Personalized Medicine and Clinical Trials

Michael R. Kosorok

University of North Carolina at Chapel Hill

Collaborators: Bert O'Neil, Mark Socinski, Yiyun Tang, Jen Jen Yeh, Donglin Zeng and Yufan Zhao, University of North Carolina at Chapel Hill

Introduction

Discovering effective therapeutic regimens for life-threatening diseases is a central goal of medical research.

The prevailing approach:

- develop candidate therapies in the laboratory,
- test in animals,
- conduct human clinical trials.

In fact,

- Very few candidate treatments make it to human clinical trials,
- only about 10% (more like 5%) of treatments making it to human clinical trials demonstrate enough efficacy to be approved for marketing.

(Hogberg, 2005, *Drug Discovery Today;* FDA, 2004, White Paper.)

We will now consider an example in cancer and one in cystic fibrosis.

Non-Small Cell Lung Cancer

In typical regimens for advanced cancer (in breast, lung, and ovarian) patients utilize

- a single agent
- in combination with a platinum-based compound
- in multiple stages (lines) of treatment.

In non-small cell lung cancer (NSCLC), 2–3 lines of treatment increases survival.

Can we improve survival by personalizing the treatment at each decision point (at the beginning of a treatment line) based on prognostic data?

Cystic Fibrosis Example

A major challenge in patients with cystic fibrosis (CF) is lung infections caused by *Pseudomonas aeruginosa* (Pa):

- Young CF patients acquire Pa off and on, but eventually the infection does not clear up.
- After several years of chronic infection, Pa infections can transform to a severe mucoid variant, leading to death or lung transplant.
- Delaying the onset of the mucoid variant is a primary goal in CF care.

Is it possible to improve on existing treatment methods by using a personalized treatment rule—to be applied at each time a Pa infection is detected—based on prognostic data?

Drug Scheduling and Adaptive Designs

These issues have motivated a vast literature on drug-scheduling strategies (especially in cancer).

This has been accompanied by extensive research on adaptive design in clinical trials:

- Multi-course clinical trials using a play-the-winner-and-drop-the-loser strategy (Thall, et al., 2000).
- Bayesian adaptive designs, such as I-SPY 2 and BATTLE (Barker, et al., 2009; Berry, 2006; Ledford, 2010; Thall, et al., 2007).

The common goal of these adaptive designs is to identify biomarkers that predict efficacy of individual drugs.

As a result, these trials are not designed to nor are able to arrive at comprehensive drug treatment rules that simultaneously involve multiple biomarkers and multiple drugs.

Thus

- comprehensive personalized therapy rules cannot be generated,
- nor can dynamic treatment strategies be discovered.

Dynamic Treatment Regimes

In contrast with classic adaptive designs, “dynamic treatment regimes” (or “adaptive treatment strategies”) (Murphy, 2005) can allow treatment to vary with time based on individual prognostic data.

Dynamic treatment regime designs are able to provide information

- not only on the best treatment choice from the beginning
- but also on treatment choices that maximize outcomes at each new decision time, taking into account long-term affects.

Dynamic treatment regimes

- are a new paradigm for treatment and long term management of chronic disease and drug and alcohol dependency, and
- have been incorporated into new trial designs such as sequential multiple assignment randomized trials (SMART) (See Murphy, 2005, *Statistics in Medicine*, and 2007, *Drug and Alcohol Dependence*).

However, there are no clinical trial methodologies for discovering personalized treatment regimens for diseases like cancer or cystic fibrosis which have points of irreversibility.

Reinforcement Learning

In this talk, we present a general reinforcement learning framework and related statistical and computational methods for clinical research.

In previous work,

- Reinforcement learning has been applied to behavioral disorders, where each patient typically has multiple opportunities to try different treatments (Pineau, 2007, *Drug and Alcohol Dependence*).
- Murphy et al. (2007, *Neuropsychopharmacology*) suggest Q-learning, an important breakthrough in reinforcement learning, for constructing decision rules in chronic psychiatric disorders.

Clinical Reinforcement Trials

Reinforcement learning has not yet been applied to potentially irreversible diseases like cancer.

We propose a variation/adaptation of the SMART concept, “clinical reinforcement trials,” wherein:

- Each patient is randomized at each decision time to a possibly continuous range of treatment possibilities (drug, dose, timing, etc.).
- At the end of the first stage, reinforcement learning is used to estimate optimal treatment per prognostic values at each decision time.
- It is not necessary that any single patient receive the optimal therapy.
- The first stage is followed by a second, confirmatory stage.

These new designs have two special features:

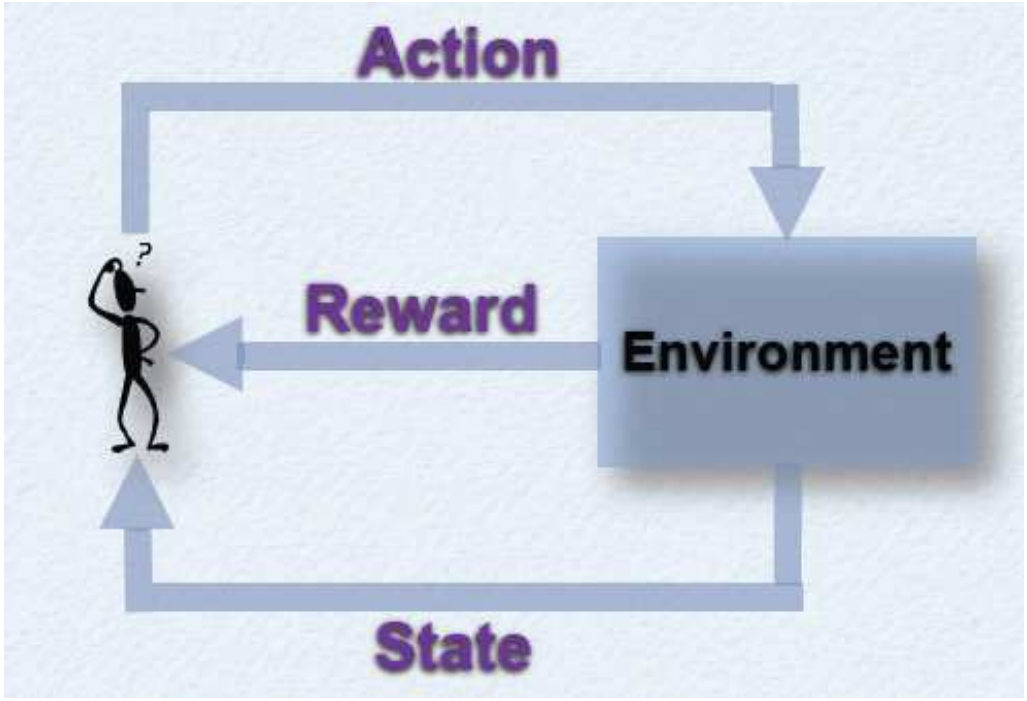
- First, without relying on pre-specified mathematical models, reinforcement learning
  - carries out treatment selection sequentially
  - with time-dependent outcomes
  - to determine which of several possible next treatments is best for which patients at each decision time and considering all future possibilities.
- Second, the proposed approach improves longer-term outcomes by considering delayed effects.

Clinical reinforcement trials can extract the optimal treatment regimen while taking into account a drug's efficacy and toxicity simultaneously.

Reinforcement Learning (Continued)

The basic process of reinforcement learning involves

- trying a sequence of actions,
- recording the consequences of those actions,
- statistically estimating the relationship between actions and consequences, and then choosing the action that results in the most desirable consequence.


Action
Reward
Environment
State

We use random variables S and A to denote, respectively, patient "state" (prognostic values and treatment history) and "actions" (treatment choice).

Specifically, define time-dependent variables $S_t = \{S_0, S_1, \ldots, S_t\}$ and $A_t = \{A_0, A_1, \ldots, A_t\}$, with lower case used for realizations.

After treatment at time step t, $t = 0, 1, \ldots, T$, the patient receives an incremental reward: $r_t = R(\mathbf{s}_t, \mathbf{a}_t, s_{t+1})$.

An important development in reinforcement learning which we will employ is Watkins' Q-learning (Watkins, 1989, 1992), which involves estimating the Q-function.

The Q-function at time t, $Q_t(s_t, a_t)$, is

- the expected total (possibly discounted) of all future incremental rewards given state s_t and action a_t at time t
- and also given that the optimal action is always applied at all future times.

One-step Q-learning is a regression on S_t and A_t with the simple recursive form

$$Q_t(S_t, A_t) \quad \leftarrow \quad r_t + \gamma \max_{a_{t+1}} Q_{t+1}(S_{t+1}, a_{t+1}), \qquad (1)$$

where $\gamma \in [0, 1]$ is the discount factor.

According to the recursive form in (1), we must estimate $\widehat{Q}_T$ from the last time point back to $\widehat{Q}_0$ at the beginning of the trajectories, where $\hat{Q}_{T+1}$ is set to zero for convenience.

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

$$\widehat{\pi}_t = arg \max_{a_t} \widehat{Q}_t(\mathbf{s}_t, \mathbf{a}_t; \theta_t),$$

where $t = 0, 1, \ldots, T$.

We thereafter use these optimal policies to evaluate in a phase III confirmatory trial.

Challenges may arise due to the complexity of the true Q-function, including:

- the non-smooth maximization operation in equation (1),
- the high-dimension of the state and action variables S and A, or
- the continuity of the action variable.

We utilize two recent nonparametric techniques from machine learning:

- support vector regression (SVR) (Vapnik, et al., 1997) and
- extremely randomized trees (ERT) (Ernst, et al., 2005; Geurts, et al., 2006).

Clinical Reinforcement Trials (Continued)

The proposed "clinical reinforcement trials" consist of three aspects:

1. A small, finite set of decision times is identified which is appropriate for the treatment process under investigation.
2. For each decision time, a set of possible treatments to be randomized is identified.
3. An incremental utility or reward function is identified which may be random.

After the optimal personalized treatment function is estimated, it is important to run a randomized, confirmatory phase III clinical trial to compare to standard of care.

Once the design has been determined:

- patients are recruited into the study and randomized to the treatment set under the protocol restrictions at each decision point;
- outcome measures used to compute patient state and incremental utility are obtained;
- each patient is followed through to completion of the protocol or until the end of the trial;
- patient data is collected and Q-learning is applied to estimate the optimal treatment rule as a function of patient variables and biomarkers, at each decision time; and
- a confirmatory phase III trial is conducted.

Three Examples

We will now present three examples, each of which involve:

1. The clinical setting.
2. The parameters of a clinical reinforcement trial.
3. Mathematical modeling of the disease process and simulated trial.
4. Simulation studies with both a clinical reinforcement trial and a phase III confirmatory trial.
5. Simulation results and conclusions.

Example 1: A Generic Cancer Setting (ZKZ)

The clinical setting:

- Artificial and not based on a specific disease.
- Cancer patients are treated monthly and followed for 6 months.
- At the beginning of each month, wellness and tumor size are assessed, and a dose of a chemotherapy agent (in the interval $[0, 1]$) is administered.
- The reward function considers (negative) wellness, tumor size and survival status.
- The goal is to find the best rule at each of months $0, 1, \ldots, 5$ which gives the best overall reward.

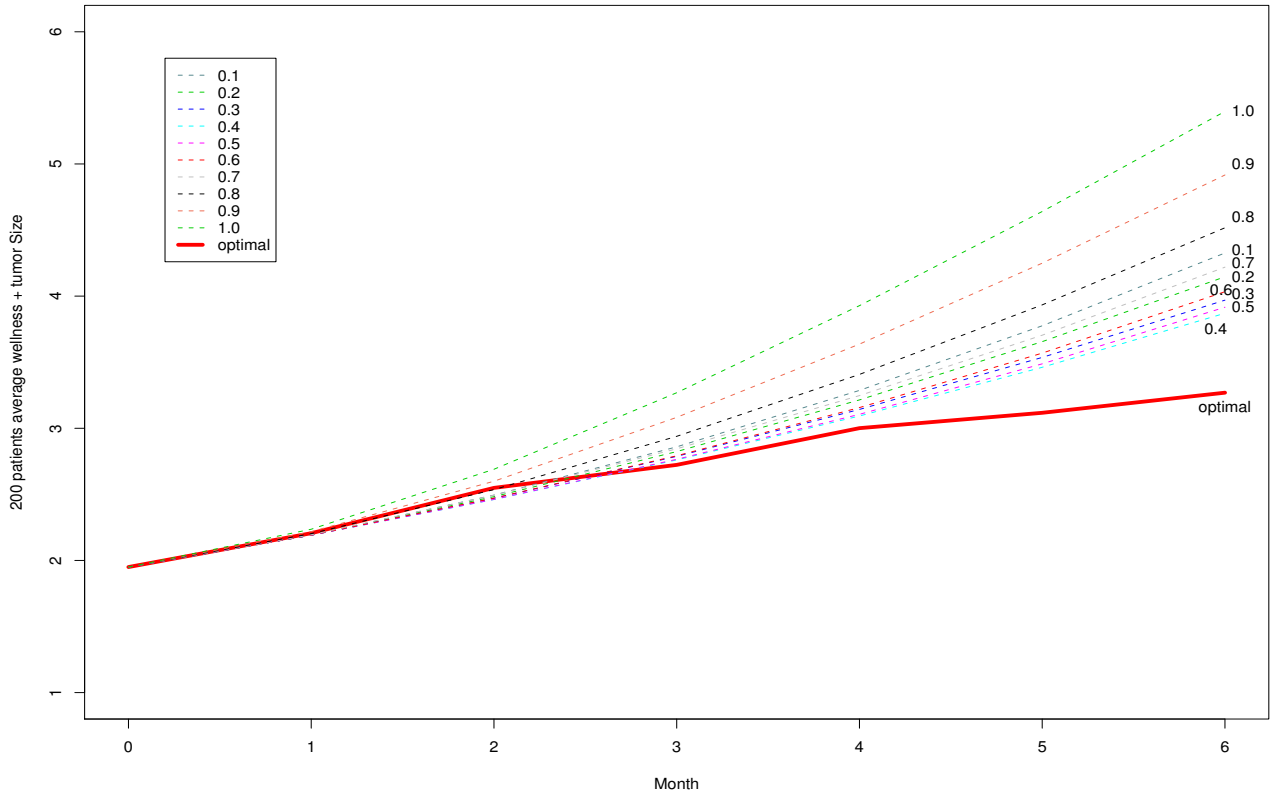
The parameters of the clinical reinforcement trial:

- 1000 patients are randomized to doses on the grid $0.0, 0.1, \ldots, 1.0$, at times (months) $0, 1, \ldots, 5$, and followed up through time $t = 6$.
- At time $t = 0$, the dose is restricted to $[0.5, 1]$ to ensure that each patient gets at least some drug initially.
- At the end of each month, the reward function is calculated based on negative patient wellness (W) and tumor size (M), change in W and M, and survival status.
- The data is collected and Q-learning (SVR and ERT) is applied to obtain optimal treatment rules as functions of status.
- A phase III clinical trial is simulated to compare the optimal treatment to each of the fixed dose treatments from the grid with 200 patients per group.

Mathematical modeling of disease process:

- The initial values of W and M are drawn from a uniform distribution.
- A simple difference equation is used to model monthly change in W and M as a function of the previous values of W and M and dose level.
- The probability of survival in a monthly interval is determined by a constant hazard function which is an exponential function of $W + M$.
- There exists an optimal solution to how to determine dose, but this is quite hard to find, and the Q-learning algorithm does not “know” the solution in advance.

Figure 1. Simulation results: Plots of average sickness (negative wellness plus tumor size) for 10 different constant-dose regimens (dashed lines) compared to the optimal regimen (solid line). The mortality rate for optimal treatment was 0.44 and 0.38 for best constant dose.



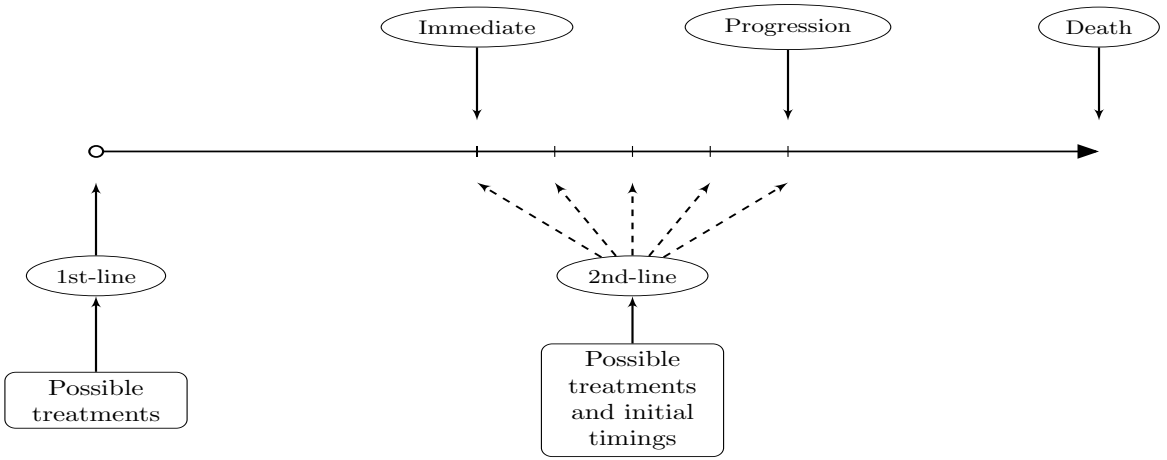
Simulation summary:

- The optimal therapy performs better in terms of the reward criteria $W + M$ as well as for overall survival.
- SVR and ERT yield very similar results, but ERT is generally more computationally intense.
- The sample size of 1000 for the clinical reinforcement trials seems to be adequate.
- This size is of the same order of magnitude as that for many phase III trials.

Example 2: NSCLC (ZZSK)

The clinical setting:

- There are two to three lines of therapy, but very few utilize three, and we will focus on two here.
- We need to make decisions at two treatment times: (1) at the beginning of the first line and (2) at the end of the first line.
- For time (1), we need to decide which of several agent options is best: we will only consider two options in the simulation.
- For time (2), we need to decide when to start the second line (out of three choices for simplicity) and which of two agents to assign.
- The reward function is overall survival which is right-censored.


Immediate
Progression
Death
1st-line
2nd-line
Possible treatments
Possible treatments and initial timings

The clinical outcome is total survival time T_D:

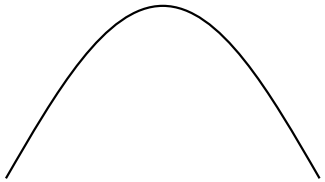
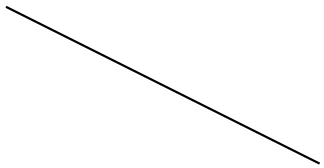
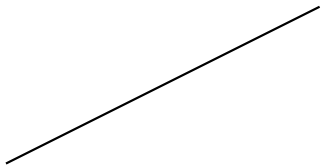
- The incremental reward for time (1) is the total time lived between start and stop of first line therapy, T_1.
- The incremental reward for time (2) is the total time lived after the end of first line therapy, T_2.
- The total overall reward is $T_D = T_1 + T_2$ (i.e., no discount) truncated at the maximum follow-up time τ.

Parameters of the clinical reinforcement trial:

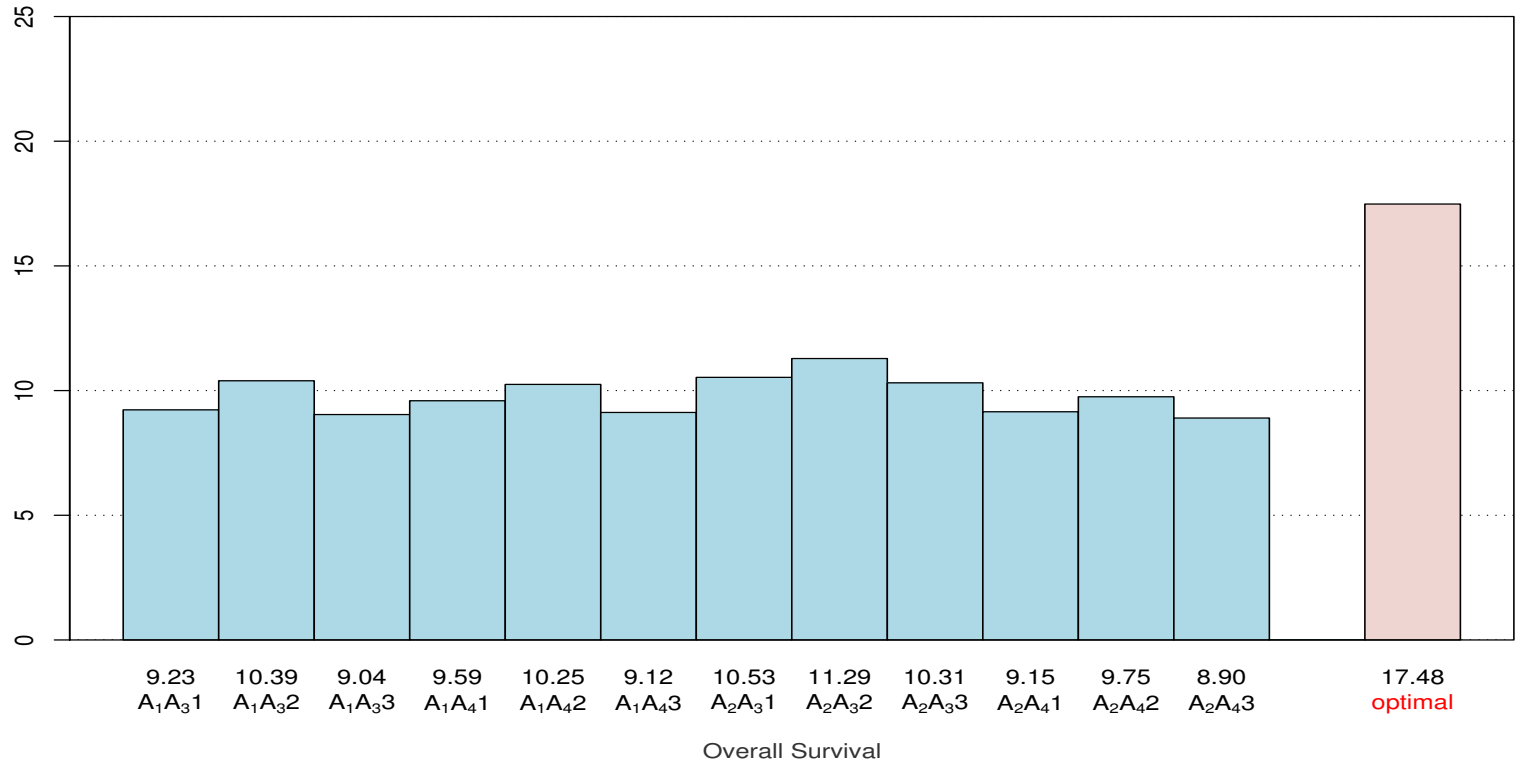
- 400 patients total: 100 each of 4 kinds of patients with different prognostic and treatment response relationships.
- Agents and start times are randomized at each decision time.
- Q-learning is used to estimate optimal personalized treatment regimens after clinical reinforcement trial completion.
- A confirmatory phase III trial is conducted with 1300 patients, 100 assigned to the optimal regimen and 100 to each of the 12 possible fixed treatments.

Mathematical modeling of disease process:

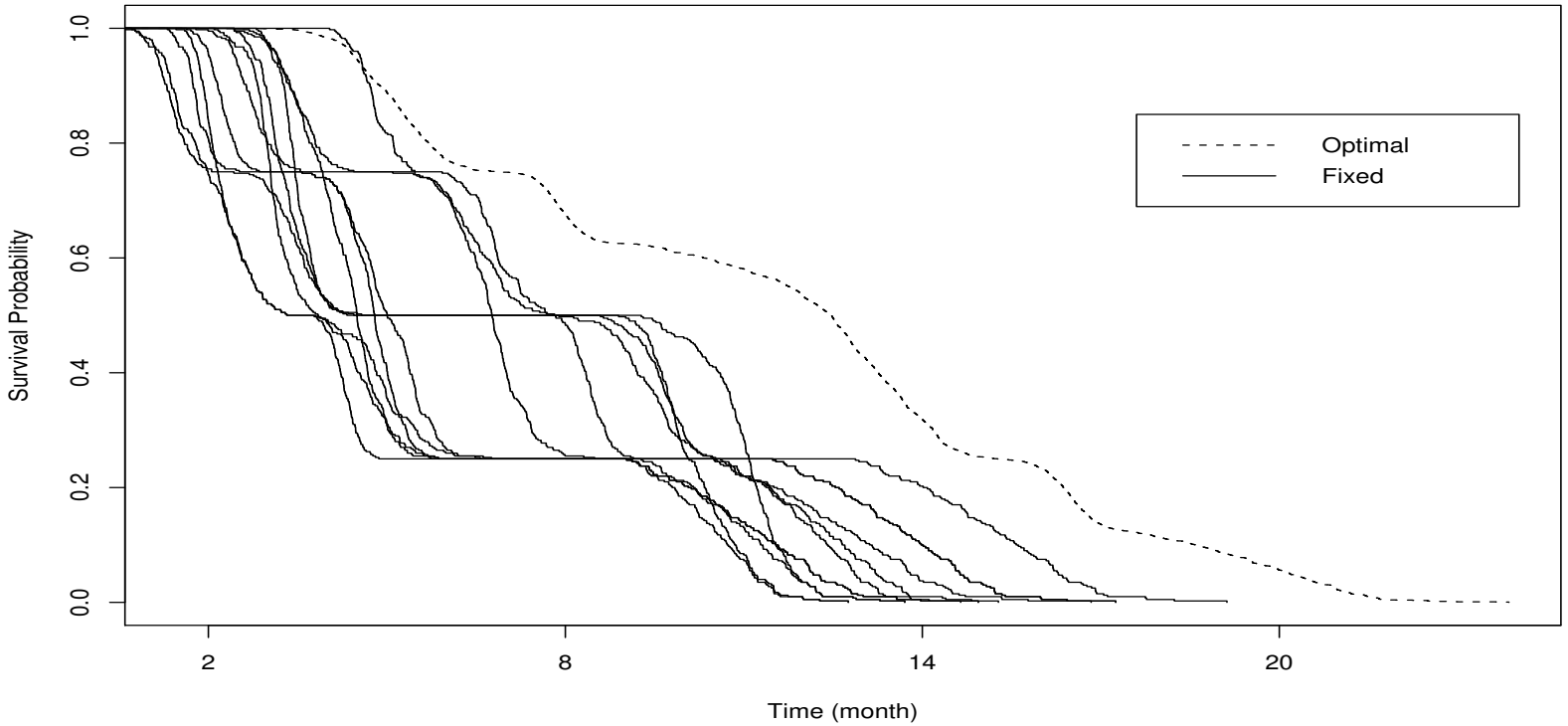
- Patients selected (randomly) from four basic treatment profiles.
- The next table gives the general features of the four corresponding models in terms of shape of optimal timing and treatment.
- W is wellness (quality of life) and M is stage of metastasis, similar to the previous example except for the sign of W.
- Reward is overall survival.
- Several levels of right censoring are considered.

Group	State Variables Status		Timing (h)	Optimal Regimen
1	$W_1 \sim N(0.25, \sigma^2)$ $M_1 \sim N(0.75, \sigma^2)$	$W_1 \downarrow M_1 \uparrow$		$A_1 A_3 2$
2	$W_1 \sim N(0.75, \sigma^2)$ $M_1 \sim N(0.75, \sigma^2)$	$W_1 \uparrow M_1 \uparrow$		$A_1 A_4 1$
3	$W_1 \sim N(0.25, \sigma^2)$ $M_1 \sim N(0.25, \sigma^2)$	$W_1 \downarrow M_1 \downarrow$		$A_2 A_3 3$
4	$W_1 \sim N(0.75, \sigma^2)$ $M_1 \sim N(0.25, \sigma^2)$	$W_1 \uparrow M_1 \downarrow$		$A_2 A_4 2$

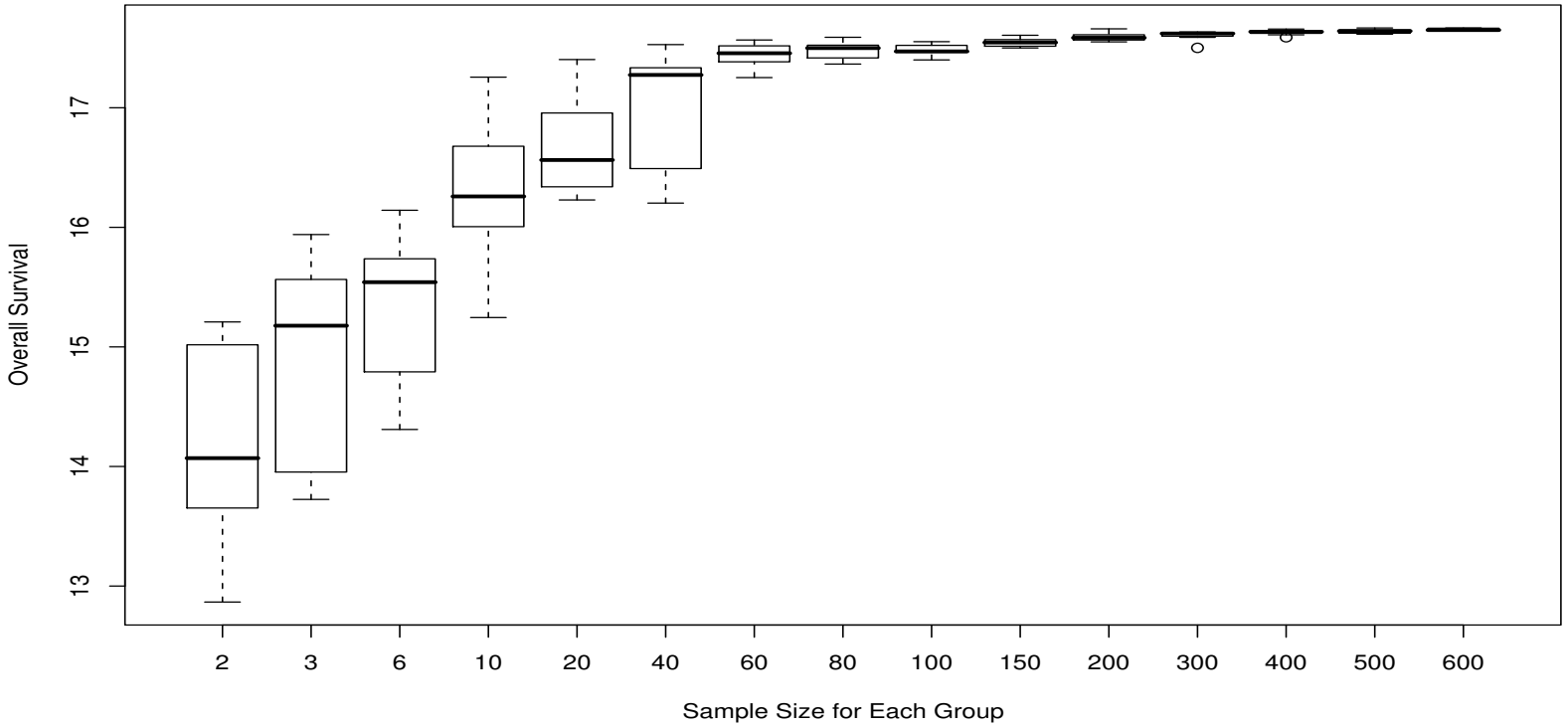
Performance of optimal personalized regimen versus the 12 fixed combinations under no censoring.



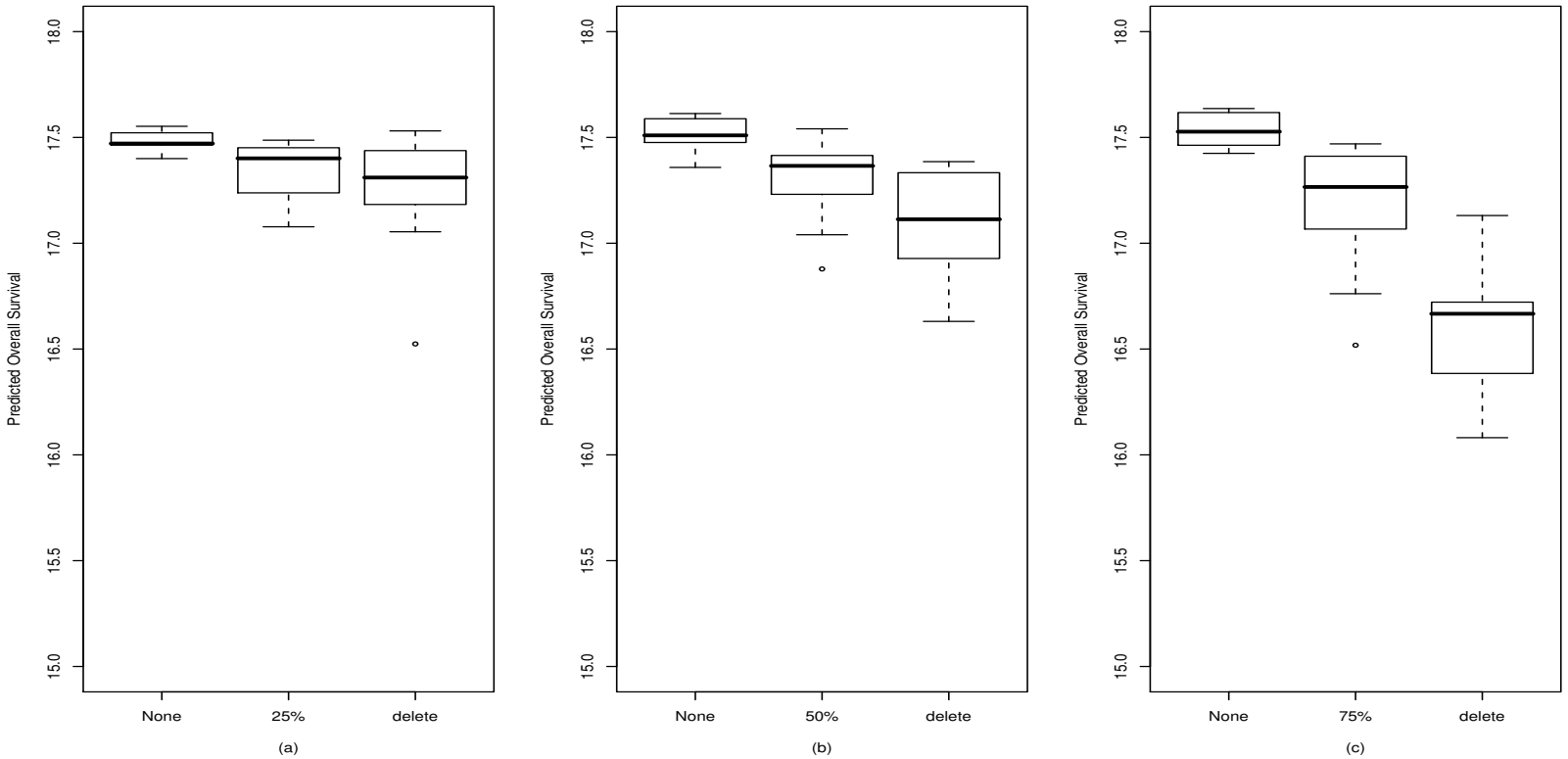
Survival functions for 13 different regimens (12 fixed plus 1 optimal regimen).



Predicted optimal treatment survival probability versus trial sample size per group.



Boxplots of the predicted survival for the optimal regimen estimated from ϵ-SVR-C by using training data with 25% (a), 50% (b), and 75% (c) censoring.



Simulation summary:

- The simulations demonstrate that the proposed approach is able to find personalized treatment regimes that are significantly better than the best of the available fixed regimes.
- The proposed ϵ-SVR-C seems to work reasonably well at finding optimal treatment regimens in the presence of censoring, but it appears that additional improvement could be helpful.
- The sample sizes needed for finding the optimal treatment seem to be not larger than 100 per group when no censoring is present: it is unclear how censoring affects this result.

Example 3: Cystic Fibrosis

The clinical setting:

- Cystic fibrosis (CF) is a homozygous recessive genetic disorder that affects predominantly the lung and digestive tract.
- Children with CF are at higher risk for lung infections and pneumonia than normal children.
- The most serious lung pathogen is Pseudomonas aeruginosa (Pa) which does not usually infect non-CF children but can have very serious consequences in CF children.

- Pa infections are usually intermittent at first but eventually become chronic, leading to mucoid Pa infection, after which lung function decline is precipitous.
- There is a belief that if Pa infections can be eradicated rapidly, then the mucoid stage can be delayed significantly.
- Our goal is to find the best choice of treatment each time a patient is infected with CF to yield the longest mucoid-free survival.

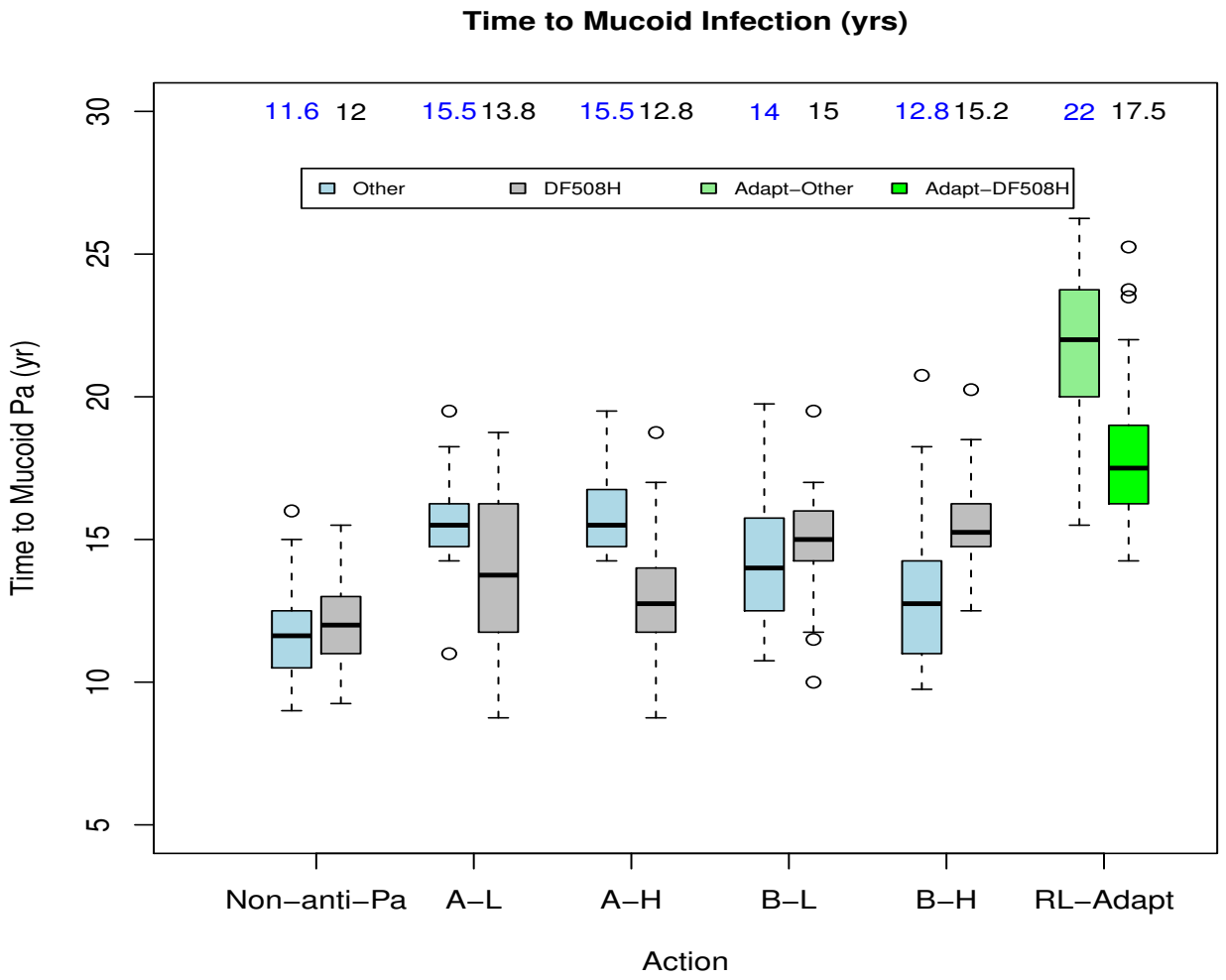
Parameters of the clinical reinforcement trial:

- We need to recruit patients with ages 0–20 years old and follow for about 2 years.
- For each episode of Pa infection, we will randomize to one of 5 treatments: placebo, AL, AH, BL and BH.
- Which treatments are acceptable will depend on patient prognostic data, including age.
- After trial completion, we will use Q-learning for an "infinite horizon" to estimate the optimal, personalized treatment choice as a function of prognostic values.
- A phase III randomize trial will then be conducted to verify superiority of the personalized treatment compared to fixed, standard-of-care approaches.

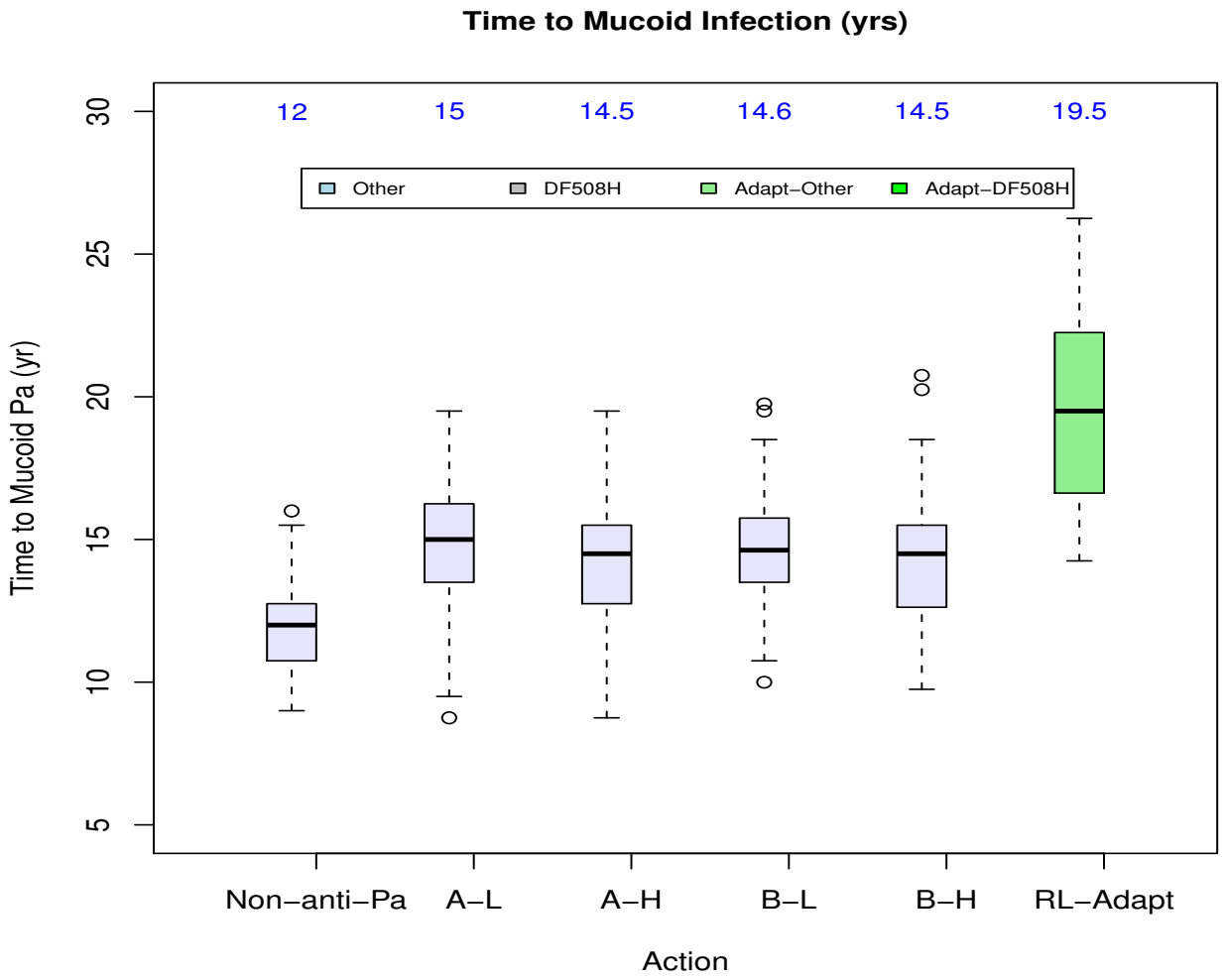
Mathematical modeling of disease process:

- A discrete, time-varying Markov transition process is used to simulate change of infection status and treatment effects.
- The discrete time steps are three-month intervals from birth.
- The three states are Pa-free, Pa positive and mucoid Pa.
- Factors in the model include age, treatment history, gender and genomics (homozygous or heterozygous for the ΔF508 mutation).
- Uncertainty in the throat culture diagnosis for Pa is factored in.

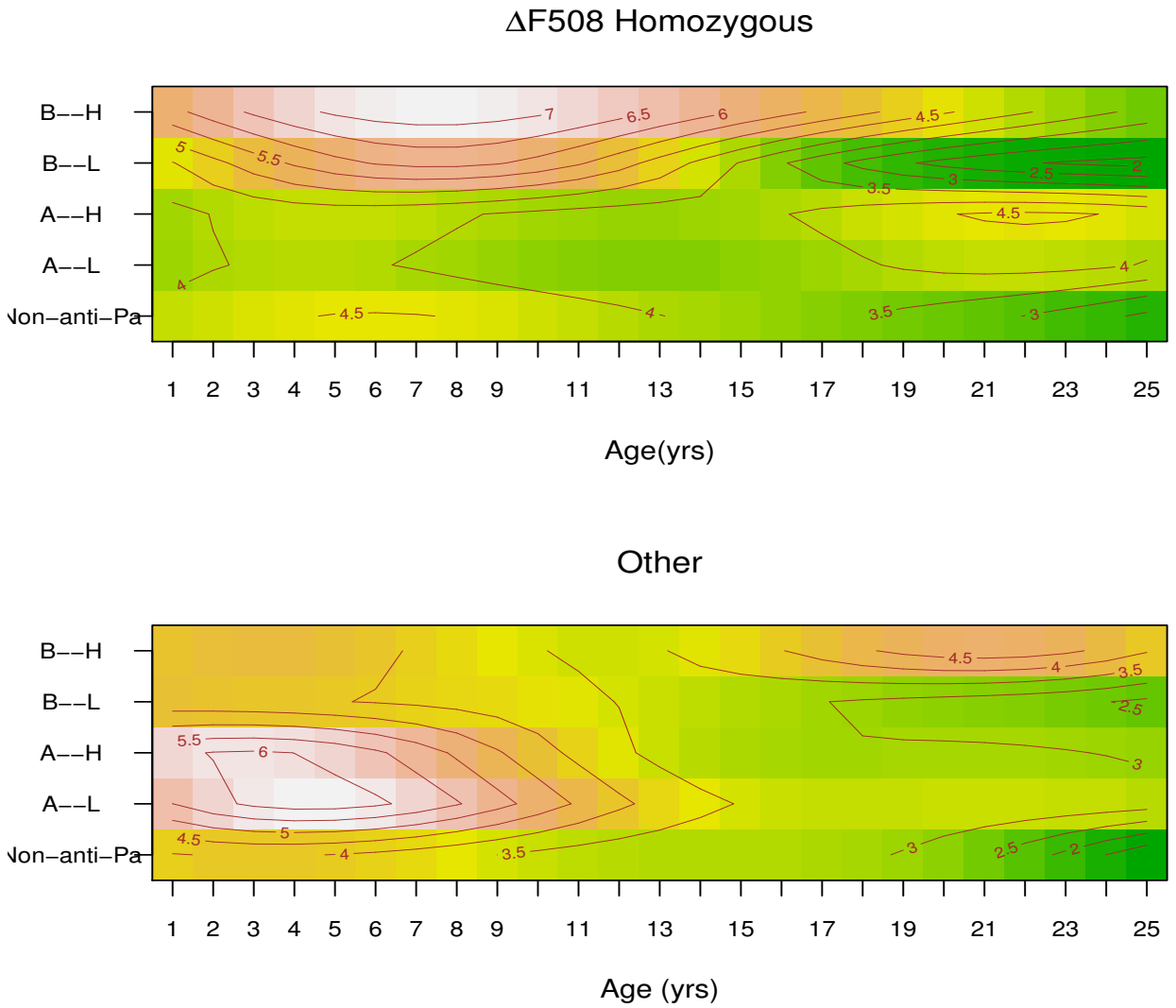
Comparison of time-to-mucoid infection between optimal personalized treatment and fixed treatments when genetics factor is included.



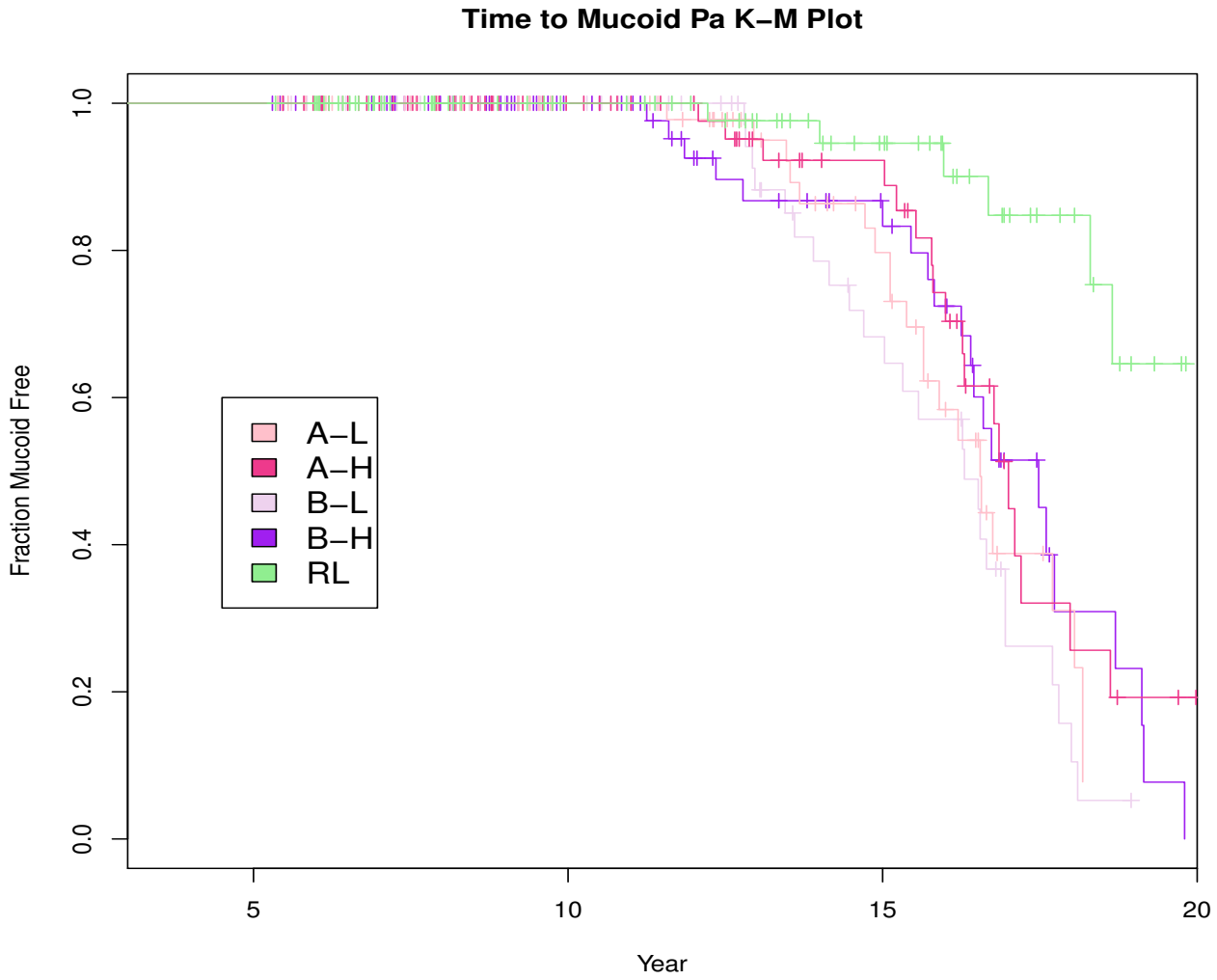
Comparison of time-to-mucoid infection between optimal personalized treatment and fixed treatments when genetics factor is pooled.



Q-functions for two genetic subgroups.



Kaplan-Meier plots from 5 year confirmatory trial of optimal versus fixed regimens.



Simulation summary:

- The procedure is able to successfully identify optimal personalized treatment regimens that are superior to fixed regimens.
- Including additional important prognostic factors, such as genetics, can further improve performance.
- We were able to tune the model parameters so that the distribution times to mucoidy approximately match published values.
- As with the previous example, a reasonably good dynamic model is needed in order to construct virtual patients and virtual trials.

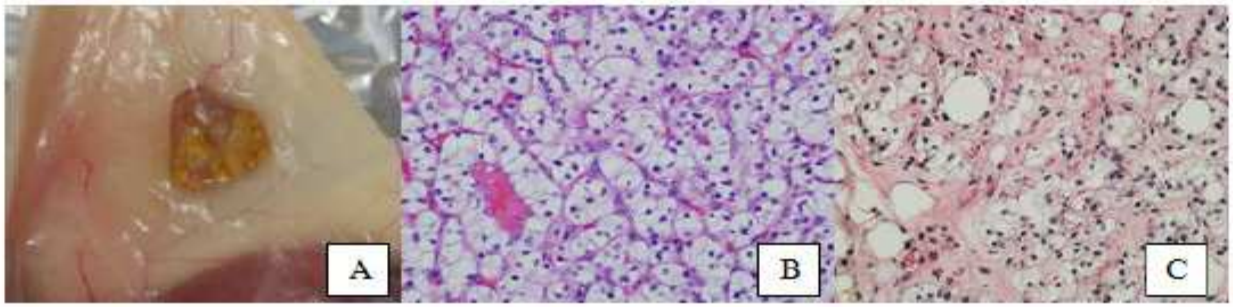
A Fourth Example: Colorectal Cancer

We have obtained a P01, in collaboration with NC State and Duke, for developing methodology for cancer clinical trials which includes two projects involving the proposed personalized medicine techniques.

Our group has also applied for additional funding to apply the proposed methodology to treating older patients with metastatic colorectal cancer (mCRC).

We will validate our methodology in actual human tumors of mCRC engrafted into mice: patient-derived xenografts (PDX).

A) Gross photomicrograph of a xenografted human renal cell carcinoma underneath the skin of a NOD/SCID mouse. B) Representative H & E stained photomicrograph of a section of ccRCC at the time of nephrectomy C) H & E stained photomicrograph of the xenograft derived from the same patient taken at the time of passage.



We will then apply the proposed statistical methods and PDX results to design a clinical trial to determine

- which elderly patients benefit from adding oxaliplatin or irinotecan to a fluoropyrimidine base,
- which patients benefit from further addition of VEGF depleting antibody bevacizumab, and
- which patients benefit from early surgical intervention.

We intend to implement our “personalized medicine trial” through a new UNC Cancer Research Network lead by Dr. O’Neil.

Discussion

- Clinical reinforcement trials can discover effective individualized regimens that improve significantly over standard approaches.
- The sample sizes required are larger than phase II trials typically are but not larger than phase III trials.
- After identifying one or more candidate individualized therapies using a clinical reinforcement trial, a phase III-like trial should be conducted to compare the discovered individualized therapies to standard therapy.
- Good dynamic models (both mechanistic and stochastic) need to be constructed based on valid clinical science in order to construct virtual patients and virtual trials in order to design clinical reinforcement trials.
- The advantage is the discovery of effective new treatments that could otherwise be missed by conventional approaches.