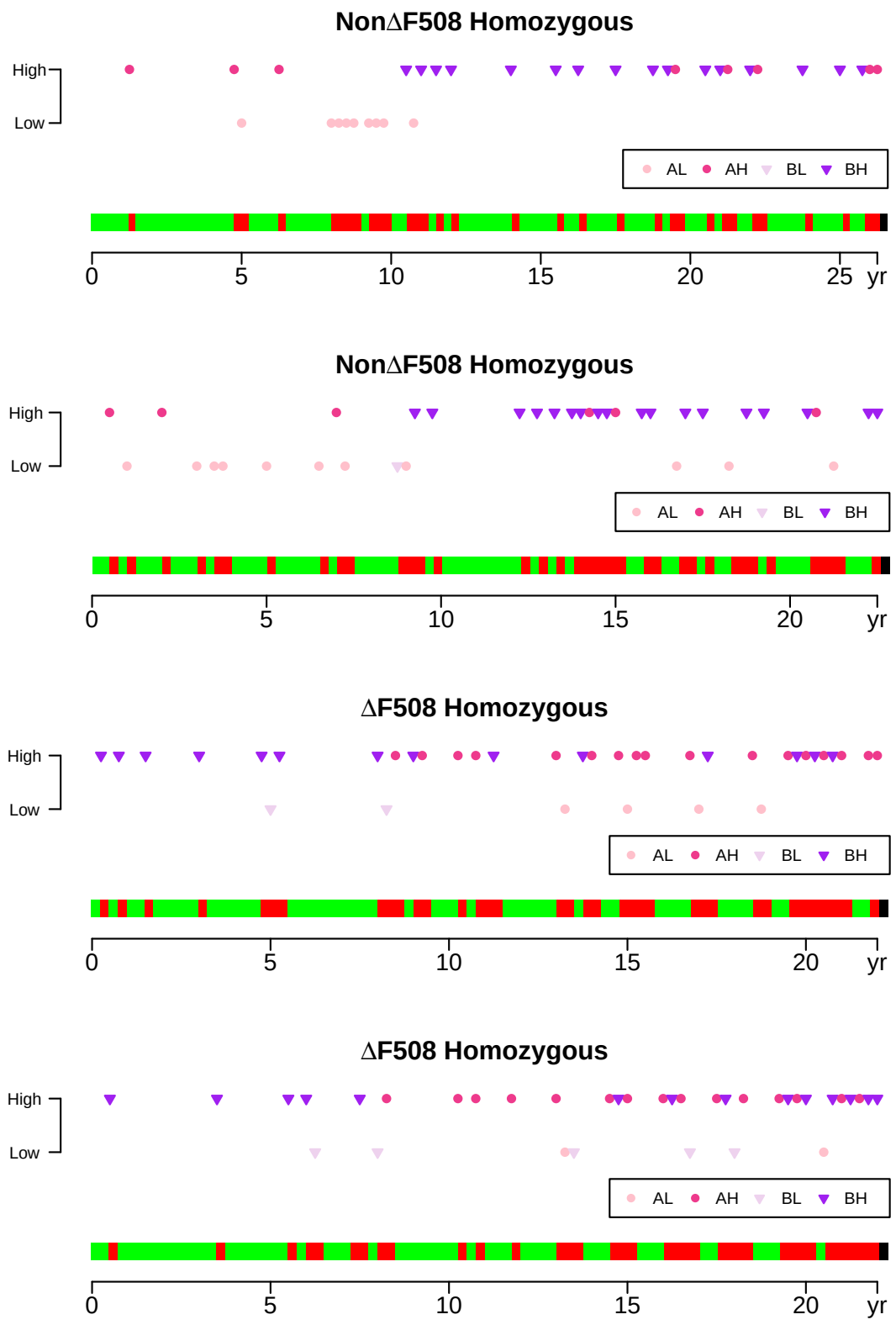
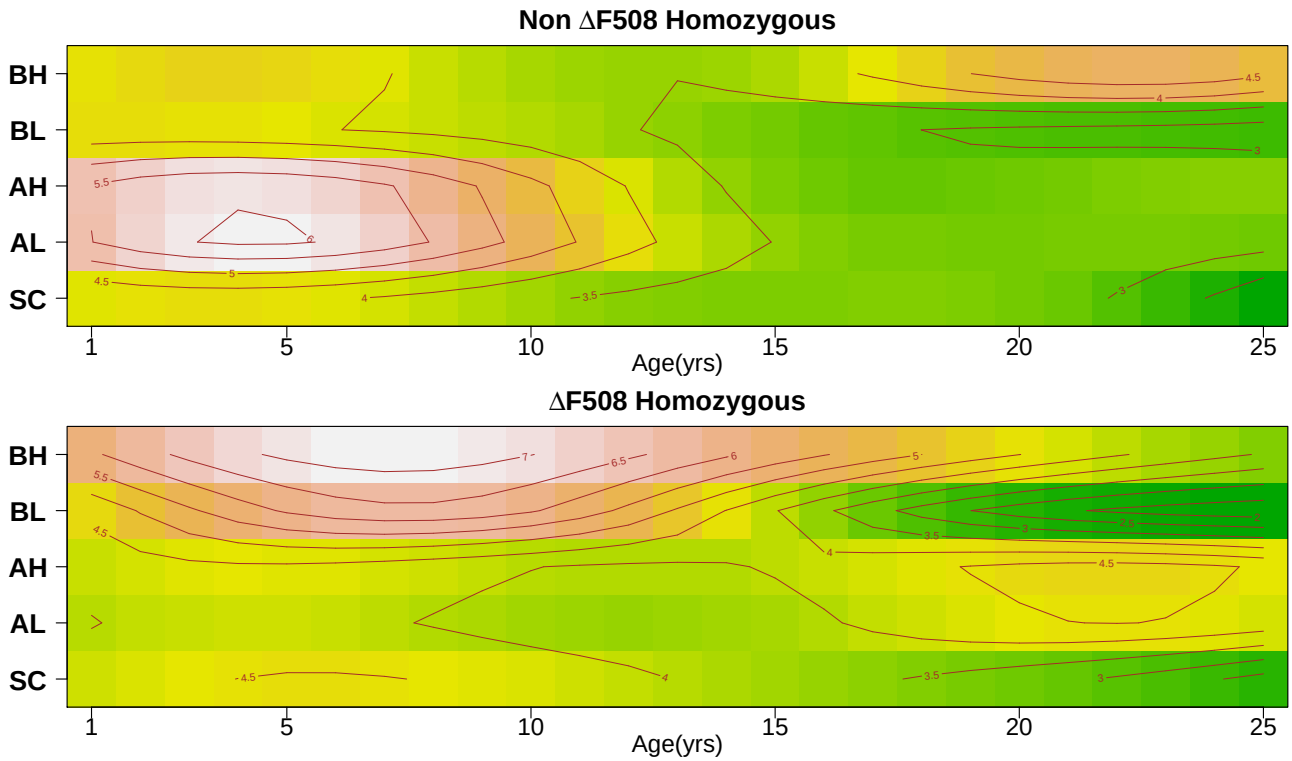


Figure 11: Cumulative reward function.



dure outperforms any fixed treatment regimen therapies because it considers the time varying treatment effect on different age specific groups and balances the trade-off between efficacy and side effects, and immediate and delayed effects, simultaneously. These findings demonstrate the reinforcement learning procedure's substantially powerful long term capabilities. Note that the reinforcement learning approach is unaware of the generative treatment model, and thus the proposed method is able to discover an optimal regimen without prior knowledge of the treatment mechanism.

Although the direct interpretation of the discovered regimen by reinforcement learning using support vector regression is challenging, we use a contour plot of the fitted Q-function by age in two subpopulations to visualize the discovered regimen in Figure 11. Here we fixed the other state variables, including intermittent severity, cumulative intensity level 10 for both antibiotics, 85% predicted $FEV_1\%$, and nonmucoid *Pa* duration as 30% of age. The estimated Q-function demonstrates differential patterns in the two subpopulations and in different age intervals. The overall trend

for patients who are not Δ F508 homozygous is to prefer antibiotic A first and then antibiotic B with low intensity level at early ages, with high intensity level for older ages. While the other group of patients with Δ F508 homozygous prefers antibiotic B first and then antibiotic A with high intensity level.

6.2.3 Testing results of study I in 4 years virtual trial

In the second testing scenario corresponding to Section 5.1, Step 5, we simulated a trial with total sample size 1000 and study duration 4 years. Figure 12 illustrates the Kaplan-Meier plot of time to mucoid Pa of the four fixed treatment regimens and the discovered personalized therapy. The analyses are based on the Cox proportional hazards model (PH), stratified Cox model (SPH), log rank test (LR) and stratified log rank test (SLR), with $\Delta F508H$ as the stratification factor. All tests show no significant treatment difference between the four fixed treatment regimens with p-values given in Figure 12, while the discovered personalized therapy is significantly superior to the other four therapies. In addition, the analysis of the proportion of Pa positive patients during the repeated measurement of culture by a GEE model using a logit link shows no significant treatment differences among the four fixed treatment regimens, while the discovered personalized therapy is significantly superior than the other four therapies.

6.2.4 Study II with MLE parameters

We generate the virtual CF trial based on the disease model in Chapter 4 with treatment effect scenarios described in Section 6.1. The parameters are estimated through maximum likelihood estimation procedure from the Wisconsin neonatal screening project (Li, et al., 2005) in Section 4.4. Similar to Section 6.2.1, the conduct of the clinical reinforcement trial follows the procedure proposed in Section 5.1 with total sample sizes $N_1 = 1000$, $N_2 = 1000$ and study durations $L_1 = 2$ years and $L_2 = 4$

Figure 12: Kaplan-Meier plot of time to mucoid *Pa* infection using different therapies in a simulated trial with 4 years of follow up.

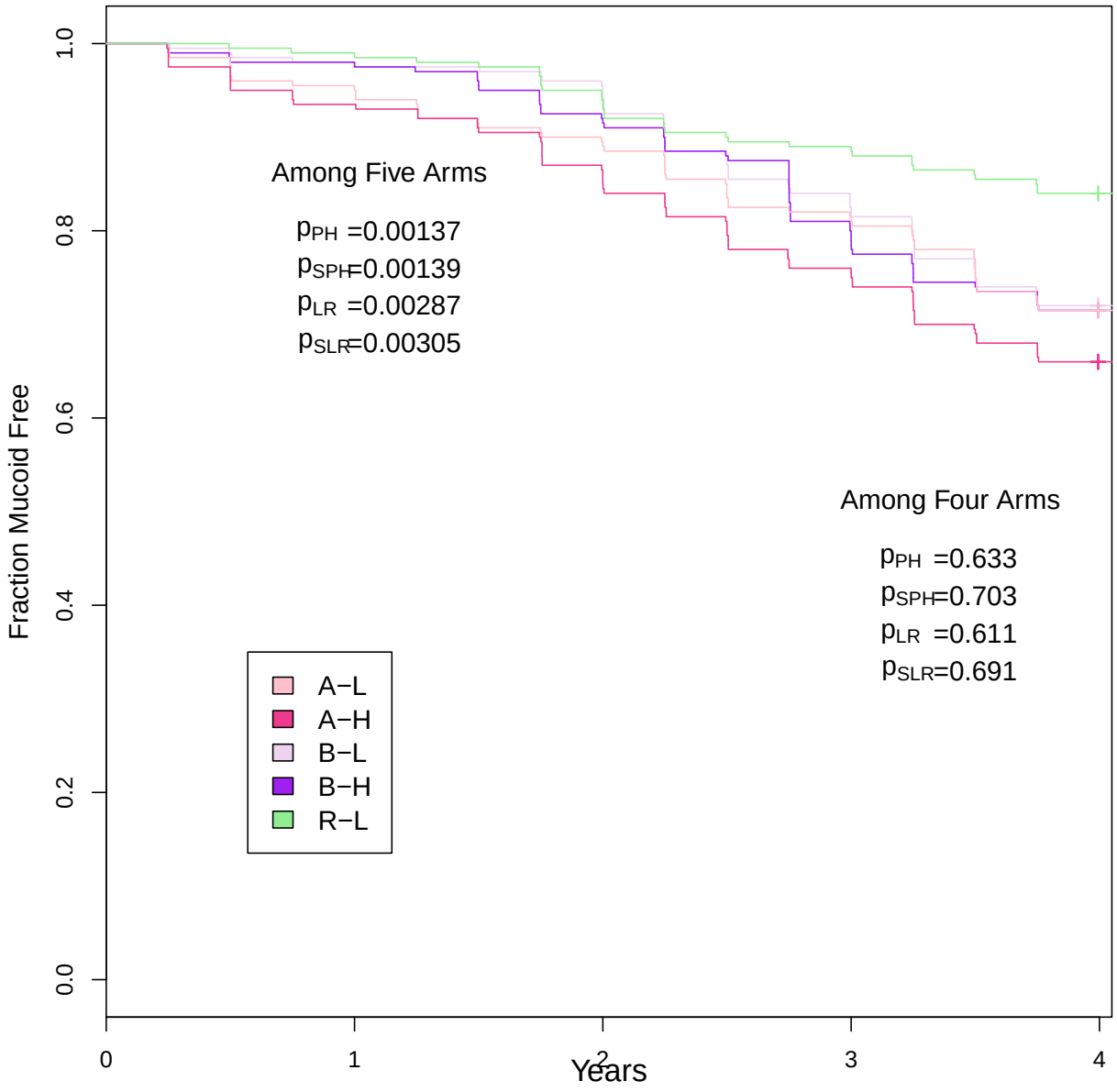


Table 7: Reward/utility function setup II

State Variables	Change	Reward
Culture/Serology	Infected to free of Pa	1
Severity of Infection	progress to mucoid	0
	intermittent to chronic/stay	0.1
	chronic to intermittent	0.2

years for the learning and confirmatory stages respectively, assuming equal numbers of patients in two subgroups defined by whether patients are $\Delta F508H$. We also have two testing scenarios including the confirmatory trial with 1000 total sample size and 4 years of follow up and the scenario with $N_2 = 1000$ where we can apply the therapies to 1000 patients from birth until a mucoid *Pa* event occurs. We use the threshold ϵ in fitted Q-iteration with stopping criteria 10^{-4} and discount factor $\gamma = 0.8$. The immediate reward function set up is in Table 7. Because the little influence of *Pa* infection state change and infection severity change on pulmonary function in the Wisconsin data, we do not assign reward based on the change of $FEV1$.

6.2.5 Testing results of study II in virtual trial from birth till mucoid infection

In this testing scenario, all the virtual patients are given one of the treatment regimes and followed up until the development of mucoid *Pa* infection. Figure 13 shows the boxplot of the time to mucoid *Pa* infection corresponding to the fixed treatment regimens *S-C*, *A-L*, *A-H*, *B-L*, *B-H* and the adaptive personalized therapy denoted *R-L*. There are 200 patients in each arm with one half being Δ F508 homozygous. The empirical performances of these treatment regimens are illustrated in Figure 13

and Table 8. In terms of time to mucoid *Pa* infection, the fixed treatment regimens *S-C*, *A-L*, *A-H*, *B-L*, *B-H* have differential but not obvious effects on the different risk groups classified by ΔF508 homozygosity. The high risk population requires higher intensity level treatment to eradicate *Pa* infection, which matches the treatment effect patterns in the clinical scenarios in Section 6.1, Table 4. But when the right drug is chosen in early childhood, the prognosis will not necessary to be better in average in both subpopulations.

In Figure 14, we illustrate results of the same study of Figure 13 in a different way by grouping patients from the two subpopulations based on Δ F508 homozygosity together. If one ignores patient heterogeneity, the benefits of treatment remain undetected among the fixed treatment therapies and are not even significantly different from standard of care. The differential treatment effects are not obvious due to the big variation. The Figure 15 shows the observed frequency of the three states (*Pa* free (in green), nonmucoid *Pa* (in red) and mucoid *Pa* (in black)) among the 200 patients in each arm over time, demonstrating a similar pattern to time to mucoid *Pa*. By optimizing the usage of these existing drugs, the discovered personalized therapy achieves superior patient outcomes than any other fixed treatment therapies even in the mixture of the two subpopulations.

Figure 13: Boxplot of distribution of time to mucoid *Pa* in study II. The gray and dark green represent patients with Δ F508 homozygosity, otherwise the colors are blue and light green.

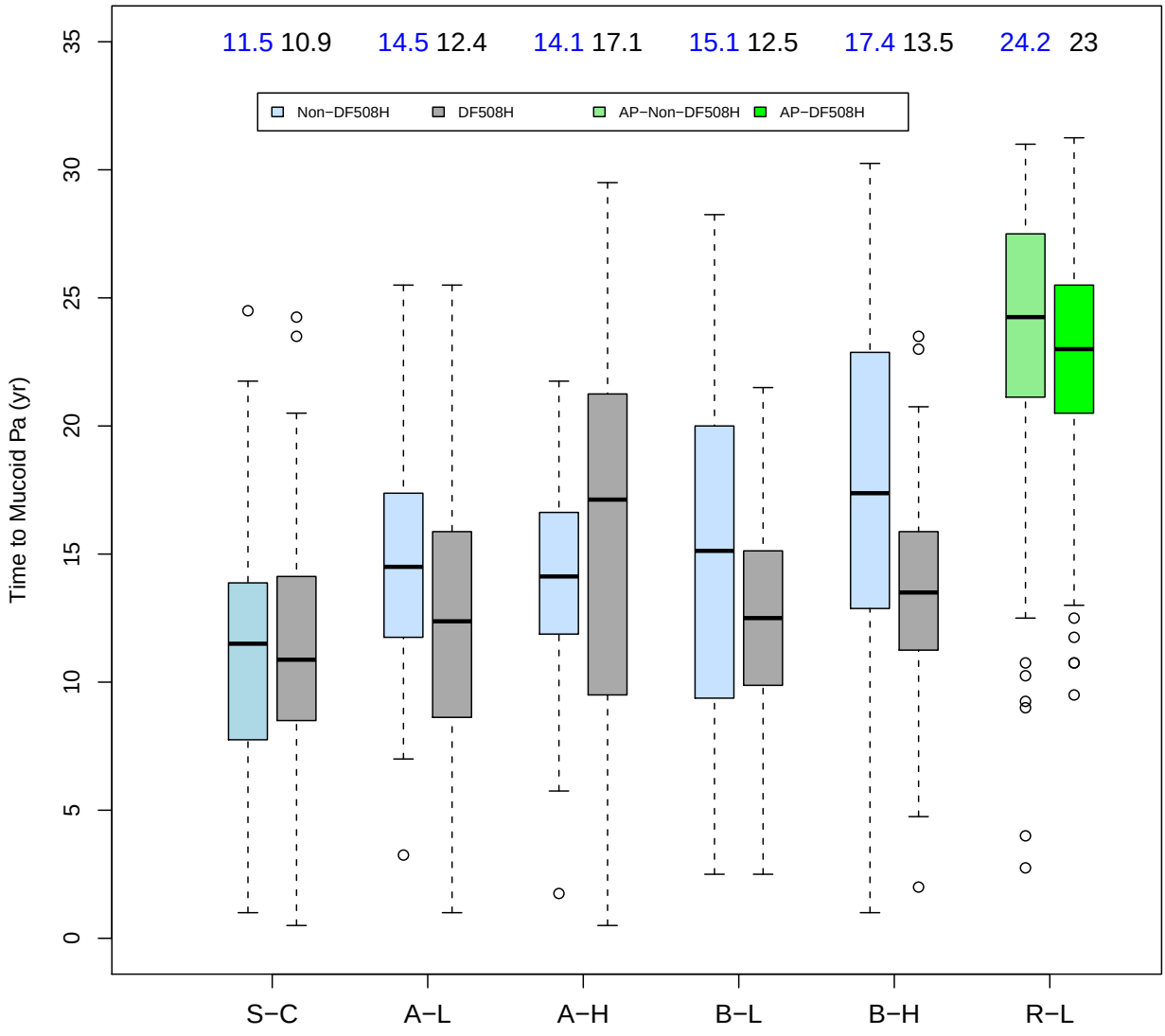


Figure 14: Boxplot of distribution of time to mucoid *Pa* grouped by two subpopulations in study II.

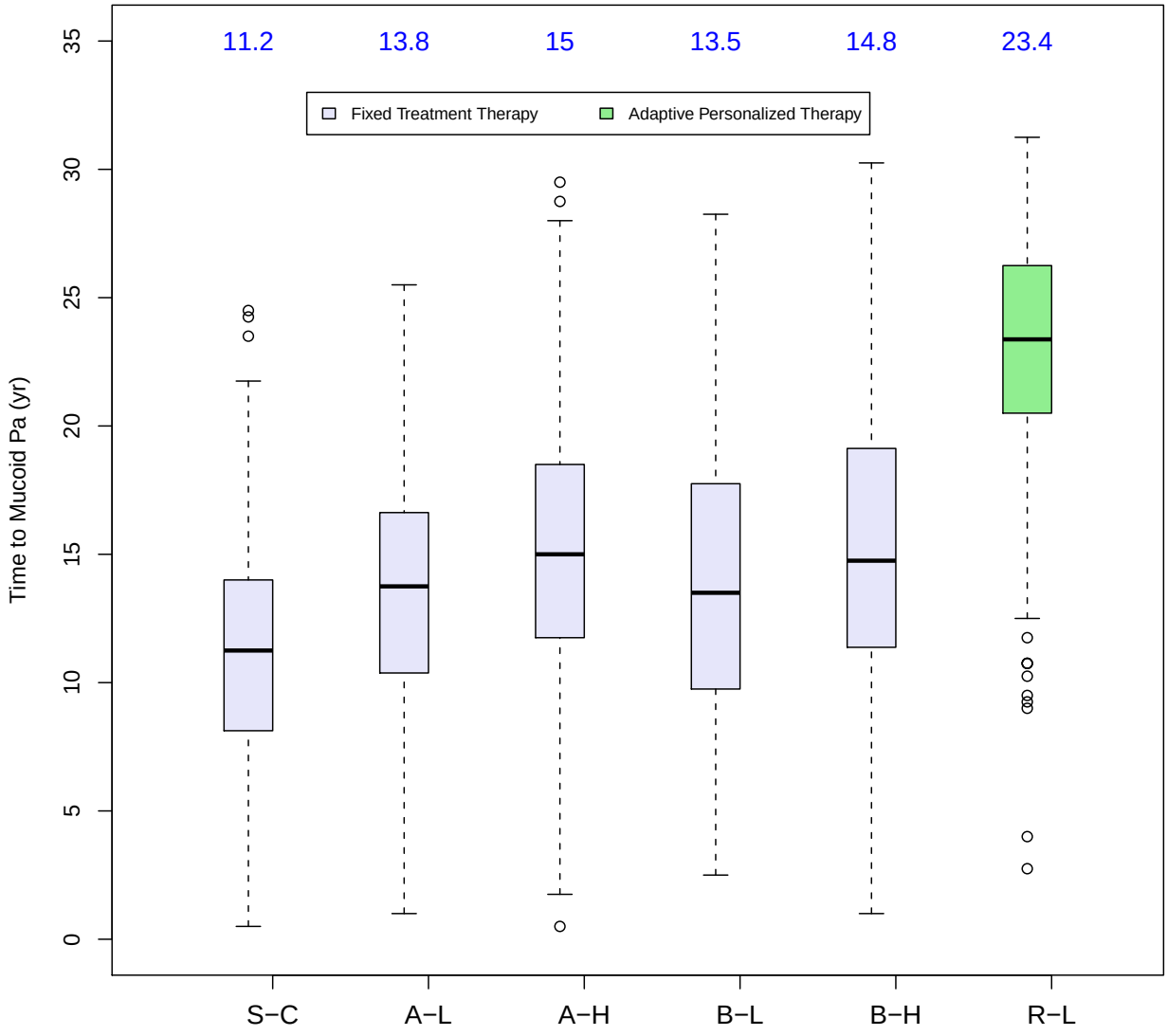


Figure 15: Barplot of *Pa* infection states average proportions over time using different therapies in a simulated trial with follow up till development of mucoid *Pa* in Study II.

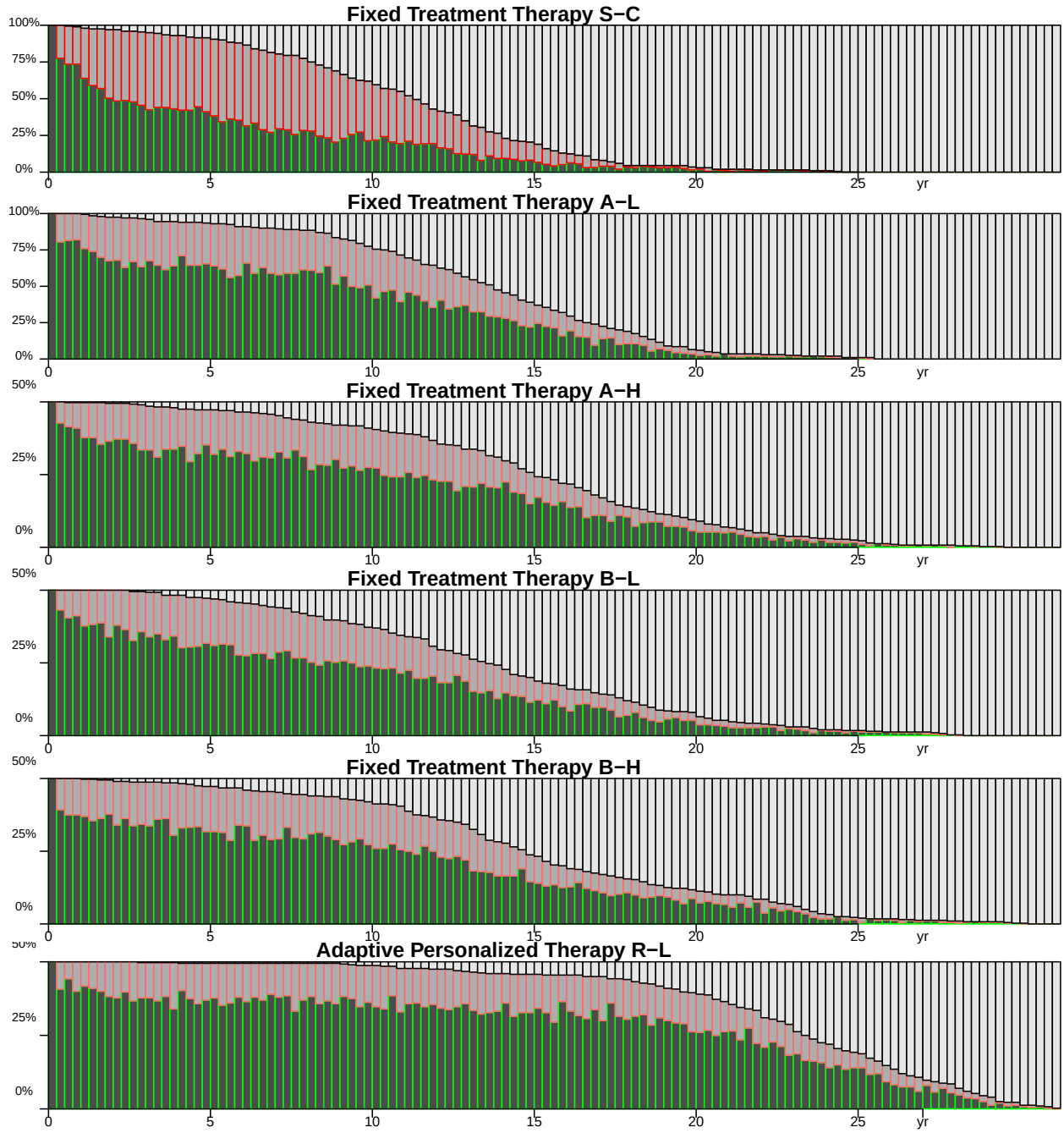


Table 8: Comparisons between fixed treatment regimens and estimated optimal therapy for time to mucoid *Pa* (year) in Study II. Each training/testing dataset is of size 100/subgroup.

Group	Non ΔF508H | ΔF508H						All					
Therapy	SC	AL	AH	BL	BH	RL	SC	AL	AH	BL	BH	RL
					Time to Mucoid *Pa* (T_2) (Yr)							
Mean	11.0|11.2	14.6|12.2	14.1|15.9	15.0|12.6	16.6|13.9	23.3|22.5	11.1	13.4	15.0	13.8	15.2	22.9
SD	4.6|4.6	3.7|5.7	3.5|7.1	6.6|4.2	7.5|3.5	5.7|4.5	4.6	4.9	5.7	5.6	6.0	5.1
Min	1.0|0.5	3.3|1.0	1.8|0.5	2.5|2.5	1.0|2.0	2.8|9.5	0.5	1.0	0.5	2.5	1.0	2.8
Median	11.5|10.9	14.5|12.4	14.1|17.1	15.1|12.5	17.4|13.5	24.3|23.0	11.3	13.8	15.0	13.5	14.8	23.4
Max	24.5|24.3	25.5|25.5	21.8|29.5	28.3|21.5	30.3|23.5	31.0|31.3	24.5	25.5	29.5	28.3	30.3	31.3
					Nonmucoid *Pa* + over T_2 (%)							
Mean	23.7|25.1	17.4|17.3	19.5|17.1	17.2|19.7	18.1|17.8	20.0|18.7	24.4	17.4	15.7	16.1	16.0	19.2

Figure 16 illustrates the discovered therapies for four individual patients, who are not Δ F508 homozygous on the top two subplots and who are Δ F508 homozygous on the bottom two subplots. The discovered regimen chooses the right antibiotic B initially, and automatically switches to the more suitable antibiotic A at the correct age of 8 years old. In this more severe high risk group, the higher intensity level is selected more frequently. When the high cumulative exposure level or continuous usage of a drug occurs, switching the drug or lowering the intensity level, alternatively, is achieved and preferable in order to lower the treatment burden and regain susceptibility.

6.2.6 Testing results of study II in 4 years virtual trial

Corresponding to Section 5.1, Step 5, we simulated a trial with total sample size 1000 and study duration 4 years. Figure 16 shows the Kaplan-Meier plot of time to mucoid *Pa* of the four fixed treatment regimens and the discovered personalized therapy. The analyses are based on the Cox proportional hazards model (PH), stratified Cox model (SPH), log rank test (LR) and stratified log rank test (SLR), with $\Delta F508H$ as the stratification factor. All tests show no significant treatment difference between the four fixed treatment regimens with p-values given in Figure 17, while the discovered personalized therapy is significantly superior than the other four therapies.

Figure 16: Representation of the optimal adaptive regimens for four individuals who are not Δ F508 homozygosity on the top and Δ F508 homozygous at the bottom in study II.

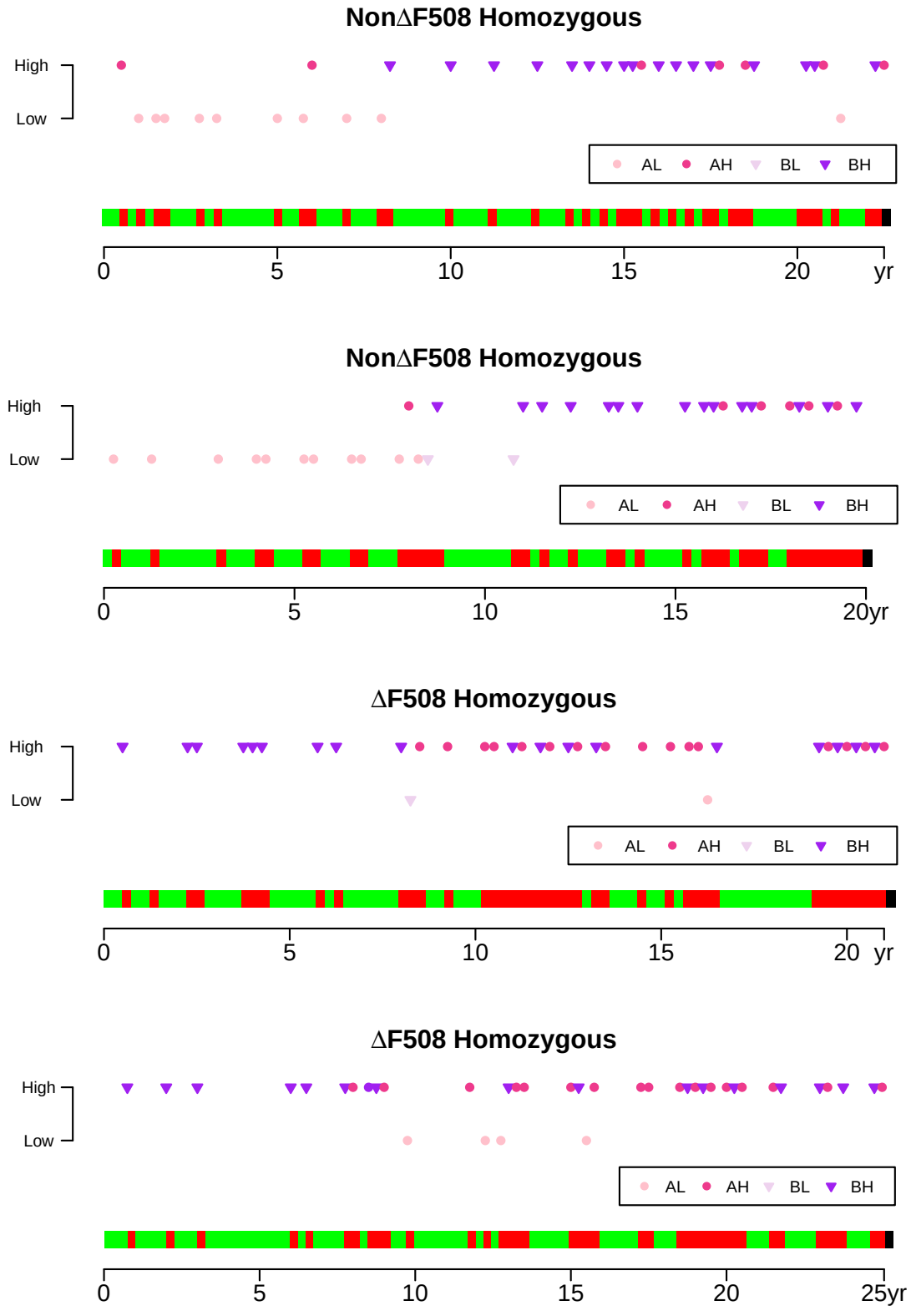
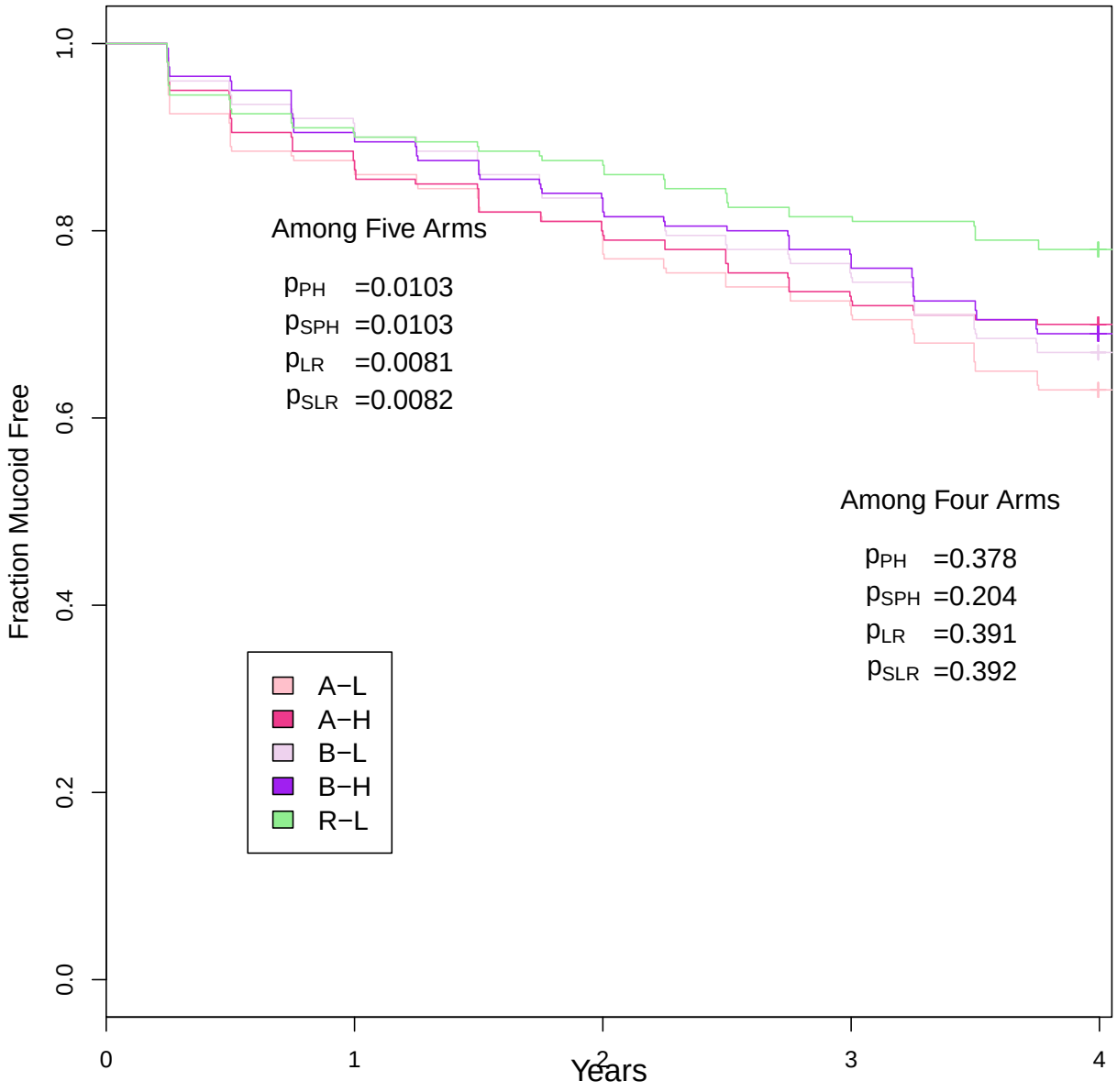


Figure 17: Kaplan-Meier plot of time to mucoid *Pa* infection using different therapies in a simulated trial with 4 years of follow up in study II.



7 CONCLUDING REMARKS

7.1 Overview

In this dissertation, we have proposed the use of a clinical reinforcement trial procedure for discovering effective personalized therapy for patients with CF. After developing a plausible multi-state Markov disease model for the underlying disease dynamics, we simulated several virtual CF trials to investigate the performance of the proposed procedure. In the simulated clinical scenario where standard one-size-fits-all and once-and-for-all approaches are ill-suited, we have shown that the proposed procedure has great potential in tailoring therapy to individual patients, optimizing the timing to switch treatment, and identifying the best suited treatment to a subpopulation. Such adaptive personalized therapies can reduce antibiotic burden while taking into account a drug's immediate and delayed toxicity.

Our research also presents a step towards a potential paradigm shift in the way antibiotic therapies for *Pa* lung infection in CF patients are conceived and evaluated, by taking the view that treatment is part of sequential decision making and thinking in the context of individualized therapies. The proposed clinical reinforcement trial could reveal the treatment effects which might be masked in the traditional clinical trial with single time point and/or ignoring patient heterogeneity.

Additionally, the proposed clinical reinforcement trial procedure has several distinct advantages, including optimizing therapy without relying on the identification of accurate mechanistic models, efficient usage of one unit time step disease transitions by fitted Q-iteration, constructing stationary personalized therapy that has high practicality as a single function representing an adaptive personalized therapy for pa-

tients at different decision time points. Also, at the same time, the therapy preserves age specific characteristics of therapy. Moreover, the cumulative reward procedure in the proposed trial not only provides a novel metric to quantify benefits and risks in the long term, but also provides a framework to integrate benefit risk assessment at the individual level and then accumulates over time to improve decision making. All these encouraging results suggest that the proposed clinical reinforcement trial and accompanying methods can be powerful tools for improving treatment strategies for long term outcomes in chronic diseases.

7.2 Future Research

There are a number of additional topics to work on and challenges we expect to address in future research. First of all, the benefit risk assessment through the reward functioning consists of the metrics and the dimension reduction to quantify the benefit and risk within patient; however, it is unclear how changing these numbers affects the resulting optimal personalized therapies identified. The sensitivity analysis of the reward function, and understanding the robustness of Q-learning to choices of numerical reward and approximation function, clearly deserves further investigation.

Secondly, the model parameters can be estimated from existing data such as the EPIC clinical trial and observational study (Treggiari, et al., 2009), along with expert judgment. The refinement of the disease model for cystic fibrosis and computer tools for evaluation of treatment and monitoring regimens can be very useful in practice to improve the design and to predict long-term health outcomes in this patient population. Refining the proposed clinical reinforcement learning trial will require close collaboration with clinical researchers to improve the practical, logistic aspects, and for actual implementation.

Thirdly, in Chapter 6, we observed that with sample size $N = 1000$ for a clinical reinforcement trial, the discovered personalized therapies are effective and confirmed.

Although the results in extensive simulation studies indicate that good performance can be achieved when the sample size is relatively small, this assumption may be violated in other settings due to the complexity associated with the performance of the approximation of the Q-function through fitted Q iteration algorithm, the higher dimensional state or action space, the estimation accuracy due to approximation through SVR. This sample size calculation is related to the statistical learning error problem. Murphy (2005) derived finite sample upper bounds in the finite horizon non Markovian setting. The finite sample bounds for the fitted Q iteration with regularized supervised learning method such as SVR as approximation methods is an interesting but potentially very difficult question. The further development of this theory can lead to better understand of how the performance of Q-learning with SVR is related to the sample size of the training data in clinical reinforcement trials.

In future research, we also plan to adapt this procedure to other antibiotic therapies for CF patients as well as other therapeutic areas; and to create user-friendly software tools to implement the proposed reinforcement learning procedure for public use.

REFERENCES

Armstrong, D. S., Grimwood, K., Carlin, J. B., Carzino, R., Olinsky, A. and Phenlan, P. D.(1996). Bronchoalveolar lavage or oropharyngeal cultures to identify lower respiratory pathogens in infants with cystic fibrosis. *Pediatr. Pulmonol.* 21(5) 267–275.

Albert, J. M., Yun, H. (2001). Statistical advances in AIDS therapy trials. *Stat. Methods Med. Res.* 10, 85-100.

Altfeld, M., Walker, B. D. (2001). Less is more? STI in acute and chronic HIV-1 infection. *Nat. Med.* 7, 881-884.

Bellman, R. E. (1957). *Dynamic programming*. Princeton University Press, Princeton.

Bertsekas, D. (2000). Dynamic Programming and Optimal Control, *Vol. 1 (Optimization and Computation Series, Athna Scientific*, 2nd edition.

Blatt, D., Murphy, S. A., and Zhu, J. (2004). *A*-learning for approximate planning. *Unpublished manuscript*.

Burns, J. L., Gibson, R. L., McNamara, S., Yim, D., Emerson, J., Rosenfeld, M., et al. (2001). Longitudinal assessment of Pseudomonas aeruginosa in young children with cystic fibrosis. *Journal of Infectious Diseases* 183, 444-452.

Collins, L. M., Murphy, S. A. Bierman, K. A., (2004). A conceptual framework for adaptive preventive interventions. *Prev. Scie* 5, 185-196.

Cystic Fibrosis Foundation. (2008). Patient Registry 2008 annual data report. *Cystic Fibrosis Foundation*.

Dennis, J. E. and Schnabel, R. B. (1983). Methods for unconstrained optimisation and nonlinear equations. *Englewood Cliffs: Prentice Hall*.

Douglas, T. A., Brennan, S., Gard, S., Berry, L., Gangell, C., Stick, S. M., Clements, B. S. and Sly, P. D.(2009). Acquisition and eradication of P. aeruginosa in young children with cystic fibrosis. *Eur. Respir. J.* 33 305–311.

Döring, G., Elborn, J. S., Johannesson, M., de Jonge H., Griese, M., Smyth, A. and Heijerman, H. for the Consensus Study Group(2007). *Journal of Cystic Fibrosis* 6 85–99.

Döring, G., Conway, S. P., Heijerman, H. G., Hodson, M. E., Høiby, N., Smyth, A. and Touw, D. J. for the Consensus Committee(2000). *Eur. Respir. J.* 16 749–767.

Döring, G., Hoiby, N. (2004). Early intervention and prevention of lung disease in cystic fibrosis: a European consensus. *Journal of Cystic Fibrosis* 3, 67-91.

Druker, B. J., et al. (2001). Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia. *New England Journal of Medicine* 344, 1031-1037.

Druker, B. J., et al. (2006). Five-year follow-up of patients receiving imatinib for chronic myeloid leukemia.*New England Journal of Medicine* 355, 2408-2417.

Ernst, D., Geurts, P. and Wehenkel, L. (2005). Tree-based batch model reinforcement learning. *Journal of Machine Learning Research* 6, 503-556.

Ernst, D., Stan, G. B., Goncalve, J., Wehenkel, L. (2006). Clinical data based optimal STI strategies for HIV: a reinforcement learning approach. *Proceeding of the 45th IEEE Conference on Decision and Control*.

Farrell, P. M., Kosorok, M. R., et al.(1997). Nutritional benefits of newborn screening for cystic fibrosis.*New England Journal of Medicine*. 337, 963-969.

Farrell P.M. (2000). Improving the health of patients with cystic fibrosis through newborn screening. *Advances in Pediatrics*. 47,79-115.

Farrell, P. M., Li, Z., Kosorok, M. R., Laxova, A., Green, C., G., Collins, J., Lai, H.C., Makholm, L. M., Rock, M. J.Splaingard, M. L. (2003). Longitudinal evaluation of bronchopulmonary disease in children with cystic fibrosis. *Pediatric Pulmonology* 36 230–240.

Flume, P. A., O'Sullivan, B. P., Robinson, K. A., Goss, C. H., Mogayzel, P. J., Willey-Courand, D. B., Dujan, J., Finder J., Lester, M. Quittell, L., Rosenblatt, R., Vender, R. L., Hazle, L., Sabadosa, K. and Marshall, B. (2007). Cystic fibrosis pulmonary guidelines: chronic medications for maintenance of lung health. *Am. J. Respir. Crit. Care Med.* 176 957–969.

Flume, P. A., Mogayzel, P. J., Robinson, K. A., Goss, C. H., Rosenblatt R. L., Kuhn, R. J. and Marshall, B. C.; Clinical Practice Guidelines for Pulmonary Therapies Committee. (2009). Cystic fibrosis pulmonary guidelines: treatment of pulmonar exacerbations. *Am. J. Respir. Crit. Care Med.* 180 802–808.

Geurts, P., Ernst, D. and Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning* 11, 3-42.

Guez, A., Vincent, R. D., Avoli, M. and Pineau, J. (2008). Adaptive Treatment of Epilepsy via Batch-mode Reinforcement Learning. *AAAI* 1671–1678.

Gibson, R. L., Burns J. L., Ramsey, B. W. (2003a). Pathophysiology and management of pulmonary infections in cystic fibrosis. *American Journal of Respiratory and Critical Care Medicine*. 168, 918-951.

Gibson, R. L., Emerson, J., McNamara, S., Burns, J. L., Rosenfeld, M., Yunker, A., et al. (2003b). Significant microbiological effect of inhaled tobramycin in young children with cystic fibrosis. *American Journal of Respiratory and Critical Care Medicine*. 167, 841-849.

Gordon, G. J. (1999). Approximate Solutions to Markov Decision Processes. PhD thesis, Carnegie Mellon University.

Hentzer, M., Teitzel GM, Balzer GJ, Heydorn A, Molin S, Givskov M, et al. (2001) Alginate overproduction affects Pseudomonas aeruginosa biofilm structure and function. *J Bacteriol* 183, 5395-5401

Kaelbling, L. P., Littman, M. L., and Moore, A. W. (1996). Reinforcement Learning: A Survey. *Journal of Artificial Intelligence Research* 4, 237-285.

Kalbfleisch, J. D. and Lawless, J. F. (1985). The analysis of panel data under a Markov assumption. *Journal of the American Statistical Association* 80(392) 863–871.

Kosorok, M. R., Zeng, L., et al.(2001) Acceleration of lung disease in children with cystic fibrosis after Pseudomonas aeruginosa acquisition. *Pediatric Pulmonology*. 32,277-287.

Langton Hewer SC, Smyth, AR (2009). Antibiotic strategies for eradicating Pseudomonas aeruginosa in people with cystic fibrosis. *Cochrane review*

Lavori, P. W., Dawson, R. (2000a). A design for testing clinical strategies: Biased adaptive within-subject randomization. *Journal of the Royal Statistical Society Seriers A* 163, 29-38.

Lavori, P. W.,Dawson, R., Rush, A.J. (2000b). Flexible treatment strategies in chronic disease: clinical and reserach implications. *Biol. Psychiatry* 48, 605-614

Lavori, P. W., Rush, A. J., Wisniewski, S. R., Alpert, J., Fava, M., Kupfer, D. J., Nierenberg, A., Quitkin, F. M., Sackeim, H. A., Thase, M. E., Trivedi, M.' (2001). Strengthening clinical effectiveness trials: equipoise-stratified randomization. *Biol. Psychiatry* 50 (10), 792-780.

Li, Z., Kosorok, M. R., Farrell, P. M., et al. (2005) Longitudinal Development of Mucoid Pseudomonas aeruginosa Infection and Lung Disease Progression in Children with Cystic Fibrosis. *Journal of the American Medical Association* 293, 581-588.

Meira-Machado, L., F., Una-Alvarez, J. D., Cadarso-Suarez, C. and Andersen, P. (2009). Multi-state models for the analysis of time-to-event data. *Stat. Method Med. Res.* 18 195–222.

Mayer-Hamblett, N., Aitken, M. L., Accurso, F. J., Kronmal, R. A., Konstan, M. W., Burns, J. L., Sagel, S. D. and Ramsey, B. W. (2007). Association between pulmonary function and sputum biomarkers in cystic fibrosis. *Am. J. Respir. Crit. Care Med.* 175 822–828.

Mayer-Hamblett, N., Ramsey, B. W. and Kronmal, R. A. (2007). Advancing outcome measures for the new era of drug development in cystic fibrosis. *Proc. Am. Thorac Soc.* 4 370–377.

Moodie, E. M., Richardson, T. S., Stephens, D. A. (2006). Demystifying optimal dynamic treatment regimes. *Biometrics* 63, 447-455.

Murphy, S. A. (2005a). An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine* 24, 1455-1481.

Murphy, S. A. (2005b). A generalization error for Q-learning. *Journal of Machine Learning Research* 6, 1073-1097.

Murphy, S. A., Lynch, K. G., Oslin, D., McKay, J. R. and TenHave, T. (2007).

Developing adaptive treatment strategies in substance abuse research. *Drug and Alcohol Dependence* 88S, S24-S30.

Murphy, S. A., Lynch, K. G., et al. (2007a). Developing adaptive treatment strategies in substance abuse research. *Drug and Alcohol Dependence* 88S, S24-S30.

Murphy, S. A., Oslin, D. W., Rush, A. J. and Zhu, J. (2007b). Methodological challenges in constructing effective treatment sequences for chronic psychiatric disorders. *Neuropsychoarmacology* 32, 257-262.

Murphy, S. A., Van der Laan, M. J., and Robins, J. M. (2001). Marginal mean models for dynamic regimes. *Journal of the American Statistical Association* 96, 1410-1423.

Ormoneit, D. and Sen, S. (2002). Kernel-based reinforcement learning. *Machine Learning*, 49(2-3), 161-178.

Oslin, D. W., Sayers, S., Ross, J., Kane, V., Have, T., Conigliaro, J., Cornelius, J., (2003). Disease management for depression and at-risk drinking via telephone in an older population of veterans. *Psychosom. Med.* 65 (6), 931-937.

O'Shea, J. C., Hafley, G. E., Greenberg, S., Hasselblad, V., Lorenz, T. J., Kitt, M. M., Strony, J., Tcheng, J. E. (2001). Platelet Glycoprotein IIb/IIIa Integrin Blockade With Eptifibatide in Coronary Stent Intervention. The ESPRIT Trial: A Randomized Controlled Trial *Journal of the American Medical Association* 285, 2468-2473.

Pedersen, S.S., Espersen, F. and Høiby, N. (1987). Diagnosis of chronic Pseudomonas aeruginosa infection in cystic fibrosis by enzyme-linked immunosorbent assay. *J. Clin. Microbiol.* 25(10) 1830–1836.

Pineau, J., Bellemare, M. G., Rush, A. J., Ghizaru, A., and Murphy, S. A. (2007). Constructing evidence-based treatment strategies using methods from computer science. *Drug and Alcohol Dependence* 88S, S52-S60.

Prince, A. S. (2002). Biofilms, antimicrobial resistance, and airway infection. *New England Journal of Medicine* 347, 1110-1111.

Putter, H., Fiocco, M. and Geskus, R. B. (2007). Tutorial in biostatistics: Competing risks and multi-state models. *Statistics in Medicine* 26 2389–2430.

Ramsey, B. W., Wentz, K. R., Smith, A. L., Richardson, M., Williams-Warren., J., Hedges, D. L., Gibson., R., Redding., G. J., Lent, K., and Harris, K. (1991). Predictive value of oropharyngeal for identifying lower airway bacteria in cystic fibrosis patients. *AM REV RESPIR DIS 1991* 144, 331-337.

Ramsey, B.W., Pepe, M.S., Quan, J. M., Otto, K. L., Montgomery, A. B.,Williams-Warren, J., Vasiljev-K, M., Borowitz, D., Bowman, C. M., Marshall, B. C., Marshall, S.and Smith, A. L. The Cystic Fibrosis Inhaled Tobramycin Study Group (1999). Intermittent administration of inhaled tobramycin in patients with cystic fibrosis. *N. Engl. J. Med.* 340 23–30.

Ratjen, F., Doring, G. (2003). Cystic Fibrosis. *Lancent* 361, 681-689.

Ratjen F., Doring G., Nikolaizik WH. (2001). Effect of inhaled tobramycin on early Pseudomonas aeruginosa colonization in patients with cystic fibrosis. *Lancent* 358, 983-984.

Retsch-Bogart, G. Z. (2009). Update on new pulmonary therapies. *Current Opinion in Pulmonary Medicine* 15, 604-610.

Rosenfeld, M., Emerson, J., Accurso, F., Armstrong, D., Castile, R., Grimwood, K., Hiatt, P., McCoy, K., McNamara, S., Ramsey, B. and Wagener, J. (1999). Diagnostic accuracy of oropharyngeal cultures in infants and young children with cystic fibrosis. *Pediatr. Pulmonol.* 28(5) 321–328.

Rosenfeld, M., Gibson, R. L., McNamara, S., Emerson, J., Burns, J. L., Castile, R., et al. (2001) Early pulmonary infection, inflammation, and clinical outcomes in infants with cystic fibrosis. *Pediatric Pulmonology* 32, 356-366.

Romond, EH., Perez, E. A., Bryant, J., et al. (2005). Trastuzumab plus adjuvant chemotherapy for operable HER2-positive breast cancer. *New England Journal of Medicine* 353, 1673-1684.

Robins, J. M. (2004). Optimal structual nested models for optimal sequential decisions. *Proceddings of the Second Seattle Symposium on Biostatistics. New York: Springer*, 189-326.

Rowe, S. M., Miller, S., and Sorscher, E. (2005). Cystic Fibrosis mechanisms of disease. *New England Journal of Medicine* 355, 2408-2417.

Rush, A. J., Crismon, M. L., Kashner, T. M., Toprac, M. G., Carmody, T. J., Trivedi, M. H., Suppes, T., Miller, A. L., Biggs, M. M., Shores-Wilson, K., Witte, B.P., Shon, S. P., Rago, W. V., Altshuler, K. Z., TMAPResearch Group (2003). Texas medication algorithm project, phase 3 (TMAP-3): rationale and study design. *J. Clin. Psychiatry* 64(4), 357-369.

Ryan, G., Mukhopadhyay, S. et al. (2009). Nebulised anti-pseudomonal antibiotics for cystic fibrosis. *Cochrane review*.

Schneider, L. S., Tariot, P. N., Lyketsos, C. G., Dagerman, K. S., Davis, K. L., Davis, S., Hsiao, J. K., Jeste, D. V., Katz, I. R., Olin, J. T., Pollock, B. G., Rabins, P. V., Rosenheck, R. A., Small, G. W., Lebowitz, B., Lieberman, J. A., et al. (2001). National Institute of Mental Health clinical antipsychotic trials of intervention effectiveness (CATIE). *J. Geriatr. Psychiatry* 9(4), 346-360.

Simon, R. (2008). Using genomics in clinical trial design. *Clinical Cancer Research* 14,5984-5993.

Slamon, D. J. , Leyland-Jones, B., Shak, S., et al. (2001). Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. *New England Journal of Medicine* 344, 783-792.

Southern, K. W., Marieke, M. E., Dankert-Roelse, J. E. and Nagelkerke, A. (2009). Newborn screening for cystic fibrosis. *Cochrane Database Syst. Rev.* 1 CD001402.

Starner, T. D., McCray, P. B. (2005). Pathogensis of early lung disease in Cystic Fibrosis: a window of opportunity to eradicate bacteria. *Annuals of Internal Medicines* 143, 816-822.

Sutton, R. S. (1988). Learning to predict by the method of temporal differences. *Machine Learning* 3(1), 9-44.

Sutton, R. S. and Barto, A. G. (1998) *Reinforcement learning: an introduction* MIT Press, Cambridge, MA.

Taccetti, G. (2005). Early eradication therapy against Pseudomonas aeruginosa in cystic fibrosis patient. *Europe Respiratory Journal* 26, 458-461.

Thall, P. F., Millikan, R. E., and Sung, H. G. (2000). Evaluating multiple treatment courses in clinical trials. *Statistics in Medicine* 19, 1011-1028.

Thall, PF., Sung, HG. and Estey, EH. (2002). Selecting therapeutic strategies based on efficacy and death in multicourse clinical trials. *Journal of the American Statistical Association* 457, 29-39.

Treggiari, M. M., Rosenfeld, M., Retsch-Bogart, G. Z., et al. (2009). Early anti-pseudomonal acquisition in young patients with cystic fibrosis: Rationale and design of the EPIC clinical trial and observational study. *Contemporary Clinical Trials* 30, 751-756.

Treggiari, M. M., Rosenfeld, M., Retsch-Bogart, G., Gibson, R. and Ramsey, B. (2007) Approach to Eradication of Initial Pseudomonas aeruginosa Infection in Children With Cystic Fibrosis. *Pediatric Pulmonology* 42(9), 751-756.

Tsitsiklis, J. N. and Van, R. B. (1996). Feature-based methods for large scale dynamic programming. *Machine Learning* 22, 59-94.

Valerius, N. H., Koch, C., Hoiby, N. (1991). Prevention of chronic Pseudomonas aeruginosa colonisation in cystic fibrosis by early treatment. *Lancent* 338, 725-726.

Vapnik, V. (1995). *The nature of statistical learning theory.* Springer, New York.

Vapnik, V., Golowich, S. and Smola, A. (1997). Support vector method for function approximation, regression estimation, and signal processing. *Advances in Neural Information Processing Systems* 9, 281-287.

Wang, S. J. (2007). Approaches to evaluation of treatment effect in randomized clinical trials with genomic subset. *Pharmaceutical Statististics* 6, 283-292.

Waters, V., Ratjen F. (2008). Combination antimicrobial susceptibility testing for acute exacerbations in chronic infection of Pseudomonas aeruginosa in cystic fibrosis. *Cochrane review*

Watkins, C. J. C. H. (1989). *Ph.D. Thesis* King's College, Cambridge, UK.

Zhao, Y., Kosorok, M. R., Zeng, D. (2009). Reinforcement learning design for cancer clinical trials. *Statistics in Medicine* 28, 3294-3315.

Zhao, Y., Zeng, D., Socinski, M. A. and Kosorok, M. R. (2010). Reinforcement Learning Strategies for Clinical Trials in Non-small Cell Lung Cancer. *Biometrics*, In revision (invited).