Southern Methodist University

SMU Scholar

Statistical Science Theses and Dissertations

Statistical Science

Spring 5-13-2023

Optimizing Tumor Xenograft Experiments Using Bayesian Linear and Nonlinear Mixed Modelling and Reinforcement Learning

Mary Lena Bleile
Southern Methodist University, mbleile@smu.edu

Follow this and additional works at: https://scholar.smu.edu/hum_sci_statisticalscience_etds

Part of the Applied Statistics Commons, Artificial Intelligence and Robotics Commons, Biostatistics Commons, Medical Biomathematics and Biometrics Commons, Statistical Methodology Commons, and the Statistical Models Commons

Recommended Citation

Bleile, Mary Lena, "Optimizing Tumor Xenograft Experiments Using Bayesian Linear and Nonlinear Mixed Modelling and Reinforcement Learning" (2023). *Statistical Science Theses and Dissertations*. 37.
https://scholar.smu.edu/hum_sci_statisticalscience_etds/37

This Dissertation is brought to you for free and open access by the Statistical Science at SMU Scholar. It has been accepted for inclusion in Statistical Science Theses and Dissertations by an authorized administrator of SMU Scholar. For more information, please visit http://digitalrepository.smu.edu.

OPTIMIZING TUMOR XENOGRAFT EXPERIMENTS USING BAYESIAN LINEAR AND NON-LINEAR MIXED MODELLING AND REINFORCEMENT LEARNING

Approved by:

Daniel F. Heitjan, Ph.D
Professor and Chair of the Department of Statistical Science, Southern Methodist University

Charles South, Ph.D
Professor of Practice in Department of Statistical Science, Southern Methodist University

Chul Moon, Ph.D
Assistant Professor in Department of Statistical Science, Southern Methodist University

Steve Jiang, Ph.D (External)
Vice Chair of Digital Health & AI and Division Chief of Medical Physics & Engineering, UT Southwestern Medical Center

OPTIMIZING TUMOR XENOGRAFT EXPERIMENTS USING BAYESIAN LINEAR AND NON-LINEAR MIXED MODELLING AND REINFORCEMENT LEARNING

A Dissertation Presented to the Graduate Faculty of the

Dedman College

Southern Methodist University

in

Partial Fulfillment of the Requirements

for the degree of

Doctor of Philosophy

with a

Major in Biostatistics

by

MaryLena Bleile

B.S., Statistical Science, Southern Methodist University
B. Mus., Cello Performance, Southern Methodist University

May 13, 2023

Copyright (2023)

MaryLena Bleile

All Rights Reserved

ACKNOWLEDGMENTS

I wish to first thank my advisors, Dr. Steve Jiang and Dr. Daniel F. Heitjan, whose expertise and advice gave this work its essence. My appreciation is also extended to my committee members and lab directors, whose feedback helped to refine it.

This work would not have been possible without the inspiration of my favourite artists: In particular, the overwhelming beauty and raw emotional energy captured by deathcore bands Lorna Shore, Shadow of Intent, Carnifex, and Whitechapel frequently carried me through the process. I thank also my friends and colleagues from SMU, UT Southwestern, and Queer in AI for discussion and support which helped me to get unstuck several times. I owe a degree of my remaining sanity, too, to the Judo coaches and my training partners at Mustang Martial Arts club and Eastside Dojo.

The cover art for this dissertation, which appears on the dedication page and as a chapter-separating asterism, was created by Aster at Gold Dust Tattoo shop.

Bleile, MaryLena B.S., Statistical Science, Southern Methodist University
B. Mus., Cello Performance, Southern Methodist University

Optimizing Tumor Xenograft Experiments Using
Bayesian Linear and Non-Linear Mixed Modelling
And Reinforcement Learning

Advisor: Daniel F. Heitjan, Ph.D

Doctor of Philosophy degree conferred May 13, 2023
Dissertation completed May 8, 2023

Tumor xenograft experiments are a popular tool of cancer biology research. In a typical such experiment, one implants a set of animals with an aliquot of the human tumor of interest, applies various treatments of interest, and observes the subsequent response. Efficient analysis of the data from these experiments is therefore of utmost importance. This dissertation proposes three methods for optimizing cancer treatment and data analysis in the tumor xenograft context. The first of these is applicable to tumor xenograft experiments in general, and the second two seek to optimize the combination of radiotherapy with immunotherapy in the tumor xenograft context.

In tumor xenograft experiments, one commonly observes that growth is exponential (log-linear) initially but later decelerates. For this reason, it is common to model tumor volume using a sigmoid growth curve such as the Gompertz, wherein growth increases in what first appears to be an exponential curve and then decelerates, eventually reaching a plateau. Scientists have advanced multiple biological hypotheses to explain this phenomenon. We propose that a contributing factor in the context of *in vivo* tumor xenograft studies may be the loss of animals whose tumors are growing most quickly. As they die or require sacrifice, we are left with only the smaller, slower-growing tumors on the remaining animals. To illustrate this point, we show *via* simulation that the performance of the Gompertz model exceeds that of the exponential when fit to the average of incomplete exponential data where larger tumors are subject to truncation. A log-linear mixed model, however, effectively recovers the individual exponential curves. We conduct an analysis of real tumor xenograft data using these models, which shows that while tumor growth *appears* Gompertz when analyzing the averages of the available tumor volumes, an exponential mixed model fits well to the individual curves.

The efficacy of a radioimmunotherapy regimen for cancer treatment is sensitive to the radiation fractionation scheme. Chapter 2 develops and evaluates a generalized, adaptive method to identify the optimal radiation regimen for use with immunotherapy in the context of a sequential tumor xenograft experiment. We use a predictive model, updated after each new observation, to forecast future tumor growth under each of a set of candidate radioimmunotherapy regimens, selecting the one that yields the best result. We evaluate and compare three versions of our method, characterized by three different predictive models used for forecasting, in a simulation experiment that models an adaptive *in vivo* tumor xenograft study. We observe that the predictive system characterized by a linear spline mixed model best balances efficiency and robustness and therefore provides the most use in practical applications.

We also develop a Reinforcement Learning system to learn and generate such personalized optimal radiotherapy regimens, which is described in Chapter 3. This model was developed based on a set of pre-clinical experimental data and can capture, in the context of combination therapy, the dependence of performance on radiotherapy scheduling. The timings chosen by the agent outperform the fixed application of the best-performing timing observed in an *in vivo* experiment to all individuals. This preliminary endeavor provides methodological foundation for a future adaptive *in vivo* tumor xenograft experiment, and potentially a subsequent human trial.

TABLE OF CONTENTS

APPENDIX

LIST OF FIGURES

1.1 As faster-growing tumors are eliminated, the growth rate of the curve of the average remaining tumor volumes declines. 3

LIST OF TABLES

Table Page

Dedicated to all Children of the Night, especially the procyon lotor and the goth community.

CHAPTER 1

Animal Sacrifice as a Potential Cause of Decelerating Growth in Xenograft Experiments

1.1 Introduction

Modelling of tumor growth in animal experiments is a pillar of oncology research (Heitjan, Manni, and Santen 1993; Demidenko 2010; Santen, Yue, and Heitjan 2012). In particular, one commonly uses statistical models to analyze data from *tumor xenograft* experiment. In a typical such experiment, one implants a set of animals with aliquots of a cancer cell line derived from a human tumor of interest, applies various treatments, then observes and compares the resulting tumor growth or decay in each treatment arm. One can fit a growth model to the data from each experimental group, drawing conclusions about treatment effects by making inferences on features of the growth models such as the overall growth rate (Heitjan, Manni, and Santen 1993).

The growth of implanted tumors — at least in the early stages of an experiment — typically resembles an exponential curve; indeed, models that assume baseline exponential growth have exhibited a moderately good fit to tumor xenograft data (Demidenko 2010;

Heitjan 1991). Closer inspection, however, reveals that the growth rate tends to decline with time, and thus models that can accommodate late-stage growth deceleration are generally preferred: Popular choices include the Gompertz and logistic curves, both of which are special cases of the *generalized logistic function* (Vaghi et al. 2020; Viossat and Noble 2021; Ghaffari Laleh et al. 2022; Hartung et al. 2014). The reason for the eventual late-stage deceleration in mean tumor growth is not generally understood. While it is plausibly a reflection of cell proliferation and diffusion mechanics and a resulting decline in carrying capacity (Frenzen and Murray 1986; Sheergojri et al. 2022), the nature of this effect remains under investigation (Yang et al. 2020; Tienderen et al. 2022; Baranowitz 2022).

Commonly in xenograft experiments some animals die spontaneously or undergo sacrifice for morbidity before the intended conclusion of observation. This removal of experimental subjects, similar to the phenomenon of *dropout* in longitudinal studies in humans, can induce a bias in graphical and numerical analyses of tumor growth. Because the larger or faster-growing tumors are removed earlier, the remaining individuals are a *biased* sample of the original cohort, consisting primarily of animals with slower-growing tumors (Figure 1.1). With the individuals bearing larger or faster-growing tumors eliminated, the cross-sectional means computed from remaining animals will therefore be lower than they would be had the animals with large tumors survived. We propose that this removal of experimental subjects from tumor xenograft experiments due to spontaneous death or compassionate sacrifice implies that one will observe a deceleration of growth, regardless of the underlying characteristics of the individual tumors.

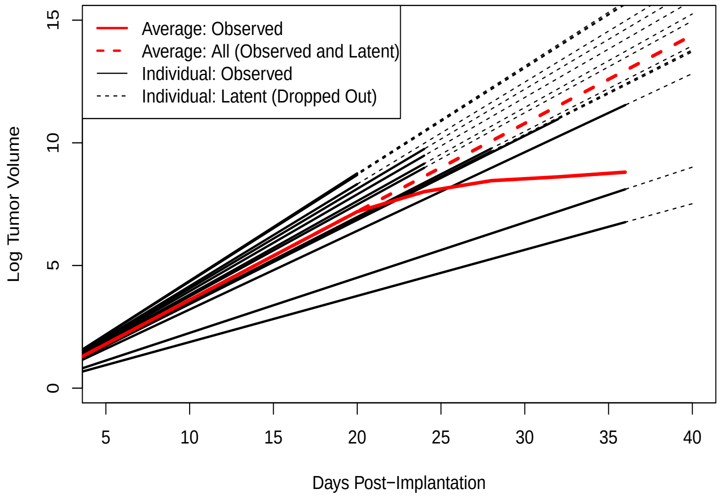

Figure 1.1: *As faster growing tumors are eliminated, the growth rate of the curve of the average remaining tumor volumes (solid red) declines.*

We illustrate the potency of this effect by a simulation experiment where we generate a series of tumor growth curves from an exponential mixed model with subject-specific slopes, assuming that subjects are terminated from the experiment with probability depending on their size at the previous observation. We then fit various models to the data aggregated two ways: Individually by subject (resulting in a *mixed model*), and averaged by time over all remaining subjects (using a classical regression model that ignores the "Animal" effect). We observe that the Gompertz model provides a better fit than the exponential to the averages of the remaining observations, whereas the exponential mixed model accurately recovers the parameters used to generate the data. This demonstration illustrates that **an aggregated curve with apparent growth deceleration can arise from underlying exponential growth curves that are subject to truncation at large volumes**. We also fit the exponential (mixed), exponential (averaged), and Gompertz (averaged) models to a real dataset, showing that whereas the Gompertz curve describes the averaged tumor volume better than the exponential, the exponential curve is preferable on the individual level. Finally, we investigate and compare growth deceleration on the subject-specific and the average level by adding a squared "time” term to the log-linear (exponential) model fit to the real dataset: While growth does still appear to decelerate on the individual level, the deceleration effect is diminished by the inclusion of subject-specific growth rates and intercepts.

1.2 Methods

1.2.1 Simulation

Data-generation model

We generated $M = 1000$ data sets consisting of synthetic log-scale tumor volumes (denoted $\ln y_{it}$) from $n = 30$ subjects, indexed $i = 1, \ldots, n$, according to the model

$$\ln y_{it} = b_{i,l0} + b_{i,l1} t + e_{it}, \tag{1.1}$$

$$e_{it} \sim N(0, \sigma^2),$$

where the subscript l is a label which stands for "linear". We generated y_{it} for subject i at times $t = 1, \ldots, 40$ per Equation (1.1), where the e_{it} are independent normal errors with common variance σ^2. For each individual i, we generated $b_{i,l0} \sim N(\beta_{l0}, \sigma^2_{\beta 0i})$, and $b_{i,l1} \sim N(\beta_{l1}, \sigma^2_{\beta 1i})$, with $b_{i,l0}, b_{i,l1}$ independent.

Next, we deleted data points according to a probabilistic removal process. Define A_{it} to be a random variable that takes the value 1 if individual i is still in the experiment at time t, and 0 otherwise, with $A_{i1} = 1$, and let this probability depend on the tumor volume at the previous measurement time:

$$\begin{aligned} p_{it} &= \Pr[A_{it} = 0 | A_{i,t-1} = 1, Y_{i,t-1} = y_{i,t-1}] \\ &= \frac{\exp(\gamma_0 + \gamma_1 y_{i,t-1})}{1 + \exp(\gamma_0 + \gamma_1 y_{i,t-1})}. \end{aligned}$$

That is, for $\gamma_1 > 0$, animals with larger tumors are more likely to undergo sacrifice, as would happen in an actual experiment. In our simulations, we set the parameter vectors to be $(\beta_{l0}, \sigma_{\beta 0i}, \beta_{l1}, \sigma_{\beta 1i}, \sigma) = (0, 0.5, 0.15, .01, 0.001)$ and $(\gamma_0, \gamma_1) = (-10, 1)$ to control the overall dropout rate and ensure realistic dropout times.

Data analysis models

We estimated three different models from each simulated dataset. The first of these assumes baseline exponential tumor growth as follows:

$$\bar{y}_t = \exp(\beta_{l0} + \beta_{l1} t + e_t). \tag{1.2}$$

In Equation (1.2), $\bar{y}_t$ is the mean volume of the remaining tumors at time t and $e_t \sim N(0, \sigma^2)$. The second model assumes that mean tumor growth is Gompertzian:

$$\bar{y}_t = \beta_{g0} \exp\left[\ln\left(\beta_{g1}/\beta_{g0}\right) \times (1 - \exp(-\beta_{g2} t))\right] + e_t, \tag{1.3}$$

where, as before, $e_t \sim N(0, \sigma^2)$ and the subscript "l" stands for "linear". Similarly, the subscript "g" on the parameters of the Gompertz model stands for "Gompertz". Equation (1.3) assumes that the average tumor growth starts at volume β_{g1}, increasing in a sigmoid to reach its limiting volume β_{g0}. Finally, we fit a linear mixed model to the log tumor volumes that matches exactly the data generation process in Equation (1.1). Because animal removal depends only on the observed data (similar to *missingness at random* in the parlance of missing data), estimation of the last model consistently recovers the original exponential parameters (Little and Rubin 2019).

1.2.2 Data

Our data consist of tumor measurements from wild-type mice implanted with Lewis Lung Cancer (LLC) cell lines. We consider here the control data from three rounds of radiation oncology experiments that investigated various combinations and timings of radiation and immunologic therapy. The sample sizes of these control groups were 5, 10, and 8, for a total of $n = 23$ mice. Moore et al (2021) analyzed Round 1 of these data in an investigation of the effect of radiotherapy schedules on the synergy between radiation and anti-PDL1 immunotherapy. They used repeated measures analysis of variance to investigate potential difference between treatment groups, concluding that the timing of the fractionated radiotherapy pulses impacts the degree of synergism between the two treatment modalities. In order to account for the pre-emptive loss of experimental animals, they excluded from analysis all data from all animals beyond the date of the first animal's death.

1.2.3 Statistical analysis

In an initial analysis, we estimated the Gompertz and exponential curves from the data averaged by measurement time, as well as the previously described exponential (log-linear) mixed model with subject-specific slopes and intercepts. Next, we investigated the potential of subject-specific nonlinearity *via* linear approximation, i.e. by adding a squared term to the mixed log-linear model as follows:

$$\ln y_{it} = b_{i,l0} + b_{i,l1}t + \beta_{l2}t^2 + e_{it} \tag{1.4}$$

In Equation (1.4), β_{l2} is a fixed quadratic coefficient; all other symbols are as defined in Equation (2.1). If $\beta_{l2} < 0$, then the individual curves manifest late-stage growth deceleration as suggested by Vaghi, et al (2020). Note, however, that some degree of subject-level late-stage growth deceleration does not contradict our thesis that pre-emptive animal removal *contributes to* the late-stage deceleration of average tumor volumes. We do not include random effects on β_{l2} because the addition of such effects cause issues with model convergence, indicating a poor fit to the data.

We also investigated the effect on the estimate of β_{l2} that occurs when one removes the subject-specific random effects, fixing $\sigma_{\beta_0} = \sigma_{\beta_1} = 0$: Note that this procedure is equivalent to fitting the model $\ln\bar{y}_t = \beta_{l0} + \beta_{l1}t + \beta_{l2}t^2 + e_t$. If the removal phenomenon contributes to the late-stage deceleration of average tumor growth, then one might expect that ignoring the subject-specific effects would result in a larger negative estimate of β_{l2}, representing the deceleration phenomenon.

We estimated the log-linear mixed models using the `R` function `lmer` (Bates et al. 2015). To compare two models using likelihood analysis, one must fit both models on the same scale. We therefore estimated both of the mean models to the raw values using nonlinear least squares as implemented in R function nls (R Core Team 2021). To compare non-nested models, we computed Akaike's information criterion (AIC).

1.3 Results

1.3.1 Simulation

The Gompertz model gave an overwhelmingly better fit to the averages of the available post-dropout values, achieving a lower AIC on 97% of the Monte Carlo replicates. Conversely, the exponential model fit the averages of the uncensored data better in all 1,000 Monte Carlo iterations. Figure 1.2 shows that the exponential mixed model successfully recovered the original (pre-removal) curve. Here, the grey dots represent the cross-sectional averages of the complete, pre-removal simulation data, and the dark black line represents the cross-sectional averages of the "observed" simulation data post-removal — that is, the remaining data after probabilistic removal. The three dashed lines characterize the predictions from the three models of interest, as fit to the observed (post-removal) data. In order to obtain accurate cross-sectional averages from the post-removal data at the later stages (when most individuals had been removed), we generated an additional 1.497 million individuals (for 1.5 million total) for plotting purposes, and truncated the curve corresponding to cross-sectional post-removal data (i.e. the dark black line) at the first timepoint that had fewer than 1,000 observed data points.

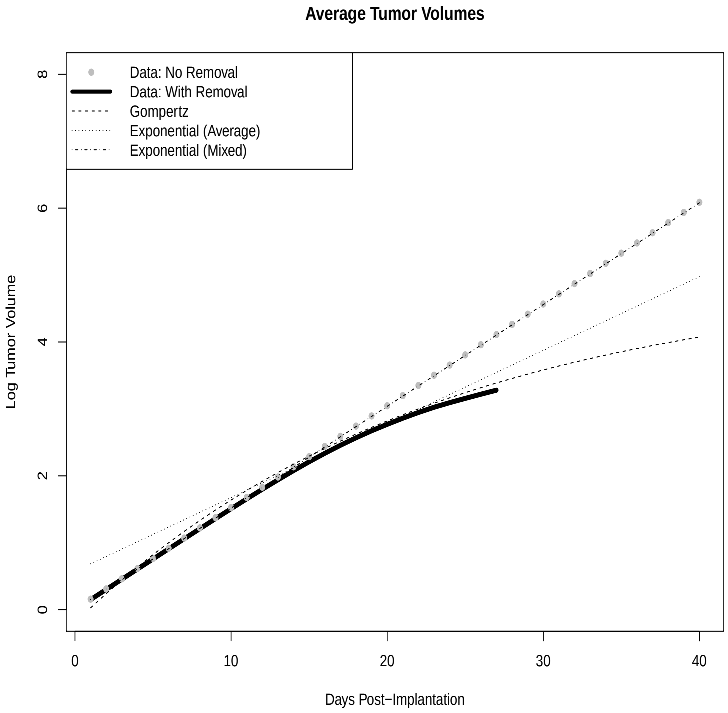

Figure 1.2: *Simulated data and predictive curves from each model. All models are estimated from the simulated data with imposed removal.*

1.3.2 Data Analysis

All of the animals in the dataset were sacrificed before the end of the experiment, either due to skin ulcer ($n_{\text{ulc}} = 6$) or tumor burden ($n_{\text{tum}} = 17$). Figure 1.3 shows predictions from each of the three models when estimated from the real data. In the left panel, the dark black dots are the cross-sectional mean observed tumor volumes, and the dashed line is the Gompertz model estimated from them. The dotted line is the exponential model fit to the average volumes, and the black dot-dashed line is the average of the individual exponential curves fit by the mixed model. The grey dot-dashed line characterizes the predictions of the mixed model characterized by Equation (1.4) (which includes the squared "time" term). Of the two models estimated from the mean volumes, the Gompertz has a better apparent fit than the exponential. The AIC on the Gompertz fit is also lower than that on the exponential model fit to the average, in agreement with the visual heuristic. The takeaway is that unlike the exponential curve, the Gompertz fits well to the tumor volumes *on average*. Moreover, when fit to the average of the data, the exponential creates predictions that substantially exceed the observed values in the later stages of the experiment: It is deficient in that it cannot capture the apparent deceleration of growth on the late-stage cross-sectional averages of observed tumor volumes.

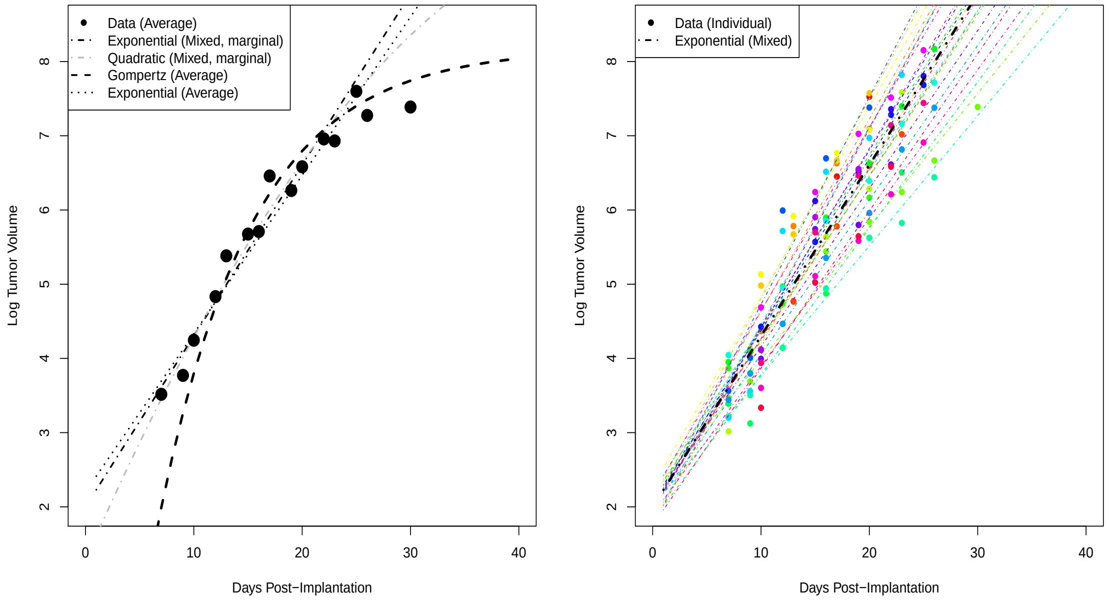

Figure 1.3: *Predictive curves from each model when estimated from the experimental data. Each color corresponds to an individual animal, and the black and grey curves/dots represent averages or marginal predictions.*

In the right panel, each set of uniquely colored dots represents the datapoints from a single animal, and the corresponding colored line represents the corresponding subject-specific predictions from the exponential mixed model for that animal. These predictions conform moderately well to reality; unlike the exponential fit to the average, the estimated exponential curves from the mixed model do not substantially exceed the observed values at the later times. This phenomenon arises numerically in the mean squared prediction error (MSPE): While the log-scale MSPE on the exponential model fit to the averages was equal to 0.17, log-scale MSPE on individual curves had an inter-quartile range of (0.02,0.06), with median 0.05. This constitutes further evidence in favor of the idea that even if each individual curve is exponential — that is, in the extreme case where the growth of individual curves does not decelerate — the average of the non-censored values may appear to decelerate. Note also that in the left panel, the marginal predictions from the exponential mixed model (characterized by the black dot-dashed line) exceed the predictions from the exponential model fit to the averages in later stages; this illustrates the potential negative bias that the removal phenomenon may cause. Estimating a mixed exponential model with subject-specific effects mitigates this bias.

The addition of a subject-specific squared term in the linear model contributed substantially to model fit, resulting in a lower AIC on the model characterized by Equation (1.4) than the model characterized by Equation (2.1). The estimated mean of the subject-specific squared coefficient was negative ($\hat{\beta}_{l2} = -0.39$), as one might expect: This analysis agrees with previous investigations such as Vaghi, et al (2020), which propose growth rates of individual curves do still decelerate in the later stages. Removing the subject-specific effects in the quadratic model, however, resulted in an estimate of $\hat{\beta}_{l2} = -0.64$, substantially exceeding in magnitude the estimate from the mixed model. Moreover, removing the subject-specific variability on the intercept and linear term in the quadratic model caused an increase in AIC, indicating that the mixed quadratic model (i.e. the model with subject-specific intercepts and initial growth rates) provides a better fit to our dataset than the quadratic model fit wholly to the averages.

Inclusion of a random effect corresponding to experiment number did not substantially improve the model fit, indicating that observations are homogeneous across rounds.

1.4 Discussion

The removal of subjects with larger or faster-growing tumors can cause decelerating growth in the cross-sectional tumor volume averaged from the remaining subjects. This removal mechanism can be so extreme as to cause a sigmoidal model to fit better than an exponential curve to cross-sectional averages of truly exponential individual curves.

We do not dispute that the best-fitting model for a set of individual tumor growth curves might still involve a late-stage deceleration in growth rate. Rather, we argue that the removal of animals with faster-growing tumors can cause an apparent deceleration of growth in the mean tumor volume curve, whatever the underlying kinetics. We have illustrated this point with both simulated and real tumor xenograft data. Our conclusion is consistent with visual inspection of the data, where we observe that the exponential model fits better on the individual curves than it does on the averages. Future studies might benefit from an investigation of the implications of this phenomenon on analysis, and the extent to which it depends on the mechanism of removal. We moreover observe that the removal of animals with larger tumors is a feature of many xenograft experiments, as tumors typically cause death or morbidity within the time frame of planned observation.

The fact that animal loss contributes to the late-stage deceleration of mean tumor growth has various implications for the analysis of data from tumor xenograft experiments. Most importantly, efforts to explain the biological mechanisms which cause this phenomenon should take into account the behavior of the individual curves. An effective way to accomplish this is to analyze data using linear or nonlinear mixed models — i.e., models that posit a basic shape for the growth curves with parameters that vary between animals (Heitjan, Manni, and Santen 1993; Heitjan 1991). Other options might include imputing the counterfactual post-death values using methods from the missing data literature; future research might include a direct comparison of a variety of methods and models to this end.

In summary, we propose here that loss of experimental animals to morbidity or sacrifice contributes to the growth deceleration commonly observed in late-stage cross-sectional averages of tumor xenograft data. Even if the individual curves exhibit deceleration, the deceleration of the average curve will be more pronounced. Among other things, this suggests that plotting average tumor volumes over time is likely to lead to an incorrect understanding of individual growth kinetics; plotting individual curves using different line types or colors will give a truer picture. Similarly, models that one estimates from data averaged by time are deficient for inference; estimating mixed models from the individual data mitigates this problem.

CHAPTER 2

Statistical and Machine Learning Methods for Adaptive Radiotherapy Treatment Scheduling

2.1 Introduction

2.1.1 Overview

Recent studies in radiation oncology have demonstrated the potential for synergistic effects of radiotherapy (RT) and immunotherapy (IO) in the treatment of solid tumors. Moore et al (2021) showed that the degree of synergism is sensitive to the timing of the RT pulses: If one applies two pulses of radiation 10 days apart, then the addition of IO suppresses tumor growth, whereas if one spaces the pulses one day apart, the IO has minimal effect. Various radiobiological hypotheses compete to explain this observation. One such hypothesis, which Moore et al (2021) termed "PULSAR”, suggests that the immune cells newly recruited by RT are more sensitive to radiation than previously existing cells. Applying the second RT pulse too soon kills the newly recruited immune cells and weakens both the secondary effect of RT and the main effect of IO.

Planning the timing of the RT pulses is challenging because i) the optimal spacing may vary among individuals due to heterogeneity in their immune systems, and ii) an individual's optimal spacing may change over the course of treatment due to immune cell depletion or a vaccine effect. A method for determining personalized, adaptive RT schedules is therefore needed. We propose here one such method.

2.1.2 Experimental context

We assume the context of an *in vivo* tumor xenograft experiment, which is a necessary step in the pre-clinical development of a treatment for a solid tumor. In a typical such experiment, one implants an aliquot of a cancer cell line into the flank of each experimental animal, applies the treatment of interest, and observes the growth of the tumor over a period of several weeks. In an experiment to study RT/IO combinations, treatment arms would likely include a double-negative control (no RT or IO), an IO-only arm, an RT-only arm, and arms representing various doses of RT and IO.

We consider here a hypothetical sequential tumor xenograft experiment aimed at identifying an optimally timed RT/IO regimen. That is, our study will involve implanting an animal with a tumor, observing a partial pre-treatment growth series, treating the tumor with one of the candidate regimens, and observing the subsequent growth of the tumor. With the data from each new animal, we will update a statistical model relating pre-treatment growth characteristics and type of treatment to an outcome that serves as a proxy for the effectiveness of the treatment. The product of this process is a treatment regimen that optimizes this outcome. Algorithm 1 describes the process.

Algorithm 1: Design for identifying an optimal radiotherapy regimen

1. For a newly implanted mouse, measure a sequence of pre-treatment tumor volumes; apply the initial dose of RT; observe the subsequent tumor volume.
2. Using the data from step 1 together with a model for tumor growth, predict the mouse's future tumor growth under each potential action.
3. Identify $\hat{a}_{\text{opt}}$, the action that gives the optimal outcome.
4. Apply $\hat{a}_{\text{opt}}$ and observe the animal's response.
5. Update the estimated parameters of the predictive model using the data collected in steps 1–4.

After application of a pulse of radiotherapy, the tumor typically shrinks for a time and then regrows. If the RT-induced tumor recession is powerful enough, the subject's immune system can — in theory — take over and drive the tumor volume to zero. With this goal in mind, we seek to minimize the *nadir*, or lowest attained value, of the post-RT curve. We therefore quantify the performance of an RT regimen by the nadir of tumor volume after application of the final pulse of RT. We define an *optimal* RT regimen as the regimen which causes, for each animal, the lowest nadir (Figure 2.1).

Others have proposed approaches similar to Algorithm 1. Kosinsky et al (2018) used a 25-parameter Bayesian non-linear mixed model of the tumor microenvironment to de-

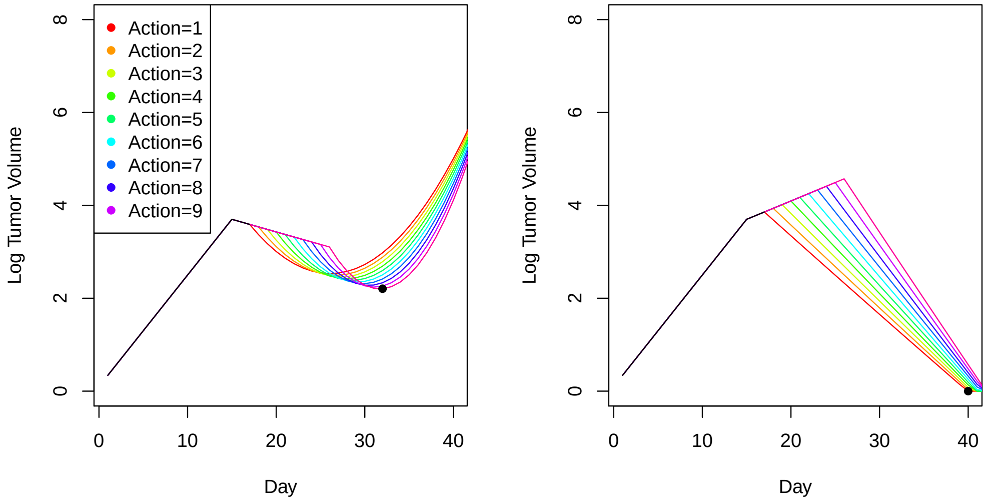

Figure 2.1: *Counterfactual curve sets produced by the spline model corresponding to regrowth (left) and cure (right). The lowest nadir is indicated on each: Here, the optimal action for the curative example is 9, whereas the optimal action for the regrowth example is 1.*

termine personalized RT-IO dose and schedule combinations, validating their results indirectly by forecasting future tumor growth. They did not use the model for treatment assignment, though they did suggest that as a potential extension. Zahid et al (2021) used a similar approach to optimize radiotherapy dose-escalation regimens (with and without chemotherapy). While these methods *do* use the mathematical model to derive future actions, they do not optimize the RT dosing schedule in light of its synergy with IO.

Our method represents a form of Reinforcement Learning called *Q-learning* (Mnih et al. 2015; Sutton and Barto 2018), which is effective for determining personalized cancer treatment in a variety of settings. These include chemotherapy (and general clinical trial) dose determination (Yauney and Shah 2018; Padmanabhan, Meskin, and Haddad 2017; Hassani and Naghibi-S 2010), optimization of a radiotherapy administration instrument called a *multi-leaf collimeter* (Hrinivich and Lee 2020), RT fractionation, and RT dose determinaton (in isolation, i.e. with no immunotherapy) (Jalalimanesh et al. 2017; Tseng, Luo, Cui, et al. 2017; Ebrahimi and Lim 2021). Various sources in the machine learning literature stress the need for adaptive simulation-based and inverse planning methods like Algorithm 1 (Tseng, Luo, Ten Haken, et al. 2018; Willcox, Ghattas, and Heimbach 2021; Enderling et al. 2019). One benefit of our method is that it is simpler to understand and apply than a general Reinforcement Learning approach, as it does not require the explicit deployment of theory from Markov Decision Processes (though one could describe our method in its parlance if desired).

As no such *in vivo* experiment has yet, to our knowledge, been attempted, we conduct the experiment *in silico* by Monte Carlo simulation, where each replication consists of running Algorithm 1 to completion (convergence). Among other things, this will allow us to evaluate the practicability of an *in vivo* sequential experiment, determine the sample size needed to obtain conclusive results, and assess effects of other factors such as model complexity and choice of prior distributions. Most importantly, the simulation will allow us to evaluate a range of possible choices of the sequential optimization method. Our work is unique in that it investigates the long-term performance of the method, treating each run from calibration to conclusion as a single Monte Carlo replicate; previous simulation-based work on adaptive cancer treatment ran the adaptive algorithm only once (Hassani and Naghibi-S 2010).

One naturally expects that a method derived from the data-generating model will give best results. But because in practice the correct model is generally unknown, it is essential to consider the robustness of specific model choices. Below, we describe and compare three different variants of the method that impose varying levels of structure on the predictive model.

We organize the remainder of this chapter as follows: First, we introduce the optimization models and outline the specifications of our simulation. We also present two nonlinear models that we used to simulate data, and interpret their parameters. We investigate the effectiveness of three versions of our method by computing the causal effect of our method on the outcome, relative to conventional methods of RT scheduling. We also compare the efficiency and robustness of the three versions by varying both the sample size and the model used to generate the simulation data, and observing the effects on each version's performance. Finally, we discuss the broader context of our method, describe strengths and limitations, and propose future research directions.

2.2 Methods

2.2.1 Conduct of the experiment

We observe the tumor volume sequence Y_i of animal i at a series of days $t = 1, \dots, t_{\text{rt0}}$. On day t_{rt0} we administer the initial RT pulse. We then observe the growth of the tumor for an additional t_w days; with these data in hand, we estimate the optimal number of days to wait to administer a second pulse. That is, we choose $a_i \in \mathscr{A} = \{1, 2, \ldots, A\}$ that gives the lowest projected post-RT nadir, and administer the second pulse at day $t_{\mathrm{rt}0} + t_w + a_i$. We identify a_i using some systematic method of prediction $Q_\theta : \mathbb{R}^{t_{\mathrm{rt}0}+t_w} \times \mathscr{A} \to \mathbb{R}$ indexed by parameter $\theta \in \Theta$ to combine the pre-decision growth curve $Y_{i,1:t_{\mathrm{rt}0}+t_w}$ with a potential action a to create a prediction of the outcome Y. Such a system might use, for example, a tumor growth model with RT pulses applied at days $t_{\mathrm{rt}0}$ and $t_{\mathrm{rt}0} + t_w + a_i$ to predict Y, or it might simply be a prediction system obtained by machine learning. In Algorithm 1, then, after observing the i^{th} growth curve up to day $t_{\mathrm{rt}0} + t_w$, one would use for prediction (and subsequent determination of a_i) an estimate $\hat{\theta}^{(i)}$, computed using the $i-1$ previous individuals as well as the partially observed sequence from subject i.

The predictive accuracy of Q only affects the action selection task inasmuch as it affects the ordering of the counterfactual nadir tumor volume predictions. For example, suppose that applying actions $a = 1, 2$ respectively to subject i will truly result in counterfactual nadir volumes of 3 and 5, respectively, so that action 1 is truly better than action 2 for that individual. Consider two predictive systems: $Q^{(1)}$, which estimates these counterfactual outcomes as $10, 50$, and $Q^{(2)}$, which estimates them as $4, 3.5$. Although $Q^{(2)}$ is more accurate in terms of mean squared error, $Q^{(1)}$ is better for selecting an optimal treatment, because it ranks the potential outcomes correctly.

The properties of an instance of Algorithm 1 depend on the form of Q. For example, one implementation could involve estimating subject-level parameters of a mechanistic model or a nonlinear growth model such as the Gompertz (Vaghi et al. 2020). Such nonlinear models, even when correct, may be difficult or impossible to estimate, especially in the early going. Conversely, a nonparametric method might approximate Q with a prediction system derived from a neural network (NN) (Kosorok and Moodie 2015), purchasing robustness at a price in efficiency. A compromise could involve a linear model that has no biological interpretation but nevertheless fits the data well. We henceforth consider three versions of Algorithm 1: A nonlinear modeling system denoted Q^{NL}, a linear model Q^{LM}, and a nonparametric neural network Q^{NN}.

2.2.2 A nonlinear model

A sum-of-exponentials model

We consider a predictive system, denoted Q^{NL}, which uses a nonlinear tumor growth model to predict Y. This model assumes exponential growth of the form $\exp(\alpha_0 + \alpha_1 t)$ for the tumor growth curve in the pre-treatment period $t < t_{\mathrm{rt}0}$ on a subject with growth parameter vector $(\alpha_0, \alpha_1) \in \mathbb{R} \times \mathbb{R}^+$. Immediately on first treatment (at day $t_{\mathrm{rt}0}$), two things happen:

1. A fraction $\rho \in (0,1)$ of the tumor cells are killed and removed with exponential decay rate $\omega > 0$.
2. The surviving fraction $1-\rho$ of the tumor continues to grow according to the original exponential growth function.

Assuming errors e_{it} that are additive on the log scale with a common standard deviation σ, we write the model as follows:

$$\ln Y_{it} = \mu(t;\alpha_0,\alpha_1) + e_{it}$$

where

$$\mu(t;\alpha_0,\alpha_1) = \begin{cases} \alpha_0 + \alpha_1 t & t < t_{\mathrm{rt0}} \\ \ln\{\rho \exp(\alpha_0 + \alpha_1 t_{\mathrm{rt0}}) \exp[-\omega(t - t_{\mathrm{rt0}})] + (1-\rho)\exp(\alpha_0 + \alpha_1 t)\} & t > t_{\mathrm{rt0}}. \end{cases}$$

We can moreover assume that the proportion of cells killed at subsequent treatments depends on proximity to the last treatment time. For example, suppose for mouse i we select action $a_i = 5$ corresponding to a 5-day waiting time between the treatment decision day $t_{\mathrm{rt0}} + t_w$ and the next pulse. Let ρ_2 denote the proportion of cells that begins to die at the day of the second pulse. Then, using $t_{\mathrm{rt1}} = t_{\mathrm{rt0}} + t_w + 5$ we can assume ρ_2 is affected by two elements: i) $t_{\mathrm{rt1}} - t_{\mathrm{rt0}}$, the time between pulses, and ii) ρ, the proportion of cells which began to die at the first pulse. The PULSAR hypothesis can thus be encoded as $\rho_2(\rho, t_{\mathrm{rt1}}, t_{\mathrm{rt0}}) = \rho(1 - \exp[-\delta(t_{rt1} - t_{\mathrm{rt0}})])$ for $\delta > 0$, where the factor of $(1 - \exp[-\delta(t_{rt1} - t_{\mathrm{rt0}})])$ penalizes re-treatment at short intervals. One could readily extend the model to multiple pulses.

Estimation

We treat a subset of the parameters as random effects that vary between individuals, and the rest as fixed effects. Denote the vector of random parameters as $\theta_{\text{rand},i} = (\rho_i, \omega_i)$, and the vector of fixed parameters as $\theta_{\text{fix}} = (\alpha_0, \alpha_1, \delta, \sigma)$, with the complete parameter set denoted $\theta_i = (\alpha_0, \alpha_1, \delta, \rho_i, \omega_i, \sigma)$. We assume that

$$\begin{pmatrix} \text{logit}(\rho_i) \\ \\ \ln(\omega_i) \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} \mu_\rho \\ \\ \mu_\omega \end{pmatrix}, \begin{pmatrix} \sigma_\rho & 0 \\ \\ 0 & \sigma_\omega \end{pmatrix}\right],$$

applying the logistic and logarithmic transformations to accommodate the restrictions on the sample spaces of ρ_i, ω_i.

Suppose we have a data set consisting of the previously collected tumor volume sequences $Y_1, \ldots Y_{i-1}$, their respective treatments, and an incomplete sequence Y_i, for which we wish to predict the final-day tumor volume for each $a \in \mathscr{A}$. We estimate the nonlinear model in three steps:

1. Use the data $Y_{1,1:t_{\text{rt0}}}, \ldots, Y_{i-1,1:t_{\text{rt0}}}$ to estimate α_0, α_1 and the residual standard error *via* the `lm()` function in R.
2. Fixing these three parameters at their estimates, use the entirety of the observed data $Y_1, \ldots, Y_i$ to estimate the remaining parameters (δ, ρ,and ω) as fixed effects.
3. Fixing θ_{fix} at its estimated value $\hat{\theta}_{\text{fix}} = (\hat{\alpha}_0, \hat{\alpha}_1, \hat{\delta})$, re-estimate θ_{rand} on the incomplete sequence only, resulting in the estimate $\hat{\theta}_{\text{rand},i} = (\hat{\rho}_i, \hat{\omega}_i)$.

We then use $(\hat{\theta}_{\text{fix}}, \hat{\theta}_{\text{rand},i})$ along with Q^{NL} to predict the outcomes for subject i for each $a \in \mathscr{A}$.

We fit the nonlinear model by hand using least squares in R (R Core Team 2021).

2.2.3 A linear model

Assume, as before, that we have observed $i-1$ individuals, and we wish to optimize treatment for individual i given $Y_{1:i-1,1:T}$, and $Y_{i,1:t_{\text{rt0}}+t_w}$ (where T is the maximum number of observation days, assumed constant between mice). We consider a system Q^{LM} that uses a mixed model encoding time and treatment as fixed-effect predictors X and random-effect predictors $Z \subseteq X$. The model used in Q^{LM} for data from subject i is as follows:

$$\ln Y_{i,1:T} = X_i\beta + Z_iG_i + E_i, \tag{2.1}$$

with

$$\begin{pmatrix} G_i \\ E_i \end{pmatrix} \sim \mathscr{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Gamma & 0 \\ 0 & \sigma^2 I \end{pmatrix}\right],$$

where Γ is the variance-covariance matrix of the random effects, and $\sigma^2 I$ is the $T \times T$ variance matrix of the residual E_i. We moreover assume independence across units.

We define X_i such that Equation (2.1) is a second-order spline model with a knot at each radiation day, allowing interaction between the terms corresponding to knots between the days (Durrleman and Simon 1989). Specifically, we set

$$X_i = \left(\underline{1}, \underline{t}, (\underline{t} - t_{rt0})_+, (\underline{t} - t_{rt0})_+^2, (\underline{t} - t_{rt1})_+, (\underline{t} - t_{rt1})_+^2\right) \tag{2.2}$$

where $(u)_+ = uI(u > 0)$, $\underline{t}$ is the vector of days (of length T for a completely observed individual), $\underline{1}$ is a vector of ones of the same length, and t_{rt0}, t_{rt1} are as defined above. We moreover allow the spline coefficients to vary between animals, defining Z_i to be

$$Z_i = \left((\underline{t} - t_{rt0})_+, (\underline{t} - t_{rt1})_+^2\right)_{T\times 2}. \tag{2.3}$$

Finally, we set $\Gamma = \begin{pmatrix} \gamma_1 & \gamma_{12} \\ \gamma_{12} & \gamma_2 \end{pmatrix}$, indicating that the linear and quadratic spline coefficients of the first and second pulses, respectively, are potentially correlated with covariance γ_{12}. Permitting nonzero γ_{12} allows the regrowth effect of the second pulse to depend on the effect of the first pulse. This facilitates prediction of the effect of the second pulse, about which the previously recorded information on the animal would otherwise be uninformative. The spline model is a flexible alternative to more complicated non-linear methods (Figure 2.1). We fit this mixed effects spline model using the function `lmer()` from package `lme4` in R (Bates et al. 2015).

2.2.4 A nonparametric model

We also tested a version of Algorithm 1 that uses a single-layer NN with 30 neurons and a linear output for the predictive model, with each neuron containing a set of weighting parameters for each of its inputs. Here, our predictive model Q^{NN} uses each raw tumor growth sequence and a potential action to predict the log tumor volume at the nadir of the post-treatment curve, and the parameter $\hat{\theta}^{(i)}$ represents weights estimated from the $i-1$ previously observed subjects. We estimate the NN parameters *via* gradient descent, updating them by one step with each new observation.

Note the difference between the inputs to Q^{NN} and Q^{LM}; whereas Q^{LM} involves preprocessing hand-engineered covariates using a quadratic action term, Q^{NN} automatically extracts the relevant information for prediction. Its performance is therefore less sensitive to the form and dependence structure of the input, although, as discussed below, it requires more data to train.

Implementing Algorithm 1 with a model that requires a large sample size poses additional difficulties due to the bias incurred by deterministic action selection. One solution to this problem is to use *ε-greedy* action selection, wherein one applies the arg max of Q with probability $1-\varepsilon$, and a random action otherwise (Sutton and Barto 2018). The value ε begins at 1 and decreases during training by some decay rate $\kappa \in (0,1)$: One selects actions randomly with a probability that converges to 0, rather than switching from 1 directly to 0 after a calibration period. We used the function `neuralnet` from the package `neuralnet` to implement the neural network (Fritsch et al. 2022).

2.3 Simulation

2.3.1 Data

We performed the simulation on data generated from two different nonlinear growth models. One of these models is characterized by a sum of exponentials described above; the other is a mechanistic model where the mean is governed by a set of recursive difference equations. Evaluating each version of our method on two different datasets allows us to investigate and compare the robustness of Q^{NL}, Q^{LM}, and Q^{NN}, as well as their data requirements under different sets of parametric assumptions.

2.3.2 A mechanistic model

In the mechanistic simulation model, the mean observed tumor size Tum_{it} on subject i at time t is governed by a set of difference equations. We generate the log tumor volume for subject i at time t per Equation (2.4), where the errors e_{it} (corresponding here to measurement error as well as other spontaneous factors unaccounted for by the deterministic part of the model), are iid normal with standard deviation σ as before:

$$\ln Y_{it} = \ln \text{Tum}_{it} + e_{it} \tag{2.4}$$

This model is based on a series of pre-clinical experiments exploring various combinations of RT and IO. The model encodes the PULSAR hypothesis; Figure 2.2 presents a diagram of the model, and Equations (2.5)–(2.10) present its defining equations. In the figure, blue arrows denote propagation, and purple lines ending in circles denote mitigation. In the equations, Latin characters denote treatments and aggregate effects; non-italicized 3-letter words represent notional tumor cell populations; and Greek letters represent parameters. Time (t) is discrete.

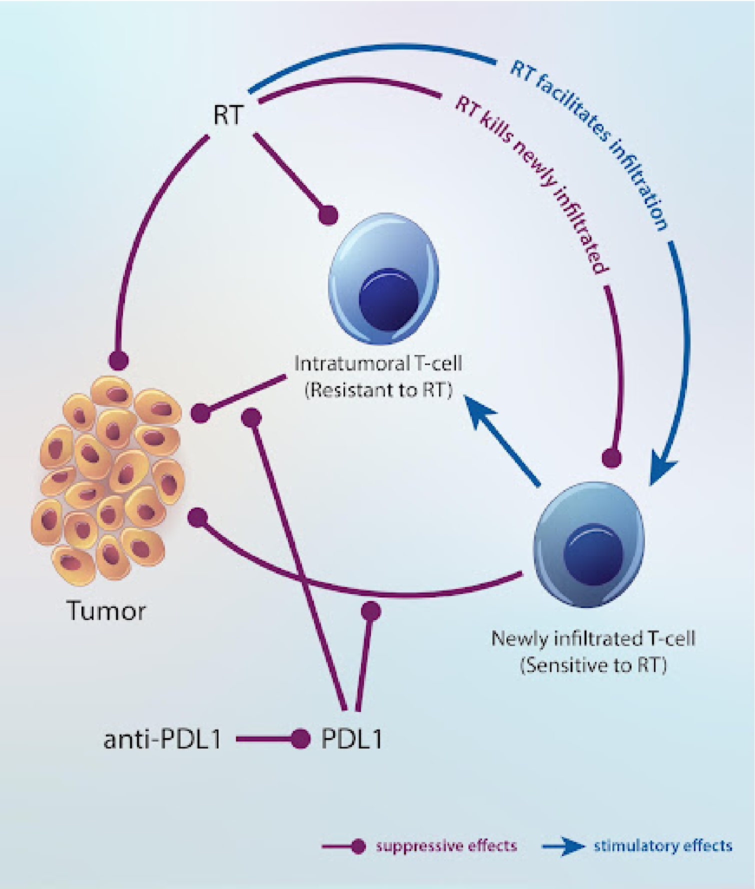

Figure 2.2: *Heuristic diagram of the interactions encoded by the difference equation model.*

$$\begin{aligned}
\mathrm{Tum}_{i,t+1} &= S_{it} \times \mathrm{Tum}_{it} \times \exp(\nu_1 - B_{it}) && (2.5)\\
S_{it} &= \exp[(1+\nu_9 R_{it})(-\nu_4 d_{it} - \nu_5 d_{it}^2)] && (2.6)\\
R_{t+1} &= \min(R_{it} - R_{it}/\nu_6 + (1 - S_{it}), 1) && (2.7)\\
B_{it} &= p1_{it}\nu_2 \mathrm{Sen}_{it}^- + p1_{it}\nu_3 \mathrm{Sen}_{i,t}^+ && (2.8)\\
\mathrm{Sen}_{i,t+1}^+ &= (1-\lambda_i)\mathrm{Sen}_{i,t}^+ \exp(-\nu_8 d_{it}) + \tau \mathrm{sign}(d_{it}) && (2.9)\\
\mathrm{Sen}_{i,t+1}^- &= \lambda_i \mathrm{Sen}_{it}^+ + \mathrm{Sen}_{it}^- \exp(-\nu_7 d_{it}) && (2.10)
\end{aligned}$$

The difference equation set involves three interacting populations: Tumor cells Tum_{it}, non-sensitive T cells Sen_{it}^-, and sensitive T cells Sen_{it}^+. We assume that the tumor would grow exponentially with rate ν_1 if left untreated. The T-cell populations — non-sensitive and sensitive — inhibit the tumor cell population with rates ν_2 and ν_3, respectively. We aggregate the two effects to produce the total immune effect B_{it}, which reduces the tumor growth rate. Note that if $B_{it} > \nu_1$, then the tumor will shrink, thereby allowing the model to accommodate a curative effect. The sensitive T cells are recruited by the radiation at rate τ, and RT dose at time t is denoted d_{it}. Here S_{it} is the fraction of tumor cells that survive the RT pulse, which we assume follows the standard linear-quadratic model. The linear term encodes the likelihood of a single DNA-strand break, while the quadratic term encodes a double-strand break, with magnitudes determined by the parameters $\nu_4 > 0$ and

$\nu_5 > 0$, respectively. R_{it} encodes the accumulated effect of RT from previous days, which decays at a rate of $\nu_6 > 0$. RT kills the sensitive T cells at rate $\exp(-\nu_7 d_{it}), \nu_7 > 0$, and non-sensitive T cells at rate $\exp(-\nu_8 d_{it}), \nu_8 > 0$.

The IO concentration at time t for subject i is denoted $p1_{it}$: Specifically, our model aims to capture the effect of anti-PDL1 immunotherapy, which permits the T cells to more effectively attack the tumor. For simplicity, we assume that the immune effect in the absence of IO is negligible: The T cells attack the tumor only in the presence of a non-zero concentration of IO. Sensitive T cells are converted to non-sensitive T cells with probability $\lambda \in (0,1)$. The parameters $\nu_1, \ldots, \nu_9$ are fixed, and τ, λ potentially vary between animals such that

$$\begin{pmatrix} \text{logit}(\lambda_i) \\ \\ \ln(\tau_i) \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} \mu_\lambda \\ \\ \mu_\tau \end{pmatrix}, \begin{pmatrix} \sigma_\lambda & 0 \\ \\ 0 & \sigma_\tau \end{pmatrix}\right].$$

Table 2.1 presents the model parameters together with their biological interpretations and population averages when fit to the data of Moore, et al (2021). This model is presented and discussed in detail in an in-progress manuscript which is anticipated to be published shortly (Xing et al. 2023).

Table 2.1: *Parameters of the non-linear model along with their respective interpretations*

Parameter	Interpretation	Population Average
v_1	Unconditional tumor aggression	0.216
v_2	Tumor control due to non-sensitive T cells	0.02
v_3	Tumor control due to sensitive T cells	0.01
v_4	Tumor control due to RT-induced single-strand DNA damage	0.024
v_5	Tumor control due to RT-induced double-strand DNA damage	0.0014
v_6	Decay rate of RT effect	8.7
v_7	Recession of non-sensitive T cells due to RT	0.05
v_8	Recession of sensitive T cells due to RT	0.96
v_9	Incremented tumor control rate due to to previously administered RT	0.88
τ	T-cell recruitment rate due to RT	1.71
λ	T cell conversion rate (sensitive $\rightarrow$ non-sensitive)	0.304

2.3.3 Model specifications

Data generation

We set $t_w = 7$ (i.e., a week-long waiting period after the first RT pulse before the treatment day) and $A = 9$ potential actions (corresponding to application of the second pulse on one of the 9 days after that). We set the date of the initial RT pulse to $t_{\text{rt0}} = 15$.

When generating data from the sum of exponentials model, we fixed the parameter vector $(\alpha_0, \alpha_1, \delta, \sigma) = (0.1, 0.216, 0.5, .001)$, and allowed ρ, ω to vary by individual as

$$\begin{pmatrix} \text{logit}(\rho_i) \\ \\ \ln(\omega_i) \end{pmatrix} \sim \mathscr{N} \left[\begin{pmatrix} \text{logit}(.975) \\ \\ \ln(1) \end{pmatrix}, \begin{pmatrix} 2 & 0 \\ \\ 0 & 0.1 \end{pmatrix} \right].$$

When generating data from the recursive model, we fixed $(\nu_1, \nu_2, \ldots, \nu_9)$ at their fitted values and set $\sigma = .001$ as before. Finally, for the i^{th} individual, we generated λ_i, τ_i as

$$\begin{pmatrix} \text{logit}(\lambda_i) \\ \\ \ln(\tau_i) \end{pmatrix} \sim \mathscr{N} \left[\begin{pmatrix} \text{logit}(0.304) \\ \\ \ln(1.707) \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ \\ 0 & 3 \end{pmatrix} \right].$$

Note that the majority of the between-subject variability is attributed to the radiation recruitment effect, τ_i. This agrees with Kosinsky, et al (2018), who found that the most important variable for explaining between-subject variability was a term which encoded the capacity of T cells to infiltrate the tumor (which in our model corresponds to entering the tumor microenvironment).

Following Moore et al (2021), we assumed that the anti-PDL1 drug was administered four times per RT dose, every two days starting two days before the RT day. We assume each anti-PDL1 dose results in an effective concentration of 1 with a 7-day linear decay; $p1_{it}$ is the aggregate of these effective concentrations for subject i at time t. To account for carrying capacity, we assumed $p1_{it} < 1.5$.

Predictive Models

For the nonlinear and the linear model, we performed an initial calibration using 30 simulated mice with actions chosen randomly. We then applied the sequential estimation/prediction procedure to a set of 30 additional training (burn-in) mice, respectively. Next, we simulated $n_t = 40$ more individuals for testing, applying an evaluation procedure described in Section 2.3.4 to each.

We trained Q^{NN} on 436 additional observations, for a total of 500 training individuals. Next, we generated 50 additional individuals, applying the testing procedure to each of these individuals as before. Note the much larger sample size required to train Q^{NN} than Q^{LM} (discussed below). We used ε-greedy action selection with decay $\kappa = .99$. We repeated the training and evaluation in $M = 1000$ Monte Carlo iterations.

2.3.4 Evaluation

For each simulated animal, we computed a predicted curve and final-day tumor volume under each potential action, and ranked them all by the post-RT nadir, with 1 corresponding to the lowest counterfactual nadir. We then ranked each action according to the predicted lowest nadir from Algorithm 1. We considered evaluating Q by computing the correlation between the projected ranks and the true ranks, but this metric has drawbacks which make it undesirable: First, since action 1 was the most frequently occurring optimum by a substantial margin, it is possible to achieve a mean correlation very close to 1 just by predicting the same ranks for every individual, thus defeating the point of a personalized radiotherapy plan. Further, since only one action can actually be applied, ranking the suboptimal actions correctly is less important than identifying the true optimal action, which the correlation coefficient does not capture: Consider the case where actions 1–9 have ranks 5,4,3,2,1,6,7,8,9, respectively (i.e. true outcome is asymmetric in the action space). Suppose we have two versions of Q: $Q^{(1)}$, under which the projected ranks are 1,2,3,4,5,6,7,8,9, and $Q^{(2)}$, under which the projected ranks are 9,8,7,6,1,2,3,4,5. Clearly, $Q^{(2)}$ is better here, because it points to a better radiotherapy schedule (the best, in fact). However, the correlation of $Q^{(2)}$'s ranks with the true ranks is $-.06$, whereas the correlation between $Q^{(1)}$'s ranks and the true ranks is 0.66.

To show that our method is effective, we wish to show that the actions chosen adaptively, conditional on the observed data, are better than the marginal optimum; that is, we want to show that our method chooses better actions (i.e. actions with lower true ranks) than any action-selection policy that treats all patients identically. We therefore selected actions under a variety of scenarios: One corresponding to the lowest predicted nadir per Algorithm 1, and nine where we applied the same action to all animals (one for each $a \in \{1,\ldots,9\}$). We then computed, for each reference, the mean difference in ranks for each individual.

For example, to compare the selections of the system Q to the fixed spacing corresponding to applying action 6 to all individuals, we compute $\zeta_{i,\mathrm{ref}=6} = 7-3 = 4$, which indicates that the action chosen by Q is 4 ranks better than action 6 for individual i. Conversely, if the fixed action was better than the selected action, then the difference in ranks would be negative. For example, suppose that in the same scenario where we apply action 6 to all individuals, 6 is the true optimal action for subject i (that is, its true rank is 1), and suppose action 1 is the worst (with rank 9). If the lowest predicted counterfactual curve under Q corresponds to action 1, then the rank difference for subject i is $\zeta_{i,\text{ref=6}} = 1 - 9 = -8$. We used the average of these rank differences across all subjects, denoted $\bar{\zeta}_{\text{ref}} = \sum_{i=1}^{n} \zeta_{i,\text{ref}}/n$, to estimate the average rank increase of the actions selected using Q relative to each reference. If $\bar{\zeta}_{\text{ref}} > 0$ for all reference actions, we conclude that Q provides a useful policy for action selection.

2.4 Results

Table 2.2 displays the average rank differences for each version of the method, as applied to each version of the simulation (sum of exponentials and recursive). When the data come from the sum of exponentials model, they are all positive, indicating that on average the action selection policy corresponding to $\arg\min Q$ produces better actions than the policy that applies any single action to all individuals. The neural network outperforms the linear model, but at great price, requiring a quantity of data that would be impractical to collect in a real animal experiment.

The nonlinear model works reasonably well when the parametric assumptions of the analysis model match those of the generating model. However, it is less robust than the

Table 2.2: *Average rank differences $\bar{\zeta}_{ref}$ for each reference action predictive system Q, and from the simulated data. Positive values indicate that the actions selected by Q were better — i.e. had lower ranks — than the corresponding reference. Parenthetical values in Q^{LM}, Q^{NN} cells were trained using* $500, 60$ *training individuals, respectively.*

	Sum of Exponentials			**Recursive**		
Reference Action	Q^{NL}	Q^{LM}	Q^{NN}	Q^{NL}	Q^{LM}	Q^{NN}
1	0.42	0.32	0.76	−0.52	0.10 (0.11)	0.27 (−0.45)
2	0.69	0.99	1.02	−0.26	0.22 (0.23)	0.38 (−0.35)
3	1.30	1.72	1.61	0.44	0.75 (0.76)	0.94 (0.20)
4	2.13	2.46	2.45	1.36	1.55 (1.57)	1.77 (1.03)
5	3.10	2.24	3.42	2.43	2.60 (2.60)	2.84 (2.09)
6	4.14	4.04	4.46	3.46	3.62 (3.60)	3.86 (3.12)
7	5.19	4.84	5.53	4.48	4.63 (4.63)	4.89 (4.13)
8	6.25	5.67	6.58	5.50	5.67 (5.66)	5.93 (5.18)
9	7.30	6.48	7.63	6.54	6.74 (6.73)	6.99 (6.24)

other two methods, even underperforming several references when the data are from the recursive model.

To highlight the efficiency/performance tradeoff between Q^{LM} and Q^{NN}, we re-ran the simulation on the recursive data using 500 observations (the number used for main results from Q^{NN}) to train Q^{LM} and 60 observations (the number used for main results from Q^{LM}) for Q^{NN} (results in parentheses in Table 2.2). Whereas Q^{NN} works well only with a large sample size, the performance of Q^{LM} is satisfactory even with a small sample size. Overall, Q^{LM} performs better on samples of a realistic size.

We also computed the average prediction error on the estimates of log tumor volumes at the post-RT nadir for all individuals used for testing, standardized by the variability of the true log tumor volumes at the true post-RT nadir. This standardized average prediction error for Q^{LM} (Q^{NN}) [Q^{NL}] was 0.75 (0.61) [0.28] when the data were generated from the sum of exponentials model, and 0.21 (0.97) when the data came from the recursive model, indicating that in all well-performing cases the between-individuals sums of squares exceeded the sum of squared prediction errors on the test set (since Q^{NL} did not perform well on the data from the recursive model, one would not expect that its predictions are accurate). When the data came from the sum of exponentials model, the error on the estimated parameters converged to zero as the sample size increased, with $\hat{\alpha}_0, \hat{\alpha}_1$ converging first (almost instantaneously). Residuals were homogeneously scattered about 0 and did not give cause for concern. The fitted coefficients on the fixed linear and quadratic terms in the spline model had differing signs (where the linear term was negative and the quadratic positive); this implies that according to the fitted spline model, after radiation the tumor growth rate decreases and then recovers, as expected.

2.5 Discussion

RT schedules adaptively chosen by Algorithm 1 generally performed better than any fixed RT schedule applied to all individuals. Q^{NL} worked well when its parametric assumptions were met, but its performance deteriorated under departures from these assumptions. Q^{NN} outperformed each reference by a greater margin than Q^{LM}, but took 500 observations (over eight times the amount of data used for Q^{LM}) to achieve this performance. The magnitude of this data requirement is not an anomaly: While the necessary sample size for stable convergence on a neural network largely depends on the size of the network and the complexity of the task, networks of similar size and structure tend to perform well on medical datasets ranging from 200 to 2000 observations (Sargent 2001). Nevertheless, 500 is an unreasonable number of observations to expect from a practical animal experiment. Because of this practical infeasibility in tandem with the difficulty in validating parametric assumptions in practice, we recommend Q^{LM} as a predictive system for the proposed adaptive experiment.

The question of efficiency is of utmost importance to the tumor xenograft experiment and to any human trial that might follow. Further, model robustness is imperative when attempting to bridge the gap from animal to human trials. Hence, a more extensive evaluation of a variety of specifications of Q is a compelling direction for future work.

There are several ways one could modify Q^{NN} in order to resolve the large data requirement problem: For example, one might apply domain randomization, a tool primarily used in robotics, to augment the real data with synthetic data simulated from a model such as the recursive model described above. Other efforts might include transfer learning from a similar task, or using a pre-trained network with the last layer replaced.

Another direction for future work might investigate ways of defining and estimating Q^{NL} to increase robustness. Candidates for the predictive system in Q^{NL} might include a Gompertz or logistic model, which have been shown to fit well to tumor growth data, or a generalized version of the sum of exponentials model described here (Vaghi et al. 2020). A potential generalization might include models that allow tumor volume to shrink to 0 instead of regrowing after some delay. Nevertheless, the efficiency of Q^{LM} relative to Q^{NN} suggests that imposing some nonlinear structure on the predictive system might repay the modeling effort, and thus a thorough investigation of nonlinear modeling techniques would be useful.

Given our stepwise approach to parameter estimation (first estimating α_0, α_1, then treating them as fixed while other parameters are estimated), the fast convergence of their estimates is encouraging. Furthermore, since these two parameters - which we treat as fixed - can be estimated from control data, and since labs often perform multiple experiments, producing realistic estimates of α_0, α_1 *in vivo* is not a concerning issue.

A limitation of our method is that it only considers point estimates of the forecasted tumor volumes. Future endeavors might benefit from an analysis of different ranking procedures which take into account the uncertainty on these predictions: For example, one might produce a confidence interval for each estimated nadir value, then choosing the one which has the lowest upper (or lower) bound. Minimizing the upper bound would correspond to a pessimistic approach, i.e. choosing the action that has the most favorable worst-possible outcome, whereas choosing the action corresponding to the lowest lower confidence interval bound would correspond to an optimistic approach, i.e. choosing the action that has the highest potential (most favorable outcome). Investigating combinations of the above methods might provide additional benefit.

2.6 Conclusion

We have presented an adaptive method for determining the optimal timing of RT in tumor xenograft experiments, demonstrating the efficacy of three different versions of the method *via* simulation. We observed that the predictive system characterized by a spline mixed model provides the best balance of efficiency and robustness. Our work, moreover, establishes precedent for not only an adaptive tumor xenograft experiment, but also a potential subsequent adaptive human trial.

CHAPTER 3

Optimizing radiotherapy delivery schedules in combination with immunotherapy using Reinforcement Learning

3.1 Introduction

Two common treatments for solid tumors are radiotherapy (RT) and immunotherapy (IO). Recent developments in radiobiology have shown the effectiveness of combination therapy including administration of both treatment modalities, wherein one attacks the tumor directly with RT while simultaneously bolstering the subject's immune system with IO (Deng et al. 2014).

The mechanistic interactions of RT with IO are not yet fully understood, which makes it difficult to determine their optimal settings (such as timing and dose) in combination therapy. Arina, et al (2019) proposed that RT has a dual effect, both killing the tumor directly and stimulating the subject's immune system by recruiting new T cells. They moreover suggest that the newly recruited T cells are more sensitive than the existing T cells to DNA damage by subsequent RT pulses; hence if one applies several RT pulses too close together, the newly recruited T cells are killed, diminishing the synergy of RT and

IO. Moore et al (2021) provided evidence for this hypothesis via an *in vivo* tumor xenograft experiment, showing that the degree of synergy between RT and IO heavily depends on the timing of the RT pulses: If one applies two pulses of RT 10 days apart, the addition of IO generally has a large effect; however, if one applies these same two pulses of RT one day apart, the effect disappears. Moreover, due to subject-specific factors such as individual variation in innate immune response, the optimal timing of the RT pulses may differ between subjects (Kosinsky et al. 2018). A system for determining personalized optimal RT schedules for use in combination with IO would therefore be of great utility.

Previous researchers have developed similar systems for personalized medicine and adaptive therapy in different contexts. For example, Hassani (2010) used reinforcement learning to develop an optimal chemotherapy schedule for patients with progressive cancer; Kosinsky (2018) used a mathematical method which involved fitting a Bayesian mixed non-linear model with random effects corresponding to individual subjects; and Yauney and Shah (2018) used Reinforcement Learning (RL) to optimize chemotherapy and clinical trial dosing regimens. All of these works use – in full or in part – the following process for developing systems for automated, adaptive regimen selection:

1. Construct some virtual environment that mimics the response-adaptive patient dynamic of interest;
2. Train an artificial intelligence (AI) agent in this virtual environment;
3. Fine-tune and test the predictions of the AI *in vivo*.

We follow this paradigm in the current work.

The remainder of this chapter is structured as follows: First, we discuss the virtual environment used to train our AI agent. Next, we introduce the reinforcement learning framework that we use for our agent training, and describe the agent's mechanisms in reinforcement learning parlance. We then discuss various methods for evaluating the agent's performance, including a virtual clinical trial. Finally, we use transfer learning to adapt our agent from the virtual world to a real-world setting, evaluating its predictive power on a real dataset with and without calibration.

3.2 Methods

3.2.1 Virtual Environment

We trained our agent in a virtual environment governed by the same difference equation model described in Chapter 2. Recall that, using the fitted parameters, the model captures the differing average effect of immunotherapy with adjustments to radiotherapy timing (Figure 3.1).

To incorporate between-individual variability, we applied noise to 3 of the 11 parameters in the model: These parameters are i) ν_1, the unconditional tumor growth rate, ii) λ, the rate at which new, sensitive T cells become non-sensitive T cells, and iii) τ, the T cell recruitment rate due to radiotherapy. The second and third of these parameters appeared to be the most important, of the 11 model parameters, for determining the optimal scheduling: Adjustments to these two parameters caused the optimal timing to move. This result

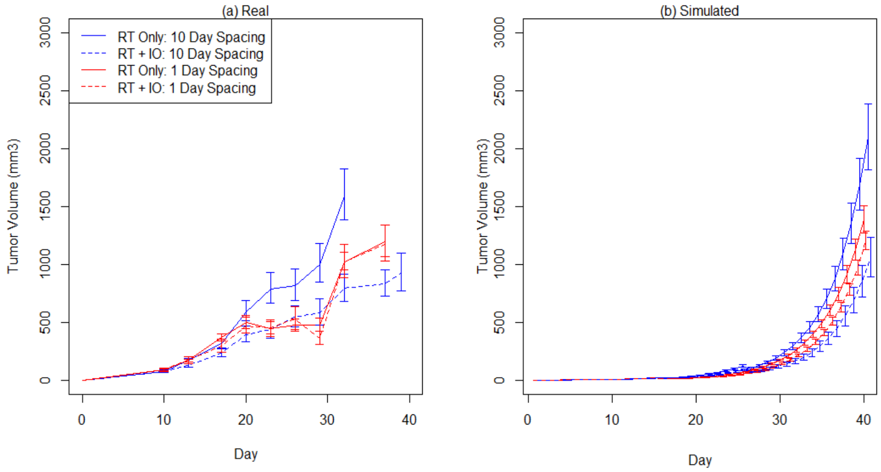

Figure 3.1: *Groupwise plots of the real (a) and simulated (b) tumor growth data under various conditions: Two pulses of RT applied 1 and 10 days apart (characterized by red and blue color, respectively), with and without IO (characterized by dashed and solid line types, respectively).*

agrees with Kosinsky (2018); in their model, the most important parameter for explaining the between-animal variability corresponded to the ability of T cells to infiltrate the tumor. In our model, this effect corresponds to the T cell recruitment rate.

Because the model is non-linear and defined through a set of difference equations, estimating the variability on these parameters from the data is challenging. In this project we used a heuristic approach, visually comparing the groupwise variances for the real and simulated data and selecting the values which resulted in similar variances (Figure 3.1). Each animal in the virtual world, therefore, is characterized by a unique set of parameters. The data visualized in Figure 3.1-(b) were generated from the recursive model with distributions placed on v_1, λ, τ s.t.

$$\begin{pmatrix} \text{logit}(\nu_1) \\ \\ \text{logit}(\lambda) \\ \\ \ln(\tau) \end{pmatrix} \sim \mathscr{N}\left(\begin{pmatrix} \text{logit}(\hat{\nu}_1) \\ \\ \text{logit}(\hat{\lambda}) \\ \\ \ln(\hat{\tau}) \end{pmatrix}, \text{diag}_{3\times 3}(0.05, 1, 1)\right) \tag{3.1}$$

where $\hat{\nu}_1, \hat{\lambda}, \hat{\tau}$ are the marginal estimates of ν_1, λ, τ when fit to the data of Moore, et al (2021). Here, we assume normality on the logit and log scales to bound the parameters within their respective spaces.

Following Moore, et al (2021), we fixed the timing of the first RT pulse at 15 days post-implantation. Because our model does not account for toxicity, we constrained the number of RT pulses to 2; our task, therefore, reduces to the selection of a day for the second pulse. For consistency with the data of Moore et al (2021), we considered days 16 through 20 for the second pulse, corresponding to spacings between 1 and 10 days.

We first applied an exhaustive search to determine the most frequently occurring optimal spacings, defining an "optimal" spacing as the one that maximizes synergy between the two treatment modalities: For each set of parameters and each potential day for the second pulse of RT, we computed two counterfactual curves, one where immunotherapy is present, and one where it is absent. We then approximated the area between the curves (shaded in gray) in Figure 2 by computing the sum of the difference in tumor volumes over all days beyond the date of the first immunotherapy treatment. We deem a radiotherapy regimen to be optimal if and only if it maximizes this sum of differences.

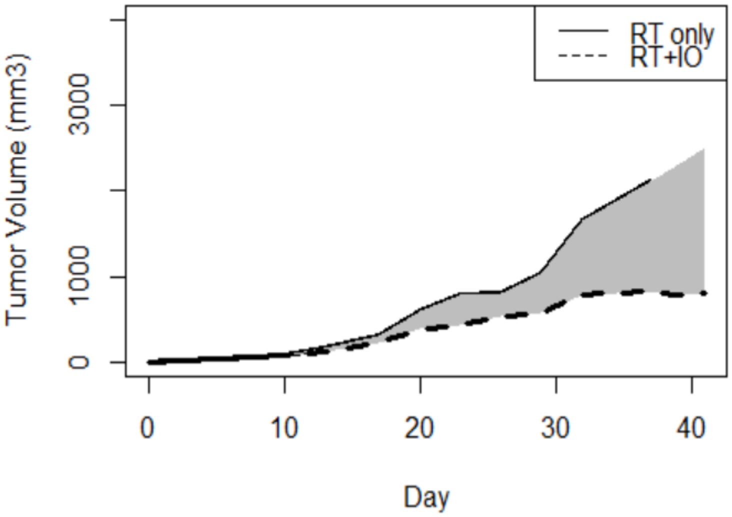

Figure 3.2: *Diagram illustrating the evaluation metric, using the data from two experimental groups investigated by Moore, et al (2021): Both groups received 10Gy radiation 10 days apart. The solid and dashed lines represent cross-sectional averages from the treatment groups that received RT only, and RT as well as IO, respectively. The shaded area, therefore, represents the effect of IO and is what we wish to maximize by optimizing the timing of the second radiotherapy pulse.*

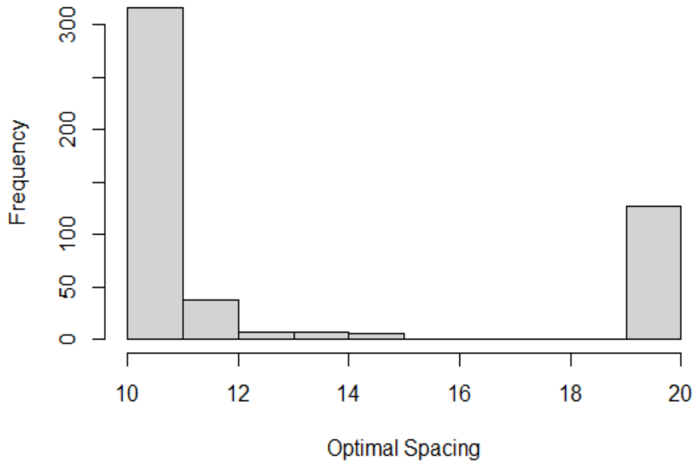

Figure 3.3: *Frequencies of optimal time between two pulses of 10 Gy radiation, per the simulated tumor microenvironment governed by a set of differential equations with noise applied to its parameters.*

Figure 3.3 presents the results of the exhaustive search: In 500 simulated test subjects, the 10-day spacing was optimal in the majority of cases, providing the maximum added benefit of IO 42.8% of the time (214 cases). We therefore trained our agent to select among 11 actions, corresponding to 10-day through 20-day spacing (inclusive).

The optima in Figure 3.3 do not follow a Gaussian distribution, possibly because the governing mathematical model is nonlinear, with partial derivatives/differentials which are not always well-defined.

3.2.2 Reinforcement Learning

In order develop an AI agent capable of obtaining optimally synergistic RT spacings, we use Reinforcement Learning (RL) (Kosorok and Moodie 2015; Yauney and Shah 2018; Hassani and Naghibi-S 2010); specifically, we use Q-learning with neural network-based function approximation as described by Mnih, et al (2015). Application of RL requires framing the problem as a Markov Decision Process (MDP). In this paradigm, the current model is an episodic MDP with one action per episode: The subject is the episode, and the selection of the day of the second RT pulse is the action. In our virtual environment, each subject has a unique set of parameters in the difference equation model.

We define a state s as a matrix of three column vectors: 1) previously observed tumor volumes at each time, 2) previously administered radiotherapy doses at each time, and 3) previously administered immunotherapy concentration at each time. Our action space $\mathscr{A}$ is the set $\{10,11,12,13,14,15,16,17,18,19,20\}$ of ten potential spacings between the two RT pulses, and the reward is the aggregate causal effect of immunotherapy on log tumor volume over the last three days of observation. That is, for each individual, we consider two counterfactual curves (i.e., vectors of log tumor volumes): One where the subject receives immunotherapy, and one where it does not. Denote these two vectors of values as $Y^{\text{io}+}, Y^{\text{io}-}$, respectively, with elements y indexed by a time variable $t = 1 \ldots 40$. We define our reward as $r = \left(\sum_{t=13}^{4} 0(Y^{\text{io}-} - Y^{\text{io}+})\right) I(s\text{is terminal})$, where I is the indicator function and the summation begins at day 13 because the first dose of immunotherapy is applied 13 days after implantation. The state transition probabilities are governed by the virtual environment characterized by our difference equation model. High rewards therefore correspond to states where the addition of immunotherapy causes a large reduction in tumor growth.

We train our agent using Q-learning, wherein one treats each subject according to the best predicted action per some function Q, whose input is the current state Q and a potential action Q, and whose output is the long-term reward. We use a neural network to approximate Q, as discussed by Mnih, et al (2015). After random initialization or initialization on a preliminary calibration set, one updates the parameters of Q by one step of gradient descent after each observation until convergence. We consider two versions of Q-learning: An offline version, where actions are chosen randomly before convergence, and an online version, where all post-initialization actions are chosen to maximize $\hat{Q}$, the estimate of Q (Riedmiller 2005). For example, suppose Q requires $n_q > 10$ observations in order to achieve stable convergence, and we use 5 observations for initial calibration. Let a_{10} denote the 10th action applied. The offline version of the agent would draw a_{10} randomly from the set of potential actions: $a_{10} \sim DU(\mathscr{A})$. Conversely, the online version would select the action which its internal Q function predicts will be the best: $a_{10} =$ max $\text{arg}_{\mathscr{A}} Q(s_{10}, a)$, where s_{10} is the 10th observed state.

We implemented our RL framework by hand in the programming language R (R Core Team 2021), using the function `neuralnet` from the package `neuralnet` to fit the Q function (Fritsch et al. 2022).

3.2.3 Evaluation

Primary Evaluation

We evaluated the performance of our agent on a testing set of $n_t = 100$ simulated individuals, indexed $i = 1, \dots, n_t$. Recall that each simulated individual is characterized by a set of parameters to the difference equation model. We can therefore simulate what would happen to that individual under a variety of different scenarios corresponding to different days for the second RT pulse (only one of which, of course, could actually be applied). For each subject, then, we first computed the set of counterfactual rewards, each of which corresponds to a different space between the two RT pulses (considering, as before, 1–20 potential days between these two pulses). From the 20 potential rewards, we retrieved the true optimal action, i.e. the action that maximizes the added benefit of immunotherapy; for subject i, denote this optimal action $a_{i,\text{opt}}$. To evaluate the quality of an action applied to a given individual, we compute the difference (in days) between that action and $a_{i,\text{opt}}$, and average these scores across all individuals in the testing set to evaluate whatever policy was used to generate those actions. Let $a_{i,\pi}$ be the action selected by some policy π for subject i. We refer to the difference between each selected and optimal action, denoted $\Delta_i = a_{i,\pi} - a_{i,\text{opt}}$, as the *optimal-day miss difference*, because it is the difference between the selected day and the optimal day, i.e. the number of days by which the action selection policy "missed" the optimum. The optimal-day miss difference can be positive or negative: Negative values indicate that the day selected for the second pulse of RT was too soon, whereas positive values indicate that the selected day was too late. The optimal-day miss differences can also be squared and averaged across the testing set to obtain an MSE-type value for policy π: We consider here the quantity $\text{RMSE}(\pi) = \sum_{i=1}^{n_t} \Delta_i^2$ as a performance metric for policy π. We computed the optimal-day miss differences across all observations in the testing set for a variety of different action selection policies:

1. $\pi =$online: A policy where actions were selected by the online version of our agent
2. $\pi =$offline: A policy where actions were selected by the offline version of our agent
3. $\pi =$random: A random action selection policy (where the day of the second pulse was chosen randomly from 1–20 days)
4. $\pi = a, a \in \mathscr{A}$: 20 policies, each corresponding to uniform application of a spacing to all individuals in the testing set

Note that some of the policies described in points iii-iv involve the application of RT spacings unselectable by either version of the agent: Our agents wait 9 days after the first pulse of RT to decide on the timing of the next pulse. Hence, for example, the 1-day spacing is not selectable by the policies $\pi =$online, offline (corresponding to actions selected by the online and offline versions of the agent).

Virtual Clinical Trial

We also evaluated our agent by performing a virtual "clinical trial", where 1,000 simulated animals are randomized to one of four treatment arms: An arm where actions are chosen randomly, an arm where actions are chosen by the online agent, an arm where actions are chosen by the offline agent, and an arm taking action 10, the best action per the data of Moore, et al (2021). After selecting actions according to the mechanism indicated by the individual's corresponding treatment arm, we computed the difference between the selected action and the optimal action (described above) for each arm. We also investigated the difference in tumor growth between groups by fitting a simple linear mixed model to the tumor growth on the last five days, after all treatments have been applied:

$$\begin{aligned} \ln Y_{it} &= b_{0i} + \beta_{0,\text{offline}} I(\text{offline}) + \beta_{0,\text{online}} I(\text{online}) + \beta_{0,\text{random}} I(\text{random}) \qquad (3.2) \\ &+ (b_{1i} + \beta_{1,\text{offline}} I(\text{offline}) + \beta_{1,\text{online}} I(\text{online}) + \beta_{1,\text{random}} I(\text{random}))t \\ &+ e_{it} \end{aligned}$$

In Equation (3.2), Y_{it} is the tumor volume from subject i at time t days post-implantation, $I(A)$ is the indicator function of treatment assignment to arm A, and $e_{it} \sim N(0, \sigma^2)$. Each Greek letter represents a scalar coefficient to be estimated. We assume subject-specific effects on the intercepts and unconditional growth rates, denoted $b_{i,0}, b_{i,1}$ respectively, with $b_{i,0} \sim N(\beta_0, \sigma^2_{b_0}), b_{i,1} \sim N(\beta_1, \sigma^2_{b_1})$, and $b_{i,0}, b_{i,1}$ independent. If our agent works, and if the model captures the mechanism of interest, one expects the agent-treated tumors to grow more slowly in the final days of the experiment, i.e., $\beta_{1,\text{online}}, \beta_{1,\text{offline}} < 0$.

Indirect Validation on Real Data

We tested our agent indirectly by evaluating its predictive performance on the data of Moore, et al (2021). Of course, in reality, each individual can only receive one treatment, so we do not directly observe the outcome that we desire to optimize (i.e., the space between the curves corresponding to IO and no IO for each individual). Instead, we used techniques from causal inference to impute these counterfactual values for each mouse (see Appendix A). For the mice who received immunotherapy, the counterfactuals correspond to the values which theoretically would have occurred had they not received immunotherapy. Conversely, for the mice who did not receive immunotherapy, we generated similar values corresponding to their outcomes had they received immunotherapy. The counterfactual values, moreover, can be thought of as the tumor growth values from a "digital twin" which received the other treatment. For simplicity, we focus only on the groups which received two pulses of RT, each with the same dose (10Gy).

With these imputations in hand, we can compute, for each individual, the sequence of log tumor volume differences $Y^{\text{io}-} - Y^{\text{io}+}$, where $Y^{\text{io}-}$ corresponds to the sequence of log tumor volumes for the case where that individual received RT only, and $Y^{\text{io}+}$ corresponds to the sequence from the case where that individual received IO as well as RT. If the individual truly received both treatments, then the values of the vector beyond the first date of IO application are counterfactual (imputed mathematically using the method described in Appendix A); conversely, if the individual received radiation only, then the values beyond the date of first IO treatment are counterfactual. As before, we used the summation of the difference vector $Y^{\text{io}-} - Y^{\text{io}+}$ to approximate the area indicated in Figure 3.2.

In order to bridge the gap from simulation to reality, we calibrated our model's predictions on a subset of the real data, fitting a simple linear model which relates the predicted outcomes (using the online agent) to the real values. We tested the performance on the remainder of the data. To measure the efficacy of this calibration, we compared the prediction error on the testing set between the calibrated and the uncalibrated versions of the neural network. We evaluated the overall predictive accuracy of our agent by comparing the average prediction error (MSE) to the variance on the true values.

3.3 Results

3.3.1 Primary Evaluation

Figure 3.4 shows the empirical cumulative densities of $\Delta_{i,\pi}$(i.e. across all subjects in the testing set), under a variety of different policies π: These policies are a representative subset of all of the tested policies discussed above. The numbers in the legend correspond to the RMSE(π) values described above. Note that these RMSE values are much lower under the agent-driven (online and offline) action selection policies: This indicates that on average, the online and offline agents pick actions closer to the true optima than one could achieve by applying the same treatment to all individuals, or by picking actions randomly. For simplicity, we show here only four fixed curves: Day 1 and Day 20; the extreme values

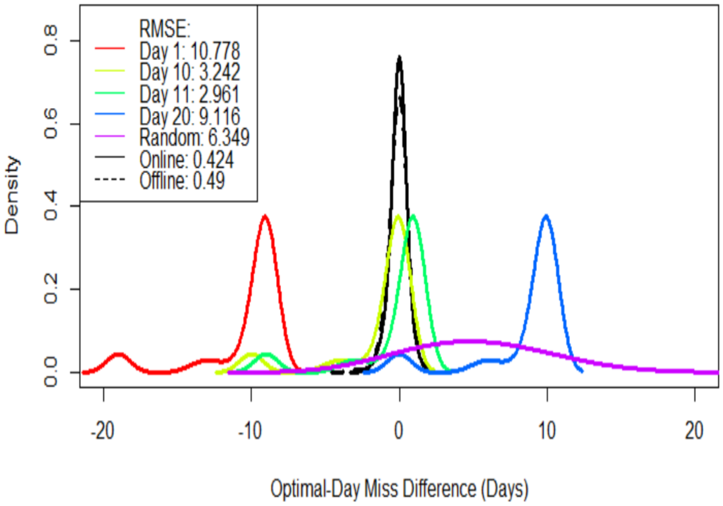

Figure 3.4: *Empirical density plots (smoothed histograms) of optimal-day miss differences – i.e. the difference between selected action and true optimal action – with actions chosen under a variety of different policies. The solid and dashed black curves characterize the densities of optimal-day miss differences using actions generated by the online and offline versions of the agent, respectively. The curve labeled "random" characterizes the optimal-day miss differences under random action selection, and the remaining curves characterize optimal-day miss differences from uniform action application corresponding to 1, 10, 11, and 20-day RT spacings. The numbers in the legends correspond to the resulting RMSE of the optimal-day miss difference under each action selection policy.*

from Moore, et al, Day 10; the best-performing spacing observed in an *in vivo* experiment per Moore et al,(2021) and Day 11; the most frequently occurring true optimum per the results of the exhaustive search excluding days already included as references. Curves using other fixed days as references looked similar.

We also investigated the efficiency and required sample size of both versions of our agent. Figure 3.5 shows the optimal-day miss difference RMSE of each version of the agent with increasing sample size: After the 700-observation mark, all subsequent aver-

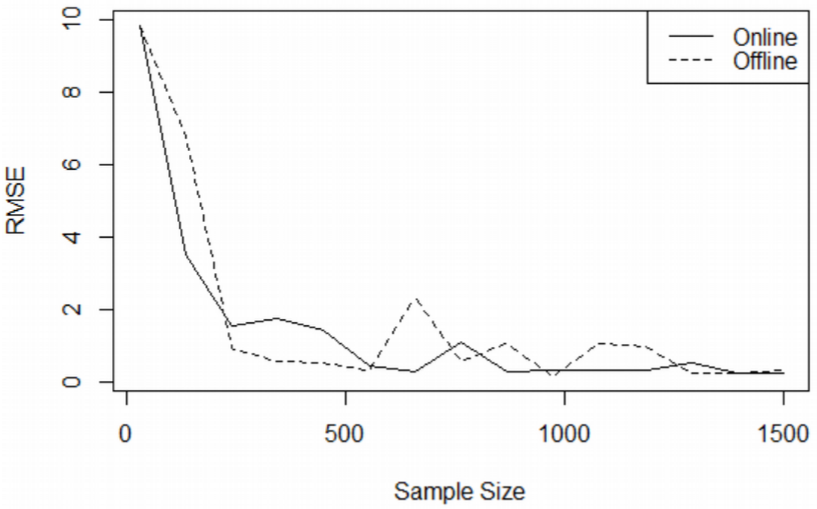

Figure 3.5: *Performance of the online and offline versions of the agent with increasing training samples. RMSE values are computed by using the online and offline versions of the agent to select actions applied to a testing set of 100: The y-axis corresponds to the RMSE between the selected and the true optimal days for the second pulse of radiation across all samples in the test set.*

age loss values were less than 1, indicating that on average, each version of the agent consistently picked the best actions on the test set when trained with $n_t = 700$.

3.3.2 Virtual Clinical Trial

Figure 3.6 shows the results from the virtual clinical trial: Note that both versions of the agent select actions closer to the true optimum than fixed spacing or random chance.

Estimates of $\beta_{1,\text{offline}}, \beta_{1,\text{online}}$ from Equation 3.2 were both significantly less than zero at $-.033$, $-.030$, respectively. This indicates that individuals randomized to treatment arms where actions were selected by our agent had slower-growing tumors post-treatment than the individuals randomized to the best-performing treatment arm from Moore, et al (2021).

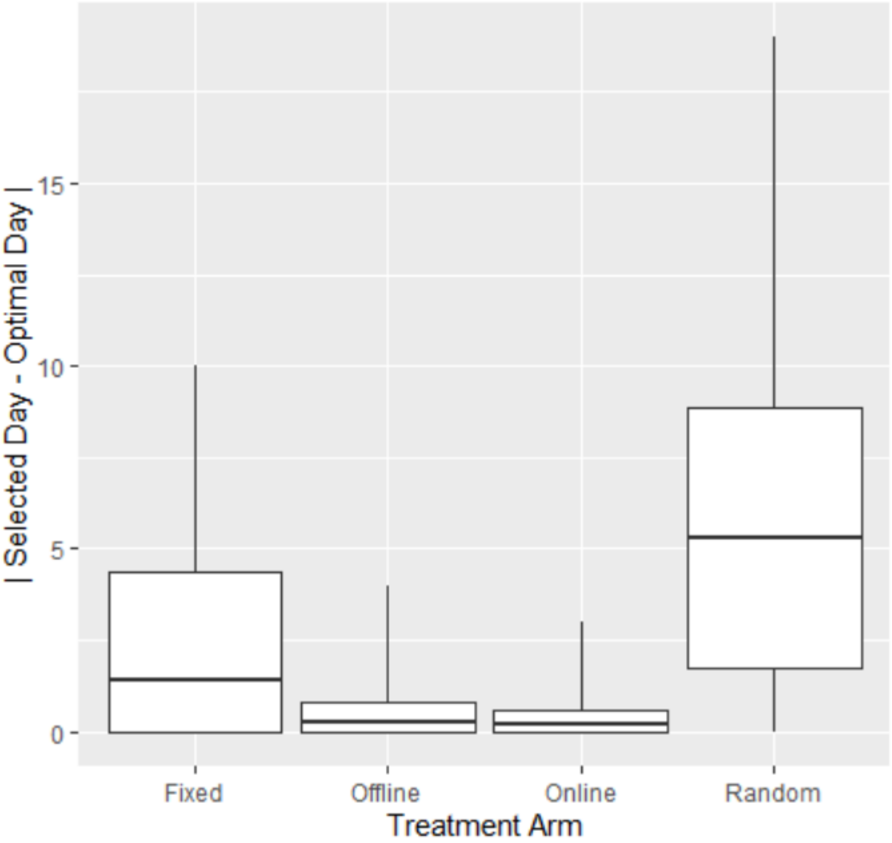

Figure 3.6: *Results from the virtual clinical trial. The black lines indicate the means. "Fixed" spacing corresponds to the application of 10-day spacing (the best-performing spacing from Moore, et al (2021)) uniformly to all individuals. Individuals randomized to "offline" were treated with actions selected by the offline version of the agent, and individuals randomized to "online" were treated with actions selected by the online version of the agent. Individuals randomized to the "random" arm were treated with randomly selected actions. Each action corresponds to a day selected for the second RT pulse.*

This is notable because both versions of the agent were trained to maximize synergy between the two treatment modalities – i.e., the reward is the sum of differences between the two counterfactuals corresponding to IO+RT and RT only; they were not trained to directly minimize tumor volume. Nevertheless, as anticipated, the maximization of synergy indirectly causes the tumors to grow slower in our experiment, providing compelling evidence for continued investigation into this line of combination therapies.

3.3.3 Indirect Predictive Evaluation

Figure 3.7 shows predictive error using the uncalibrated vs the calibrated version of our agent. Testing values were not used for calibration. Note that the predictive error is substantially lower under the calibrated model, indicating that the calibrated version produces better predictions.

The standard deviation of the residuals from the calibrated model (RMSE of model prediction – true value) was less than the between-subject standard deviation on the true values, indicating that the model accounts for a substantial portion of the variability on the testing set, as desired.

3.4 Discussion

We have trained and tested an AI agent *in silico* to find RT schedules that maximize the synergy between RT and IO, demonstrating its effectiveness in a virtual environment gov-

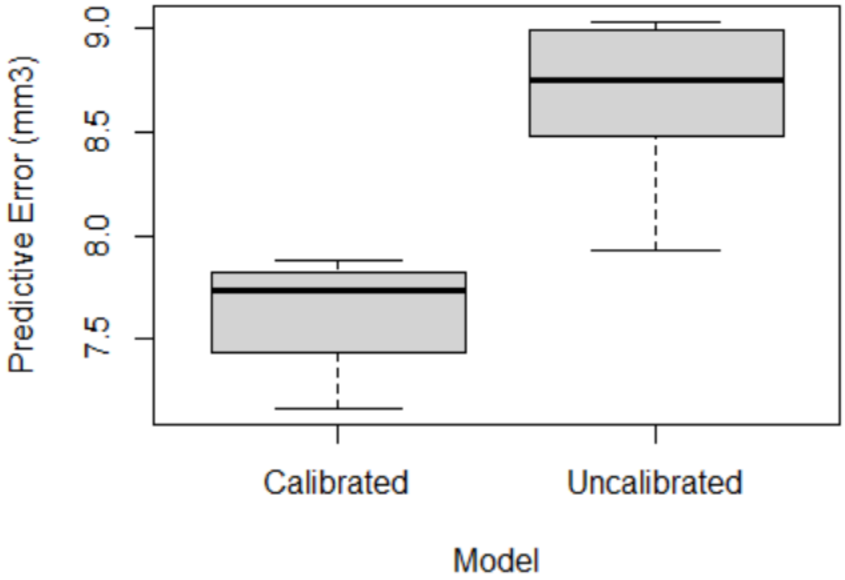

Figure 3.7: *Difference between predicted and actual synergy between RT and IO, acquired using counterfactual outcome prediction, on the testing set of real data.*

erned by a mechanistic model. The actions chosen by our agent were closer, on average, to the actions that yield the greatest synergy between treatment modalities than actions chosen at random, or any fixed spacing when applied to all subjects. Moreover, in a virtual clinical trial, individuals randomized to treatment by our agent had slower-growing tumors in the post-treatment stage than individuals randomized to the best-performing spacing of Moore, et al (2021). This result is notable, because both versions of the agent were trained to maximize synergy between the treatment modalities, as opposed to directly minimizing tumor volume.

Using the data of Moore, et al (2021), we have also shown the effectiveness of a simple calibration method to bridge the domain shift from the virtual world to a real tumor xenograft experiment. This preliminary endeavor appears to be effective; however, future research might benefit from a thorough exploration and application of more advanced methods from the Sim2Real literature. Such methods, while they have not yet been applied to radiotherapy regimen selection, have been shown to be effective in other fields, such as autonomous robotics and self-driving cars (Höfer et al. 2020). This approach would lend credibility to our method when considering a future adaptive *in vivo* experiment, and subsequent human trial.

In *in vivo* experiments, efficiency is of utmost importance. Consequently, other future studies could investigate the sample size required to calibrate these agents, and explore methods to improve efficiency.

3.5 Conclusion

We have developed a reinforcement learning agent to identify the optimal combination of RT and IO in a tumor xenograft experiment. The agent selects RT regimens that outperform the best experimentally observed regimen and random selection of the regimen. An exhaustive search performed in a virtual environment suggests that when applying two pulses of RT, 10-day spacing is best on average, and that calibrating the model on real data substantially improves its predictive accuracy. Future directions include rigorous estimation of the variability of each of the model parameters, and evaluation of the system in virtual environments that correspond to different radiobiological hypotheses. Ultimately, one could verify the results in a real adaptive tumor xenograft experiment with actions generated by the agent.

Appendix A

Imputation Procedure

Dropout is a phenomenon well-studied in the statistical literature on missing data which arises in follow-up studies when an experimental unit on which one is taking serial measurements becomes unavailable for evaluation prior to the planned end of follow-up. We consider each animal in the immunotherapy group as a dropout from the radiation-only group beyond the first date of immunotherapy administration: Subsequent values are not strictly “missing” from the radiation-only group (since these values are counterfactual rather than lost), nevertheless, it can be useful to consider them as such under the premise that we are able to recover the values that would have occurred had each of these subjects not received immunotherapy. The converse premise is identically applicable to the individuals who received radiation only. We use here established statistical techniques from the missing data literature to impute these missing values, thereby obtaining two completed datasets consisting of values from the same individuals under the two different conditions: Radiation only, and radiation in combination with immunotherapy. We then subtract the values in the combination treatment dataset from the values in the radiation only dataset to obtain estimates of the causal immunotherapy effect illustrated in Figure 3.2.

We describe the series of log tumor volumes for a mouse as a p-vector Y of observations that follows the multivariate Gaussian distribution MVN(μ, Σ). Here μ is a p-vector mean and Σ is a $p \times p$ variance-covariance matrix. We further denote $Y = \begin{pmatrix} Y_\mathrm{o} \\ \\ Y_\mathrm{m} \end{pmatrix}$, where Y_o is the p_o-vector of tumor log volume values that are observed, and Y_m is the p_m-vector of tumor log volume values that are considered missing, with $p = p_\mathrm{o} + p_\mathrm{m}$. We partition the mean of Y as $\mu = \begin{pmatrix} \mu_\mathrm{o} \\ \\ \mu_\mathrm{m} \end{pmatrix}$ and the variance-covariance matrix as $\Sigma = \begin{pmatrix} \Sigma_\mathrm{oo} & \Sigma_\mathrm{om} \\ \\ \Sigma_\mathrm{mo} & \Sigma_\mathrm{mm} \end{pmatrix}$ where, as before, the subscript “o” (“m”) refers to the observed (missing) portion of Y.

We applied a simple imputation procedure that consists of i) estimating the model parameters μ and Σ and ii) imputing individual tumor log volumes conditionally on the estimated parameter values and Y_o. We based our imputations on a mixed linear model, as implemented in the R package `lme`. This procedure fits models of the type

$$Y \sim \mathrm{MVN}(X\beta, \Sigma(\theta)),$$

where X is a matrix of predictors, β is a regression coefficient, and θ is a vector of parameters governing the variance matrix. A simple but often realistic model assumes that the error consists of two components: A mouse-specific random intercept plus an element of white noise.

Although `lme` does not directly impute missing observations, for some specifications it can draw samples from the Bayesian posterior distribution of the model parameters; with these in hand, it is straightforward to impute the missing portion conditional on the observed portion. Specifically, when the variance model is of the variance components type (including, as a special case, the random intercept model mentioned above), `lme` creates parameter estimates which characterize the posterior distribution of (β,θ). Denote one such sample as $(\tilde{\beta},\tilde{\theta}$ and $\tilde{\Sigma}=\Sigma(\tilde{\beta})$, and set $\tilde{\mu}=\begin{pmatrix}\tilde{\mu_{\mathrm{o}}}\\ \\ \tilde{\mu_{\mathrm{m}}}\end{pmatrix}$ with variance $\tilde{\Sigma}=\begin{pmatrix}\Sigma_{\tilde{\mathrm{oo}}} & \Sigma_{\tilde{\mathrm{om}}}\\ \\ \Sigma_{\tilde{\mathrm{mo}}} & \Sigma_{\tilde{\mathrm{mm}}}\end{pmatrix}$ Then, by standard normal theory, we have $Y_{\mathrm{m}}|y_{\mathrm{o}},\tilde{\beta},\tilde{\Sigma}\sim\mathrm{MVN}(\mu_{\tilde{\mathrm{m|o}}},\Sigma_{\tilde{\mathrm{m|o}}})$, where $\mu_{\tilde{\mathrm{m|o}}}=\tilde{\mu_{\mathrm{m}}}+\Sigma_{\tilde{\mathrm{mo}}}\Sigma_{\tilde{\mathrm{oo}}}{}^{-1}(y_{\mathrm{o}}-\tilde{\mu_{\mathrm{o}}})$ and $\Sigma_{\tilde{\mathrm{m|o}}}=\Sigma_{\tilde{\mathrm{mm}}}-\Sigma_{\tilde{\mathrm{mo}}}\Sigma_{\tilde{\mathrm{oo}}}{}^{-1}\Sigma_{\tilde{\mathrm{om}}}$. See also Tan et al.(2002)

Appendix B

Poetical Supplement

The schedule best for RT
Various surely will be
So spacings should switch
Depending on which
Initial response that we see!

And we have to assume that the end
On two pulses will jointly depend
Else prior pulse two
We don't have a clue
What effect the second will lend!

But to benefit from Machine Learning
We need sample size fit for a king
So classic statistics
Yield far more realistic
Methods for scheduling!

References

Ali, Akhtar et al. (2021). “Numerical simulations and analysis for mathematical model of avascular tumor growth using Gompertz growth rate function”. In: *Alexandria Engineering Journal* 60, pp. 3731–3740.

Arina, Ainhoa et al. (2019). “Tumor-reprogrammed resident T cells resist radiation to control tumors”. In: *Nature communications* 10.1, pp. 1–13.

Baranowitz, Steven (2022). “Gompertz Kinetics in Developmental Fields: An Information Theory Approach”. In: *ArXiV*.

Bates, Douglass et al. (2015). “Fitting Linear Mixed-Effects Models Using lme4”. In: *Journal of Statistical Software* 67.1, pp. 1–48. DOI: 10.18637/jss.v067.i01.

Demidenko, Eugene (2010). “Three endpoints of *in vivo* tumour radiobiology and their statistical estimation”. In: *International Journal of Radiation Biology* 86, pp. 164–173.

Deng, Liufu et al. (2014). “Irradiation and anti–PD-L1 treatment synergistically promote antitumor immunity in mice”. In: *The Journal of clinical Investigation* 124.2, pp. 687–695.

Durrleman, Sylvain and Richard Simon (1989). “Flexible regression models with cubic splines”. In: *Statistics in Medicine* 8, pp. 551–561.

Ebrahimi, Saba and Gino J Lim (2021). “A reinforcement learning approach for finding optimal policy of adaptive radiation therapy considering uncertain tumor biological response”. In: *Artificial Intelligence in Medicine* 121, pp. 102–193.

Enderling, Heiko et al. (2019). “Integrating mathematical modeling into the roadmap for personalized adaptive radiation therapy”. In: *Trends in Cancer* 5, pp. 467–474.

Frenzen, Chistopher and James Murray (1986). “A cell kinetics justification for Gompertz’ equation”. In: *SIAM Journal on Applied Mathematics* 46, pp. 614–629.

Fritsch, Stefan et al. (2022). *Training of Neural Networks*. URL: https://cran.r-project.org/web/packages/neuralnet/neuralnet.pdf.

Gallaher, Jill A et al. (2018). “Spatial heterogeneity and evolutionary dynamics modulate time to recurrence in continuous and adaptive cancer therapies”. In: *Cancer Research* 78, pp. 2127–2139.

Ghaffari Laleh, Narmin et al. (2022). “Classical mathematical models for prediction of response to chemotherapy and immunotherapy”. In: *PLoS Computational Biology* 18, e1009822.

Harrison, Richard and Richard Frolloni (2018). *On not losing my father’s ashes in the flood.* ’round midnight. ISBN: 9788898749256. URL: `https://books.google.com/books?id=K9EOuAEACAAJ`.

Hartung, Niklas et al. (2014). “Mathematical modeling of tumor growth and metastatic spreading: Validation in tumor-bearing mice”. In: *Cancer Research* 74.22, pp. 6397–6407.

Hassani, Amin and Mohammad Bagher Naghibi-S (2010). “Reinforcement learning-based control of tumor growth with chemotherapy”. In: *Proceedings of the 2010 International Conference on System Science and Engineering.* IEEE, pp. 185–189.

Heidari, Hossein, Mahdi Rezaei Karamati, and Hossein Motavalli (2022). “Tumor growth modeling via Fokker-Planck equation”. In: *Physica A: Statistical Mechanics & its Applications* 596, p. 127168.

Heitjan, Daniel F (1991). “Generalized Norton-Simon models of tumour growth”. In: *Statistics in Medicine* 10, pp. 1075–1088.

Heitjan, Daniel F, Andrea Manni, and Richard J Santen (1993). “Statistical analysis of *in vivo* tumor growth experiments”. In: *Cancer Research* 53, pp. 6042–6050.

Hernán, Miguel A and James M Robins (2020). *Causal Inference: What If.* CRC.

Höfer, Sebastian et al. (2020). “Perspectives on sim2real transfer for robotics: A summary of the r: Ss 2020 workshop”. In: *ArXiV.*

Hrinivich, William Thomas and Junghoon Lee (2020). “Artificial intelligence-based radiotherapy machine parameter optimization using reinforcement learning”. In: *Medical Physics* 47.12, pp. 6140–6150.

Jalalimanesh, Ammar et al. (2017). “Simulation-based optimization of radiotherapy: Agent-based modeling and reinforcement learning”. In: *Mathematics & Computers in Simulation* 133, pp. 235–248.

Kosinsky, Yuri et al. (2018). “Radiation and PD-(L) 1 treatment combinations: Immune response and dose optimization via a predictive systems model”. In: *Journal for Immunotherapy of Cancer* 6, pp. 1–15.

Kosorok, Michael R and Erica EM Moodie (2015). *Adaptive Treatment Strategies in Practice: Planning Trials and Analyzing Data for Personalized Medicine*. SIAM.

Little, Roderick JA and Donald B Rubin (2019). *Statistical Analysis with Missing Data*. John Wiley & Sons.

Lv, Huijun et al. (2022). “Stochastic behaviors of an improved Gompertz tumor growth model with coupled two types noise”. In: *Heliyon* 8, e11574.

Mastri, Michalis, Amanda Tracz, and John ML Ebos (2019). *Tumor growth kinetics of human LM2-4LUC+ triple-negative breast carcinoma cells*.

Mnih, Volodymyr et al. (2015). “Human-level control through deep reinforcement learning”. In: *Nature* 518, pp. 529–533.

Moore, Casey et al. (2021). “Personalized ultrafractionated stereotactic adaptive radiotherapy (PULSAR) in preclinical models enhances single-agent immune checkpoint blockade”. In: *International Journal of Radiation Oncology* Biology* Physics* 110, pp. 1306–1316.

Moreau, Grégoire et al. (2021). “Reinforcement Learning for Radiotherapy Dose Fractioning Automation”. In: *Biomedicines* 9.2, p. 214.

Padmanabhan, Regina, Nader Meskin, and Wassim M Haddad (2017). “Reinforcement learning-based control of drug dosing for cancer chemotherapy treatment”. In: *Mathematical Biosciences* 293, pp. 11–20.

Phan, Tuan Anh, Shuxun Wang, and Jianjun Paul Tian (2022). “Analysis of a new stochastic Gompertz diffusion model for untreated human glioblastomas”. In: *Stochastics and Dynamics*, p. 2250019.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: `https://www.R-project.org/`.

Riedmiller, Martin (2005). “Neural fitted Q iteration–first experiences with a data efficient neural reinforcement learning method”. In: *European conference on machine learning*. Springer, pp. 317–328.

Santen, Richard J, Wei Yue, and Daniel F Heitjan (2012). “Modeling of the growth kinetics of occult breast tumors: Role in interpretation of studies of prevention and menopausal hormone therapy”. In: *Cancer Epidemiology, Biomarkers & Prevention* 21, pp. 1038–1048.

Sargent, Daniel J (2001). “Comparison of artificial neural networks with other statistical approaches: results from medical data sets”. In: *Cancer: Interdisciplinary International Journal of the American Cancer Society* 91.S8, pp. 1636–1642.

Sheergojri, Aadil Rashid et al. (2022). “Uncertainty-based Gompertz growth model for tumor population and its numerical analysis”. In: *An International Journal of Optimization and Control: Theories & Applications* 12, pp. 137–150.

Sutton, Richard S and Andrew G Barto (2018). *Reinforcement Learning: An Introduction.* MIT Press.

Taib, Siti Farzana and Syahira Binti Mansur (2022). “Mathematical modeling in tumor growth using Gompertz model”. In: *Enhanced Knowledge in Sciences and Technology* 2, pp. 481–490.

Tan, Ming et al. (2002). “Small-sample inference for incomplete longitudinal data with truncation and censoring in tumor xenograft models”. In: *Biometrics* 58.3, pp. 612–620.

Tienderen, Gilles et al. (2022). “Extracellular matrix drives tumor organoids toward desmoplastic matrix deposition and mesenchymal transition”. In: *Acta Biomaterialia.*

Troxel, Andrea B, Guoguang Ma, and Daniel F Heitjan (2004). “An index of local sensitivity to nonignorability”. In: *Statistica Sinica*, pp. 1221–1237.

Tseng, Huan-Hsin, Yi Luo, Sunan Cui, et al. (2017). “Deep reinforcement learning for automated radiation adaptation in lung cancer”. In: *Medical Physics* 44.12, pp. 6690–6705.

Tseng, Huan-Hsin, Yi Luo, Randall K Ten Haken, et al. (2018). “The role of machine learning in knowledge-based response-adapted radiotherapy”. In: *Frontiers in Oncology* 8, p. 266.

Vaghi, Cristina et al. (2020). “Population modeling of tumor growth curves and the reduced Gompertz model improve prediction of the age of experimental tumors”. In: *PLoS Computational Biology* 16, e1007178.

Viossat, Yannick and Robert Noble (2021). “A theoretical analysis of tumour containment”. In: *Nature Ecology & Evolution* 5, pp. 826–835.

Willcox, Karen E, Omar Ghattas, and Patrick Heimbach (2021). “The imperative of physics-based modeling and inverse theory in computational science”. In: *Nature Computational Science* 1, pp. 166–168.

Xing, Yixun et al. (2023). “Mathematical Modeling of the Synergetic Effect between Cancer Radiotherapy and Immunotherapy”. In: *Unpublished Manuscript.*

Yang, Dongdong et al. (2020). “Gompertz tracking of the growth trajectories of the human-liver-cancer xenograft-tumors in nude mice”. In: *Computer Methods and Programs in Biomedicine* 191, p. 105412.

Yauney, Gregory and Pratik Shah (2018). "Reinforcement learning with action-derived rewards for chemotherapy and clinical trial dosing regimen selection". In: *Machine Learning for Healthcare Conference*. PMLR, pp. 161–226.

Zahid, Mohammad U et al. (2021). "Dynamics-adapted radiotherapy dose (DARD) for head and neck cancer radiotherapy dose personalization". In: *Journal of Personalized Medicine* 11, pp. 11–24.