DEVELOPING ADAPTIVE PERSONALIZED THERAPY FOR CYSTIC FIBROSIS USING REINFORCEMENT LEARNING

BY YIYUN TANG^{*}, MICHAEL R. KOSOROK^{*}

University of North Carolina at Chapel Hill

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 patient. In this paper, we examine the problem of computing optimal adaptive personalized 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 conduct a simulation study of virtual cystic fibrosis patients with *Pseudomonas aeruginosa* infection and antibiotic therapy with parameters tuned to approximately match published data from CF patients. Our simulation results indicate that reinforcement learning can be an effective tool in developing personalized therapy which optimises the benefit-risk trade off in multi-stage decision making and improves long term outcomes in chronic diseases.

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 [4]. 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 [41]. 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 [5, 12]. Essentially, these treatment decisions are made based on clinical

^{*}To whom corresponds should be addressed.

Keywords and phrases: Reinforcement Learning, Cystic Fibrosis, Dynamic Treatment Regime, Personalized Medicine, Clinical Trial, Multi-stage Decision Making

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 [41]. Second, the frequent infection relapse and progression require timely treatment modification [29]. Third, the chronic nature of CF leads to repeated courses of potentially toxic drugs for many years, increasing risk of cumulative sideeffect, such as drug resistance, impairment of renal function and hearing [5, 6, 12]. 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 [7, 20, 37, 42, 54] 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 evaluating 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 [46, 48, 49]; specifically, the strategy of intermittent administration of inhaled tobramycin may reduce the risk of resistance development [35]. 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 article, we present a "clinical reinforcement trial" procedure to discover optimal personalized therapy for CF which seeks to address the above questions and to tailor therapy to patients' inherent characteristics and adapt to time varying factors in the disease process in order to improve the entire decision-making process. The discovery of optimal therapy in this approach is based on a reinforcement learning method, called Q-learning, which obtains patient responses to different regimens and maximizes the average long term outcomes as a function of patients' clinical status and multi-stage regimens using backward and/or recursive algorithms. The clinical reinforcement trial approach based on Q-learning for discovering effective regimens was first introduced for potentially irreversible diseases such as cancer in [55]. This framework was further refined for clinical trials in non-small cell lung cancer after adaptation to handle right-censored survival data [56]. This clinical reinforcement trial framework is an extension and melding of earlier work on dynamic treatment regimens in counterfactual frameworks [26, 28, 40] and sequential multiple assignment randomized trials (SMART) [25, 47] which have been applied to behavioral and psychiatric disorders [27, 32]. 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 [43], acquire Pa infection in early childhood, and experience frequent reinfection [19, 22]. 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 [4]. 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 [22]. 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 consisting of exploratory and confirmatory stages. In the exploratory stage, we utilize our prior knowledge and historical data of this disease, such as pathogenesis and age-specific feature of Pa infection to design the multiple courses trial involving a fair randomization of patients among different possible treatment options as well as the collection of clinical relevant outcomes and biomarkers at each time point, usually every 3-month in CF clinical practice. Based on the resulting longitudinal data from the proposed sequential randomization trial, we propose to estimate a personalized therapeutic regimen which synthesizes patient information on all aspects 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 the confirmatory stages, we will demonstrate that the optimal personalized therapy identified in the learning stage is superior to fixed treatment regimens in prolonging time to mucoid Pa in patients with CF by a conventional randomized control trial. The data and positive results from both stages will establish the clinical utility of the optimal personalized therapy for future CF patients.

In the clinical trials, patients at various age ranges are enrolled with long follow-up period in order to capture the known age-specific feature of Pa infection, including patients diagnosed at birth, followed up life time, disease progressing at different age, physician visit at least every 3 months. Age services as both decision time point and important factor that might dictate the optimal treatment option. The wide age-range enrollment and long follow-up time aim to obtain representative samples and capture the temporary impact of the treatment effects.

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 [9] in reinforcement learning (RL) [18] to estimate the optimal therapy. In some applications of RL to inform multi-stage therapies, such as STI strategies for HIV [10], 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 [11], and so we propose instead to conduct a second, confirmatory trial to validate the estimated optimal therapy by comparing to existing 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 [22]. 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 Painfection, and mucoid Pa.

In Section 2, we formulate the problem within a reinforcement learning context, specifically Q-learning, followed by the fitted Q-iteration algorithm for estimating the required Q-functions without the time index. In Section 3, we provide details on disease progression and propose a discrete time non-homogeneous multistage CF disease model as a generative model for the simulation studies. The CF trial conduct, and related computation and validation of the optimal therapy are presented in Section 4. We apply the proposed procedure in a simulation study in Section 5. We close with a discussion in Section 6.

2. Technical Background.



FIG 1. Reinforcement learning in anti-Pa therapy treating lung infection for cystic fibrosis

2.1. Reinforcement Learning in Medical Decision Making. Reinforcement learning (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 [17, 45].

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. In the medical decision making setting, the state S_t corresponds 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. Let $\bar{S}_t = (S_0, ..., S_t)$ and $\bar{A}_t = (A_0, ..., A_t)$ represent histories of state and action. The reward is defined as a function of action and state, i.e., $R_t = r_t(\bar{S}_{t+1}, \bar{A}_t)$, which 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. 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

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

The "policy" $\pi_t(\bar{s}_t, \bar{a}_{t-1}) = a_t, t = 0, ..., T$ maps from the state-action history to the next action. The resulting action a_t from π_t depends on early action sequential $\bar{a}_{t-1} = (a_0, ..., a_{t-1})$ through state sequential $\bar{s}_t = (s_0, ..., s_t)$.

The goal of reinforcement learning is to find the optimal policy resulting in the maximum expected discounted cumulative return given by $\sum_{t=0}^{T} \gamma^t r_t$, which corresponds to our aim to discover the optimal personalized therapy which achieves the most beneficial ultimate outcome in the long run. The discount rate γ (0 < γ < 1) for each time unit balances the weights of immediate rewards and future rewards.

To accomplish this goal, we utilize Q-learning [53], one of the most widely used reinforcement learning methods. The direct relationship between the optimal "value function" V_t^* , the optimal value function over state-action pairs (the "Q-function") Q_t^* , and the optimal policy are given by

(2.1)

$$Q_{t}^{*}(\bar{S}_{t}, \bar{A}_{t}) = E \Big[R_{t} + V_{t+1}^{*}(\bar{S}_{t+1}, \bar{A}_{t}) | \bar{S}_{t}, \bar{A}_{t} \Big],$$

$$V_{t}^{*}(\bar{S}_{t}, \bar{A}_{t-1}) = \max_{a_{t}} Q_{t}^{*}(\bar{S}_{t}, \bar{A}_{t-1}, a_{t}),$$

$$\pi_{t}^{*}(\bar{s}_{t}, \bar{a}_{t-1}) = \arg\max_{a_{t}} Q_{t}^{*}(\bar{s}_{t}, \bar{a}_{t-1}, a_{t}).$$

Based on Bellman equation, the sequence of Q-functions satisfy the recurrence equation [2, 26, 45]

(2.2)
$$Q_t^*(\bar{S}_t, \bar{A}_t) = E\Big[R_t + \gamma \max_{a_{t+1}} Q_{t+1}^*(\bar{S}_{t+1}, \bar{A}_t, a_{t+1}) | \bar{S}_t, \bar{A}_t\Big].$$

The value of the action and state at time t is the sum of expected immediate reward at t and the expected future value if making the optimal decision from time t+1 till the end. Backward induction is the key point of optimization, which starts from the end, i.e. the immediate reward of the last treatment at t = T, and works backward through time t = T - 1, ..., 0 sequentially, until Q_0 for the initial action and state.

The Q-learning algorithm combined with supervised learning approximates the sequences of Q-functions by \hat{Q}_t , which determines the optimal policy directly as given in (2.1).

The most suitable type of Q-learning for our setting is model-free Qlearning in batch mode with an approximation algorithm [53]. 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. 2.2. Fitted Q-Iteration Algorithm. The fitted Q-iteration algorithm [9] 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 (2.2) in the discrete MDP problem becomes:

(2.3)
$$Q_N^{\star}(S_t, A_t) = E\Big[R(S_t, A_t) + \gamma \max_a Q_{N-1}^{\star}(S_{t+1}, a)|S_t, A_t\Big], \forall N > 1$$

with $Q_1^{\star}(S_t, A_t) = R(S_t, A_t)$. As number of one-step transition N increases, i.e. the total patient sample size and/or the stage number of single patient trajectory increase, the sequence converges in infinity norm to the optimal stationary Q-function in a stationary process. The resulting optimal policy is $\pi_N^*(s) = \arg \max_a Q_N^*(s, a)$.

At each iteration, using the empirical r_t , the approximation (2.3) can be formulated as a sequence of standard supervised learning steps on the kth training sample, taking the form

$$\mathcal{TS}_{k} = \{(s_{t}^{l}, a_{t}^{l}, r_{t}^{l} + \gamma max_{a}\hat{Q}_{k-1}(s_{t+1}^{l}, a), s_{t+1}^{l})\}_{l=1}^{\#\mathcal{F}}, \forall k > 1$$

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

(2.4)
$$\hat{\pi}_N(s) = \arg \max_a \hat{Q}_N(s,a).$$

Hence, fitted Q-iteration can be combined with any regression algorithm to fit the Q-function with the property of consistency [9]. The diagram and realization of this algorithm in estimating optimal CF therapy will be provided in Section 4.1 step 4. The extensive testing of fitted Q-iteration in standard RL simulation [9] and in clinical applications in HIV [10] and epilepsy [16] 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, which lead to an appropriate Markovian working assumption 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.

2.3. Support Vector Regression. 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) [52] as the main approximation method for fitting the Q-functions.

As one of the most popular extensions of support vector machine, SVR is a more general and flexible approach compared to competing methods



which handle potentially complex nonlinear relationship between rewards and state-action pairs, because ϵ -insensitive loss function defines errors within deviation ϵ as acceptable and the data is mapped through the nonlinear transformation into a feature space within the reproducing kernel Hilbert space (RKHS) context. Also, overfitting training data can be avoided in SVR through regularization term. Hence, fast and high quality performance can be archieved.

SVR performs with similar or better reproducibility in clinical research settings [55] as extremely randomized trees [13], a popular, more computationally intense alternative also used for fitted Q-iteration.

3. Disease Dynamics.

3.1. 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 [19, 22, 46, 49]. Mucoid Pa, a mutant phenotype of Pa, develops at later stages, and lives in a defensive mode of growth called a biofilm [33]. Hence it confers resistance to phagocytosis and antibiotics and is much more difficult to treat and eradicate [14]. Therefore, there are three phenotypically distinguishable states: free of Pa (state 1), nonmucoid Pa (state 2), and irreversible mucoid Pa (state 3), as illustrated in Figure 2.

There are three major classes of endpoints in CF trials. First, one of the most established sets of biomarkers in CF is microbiological parameters relating to Pa [23]. Secondly, the FDA defines forced expiratory volume FEV_1 , the maximum amount of air expired in one second, and rate of decline as surrogate endpoints because they are well established predictors of survival. Thirdly, pulmonary exacerbation is a clinical efficacy measure for definitive

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

 TABLE 1

 Patient outcomes and biomarkers collected in regular study visits

clinical trials. 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 [19, 22], 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 [15]. 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.

3.2. 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 [18, 21, 34]. For non-homogeneous processes, the model is either reduced to the homogeneous case or fitted through piecewise constant transition intensities between different time points [21]. The time-homogeneous Markovian assumption is a working assumption for modeling, which does not need to be true for the proposed methods to work.

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

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

Note that the zero values of p_{13} , p_{31} and p_{32} reflect the disease progression nature in Figure 2, where patient will experience nonmucoid Pa (state 2) before progressing to irreversible stage mucoid Pa (state 3).

Based on longitudinal studies of *Pa* development [3, 14, 19, 22, 38, 44], $\eta_{ii}(t, \mathbf{Z}(t))$ is related to individual characteristics through the time-dependent covariates $\mathbf{Z}(t)$, 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. ΔF 508*H* indicates ΔF 508 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. Once nonmucoid Pa infection is detected, treatment aims to eradicate bacteria and change patient back to free of *Pa*. In order to exam the capacity of the reinforcement learning procedure to discover the optimal therapy in such disease dynamics, we simulate different treatment effect scenarios with time-varying efficacy and toxicity through parameter $\beta_{21_2}(t)$. Details will be provided in Section 3.3. Thirdly, the probability of progression to mucoid Pa infection, $p_{23}(t)$, is modeled through the residual probability $1 - p_{21}(t)$ of failing to eradicate Pa, which 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

$$\eta_{12}(t, \mathbf{Z}(t)) = \beta_{12_0} + \beta_{12_2}t + \beta_{12_3}\Delta F508H,$$

10

Patient Outcomes	Literature	Model Output
Time to first acquisition of nonmucoid Pa (yr)	1.0 (0.5 - 1.5) [22]	1.0 (0.5-2.5)†
Time to mucoid Pa (yr)	13.0(10.0-14.9)[22]	13.6 (9.8-17.5)†
Pa free duration after eradication (month)	8(3,25)[8],18(4,80)[44]	$9 (1.5, 39)^*$
$\Delta FEV_1 \cdot yr^{-1}$ standard care	$-4.69 \pm 2.95\%$ [44]	$-4.42 \pm 12.2\%$ ‡
$\Delta FEV_1 \cdot yr^{-1}$ some anti-Pa treatment	$-1.63 \pm 1.60\%$ [44]	$-1.5 \pm 8.1\%$ ‡
Sputum culture sensitivity	83% [39]	85%
Serology markers sensitivity	93% [31]	93%

TABLE 2									
Literature	and	model	generative	patient	outcomes				

[†] Median (95%CI), ^{*} Median (Range), [‡] Mean \pm SD

(3.1)
$$\eta_{21}(t, \mathbf{Z}(t)) = \beta_{21_0} + \beta_{21_1}t + \beta_{21_2}(t)Trt(t)^{Cul(t) \times Ser(t)} + \beta_{21_4}\Delta F 508H,$$

$$\eta_{23}(t, \mathbf{Z}(t)) = \beta_{23_0} + \beta_{23_1}t + \beta_{23_2}D_2(t) + \beta_{23_4}\Delta F508H.$$

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

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

(3.2)
$$p_{21}(t, t+1, \mathbf{Z}(t)) = \frac{exp(\eta_{21}(t, \mathbf{Z}(t)))}{1 + exp(\eta_{21}(t, \mathbf{Z}(t)))}$$

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

The formulation in (3.1) and (3.2) also accommodates an arbitrary number of treatment courses as well as options for either discrete or continuous time. Because we aim at optimizing the maintance therapy of Pa infection, the patient outcomes simulated include the observed Pa infection state, severity, and FEV_1 based on the underlying true state.

We tune the model so that when under standard care or anti-Pa antibiotic treatment, the descriptive statistics of patient outcomes are comparable to those in prior clinical studies [1, 8, 13, 15, 19, 22, 23, 31, 39, 46]. We list the important clinical outcomes in Table 2, including time to first acquisition and progression to mucoid, pulmonary function and sensitivity of culture and serology tests, etc. The age prevalences given in [22] and in simulations 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.



FIG 3. Age-specific prevalence of no, nonmucoid, and mucoid Pa from birth to age 16 years in the literature [22] and in 1000 simulations.

 $\begin{array}{c} {\rm TABLE \ 3} \\ {\it Efficacy \ and \ side-effects \ of \ treatment \ regimens} \end{array}$

	Mutation	Age Range	Population	Antibiotics	8	Intensity	Effect	Scenario
Immediate	$Non\Delta F508H$	I Age≤8	1.early	А		L	High	1
Efficacy			1.early	А		Н	High	2
U U			1.early	В		\mathbf{L}	Low	3
			1.early	В		Н	Low	4
		Age > 8	1.later	А		L	Low	5
			1.later	А		Η	Low	6
			1.later	В		L	Medium	7
			1.later	В		Н	High	8
	$\Delta F508H$	$Age \leq 8$	2.early	А		L	Low	9
		0 _	2.early	А		Η	Low	10
			2.early	В		\mathbf{L}	Medium	11
			2.early	В		Η	High	12
		Age>8	2.later	А		L	Low	13
			2.later	А		Н	Medium	14
			2.later	В		\mathbf{L}	Low	15
			2.later	В		Η	Low	16
			all	Standard of	of care	(SC)	Low	0
Delayed			all	No Off-drug Cycle			Susceptibility ↓	
Side-effects		all	Life-time $Exposure > 20$		> 20	Eradication \downarrow		
Abbreviatio	ns A-L A-H B-L B-H	antibiotics A in antibiotics A in antibiotics B in antibiotics B in	n low intensit n high intensi n low intensit n high intensi	optimal for 1.early,lower burden optimal for 1.early,2.later suboptimal for 1.later,2.early,lower burden optimal for 1.later,2.early				

3.3. Clinical Scenarios. We evaluate the design under the realistic clinical scenarios [12, 35, 37, 38, 46, 48–50] described in Table 3, where the treatment options consist of antibiotic (A or B) and intensity level(H or L) choice and standard of care for drug holiday. There are differential treatment efficacy and side effects in terms of probability of successful eradication for patients with different mutation class and age range. For simplicity, we denote the different population and age range by population 1.early, 1.later, 2.early and 2.later respectively. The different optimal treatment options in the four populations are presented and boxed in Table 3.

For the low risk population 1, both high and low intensity A treatments are preferable for the patient is under 8 years old (population 1.early); while high intensity B is best when the patient is older than 8 years old (population 1.later). For the high risk population 2, high intensity B is preferable when the patient is under 8 years old (population 2.early); while antibiotic A is best when the patient is older than 8 years old (population 2.later); the higher intensity level regimens are more successful for bacteria eradication. Patients who are Δ F508 homozygous are a high risk population, generally more severe, more easily acquire mucoidy and have greater difficulty eradicating Pa infection, hence, the treatments have lower efficacy compared to low risk population 1.

In the middle panel of Table 3, 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) [12, 15, 16, 35, 36]. Standard of care is the optimal treatment option in drug holiday to lower cumulative burden.

4. Clinical Reinforcement Trials. 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 [55, 56] and for SMART designs can be found in [25, 27, 47]. Based on the published results from previous CF trials [7], the CF neonatal screen project [14, 46], and a contemporary CF trial [48], we develop a virtual clinical reinforcement trial that provides a realistic approximation to a potential real CF trial.

4.1. Virtual CF Trial Conduct.

1. Learning stage trial design.

In a randomized trial in CF for patients 1–15 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 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. The conceptional overview is given in Figure 1.

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 insensitive 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 [5, 48]. The treatment S-C represents stand of care without targeted anti-Pa antibiotics, which could be applied in drug holiday to lower burden and avoid resistence development.

2. Learning stage rationale and goal.

The rationale of culture-based therapy is based on the clinical guidance for Pa infection in CF patients [5, 6, 12]. 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 level such as IV anti-Pa treatments are required [12]. 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

14

State Variables	Definition	Change	Reward R_t
Sputum/Serology	Pa + for one visit	Infected to free of Pa	0.9
Lung Function	Predicted FEV_1 change	$\downarrow \Delta \leq -10\%$	-0.1
0	between adjacent visits	$-10\% < \Delta \le 10\%$	0
		$\uparrow \Delta > 10\%$	0.1
Severity of Infection	mucoid Pa	progress to mucoid	-0.1
	non-mucoid Pa:free/intermittent/chronic	free to intermittent	0
	free infection: $0\%Pa$ + in 12 months	stay as intermittent	0
	intermittent infection: $<50\%Pa + in 12$ months	intermittent to chronic	0
	chronic infection: $\geq 50\% Pa + in 12$ months	chronic to intermittent	0.1

TABLE 4Reward/utility function setup

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 [20]. Specifically, Table 4 is our reward function for the simulation study of Section 5. 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: $Q_N^*(s, a)$ sequence in (2.3) is fitted by Q-iteration:
 - **repeat** at each iteration $k, k \ge 1$
 - 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_E, ζ) are selected by grid search over cost parameter $C_E = 2^{-5}, 2^{-3}, \ldots, 2^{15}$ and scale parameter $\zeta = 2^{-15}, 2^{-9}, \ldots, 2^3$ in 10-fold cross-validation.

- (d) Output: Personalized therapy $\hat{\pi}^{\star}(s) = \arg \max_{a} \hat{Q}_{N}^{\star}(s, a)$
- 5. Confirmatory stage design.

In a separate, potentially longer duration L_2 years trial, participants are only randomized at enrollment to either one of the four fixed therapies A-L, A-H, B-L, B-H or the new arm R-L with equal allocation N_2 pre arm 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 five fixed treatment options. The standard of care arm is not included at this stage due to ethical consideration.

4.2. Consistency of Estimating Optimal Therapy. Using the notation in Section 2.2 and Section 4.2 Step 4, we denote a_t , a_{t+1} as the decision at time points t and t + 1, respectively.

Moreover, we let $Q_N^{\star}(S_t, A_t)$ be the potential outcome as the Nth Q-function value of the sequence by fitted Q-iteration in (2.3), after time t and before t + 1. We also let $Q_{N-1}^{\star}(S_{t+1}, A_t, A_{t+1})$ be the potential outcomes from the N - 1th Q-function in fitted Q-iteration.

Additionally, S_t denotes the states at time t and $S_{t+1}(a_t)$ denotes the state at time t+1, after policy a_t and independent of any other previous action sequence (due to the Markov assumption). Under counterfactual framework, which assumes relations between observed and unobserved (factual and counterfactual) random variables to determine causality and to find optimal treatment strategy from correlations in longitudinal data, we need to maximize the value state function

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

Specifically, the optimal policy in (2.4) can be obtained via the Q-iteration algorithm:

(4.1)
$$\pi_{N-1}^{*}(\pi_{N}) = \arg \max_{a_{t+1}} E_{\pi_{N}, a_{t+1}} \left[Q_{N-1}^{*}(\pi_{N}, a_{t+1}) \middle| S_{t+1}(\pi_{N}) \right]$$

(4.2)
$$\pi_N^* = \arg \max_{a_t} E_{a_t} \left[R(S_t, a_t, S_{t+1}(a_t)) + \gamma E_{a_t, \pi_{N-1}^*} \left[Q_{N-1}^\star(a_t, \pi_{N-1}^*) | S_{t+1}(a_t) \right] \Big| S_t \right]$$

We now justify that the above functions of the potential outcomes in (4.1) and (4.2) can be estimated via the observed data under our sequential randomized designs. We assume stable unit treatment value assumptions (SUTVA), which implies that one patient's counterfactual outcomes do not depend on the treatment received by any other patients, i.e. "no interaction between subjects", Q_N^{\star} and Q_{N-1}^{\star} satisfy

(4.3)
$$Q_N^{\star}(S_t, A_t) = \sum_{a_t} Q_N^{\star}(S_t) I(A_t = a_t),$$

$$(4.4) Q_{N-1}^{\star}(S_{t+1}, A_t, A_{t+1}) = \sum_{a_t, a_{t+1}} Q_{N-1}^{\star}(S_{t+1}, a_t, a_{t+1}) I(A_t = a_t, A_{t+1} = a_{t+1}).$$

The sequential randomized assumption and (4.3) implies [25, 56] that

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

This justifies the fact that the function of the potential outcomes on the right hand side of (4.2) can be estimated via estimating $E[Q_{N-1}^{\star}|S_{t+1}, A_t, A_{t+1}]$. Similarly, based on (4.3) and (4.4), we have

$$E_{a_{t}} \left[R(S_{t}, a_{t}, S_{t+1}(a_{t})) + \gamma E_{a_{t}, \pi_{N-1}^{*}} \left[Q_{N-1}^{*}(a_{t}, \pi_{N-1}^{*}) | S_{t+1}(a_{t}) \right] | S_{t} \right]$$

$$= E_{a_{t}} \left[R(S_{t}, a_{t}, S_{t+1}(a_{t})) + \gamma \max_{a_{t+1}} E \left[Q_{N-1}^{*}(A_{t}, A_{t+1}) | S_{t+1}, A_{t} = a_{t}, A_{t+1} = a_{t+1} \right] | S_{t} \right]$$

$$= E \left[R(S_{t}, a_{t}, S_{t+1}(a_{t})) + \gamma \max_{a_{t+1}} E \left[Q_{N-1}^{*}(A_{t}, A_{t+1}) | S_{t+1}, A_{t} = a_{t}, A_{t+1} = a_{t+1} \right] | S_{t}, A_{t} = a_{t} \right]$$

$$= E \left[R(S_{t}, A_{t}, S_{t+1}) + \gamma \max_{a_{t+1}} E \left[Q_{N-1}^{*}(A_{t}, A_{t+1}) | S_{t+1}, A_{t} = a_{t}, A_{t+1} = a_{t+1} \right] | S_{t}, A_{t} = a_{t} \right].$$

Therefore, the function regarding the potential outcome on the right hand side of (4.2) can be estimated via estimating $E[Q_N^*|S_t, A_t]$.

5. Simulation Studies.

17

FIG 4. Boxplots of distribution of time to mucoid Pa and barplots of Pa infection state average frequencies over time using different therapies in a simulated trial with follow up until development of mucoid Pa. In the boxplots, the gray and dark green represent patients with Δ F508 homozygosity, otherwise the colors are blue and light green. In the barcharts, the green, red, gray colors represent patients in state 1 (free of Pa), state 2(nonmucoid Pa) and state 3(mucoid Pa) respectively; x-axis represents age and each bar corresponds to 3 months interval.



PERSONALIZED THERAPY FOR CYSTIC FIBROSIS

TABLE 5

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

Summary	By Non Δ F508H versus Δ F508H					Overall All						
Therapy	SOC	AL	AH	BL	BH	RL	SOC	AL	AH	BL	BH	RL
Scenarios	0	1&5	2&6	3&7	4&8	1 - 8	0	1&5	2&6	3&7	4&8	1 - 8
Population	1 2	1 2	1 2	1 2	1 2	1 2	1&2	1&2	1&2	1&2	1&2	1&2
Time to Mucoid $Pa(T_2)$ (Yr)												
Mean	11.7 12.0	15.6 13.8	15.7 12.9	14.0 15.0	13.0 15.5	21.4 17.8	11.8	14.7	14.3	14.5	14.3	19.6
SD	1.6 1.5	1.4 2.6	1.3 1.9	2.2 1.4	2.3 1.4	2.8 2.6	1.4	2.3	2.1	1.9	2.2	3.3
Min	9.1 9.3	11.0 8.8	14.3 8.8	10.8 10.0	9.8 12.5	15.5 14.3	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	16.0 15.5	19.5 18.8	19.5 18.8	19.8 19.5	20.8 20.3	26.3 25.3	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
\downarrow 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

5.1. Simulation Results. We generate a virtual CF trial based on the disease model with treatment effect scenarios in Section 3. The conduct of the clinical reinforcement trial follows the procedure proposed in Section 4.1 with total sample sizes $N_1 = 1000$, $N_2 = 200$ pre arm 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 4 years of follow up, we examine the procedure in the scenario where we can apply the therapies 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 $\gamma = 0.5$, which implies that the impact of the immediate reward will be less than 10% after 12 months and less than 1% after 2 years. We explore the effect of γ on the performance of discovered therapy *R-L*. The larger γ , the less sensitive R-L therapy to capture the immediate treatment efficacy pattern and, as expected, the slower in convergence.

5.1.1. Results in testing scenario I: from birth until a mucoid Pa. To evaluate the empirical performances of fixed treatment regimens S-C,

19

A-L, A-H, B-L, B-H and the adaptive personalized therapy denoted R-L, we follow up 200 virtual patients in each arm with one half being Δ F508 homozygous until their development of mucoid Pa infection (Figure 4 and Table 5).

Time to mucoid Pa infection of Δ F508 homozygosity A-L, A-H, B-L, B-H, R-L treated patients, calculated from the date of birth, has longer median than that of patients treated by S-C as shown in Figure 4.B. It is potentially complicated to compare the outcomes among four fixed anti-Pa antibiotics regimens A-L, A-H, B-L, B-H. The time-varying and subpopulation specific treatment effect in Table 3 might remain undetected. For example, drug A's benefit to population 1.early is masked and "averaged out" by outcomes for patients in population 2 and population 1.later.

We therefore examine the outcomes among patients treated by the identified personalized therapy R-L, where we optimize the usage of the existing drugs A, B and standard of care. The R-L achieves superior patient outcomes to any other fixed treatment regimens even in the mixture of the two subpopulations 1&2 (Figure 4.A, .B). The superior treatment benefits of these drugs might be missed in a traditional, single-decision point clinical trial.

Other endpoints, the observed frequency of the three states (Figure 4.C), nonmucoid Pa infection proportion, predicted $FEV_1\%$ and change per year, all demonstrate the same patient outcome patterns (Table 5). For simplicity, Table 5 presents the outcomes in two formats, two subpopulations side by side and overall, where one can see the performances of different treatment options in subpopulations and general population, which are consistent with the treatment effect scenarios (Table 3). In short, the right drug for the subpopulation chosen in early childhood improves prognosis and the high risk population 2 requires higher intensity level treatment to eradicate Pa infection.

We next illustrate the discovered therapies R-L for two individual patients who are in population 1 (Figure 5.A,B). When a patient is under 8 years old, the right antibiotic A is often chosen at the effective and lower intensity level. After 8 years old, the right antibiotic B is chosen more frequently and with higher intensity level. However, the discovered therapies for two individual patients who are in population 2(Figure 5.C,D) choose the right antibiotic B initially, and automatically switches to the more suitable antibiotic A at the correct age. In this more severe population 2, high intensity level is chosen more often than low intensity level. For both subpopulations 1&2, the alternating



FIG 5. 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.



FIG 6. Kaplan-Meier plot of time to mucoid Pa infection using different therapies in a simulated trial with 4 years of follow up.

patterns to switch the drug or lower the intensity level, are achieved in order to avoid resistance development, regain susceptibility and lower the cumulative toxicity burden.

The discovered adaptive personalized therapy by the reinforcement learning procedure outperforms any fixed treatment regimen therapies because it considers the time varying treatment effect on different age specific groups and balances the trade-offs 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 does not require the generative treatment model, and thus the proposed method is able to discover an optimal regimen without prior knowledge of the treatment mechanism.

5.1.2. Results in testing scenario II: separate confirmatory trial. In the second follow up scenarios, we simulated trials with study durations of 4 years. Figure 6 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 6, 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. Discussion. 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.

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 patients 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.

There are a number of additional topics to work on and challenges we expect to address in future research. First, 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, for example, the Wisconsin CF neonatal screening project [11, 22] 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. Thirdly, refining the proposed clinical reinforcement learning trial will require close collaboration with clinical researchers to improve the practical, logistic aspects, and for actual implementation.

Acknowledgements. The authors would like to thank Dr. George Retsch-Bogart and Dr. Richard Boucher for helpful discussions on cystic fibrosis research and the Reinforcement Learning Group at the University of North Carolina for many stimulating exchanges. The research was funded in part by the Biometric Consulting Laboratory at the University of North Carolina and NIH grant RR025747 which funds the North Carolina Translational and Clinical Sciences Institute.

References.

- ARMSTRONG, D. S., GRIMWOOD, K., CARLIN, J. B., CARZINO, R., OLIN-SKY, 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.
- BELLMAN, R. E. (1957). Dynamic programming, Princeton University Press, Princeton.
- BURNS, J. L., GIBSON, R. L., MCNAMARA, S., YIM, D., EMERSON, J., ROSEN-FELD, M., HIATT., P., MCCOY, K., CASTILE, R., SMITH, A. L. and RAMSEY, B. W. (2001). Longitudinal assessment of Pseudomonas aeruginosa in young children with cystic fibrosis. *Journal of Infectious Diseases* 183 444–452.

- [4] CYSTIC FIBROSIS FOUNDATION. (2008). Patient Registry 2008 annual data report. Cystic Fibrosis Foundation.
- [5] DÖRING, G., CONWAY, S. P., HEIJERMAN, H. G., HODSON, M. E., HØIBY, N., SMYTH, A. and TOUW, D. J. for the Consensus Committee (2000). Antibiotic therapy again Pseudomonas aeruginosa in cystic fibrosis: a European consensus. *Eur. Respir. J.* **16** 749–767.
- [6] DÖRING, G. and HØIBY, N. for the Consensus Study Group (2004). Early intervention and prevention of lung disease in cystic fibrosis: a European consensus. *Journal of Cystic Fibrosis* 3 67–91.
- [7] DÖRING, G., ELBORN, J. S., JOHANNESSON, M., DE JONGE H., GRIESE, M., SMYTH, A. and HEIJERMAN, H. for the Consensus Study Group (2007). Clinical trials in cystic fibrosis. *Journal of Cystic Fibrosis* 6 85–99.
- [8] 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.
- [9] ERNST, D., GEURTS, P. and WEHENKEL, L. (2005). Tree-based batch model reinforcement learning. J. Mach. Learn. Res. 6 503–556.
- [10] ERNST, D., STAN, G. B., GONCALVE, J. and 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* 667–672.
- [11] FARRELL, P. M., LI, Z., KOSOROK, M. R., LAXOVA, A., GREEN, C., G., COLLINS, J., LAI, H.C., MAKHOLM, L. M., ROCK, M. J. and SPLAINGARD, M. L. (2003). Longitudinal evaluation of bronchopulmonary disease in children with cystic fibrosis. *Pediatric Pulmonology* **36** 230–240.
- [12] FLUME, P. A., O'SULLIVAN, B. P., ROBINSON, K. A., GOSS, C. H., MO-GAYZEL, 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.
- [13] GEURTS, P., ERNST, D. and WEHENKEL, L. (2006). Extremely randomized trees. *Machine Learning* 11 3–42.
- [14] GIBSON, R. L., BURNS J. L. and RAMSEY, B. W. (2003). Pathophysiology and management of pulmonary infections in cystic fibrosis. Am. J. Respir. Crit. Care Med. 168 918–951.
- [15] GIBSON, R. L., EMERSON, J., MCNAMARA, S., BURNS, J. L., ROSENFELD, M., YUNKER, A., HAMBLETT, N., ACCURSO, F., DOVEY, M., HIATT, P., KON-STAN, M. W., MOSS, R., RETSCH-BOGART, G., WAGENER, J., WALTZ, D., WILMOTT, R., ZEITLIN, P. L. and RAMSEY, B. W.; Cystic Fibrosis Therapeutics Development Network Study Group. (2003b). Significant microbiological effect of inhaled tobramycin in young children with cystic fibrosis. Am. J. Respir. Crit. Care Med. 167 841–849.
- [16] GUEZ, A., VINCENT, R. D., AVOLI, M. and PINEAU, J. (2008). Adaptive Treatment of Epilepsy via Batch-mode Reinforcement Learning. AAAI 1671– 1678.
- [17] KAELBLING, L. P., LITTMAN, M. L. and MOORE, A. W. (1996). Reinforcement Learning: A Survey. Journal of Artificial Intelligence Research 4 237–285.
- [18] 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.

- [19] KOSOROK, M. R., ZENG, L., WEST, S. E., ROCK, M. J., SPLAINGARD, M. L., LAXOVA, A., GREEN, C. G., COLLINS, J. and FARRELL, P. M. (2001). Acceleration of lung disease in children with cystic fibrosis after Pseudomonas aeruginosa acquisition. *Pediatr. Pulmonol.* **32** 277–287.
- [20] LANGTON HEWER, S. C. and SMYTH, A. R. (2009). Antibiotic strategies for eradicating Pseudomonas aeruginosa in people with cystic fibrosis. *Cochrane Database Syst. Rev.*4 CD004197.
- [21] LMEIRA-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.
- [22] LI, Z., KOSOROK, M. R., FARRELL, P. M., LAXOVA, A., WEST, S. E. H., GREEN, C. G., COLLINS, J., JOCK, M. J. and SPLAINGARD, M. L. (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.
- [23] 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.
- [24] 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.
- [25] MURPHY, S. A. (2005). An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine* 24 1455–1481.
- [26] MURPHY, S. A. (2005). A generalization error for Q-learning. J. Mach. Learn. Res. 6 1073–1097.
- [27] 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.
- [28] 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.
- [29] FLUME, P. A., O'SULLIVAN, B. P., ROBINSON, K. A., GOSS, C. H., MO-GAYZEL P. J., WILLEY-COURAND, D. B., BUJAN, J., FINDER, J., LESTER, M., QUITTELL, L., ROSENBLATT, R., VENDER, R. L., HAZLE, L., SABADOSA, K. and MARSHALL, B. C.; Clinical Practice Guidelines for Pulmonary Therapies Committee. (2007). Cystic fibrosis pulmonary guidelines: chronic medications for maintenance of lung health. Am. J. Respir. Crit. Care Med. **176** 957–969.
- [30] FLUME, P. A., MOGAYZEL, P. J., ROBINSON, K. A., GOSS, C. H., ROSEN-BLATT 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.
- [31] 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.
- [32] 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.
- [33] PRINCE, A. S. (2002). Biofilms, antimicrobial resistance, and airway infection.

N. Engl. J. Med.347 1110–1111.

- [34] PUTTER, H., FIOCCO, M. and GESKUS, R. B. (2007). Tutorial in biostatistics: Competing risks and multi-state models. *Statistics in Medicine* 26 2389–2430.
- [35] 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.
- [36] RATJEN, F., DÖRING, G. and NIKOLAIZIK, W. H. (2001). Effect of inhaled tobramycin on early Pseudomonas aeruginosa colonization in patients with cystic fibrosis. *Lancent* 358 983–984.
- [37] RETSCH-BOGART, G. Z. (2009). Update on new pulmonary therapies. Current Opinion in Pulmonary Medicine 15 604–610.
- [38] ROSENFELD, M., GIBSON, R. L., MCNAMARA, S., EMERSON, J., BURNS, J. L., CASTILE, HIATT, P., MCCOY, K., WILSON, C. B., INGLIS, A., SMITH, A., MARTIN, T.R. and RAMSEY, B.W. (2001). Early pulmonary infection, inflammation, and clinical outcomes in infants with cystic fibrosis. *Pediatr. Pulmonol.* 32 356–366.
- [39] 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.
- [40] ROBINS, J. M. (2004). Optimal structual nested models for optimal sequential decisions. *Proceedings of the Second Seattle Symposium on Biostatistics*, Springer, New York, 189–326.
- [41] ROWE, S. M., MILLER, S. and SORSCHER, E. J. (2005). Cystic Fibrosis mechanisms of disease. N. Engl. J. Med. 355 2408–2417.
- [42] RYAN, G., MUKHOPADHYAY, S. and SINGH, M. (2000). Nebulised antipseudomonal antibiotics for cystic fibrosis. *Cochrane Database Syst. Rev.* 2 CD001021.
- [43] SOUTHERN, K. W., MARIEKE, M. E., DANKERT-ROELSE, J. E. and NAGELK-ERKE, A. (2009). Newborn screening for cystic fibrosis. *Cochrane Database Syst. Rev.*1 CD001402.
- [44] STARNER, T. D. and 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.
- [45] SUTTON, R. S. and BARTO, A. G. (1998). Reinforcement learning: an introduction, MIT Press, Cambridge, MA.
- [46] TACCETTI, CAMPANA1, S., FESTINI1, F., MASCHERINI, M. and DÖRING, G. (2005). Early eradication therapy against Pseudomonas aeruginosa in cystic fibrosis patient. *Eur. Respir. J.* 26 458–461.
- [47] THALL, P. F., SUNG, H. G. and ESTEY, E. H. (2002). Selecting therapeutic strategies based on efficacy and death in multicourse clinical trials. *Journal of* the American Statistical Association 457 29–39.
- [48] TREGGIARI, M. M., ROSENFELD, M., RETSCH-BOGART, GIBSON, R. L., WILLIAMS, J., EMERSON, J., KRONMAL, R. A. and RAMSEY, B. W. EPIC Study Group (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.
- [49] TREGGIARI, M. M., ROSENFELD, M., RETSCH-BOGART, G., GIBSON, R. L.

and RAMSEY, B. W. (2007). Approach to Eradication of Initial Pseudomonas aeruginosa Infection in Children With Cystic Fibrosis. *Pediatr. Pulmonol.* **42(9)** 751–756.

- [50] VALERIUS, N. H., KOCH, C. and HØIBY, N. (1991). Prevention of chronic Pseudomonas aeruginosa colonisation in cystic fibrosis by early treatment. *Lancent* 338 725–726.
- [51] VAPNIK, V. (1995). The nature of statistical learning theory, Springer, New York.
- [52] 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.
- [53] WATKINS, C. J. C. H. (1989). Learning from delayed rewards. *Ph.D. Thesis*, King's College, Cambridge, UK.
- [54] WATERS, V. and RATJEN F. (2008). Combination antimicrobial susceptibility testing for acute exacerbations in chronic infection of Pseudomonas aeruginosa in cystic fibrosis. *Cochrane Database Syst. Rev.* 3 CD006961.
- [55] ZHAO, Y., KOSOROK, M. R. and ZENG, D. (2009). Reinforcement learning design for cancer clinical trials. *Statistics in Medicine* 28 3294–3315.
- [56] 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).