Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/mbe

MBE, 20(10): 17986–18017.
DOI: 10.3934/mbe.2023799
Received: 14 June 2023
Revised: 24 August 2023
Accepted: 10 September 2023
Published: 19 September 2023

Research article

An adaptive information-theoretic experimental design procedure for high-to-low fidelity calibration of prostate cancer models

Heyrim Cho[1,†], **Allison L. Lewis**[2,†,*], **Kathleen M. Storey**[2,†] **and Anna C. Zittle**[2]

[1] Department of Mathematics, University of California, Riverside CA, USA

[2] Department of Mathematics, Lafayette College, Easton PA, USA

† The authors contributed equally to this work.

* **Correspondence:** Email: lewisall@lafayette.edu; Tel: +16103305276.

Abstract: The use of mathematical models to make predictions about tumor growth and response to treatment has become increasingly prevalent in the clinical setting. The level of complexity within these models ranges broadly, and the calibration of more complex models requires detailed clinical data. This raises questions about the type and quantity of data that should be collected and when, in order to maximize the information gain about the model behavior while still minimizing the total amount of data used and the time until a model can be calibrated accurately. To address these questions, we propose a Bayesian information-theoretic procedure, using an adaptive score function to determine the optimal data collection times and measurement types. The novel score function introduced in this work eliminates the need for a penalization parameter used in a previous study, while yielding model predictions that are superior to those obtained using two potential pre-determined data collection protocols for two different prostate cancer model scenarios: one in which we fit a simple ODE system to synthetic data generated from a cellular automaton model using radiotherapy as the imposed treatment, and a second scenario in which a more complex ODE system is fit to clinical patient data for patients undergoing intermittent androgen suppression therapy. We also conduct a robust analysis of the calibration results, using both error and uncertainty metrics in combination to determine when additional data acquisition may be terminated.

Keywords: data-driven mathematical oncology; Bayesian inference; sequential experimental design; prostate cancer models; personalized medicine

1. Introduction

In recent decades, mathematical modeling has frequently been used to advance our understanding of tumor evolution [1–5]. Modeling of cancer can be performed from the complex, highly-refined cellular level to a more "macro" level view, where we assume that the tumor acts as a mass of homogeneous tissue. Estimating the parameter values of such models requires detailed data, which may take many forms [6, 7]. The models can then be used to make predictions about the evolution of the tumor and its response to various treatment modalities, including radiotherapy, chemotherapy, immunotherapy, and viral therapy, among others. Recent technological advances have made it possible to collect a wide variety of data describing tumors, from the molecular level to the tissue level. Collecting data at multiple time points can aid in the calibration of mathematical models, which can be tailored to incorporate the available data. However, some data collection can be prohibitively expensive or invasive; this raises questions about how much data—and of what type—is needed to make accurate clinical predictions using mathematical models, and when this data should be collected.

In the age of personalized medicine, clinicians are turning to individualized treatment protocols, each tailored to the unique patient. Mathematical modeling can play a significant role here; given data from an individual tumor, we can calibrate a model and determine patient-specific parameter values which may give insight into the efficacy of the proposed treatment regimen for that individual. However, it is important that we bridge the gap between the idealized math modeling framework and the clinical constraints. While highly complex models can be insightful as far as determining the underlying mechanisms of the tumor and predicting how different cell populations might interact, at the clinical level, we are very constrained in the level of detail that might be inferred from the available data. The question then is: can an inherently simplistic model calibrated solely from a very small budget of crude data (i.e., estimated tumor volumes from MRI scans or estimates of a tumor biomarker in the bloodstream) still yield useful information regarding predicted response to treatment? Our work adds to a growing collection of literature that aims to inform data collection schedules in clinical oncology [8–10].

Because data collection in a clinical oncology setting is both expensive and potentially invasive for the patient, clinicians are constrained to a very sparse budget of measurements. In practice, a clinician might collect one or two tumor volume scans prior to treatment, and then forego measuring again until the treatment period has ended [11–13]. Due to the expense of data collection, it is imperative that we optimize the information content from those scans that can be collected. A variety of methods for optimal experimental design have been utilized in previous work, including the use of profile likelihood to resolve practical nonidentifiability in [14], parallel tempering and LASSO regression methods in [15], and structural identifiability and sensitivity analysis in [16]. In this work, we utilize a high-to-low fidelity model calibration framework that chooses the set of high-fidelity data points to be collected in such a way as to maximize their information content with respect to inferring parameters of a lower-fidelity model, which can then be used to make predictions about future outcomes [17]. In general, the high-fidelity setting may represent computationally expensive models such as multiscale models or agent-based models, or may represent the acquisition of experimental or clinical data, which is physically expensive to obtain. On the contrary, low-fidelity/low-cost models such as spatially-averaged differential equation models can be easily evaluated once validated [18, 19].

In [20], the authors employed such a sequential experimental design framework to determine an optimal selection of temporal data for model calibration. In that work, a score function was proposed as a means of adapting the pre-existing sequential design framework to handle temporal data, as opposed to other studies [21, 22] which dealt solely with non-temporal data (i.e., spatial design conditions). In addition to trying to maximize the reduction in parameter uncertainty through the choice of a highly informative data point, we also sought to penalize the algorithm for skipping too many data points, since the temporal data framework does not allow for those points to be subsequently collected at a later date. In [20], this penalization step relied upon a penalization parameter k, which was varied over the interval $[0, 1]$ in an attempt to optimize the efficiency and accuracy of the model calibration. The previous study tested this algorithm on three sets of synthetic data of varying radiation response types, and concluded that the optimal k value varies depending on the strength of patient response to the radiotherapy treatment. For instance, in scenarios where the tumor was highly sensitive to radiation, the model calibration procedure benefited most from the use of a k value near or at 1. Scenarios with data that was less responsive tended to favor k values in the low-to-middle spectrum, i.e., $k = 0$ to $k = 0.3$.

Although this framework was demonstrated to be effective in determining which scans to select for model calibration, the previous study did have several weaknesses. Most notably, the reliance of the choice of parameter value k upon the strength of the patient's response to therapy was constrictive; an optimal k value could not be determined until the general shape of the data could be assessed, which required at least several data points. In a highly restrictive scan budget scenario—i.e., in the clinical scenarios we are attempting to mimic—this means that an optimal k value realistically cannot be determined in time to have a positive impact on the algorithm efficiency. In this work, we propose a new adaptive score function, which adjusts the penalty at each step of the algorithm based on the anticipated final measurement—that is, we optimize the penalty term in accordance with how much further change is anticipated in the system dynamics over the remainder of the treatment period—in place of using a static parameter k.

We conduct an analysis of the model calibration resulting from this new adaptive score function with mean-square error, as was used in [20]. Additionally, we supplement this with uncertainty-based analysis, using credible intervals constructed by propagating parameter posterior distributions through the model to assess the level of certainty in the resulting model trajectory. The uncertainty analysis relies solely on the data that has been collected up to the current day, so it provides a more practical assessment of confidence in the model predictions for use in a clinical setting.

In addition to testing the algorithm on synthetic data generated from a cellular automaton model with an imposed radiotherapy treatment protocol, as in [20], we also test our algorithm on clinical patient data from prostate cancer patients, which includes measurements of both the prostate-specific antigen (PSA) tumor biomarker and serum androgen levels [23]. We calibrate a subset of the parameters in an ODE system [24] to an early subset of the PSA data using our adaptive score function, in order to assess how well the inferred model can predict future behavior. We also test our algorithm using multiple metrics, by making both PSA and androgen levels available to be chosen at each step of the sequential design procedure. In both cases, the resulting calibration and prediction results are compared to two potential pre-determined data allocation designs, and the sequence of data points chosen by our algorithm is demonstrated to yield the best mean-squared-error both for the inference period as well as for an extrapolated prediction period.

We begin in Section 2 by describing the algorithm development, including the necessary background in Bayesian parameter estimation and sequential design and the formulation of the adaptive score function. Our metrics for model assessment are discussed in Section 3. Section 4 describes the low-fidelity ordinary differential equation models and corresponding high-fidelity data for both applications that we use to illustrate the algorithm. The first application is radiotherapy treatment of prostate cancer, using synthetic data obtained from a cellular automaton model; the second is use of intermittent androgen suppression therapy (IAS) to treat prostate cancer, using clinical data from [23]. Section 5 presents the results obtained by applying our adaptive algorithm to the synthetic radiotherapy data and to the clinical IAS data. We demonstrate the benefit of our adaptive information-based design procedure by comparing our results with those obtained using alternative pre-determined design schemes. Section 6 summarizes the findings of the investigation and discusses their implications.

2. Methodology for optimal data collection

In this section we introduce the methodology used to determine optimal experimental designs for high-fidelity data collection in order to best inform our low-fidelity model parameters. The methodology is based upon a sequential experimental design procedure that employs Bayesian inference at each step. As the general procedure formed the basis of a previous study, we direct the reader to [20] for further details, and focus here on the introduction of our new adaptive score function for balancing the choice of an informative data point with the need to gather data early during the treatment period, so that treatment protocols can be altered mid-course in the event that the model trajectory predicts an undesirable outcome.

2.1. Overview of sequential experimental design procedure

Recall that our overarching goal is to accurately calibrate our model parameters using as little data as possible, ideally finishing early on during the treatment phase so as to allow for modified treatment if the predicted outcome is not ideal. As such, we need a way to determine which potential data points will be most informative for our parameter set; that is, given a choice of potential days at which to collect data and a choice of quantities that can be measured, which collection of measurements will maximize the reduction in uncertainty of the low-fidelity model parameters? To answer this question, we utilize a sequential experimental design framework, in which data points are acquired one-by-one and parameter estimates are updated between each data acquisition using a Bayesian inference framework. Because the Bayesian perspective assumes that parameters are random variables with associated densities that can be repeatedly updated to reflect information from newly acquired data, this method is ideally suited for our sequential framework.

Assuming a current data set $D_{n-1} = \{\tilde{d}_1, \tilde{d}_2, \ldots, \tilde{d}_{n-1}\}$, consisting of high-fidelity measurements (i.e., synthetic or experimental data), and a set of possible design conditions Ξ, we select the design condition $\xi_n \in \Xi$ that will maximize the reduction in uncertainty of the low-fidelity model parameters θ when $\tilde{d}_n$—the data point resulting from collecting experimental (or synthetic) data at condition ξ_n—is added to the existing data set. We can quantify the predicted information contribution of design ξ_n upon

parameter set θ by computing the mutual information, denoted by I, between these two quantities,

$$I(\theta; d_n \mid D_{n-1}, \xi_n) = \int_{\mathcal{D}} \int_{\Omega} p(\theta, d_n \mid D_{n-1}, \xi_n) \log \frac{p(\theta, d_n \mid D_{n-1}, \xi_n)}{p(\theta \mid D_{n-1}) p(d_n \mid D_{n-1}, \xi_n)} \, d\theta \, dd_n, \tag{2.1}$$

where d_n represents the predicted value of $\tilde{d}_n$ using our low-fidelity model, $\mathcal{D}$ is the full set of all unknown future observations, and Ω describes the admissible parameter space. The mutual information provides a measure of parameter uncertainty reduction; a larger MI value indicates potential for a greater acquisition of knowledge about the parameter values than a small MI value, given the inclusion of the data point(s) corresponding to the experimental design under scrutiny. In practice, the high-dimensional integral in (2.1) is typically estimated via the kth-nearest neighbor (kNN) method proposed by Kraskov et al. [25]. For more details on the derivation of Eq (2.1) and computational methods used to estimate it in practice, we point the reader to [20–22, 25].

In a standard non-temporal sequential design framework utilizing MI as a metric to be optimized, one would compute the MI for each of the potential design conditions and choose the condition which maximizes the MI as the next condition for experimental or synthetic evaluation. After evaluation of this data point, the parameter set is re-calibrated using a Bayesian Metropolis algorithm and the computation of MI begins anew for all remaining design conditions. The algorithm can be terminated when either (a) a user-defined threshold for model accuracy or uncertainty is achieved, or (b) a pre-defined data allocation budget is exhausted. A visual outline of the standard sequential design procedure for high-to-low fidelity model calibration is given in Figure 1, where $d_\ell(\theta, \xi_n)$ denotes the low-fidelity model evaluated at parameter set θ and design condition ξ_n, while $d_h(\xi_n)$ denotes the corresponding high-fidelity model (or data) evaluation at that same design condition.

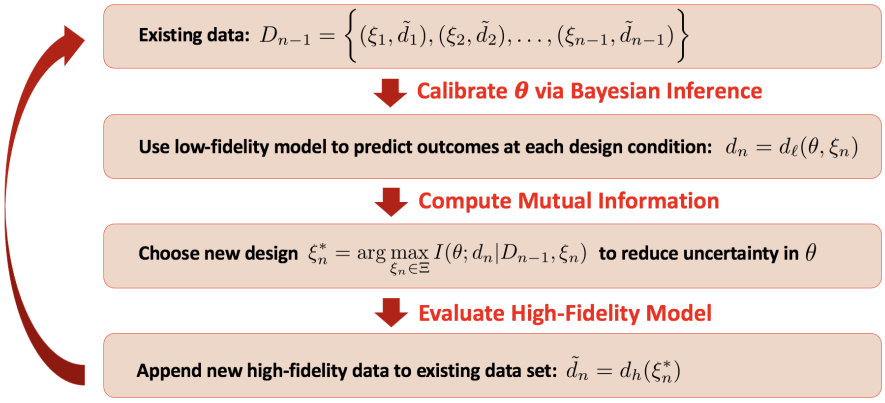

Figure 1. Multi-fidelity framework for sequential data collection and model calibration.

2.2. *Score function*

For a scenario such as this investigation, in which design conditions represent days at which data can be collected, we require an adaptation to the standard methodology. Because collecting data at time t_n precludes the collection of data for all times t_i with $i < n$, we must account for the potential loss of information from skipped data points. Additionally, skipping far ahead in time to obtain an informative data point towards the end of the treatment schedule impedes our goal of calibrating the

model parameters early in time, to allow for potential modification of the treatment regimen. In a previous study [20], we proposed an extension to the sequential MI framework using a score function that would reward a user for choosing a point with a large MI but simultaneously penalize them for skipping too many intermediate days. This score function contained a penalization parameter, k, which could be tuned via visual inspection of the data to increase or decrease the penalty term as appropriate; for patients whose data was highly dynamic, a steeper penalty could be imposed via a larger k value, in order to force the algorithm to choose designs near in time to the recently collected data. On the contrary, for patients whose trajectories were relatively stable, the penalty term could be adjusted so that data collection would be more sparse, enabling calibration to be performed on a smaller budget with negligible loss of information.

While the previously employed score function successfully identified appropriate data collection protocols for patients with various responses to treatment, the reliance of the procedure on identifying an optimal penalization parameter value k a priori was constrictive, especially since that parameter value could not be optimized without some prior knowledge of how the patient would respond to treatment. By testing a variety of different k values on the interval $[0, 1]$, it was found that for tumors that were highly sensitive to treatment, larger values of k were optimal, while the algorithm favored smaller values of k for those patients that were less responsive, as expected. Exploiting this observation, we now amend this score function to allow for better optimization of our algorithm without the need for an additional penalization parameter or a priori knowledge of how a patient will respond.

Suppose that the current data set D_r consists of high-fidelity observations $D_r = \{\tilde{d}_1, \tilde{d}_2, \ldots, \tilde{d}_r\}$, where $\tilde{d}_r = \tilde{d}(t_r)$ is the most recently appended data point, collected at time t_r. Among all possible future data points, $\tilde{d}_{r+1}, \tilde{d}_{r+2}, \ldots \tilde{d}_N$, we wish to determine which design will yield the most information about our model parameters. For each of the remaining designs, we calculate the mutual information at step r, denoted $I(\theta; d(t_i) \mid D_r)$, for $i = r + 1, r + 2, \ldots, N$. We then rescale the mutual information values to adhere to the interval $[0, 1]$, so that the information quantity will be on the same order of magnitude as the forthcoming penalty term. We denote these rescaled MIs at step r as $\{R(i, r)\}_{i=r+1}^{N}$. In the standard sequential design procedure where choosing one design condition does not preclude subsequent collection of others, this would be the quantity that we would seek to maximize.

To create the penalty term for temporal data collection, we begin by summarizing the potential information loss from skipping points $r + 1$ through $i - 1$ using the ratio

$$\text{Information Loss Ratio} = \frac{\sum_{j=r+1}^{i-1} R(j, r)}{\sum_{\ell=r+1}^{N} R(\ell, r)},$$

which totals the rescaled mutual information of all skipped points and divides by the sum of the rescaled mutual information across all possible remaining design conditions. In the previous study [20], this ratio was appended to the penalty weight parameter k and subtracted from the mutual information to yield the employed score function.

Rather than rely on an unknown weight parameter in this study, we replace k with an adaptive penalty coefficient designed to alter the penalty term in one of two ways: (a) to increase the penalty when the current dynamics still differ drastically from the expected ending point and (b) to decrease the penalty when it is suspected that the final outcome of the model trajectory is close to being established.

We employ a symmetric absolute error that compares the previously chosen high-fidelity measurement, $\tilde{d}_r$, to the anticipated final measurement based on the low-fidelity model prediction (i.e., the low-fidelity model prediction for the final day of treatment), d_N, calculated as

$$\text{Penalty Coefficient} = \left|\frac{\tilde{d}_r - d_N}{\tilde{d}_r + d_N}\right|.$$

This quantity is bounded on the interval $[0, 1]$, provided that $\tilde{d}_r$ and d_N are not both simultaneously zero. (Computationally, if both terms are zero, there is no further dynamical change expected, and so we set the penalty coefficient equal to zero.) In practice, this means that for a quantity that is still expected to undergo a lot of change over the course of the remaining treatment period (as assessed by using a prediction from the current calibrated model trajectory), the penalty for skipping data points is relatively high; for quantities that appear to have stabilized in size near the expected final outcome, the penalty is close to zero, allowing for sparser data collection to minimize costs. In the event that measurements for multiple quantities are able to be collected, as in Section 5.2, we calculate the penalty coefficient for each measurement type separately and use the maximum symmetric absolute error across all quantities, to favor a more conservative design scheme in terms of choosing a point closer in time to those already observed. The specific form of the penalty coefficient was chosen in order to meet our objective of quantifying the distance from the expected final measurement while also remaining bounded on $[0, 1]$, such that the magnitude of the penalty is comparable to the magnitude of the term to which it is applied. Certainly, other choices which meet these criteria—such as a squared symmetric error—could be considered.

Combining the three relevant quantities, we obtain our finalized score function for design condition i at step r:

$$S(i, r) = \underbrace{R(i, r)}_{\text{Rescaled MI}} - \underbrace{\left|\frac{\tilde{d}_r - d_N}{\tilde{d}_r + d_N}\right|}_{\text{Penalty Coefficient}} \cdot \underbrace{\left(\frac{\sum_{j=r+1}^{i-1} R(j, r)}{\sum_{\ell=r+1}^{N} R(\ell, r)}\right)}_{\text{Information Loss Ratio}}, \quad \text{for } i = r+1, \ldots, N. \tag{2.2}$$

Throughout the remainder of this investigation, Eq (2.2) will be used to select the next design condition for data acquisition at each cycle of the algorithm.

3. Analyzing model accuracy and uncertainty

Section 2 outlined a procedure for choosing optimal design conditions at which to collect data in a sequential manner. But when should we terminate the algorithm? The user has two options. If there are no constraints on the data collection budget, the user might define a goal that they wish to meet in terms of model accuracy or reduction of uncertainty; this might take the form of a user-defined error or uncertainty threshold. Once this goal is attained, the user may terminate the algorithm. In a more likely scenario, there are significant constraints on the data collection budget due to expensive or invasive scanning or sampling procedures that will force termination of the algorithm. For this scenario, the user must determine whether an adequate reduction in uncertainty or error has been achieved. This

analysis will assist the user in deciding whether the resulting model is reliable enough to be used for decision-making at the clinical level.

Though the previous study in [20] relied solely on error analysis to determine the predictive power of the final model, here we expand our model assessment to include an uncertainty analysis component. This is a more suitable assessment method in practice, since a user can measure the level of uncertainty in the model at any point using only the data collected so far, but cannot measure the full error in the model until after data collection has ceased, using the metric as defined.

3.1. Error analysis

Our previous study [20] relied solely on error analysis as a means of determining the number of scans required to achieve model accuracy. We conduct that analysis here again, and compare how the goals of achieving model accuracy (i.e., error reduction) and model certainty (i.e., uncertainty reduction) align.

To assess model error, we calculate the mean-square-error (MSE) between the low-fidelity model predictions and the high-fidelity synthetic or experimental data for all possible scan choices, given by

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i;\theta))^2,$$

where y_i represents the high-fidelity synthetic data measurement on day i and $f(x_i;\theta)$ represents the low-fidelity model prediction at day i given current estimated parameter set θ, for each cycle of the algorithm. We use this metric to demonstrate how the low-fidelity model trajectory converges toward the "truth" data as the scan number increases—that is, the model fit to the subset of chosen data tends to improve in fit to the full set of possible high-fidelity data points over time.

The drawback to using this metric for model assessment is that in practice, one would not have access to all of these high-fidelity data evaluations for computation. Given only the data points about which the user is actually aware, it is difficult to assess whether the model parameters have converged to the values that will create the idealized model fit across the entire data regime. Using only the selected scans, a final error could potentially be informative, but this is in essence a "hindsight" analysis; we cannot compute this error until all data have been collected. Thus, we supplement our error analysis in this investigation with an uncertainty analysis, and investigate how either one might be used to assess convergence towards the ideal model fit.

3.2. Uncertainty analysis

The use of Bayesian methods for parameter estimation provides an ideal setting for performing uncertainty quantification. The posterior distributions for each of the parameter values can be propagated through the model to simulate the full array of resulting trajectories that might arise. In essence, one can directly observe the uncertainty in the model output that arises from the uncertainty in the parameter inputs. In this study, we construct 95% credible intervals for the model output by propagating the parameter posterior density chains through the model, then plotting the middle 95% of trajectories to accompany the chosen model fit (the fit utilizing the set of parameter values which is found to maximize the likelihood function).

As a metric for quantifying the amount of uncertainty in our model predictions, we estimate the area of the credible interval after each data acquisition. As the uncertainty in the model parameters is

generally reduced with each new added data point, this manifests as a tighter credible interval about the fitted model trajectory; we can observe how the area of the interval trends generally downward with each new data collection.

The major benefit to conducting uncertainty analysis, as opposed to error analysis, is that it can be considered "foresight analysis." Given only the data that we have already collected, we can measure the uncertainty in the model trajectory for future times and assess whether this uncertainty has been reduced to an acceptable level to allow for decision-making based on the model. However, it should be noted that just because the model uncertainty has been reduced to an acceptable level does not guarantee that the model fit to future data will be decent. We recommend a two-fold approach: waiting for the model uncertainty to be reduced while also checking that the model trajectory has stabilized across the previous few data additions; observing the trajectory fluctuating with each added point suggests that the model may require additional data in order to settle upon a best fit.

4. High-to-low fidelity model systems for demonstration

In this section, we present two high-to-low fidelity model systems to demonstrate the effectiveness of the proposed data collection framework. We consider two different treatments for prostate cancer: radiotherapy and intermittent androgen suppression therapy. In both applications, the low-fidelity models are ordinary differential equation systems, presented in Sections 4.1.1 and 4.2.1, respectively. For high-fidelity data, we consider the use of both synthetically generated data (Section 4.1.2) as well as clinical data (Section 4.2.2).

4.1. Prostate cancer with radiotherapy treatment

The first application we consider is radiotherapy treatment of prostate cancer. We assume a scenario of collecting tumor volume data to monitor the treatment response; thus, an ODE model describing the tumor volume dynamics is used as the low-fidelity model. For the purpose of a proof-of-concept investigation that requires comparing errors and uncertainties across different collections of data measurements that may occur as often as once per day, we make use of synthetic high-fidelity data generated from a more complex *in silico* cellular automaton (CA) model.

4.1.1. Radiotherapy: low-fidelity differential equation model

The low-fidelity model that we use for calibration is an ODE model that governs the total tumor volume over time. The model describes the time evolution of the total tumor volume, $V(t)$, using a logistic growth model with an effective growth rate A and carrying capacity B:

$$\frac{dV}{dt} = AV\left(1 - \frac{V}{B}\right). \tag{4.1}$$

We denote them as effective parameters, since the effective growth rate A includes the net effect of proliferation and natural cell death, and the carrying capacity is scaled accordingly—see [26] for details. In this simple low-fidelity model, we assume that any cells that transition to a necrotic state due to sustained oxygen or nutrient deprivation are removed from the tumor instantaneously; that is, we view the tumor as a homogeneous mass of proliferating, viable cells.

The radiotherapy (RT) treatment protocol is modeled using the linear-quadratic model [27, 28] to account for the effects of RT, which is a reasonable choice of model for fast-growing tumors [29]. In this model, the fraction of cells that survive exposure to a single administered dose d of RT is given by

$$\text{Survival fraction,} \quad SF = e^{-\alpha d - \beta d^2}, \tag{4.2}$$

where α and β represent tissue-specific radiosensitivity parameters (for single-strand and double-strand DNA breaks, respectively), and d is the radiotherapy dosage. We incorporate a typical radiotherapy regimen for solid tumors, with daily doses of 2 Gy administered Monday through Friday for six consecutive weeks, initiated following a two-week growth observation period. We assume that the irradiated cell fraction is removed immediately from the tumor volume, similarly to the instantaneous removal of necrotic cells in our base model. We reformulate the low-fidelity model with radiotherapy under these assumptions, as:

$$\begin{cases} \frac{dV}{dt} = AV\left(1 - \frac{V}{B}\right), & \text{for } t_i^+ < t < t_{i+1}^-, \\ \\ V(t_i^+) = \exp(-\alpha d - \beta d^2)\, V(t_i^-), \end{cases} \tag{4.3}$$

where t_i (for $i = 1, 2, \ldots, n$) denote the times at which an RT dose is delivered, and $V(t_i^{\pm})$ denote the tumor volume just before and after radiotherapy is administered. Previous work [26] has illustrated that the full parameter set $[A, B, \alpha, \beta]$ is unidentifiable in the sense that multiple sets of parameters may yield the same model output. As such, we use two pre-treatment data points at days 8 and 15 to estimate and subsequently fix parameters A and B, and fix the α/β ratio to be 1.5, a typical value for prostate cancer [30]. We then initiate the algorithm with a single data point during the treatment phase (day 19, the first Friday of RT), and estimate β only, which serves as a measure of the strength of the patient response to treatment. We employ a flat prior distribution of $\mathcal{U}(0, 1)$ for β.

4.1.2. Radiotherapy: high-fidelity synthetic data

In place of experimental data for this proof-of-concept application, we generate high-fidelity synthetic data using the hybrid cellular automaton (CA) model from [20, 26], adapted from the model described in [31]. We use this model to generate a series of synthetic data representing virtual patients, consisting of tumor spheroids that differ in their response to radiosensitivity. In the model, cells are arranged on a discrete lattice representing a two-dimensional square cross-section through a three-dimensional cancer spheroid *in vitro*. Notable features of the model include a heterogeneous cancer population and stochastic cell cycle, coupled with spatially heterogeneous oxygen levels modeled by a reaction-diffusion equation. The surrounding oxygen levels determine whether each cell is in a proliferating, quiescent, or necrotic state. See [20, 31] for a detailed description of the CA model.

The baseline parameter values that are used to generate data using the CA model are shown in Table A1 of the supplementary material, while detail about the key parameters that are varied to generate distinct synthetic spheroids is provided below. These parameter values are estimated using experimental data from the prostate cancer cell line, PC3, in [31]. Radiotherapy in the CA model is implemented using the linear-quadratic model detailed in Eq (4.2). The probability of survival for all

proliferating cells in the CA model is identical to the survival fraction in the low-fidelity model. In the CA model, the quiescent cells are $\frac{2}{3}$ times as likely as the proliferating to be irradiated, in order to reflect the lower sensitivity to radiation-induced DNA damage for quiescent cells, in comparison to proliferating cells.

We generate a virtual cohort of 27 tumor spheroids using the CA model, for calibration testing. In order to generate spheroids with a range of responses to radiotherapy, we vary the mean cell cycle time, $\bar{\tau}_{cycle}$ (in hours), the radiosensitivity parameter α, and the ratio α/β. We generate one virtual spheroid with each combination of parameter values listed in the ranges below, while fixing all other parameter values at the values listed in Table A1.

$$\begin{aligned}
&\bar{\tau}_{cycle} \in [15,\ 22,\ 30],\\
&\alpha \in [0.014,\ 0.5,\ 0.14],\\
&\alpha/\beta \in [1,\ 1.5,\ 2].
\end{aligned}$$

Next, by visually inspecting the simulation results, we separate the 27 virtual spheroids into three categories: non-responders, weak responders, and strong responders. We observe similar patterns among the spheroids in each category with respect to the quality of fits and to the timing of and number of scans chosen using the score function in Eq (2.2). For simplicity, we choose one representative from each category to present in Section 5.1, each of which is reflective of the average behavior across the simulations in its category. Our chosen representative non-responder is generated using the parameter values $\bar{\tau}_{cycle} = 22$, $\alpha = 0.014$, and $\alpha/\beta = 1$. Our chosen representative weak responder is generated using the parameter values $\bar{\tau}_{cycle} = 22$, $\alpha = 0.05$, and $\alpha/\beta = 1.5$, and our chosen representative strong responder is generated using the parameter values $\bar{\tau}_{cycle} = 15$, $\alpha = 0.14$, and $\alpha/\beta = 1$. The synthetic tumor volume data of the three representative virtual patients are shown in Figure 2.

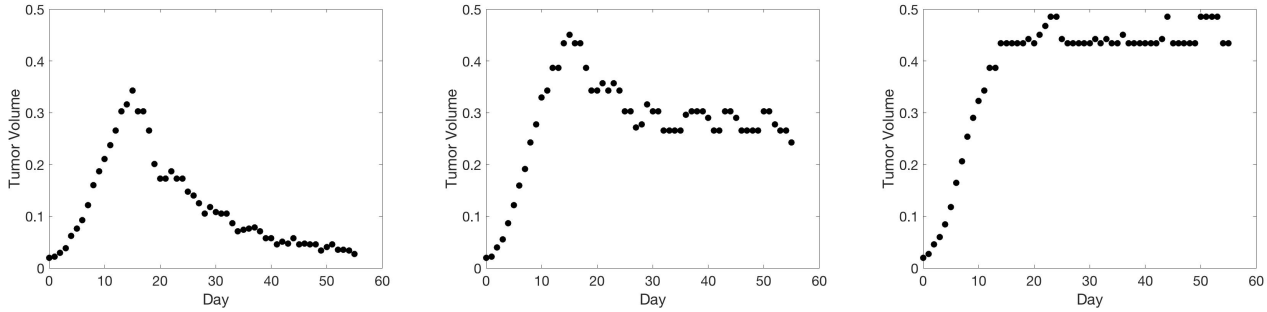

Figure 2. High fidelity synthetic data of tumor volume trajectory subject to radiotherapy for the three chosen representatives: strong responder (left), weak responder (middle), and non-responder (right). The data is generated by the hybrid CA model using parameters described in Section 4.1.2 and Table A1.

4.2. Prostate cancer with intermittent androgen suppression therapy

Our second application problem applies our approach to real clinical data, in which prostate cancer patients were treated with intermittent androgen suppression therapy (IAS), a treatment developed to help reduce the side effects of continuous androgen suppression therapy (CAS) while also delaying the

transition of prostate cancer cells from androgen-sensitive to androgen-resistant [32]. Measurements of both PSA and serum androgen levels are reported for each patient at each data collection time. As our low-fidelity model, we consider an ODE system developed in [24] that tracks the dynamics of cancer cells, PSA, and androgen levels.

4.2.1. IAS therapy: low-fidelity differential equation model

The low-fidelity ODE model that we use for calibration describes the dynamics of PSA level $P(t)$, intracellular androgen level $Q(t)$, serum androgen level $A(t)$, and two prostate cancer cell populations: androgen-dependent cancer cells $x_1(t)$ and androgen-independent cancer cells $x_2(t)$ [24]. The system of equations is as follows:

$$\begin{aligned}
\frac{dx_1}{dt} &= \max\left(\mu\left(1-\frac{q_1}{Q}\right)x_1, 0\right) - dx_1(x_1+x_2) - c\frac{K}{Q+K}x_1 \\
\frac{dx_2}{dt} &= \max\left(\mu\left(1-\frac{q_2}{Q}\right)x_2, 0\right) - dx_2(x_1+x_2) + c\frac{K}{Q+K}x_1 \\
\frac{dQ}{dt} &= m(A-Q) - \frac{\mu(Q-q_1)x_1 + \mu(Q-q_2)x_2}{x_1+x_2} \\
\frac{dA}{dt} &= \gamma_1 u(t)\left(1-\frac{A}{A_0}\right) + \gamma_2 - \delta A \\
\frac{dP}{dt} &= bQ + \max\left(\sigma_1\left(1-\frac{q_1}{Q}\right)x_1, 0\right) + \max\left(\sigma_2\left(1-\frac{q_2}{Q}\right)x_2, 0\right) - \epsilon P,
\end{aligned} \tag{4.4}$$

where the intermittent treatment is controlled by the function

$$u(t) = \begin{cases} 0, & \text{on treatment} \\ 1, & \text{off treatment} \end{cases}$$

Parameter interpretations can be found in Table A2 in the supplementary material. Further details regarding the construction and interpretation of this model can be found in [24]. As reported in [24], the parameter subset $\{q_2, \gamma_1, A_0, \sigma_2\}$ is expected to contain some of the most sensitive parameters; thus, we estimate this subset and fix all other parameters at the values listed in Table A2. As in [33], the initial values $A(0)$ and $P(0)$ are taken to be the first reported values of the clinical data set. The initial population of x_2 is assumed to be some fraction of the initial x_1 population, representing the fact that we begin with mostly androgen-sensitive cells. The initial intracellular androgen level, $Q(0)$ is assumed to be a fraction of the initial serum androgen level, $A(0)$. For further discussion of the determination of appropriate initial conditions, see [33].

4.2.2. IAS therapy: high-fidelity clinical data

In this model system, we study our approach using clinical data provided in [23]. We consider a single patient (Patient 39), for whom the response to treatment is fairly representative of the full group. The PSA and serum androgen level dynamics of the patient for 3.5 cycles of treatment are displayed in Figure 3. The patient begins the treatment at the start of data collection (day 0). Once their PSA level has declined below a predetermined threshold, the hormone therapy is temporarily halted; this occurs

around day 300, as indicated by both the PSA and serum androgen levels beginning to increase. One full cycle of "on-off" treatment has concluded around day 440, at which point the cycle begins anew.

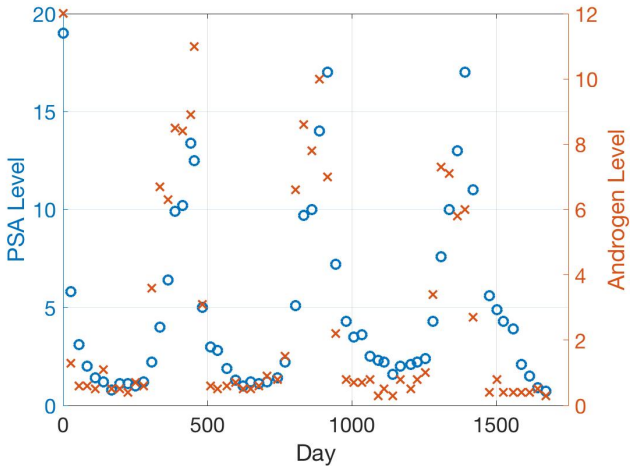

Figure 3. High fidelity clinical data for prostate cancer Patient 39 undergoing intermittent androgen suppression therapy in [23]; both PSA and serum androgen levels are included.

While PSA is the primary metric used for assessing prostate tumor growth in the clinic, we will make use of both PSA alone as well as PSA and serum androgen measurements in combination in Section 5.2, to see how acquiring multiple types of data may enhance the predictive capability of the calibrated low-fidelity model. Practically, both of these quantities can be measured using a simple blood test, and so allowing for the inclusion of androgen measurements is not a prohibitive suggestion.

5. Results

5.1. Prostate cancer with radiotherapy

We first test our algorithm on synthetic CA data using the models presented in Section 4.1, representing three different types of radiotherapy response: a strong responder, weak responder, and non-responder. In each case, data from days 8 and 15 of the growth observation period is used to estimate and fix the growth parameters, A and B, and then the algorithm is initiated using a single additional treatment data point (day 19, the first Friday of treatment), with only β being estimated. Figure 4 displays the calibration of the model to the strong responder data and corresponding credible interval after each tumor volume scan is chosen by the algorithm. In this case, the algorithm chooses 15 points over the full course of the six-week treatment, with a preference for scans near the start of each treatment week. We note that a large number of scans are chosen in this case because our score function penalizes skipping points when the current dynamics differ substantially from the expected ending behavior, which is observed for most of the treatment period for the strong responder. Figure 5 shows the posterior distributions for treatment parameter β as the model is re-calibrated when additional scans are added. We observe that the density plots become more narrow and shift to the right as more data is added to inform a more accurate estimate of the parameter value. The parameter appears to be reasonably well-informed after nine treatment scans are collected in this case (coinciding with the stabilization of the model trajectory from step-to-step), suggesting that it may not

be necessary to continue collecting data until the termination of treatment in order to establish a well-calibrated model for clinical predictions.

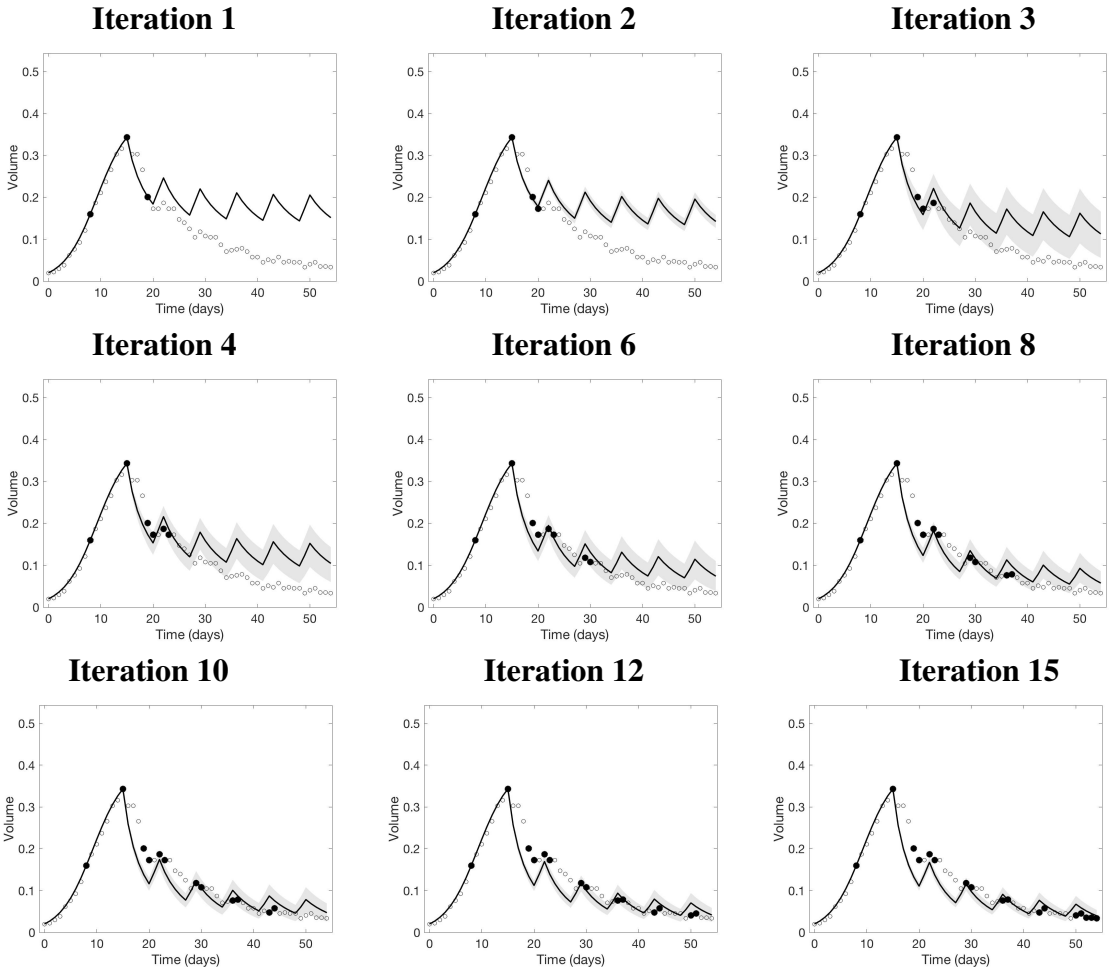

Figure 4. RT strong responder. Credible interval evolution over scan progression. The first plot shows the initial calibration using data at days 8, 15, and 19, with radiotherapy beginning at day 15 and following the treatment schedule described in Section 4.1.1. The points that are filled indicate data used for model calibration at the current iteration, whereas unfilled points represent scans that are not being used for calibration. The subsequent plots show the progression as the following scans are added. Both accuracy and uncertainty continuously generally improve as scans are added; these metrics are quantified in Figure 8.

In the weak responder case, the dynamics stabilize very quickly, so the penalty for skipping points is small after the first treatment scan is chosen. Additionally, the sawtooth behavior of the model observed during the treatment stage means that different scan choices have varying mutual information, further encouraging the algorithm to skip certain points in favor of later, more informative choices. As a result, only five scans are chosen throughout the course of the treatment period, as is shown in Figure 6. We also observe in this figure that the credible interval is small after two iterations of the algorithm, suggesting that it is sufficient to take two treatment scans, one in week 1 and one in week 3, in order to determine the level of treatment response from such a patient.

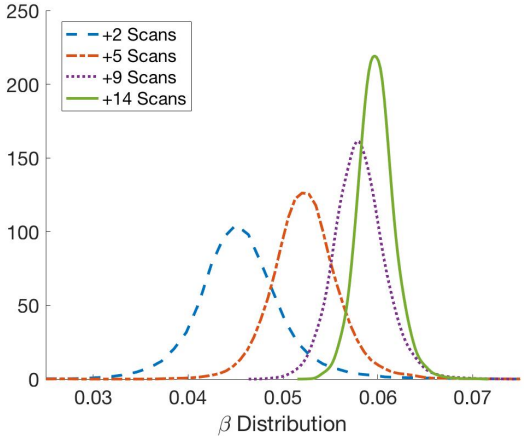

Figure 5. RT strong responder. Comparing the posterior distribution of estimated parameter β as scans are added and the parameter estimate is updated. The employed prior distribution for β is $\mathcal{U}(0, 1)$.

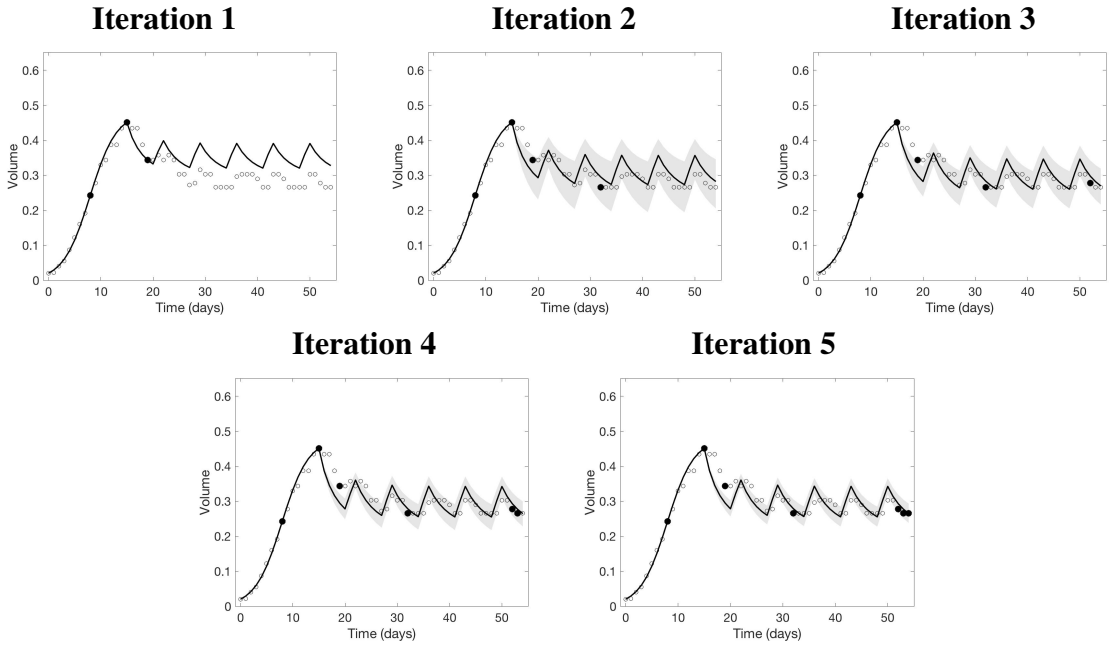

Figure 6. RT weak responder. Credible interval evolution over scan progression. The first plot shows the initial calibration using data at days 8, 15, and 19. The subsequent plots show the progression as the following scans are added. The accuracy and uncertainty are quantified in Figure 8.

A non-responder has a nearly flat trajectory, as shown in Figure 7, so all points are essentially equally informative. Due to the penalty for skipping points, the score function favors the early points over later points, so the algorithm begins its scan selection close to the start of treatment and chooses each subsequent scan to be close to the previous one. If the algorithm is allowed to proceed until it runs out of scan choices, this yields a large total number of scans. However, due to the nearly constant

dynamics in this case, only one or two early scans are needed in practice to assess such a patient's response level. In the clinic, this would be sufficient evidence to suggest that a patient will likely not respond well to this treatment and may benefit from switching to an alternative therapy modality.

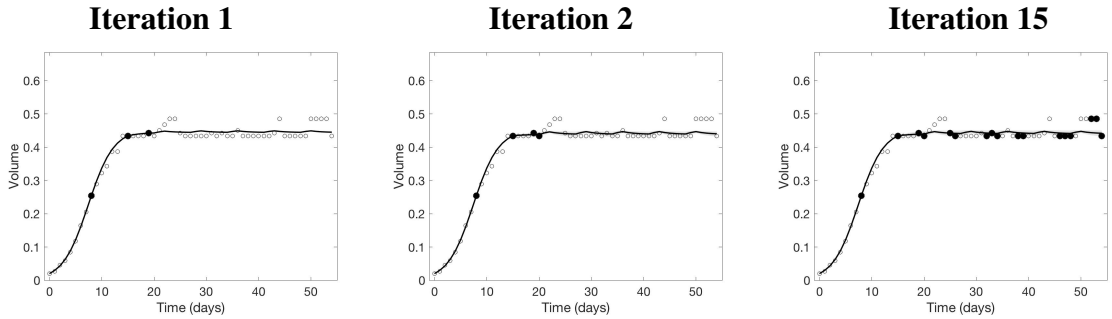

Figure 7. RT non-responder. Credible interval evolution over scan progression. The first plot shows the initial calibration using data at days 8, 15, and 19. The subsequent plots show the progression as the following scans are added. Only the initial and final iterations are shown, since intermediate iterations are similar in terms of error and uncertainty. The accuracy and uncertainty are quantified in Figure 8.

Figure 8 displays the progression of error and uncertainty as additional scans are added and the model is re-calibrated for the strong responder, weak responder, and non-responder. Note that the initial uncertainty measure—using only one treatment scan—is misleading. With only one point to fit, there is a single value of β that provides an essentially perfect fit to this data; thus, there is no uncertainty in the random variable for β at this point. In fact, this trend persists even into the second scan for the strong-responder, due to the close alignment of the first two points. Only once the third treatment point is added do we see significant noise in the data, and a resulting wider posterior distribution for β which in turn yields a wider credible interval. In general, the uncertainty metric should not be trusted when the number of data points is close to the number of parameters being estimated. Once we overcome this sparsity of data in comparison to parameters, we observe a downward trend in both error and uncertainty as scans are added for calibration in nearly all cases, outside of a small increase in uncertainty at the end of the non-responder calibration. We can also use these plots to determine when to terminate the algorithm—when the model parameter values are sufficiently informed—rather than continuing to collect data until the end of treatment. In particular, using pre-determined thresholds for error and uncertainty, the algorithm could be terminated once both the error and uncertainty for the model calibration reach a level below the corresponding threshold. For example, if we used thresholds of 2 for uncertainty and 10^{-3} for error, then we could stop the strong responder calibration after 7 additional treatment scans (i.e., after day 37). Likewise, we could terminate the algorithm for the weak responder after 4 additional scans at the end of treatment, and for the non-responder after just one additional scan on day 20.

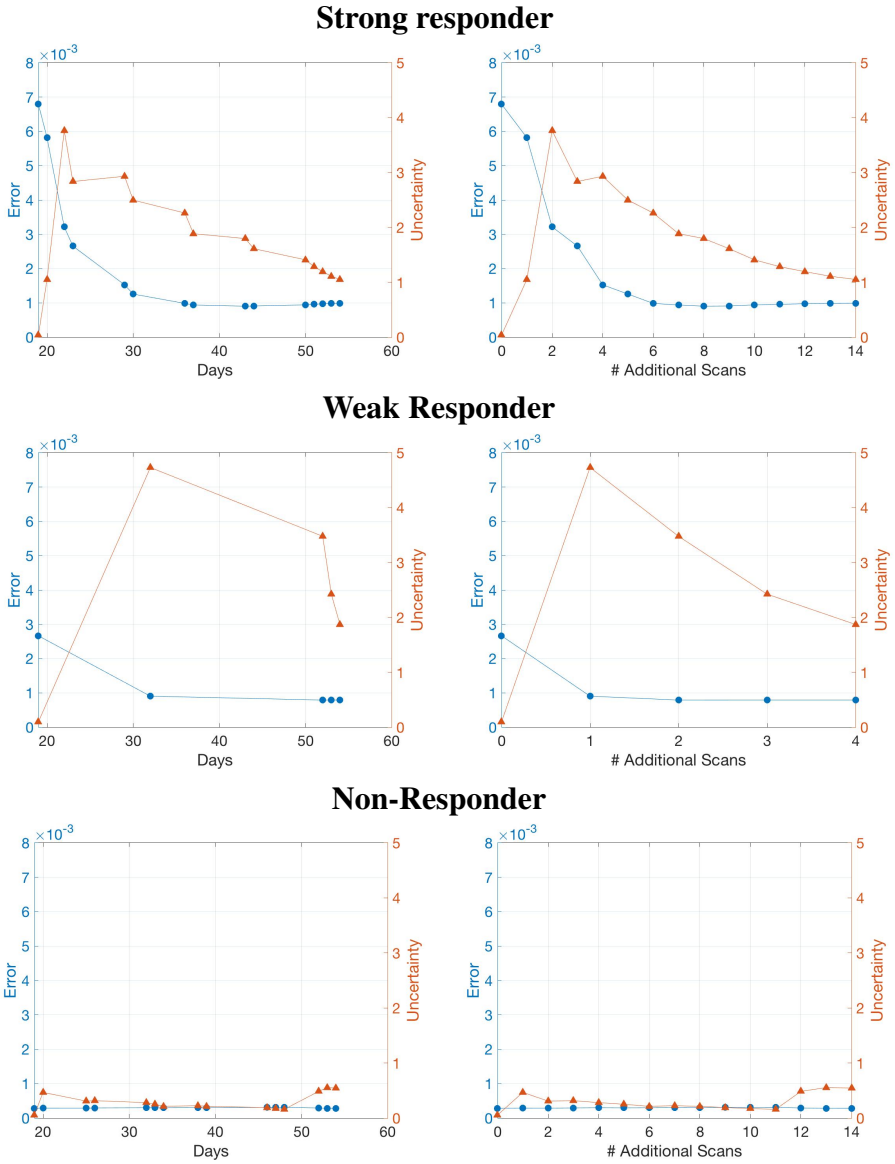

Figure 8. RT all responders. Reporting error and uncertainty by day of scan (left) and by additional scan number (right), for all responder types as shown in Figures 4, 6, and 7.

In order to assess the utility of our adaptive algorithm for choosing preferred collection days, we compare the error and uncertainty from our first six chosen scans, for the strong responder and non-responder, and the five chosen scans for the weak responder, with two potential pre-determined design schemes using six scans each. The choice of six scans is based upon a common scanning protocol for this treatment regimen, in which one scan is collected for each of the six weeks of treatment [10]. Our first design scheme for comparison is to collect tumor volume scans on the first six days of treatment, essentially front-loading the data collection; the second is the standard schedule choosing one scan per week, on each Friday, throughout the course of the six-week treatment phase. The choice of these two comparative design schemes allows us to observe how the model calibration would perform under different conditions; for instance, what would happen if we completed our

six-scan data collection as early as possible in order to allow maximum time for treatment intervention (design #1), versus utilizing the entire time frame with evenly-spaced scans in an attempt to reduce uncertainty and error over the full treatment interval (design #2). The use of our algorithm allows for the pursuit of both goals—early data collection and reduction of uncertainty and error—simultaneously. As noted previously, not all of our responders truly require six scans; however, we use six scans in this analysis to demonstrate a side-by-side comparison to the predetermined scan designs with all other variables fixed.

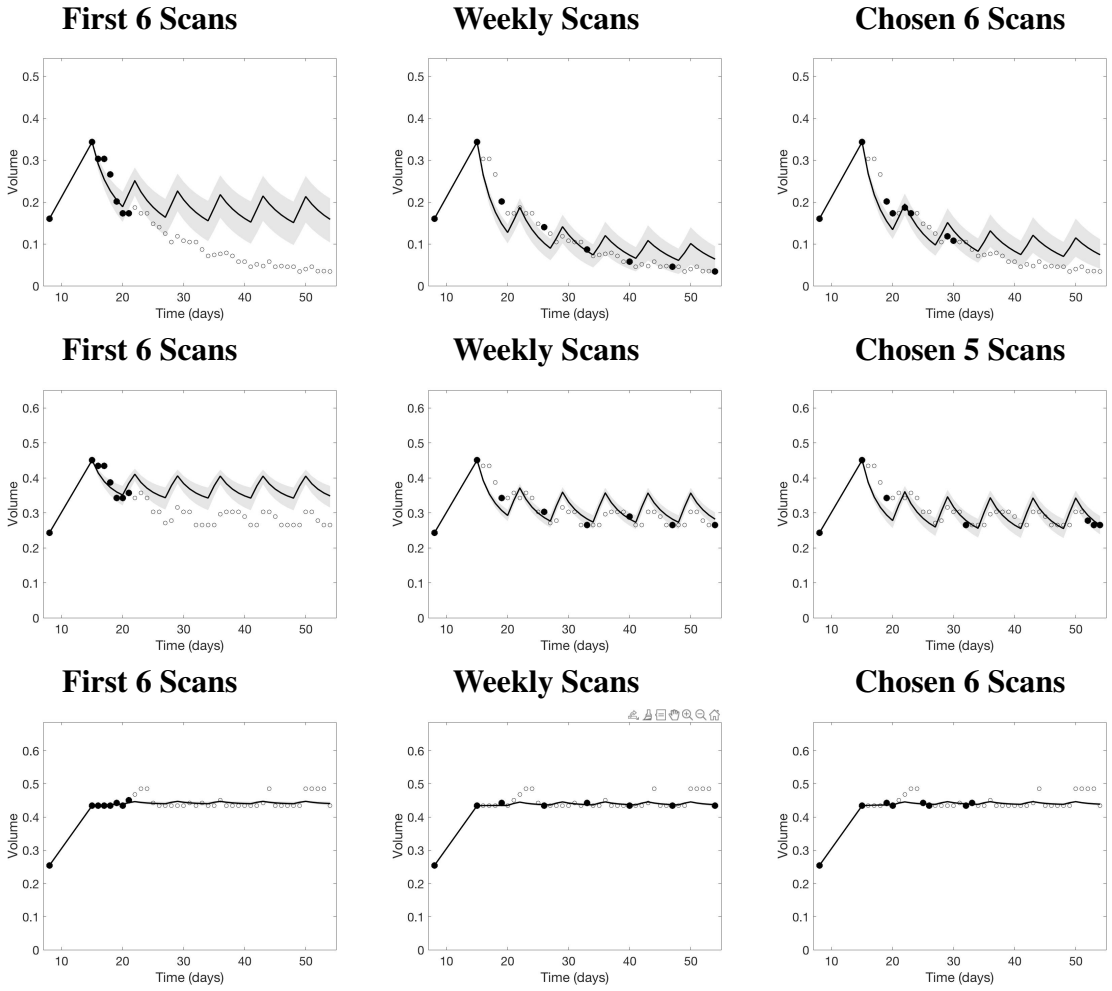

Figure 9. Comparison of design schemes: choosing the first six points (left), versus one scan per week (middle), versus algorithm choice of scans (right). The trajectories are shown for the strong responder (top), the weak responder (middle), and the non-responder (bottom). See Table 1 for corresponding values.

The top row of Figure 9 shows the fitted model trajectories and credible intervals for the three design schemes in the strong responder case. Choosing the first six scans does not provide sufficient information about the treatment response, leading to a relatively large error between the fitted model and the data. The results from both weekly scans and the six scans chosen by our algorithm are comparable in both error and uncertainty. We note that our chosen 6 scans produce a comparable result

in about half the time, around day 30, as compared to day 56 for the weekly scans. Achieving an early assessment of the treatment response increases the predictive power of the model, which would provide a significant advantage for clinical decision-making. The comparison of error and uncertainty for the three design schemes are summarized in Table 1.

Table 1. All responders. Comparing error and uncertainty metrics for three different design schemes: using the first available data points, using weekly scans, and using the points selected by the algorithm. Table results corresponds to Figure 9. Note: in all cases, six scans were used, except for the weak responder algorithm-chosen scan scenario, in which only five scans were selected.

		First Scans	Weekly Scans	Chosen Scans
Strong	Error	0.0102	0.0013	0.0017
	Uncertainty	3.5445	2.4024	2.4587
Weak	Error	0.0056	0.0012	0.0010
	Uncertainty	2.0207	1.2887	1.7847
Non	Error	3.42×10^{-4}	3.65×10^{-4}	3.53×10^{-4}
	Uncertainty	0.3412	0.2649	0.2409

The middle row of Figure 9 shows the comparison of model trajectories resulting from the three design schemes for the weak responder data. Again the model calibration using the first six scans overestimates the tumor size throughout the treatment period, while both the weekly scans and the scans chosen by our algorithm provide comparably accurate model fits, with small errors and credible interval areas. In this case, our algorithm achieves comparable results to the weekly scans using only five scans, in comparison to the six weekly scans.

The comparison of design schemes for the non-responder is shown in the bottom row of Figure 9. In this case, all three schemes produce accurate results, with almost no uncertainty. Even with using six scans (which, as previously noted, is several more than actually needed for this patient), our algorithm completes the model calibration in about half the time of the weekly-scan protocol, without any consequences for error and uncertainty. Again, the error and uncertainty values for the three design schemes are provided in Table 1.

In summary, for the *in silico* prostate cancer data, the scans chosen by our algorithm produce more accurate model trajectories for the strong and weak responders than those produced using the first six scans, as measured by both error and uncertainty metrics. Our chosen scans produce comparable results to the weekly design scheme for all three responder types, with our algorithm achieving these results either much sooner in time (strong and non-responders) or using fewer scans (weak responder).

5.2. Prostate cancer with intermittent androgen suppression therapy

We now test our algorithm on a clinical data set that measures the PSA biomarker as the primary metric for assessing tumor growth, and also measures the serum androgen level as a secondary means of assessing the efficacy of an intermittent androgen suppression treatment therapy. We begin by considering PSA data only, comparable to using only tumor volume measurements in our previous example. The low-fidelity model and high-fidelity data employed here are described in Section 4.2.

We initialize the algorithm with two PSA data points at days 0 and 28 (the first two recorded measurements for this patient), and estimate the parameter set $\{q_2, \gamma_1, A_0, \sigma_2\}$, fixing all other parameters at the values contained in Table A2. We make available to our algorithm the day choices of the PSA data from the first 1.5 cycles of treatment (i.e., an "on-off-on" sequence with regards to androgen suppression). The progression of PSA collection choices and corresponding evolution of the credible intervals are displayed in Figure 10. An additional five data points are chosen over the course of the 1.5 cycles to supplement the initial two supplied points. The resulting fit to the data is very strong; the overall trend in PSA is well-captured, including the peak that occurs around day 450. As additional scans are added, the uncertainty and error metrics improve, as shown in Figure 11. In this case, all five data points are needed to achieve the desired thresholds for these metrics and to observe stabilization of the model trajectory, such that it is no longer changing significantly at each cycle.

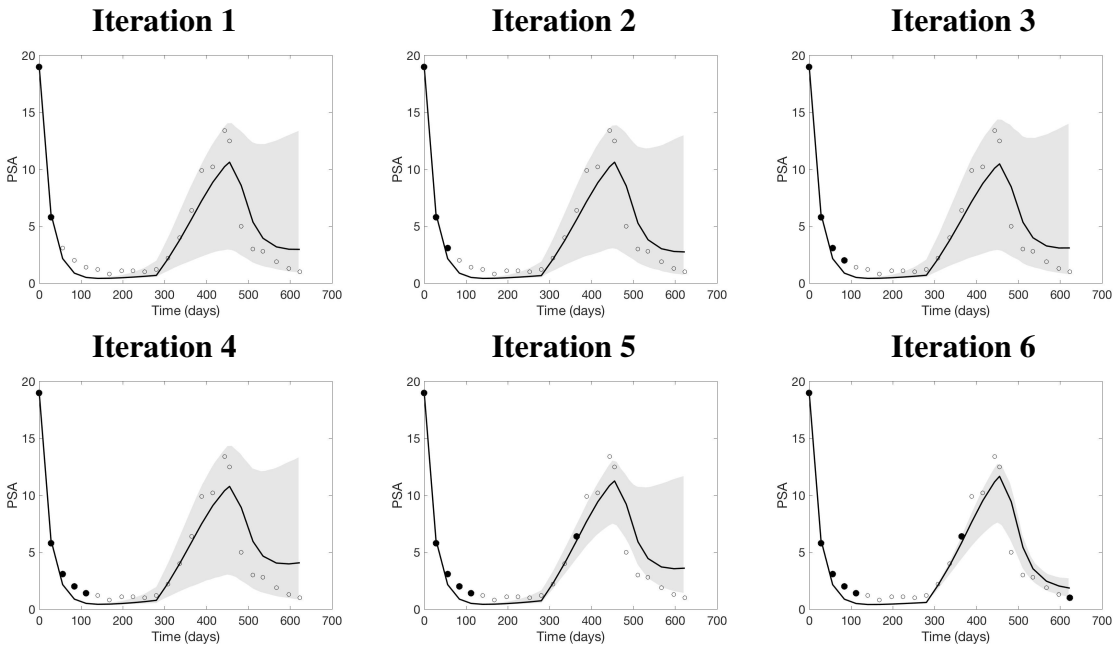

Figure 10. IAS therapy with PSA only. Credible interval evolution over scan progression, fitting to 1.5 cycles of data. The first plot shows the initial calibration using first two PSA data points. The subsequent plots show the progression as the following scans are added.

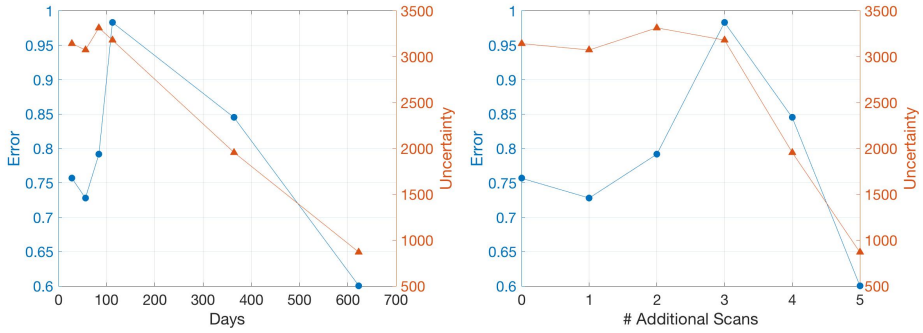

Figure 11. IAS therapy with PSA only. Reporting error and uncertainty by day of scan (left) and by additional scan number (right), fitting to 1.5 cycles of data.

As before, we compare the outcome of our algorithm with two potential pre-determined data collection designs. Our algorithm chose five scans: thus, we compare to what would happen if we used the *first* five available data points as well as if we used five evenly-spaced data points. The resulting model trajectories and credible intervals are shown in Figure 12, with numerical metrics given in Table 2. The data points chosen by our algorithm yield a much stronger mean-squared-error over all possible high-fidelity points than either of the other two designs. In particular, the first design scheme fails to capture the return to a near-zero state for PSA at the end of the 1.5 cycles, while the evenly-spaced design scheme underestimates the peak PSA value. The uncertainty is largest in the first-five-points scheme, and comparable between the other two designs.

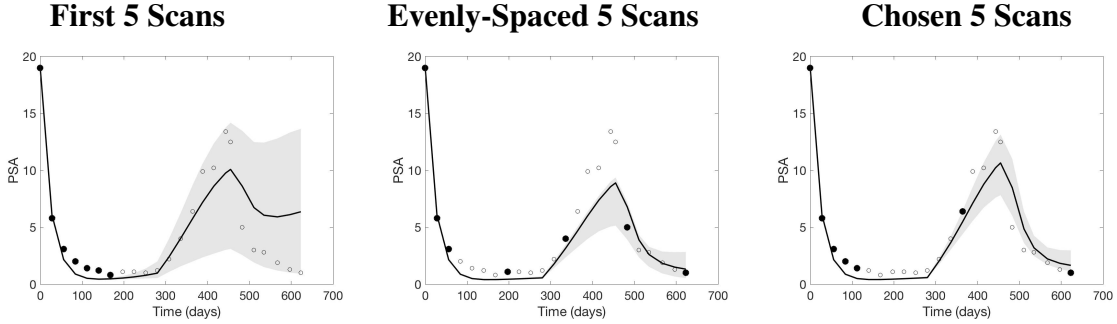

Figure 12. IAS therapy with PSA only. Comparison of design schemes: choosing first five points (left), versus five evenly spaced points (middle), versus algorithm choice of five points (right), fitting to 1.5 cycles of data. See Table 2 for corresponding values.

Table 2. IAS therapy with PSA only. Comparing error and uncertainty metrics for three different design schemes: using the first five available points, using five evenly-spaced points, and using the points selected by the algorithm, fitting to 1.5 cycles of data. Table corresponds to Figure 12.

	First Scans	Evenly-Spaced Scans	Chosen Scans
Error	5.80	3.04	1.92
Uncertainty	126.54	31.18	36.13

Of particular interest is whether or not the parameters inferred using the first 1.5 cycles of data can be used to accurately predict future behavior. We extend the timeline to include 3.5 cycles of intermittent androgen suppression therapy, and analyze how well the inferred trajectory from each of the three design schemes is able to capture future dynamics. This comparison is displayed in Figure 13. Predictably, the design scheme using the first five points is unable to adequately capture trends. In particular, it fails to capture the necessary return to a near-zero state that triggers the clinician to discontinue the suppression medication. In the other two design schemes, this return to a near-zero state is captured, but the oscillations are not maintained at a level that can capture future PSA peaks during "off" cycles. However, the oscillations from the algorithm-chosen design scheme die down more gradually, yielding a slightly better MSE over 3.5 cycles of high-fidelity data choices. Numerical values for the uncertainty and error metrics are included in Table 3.

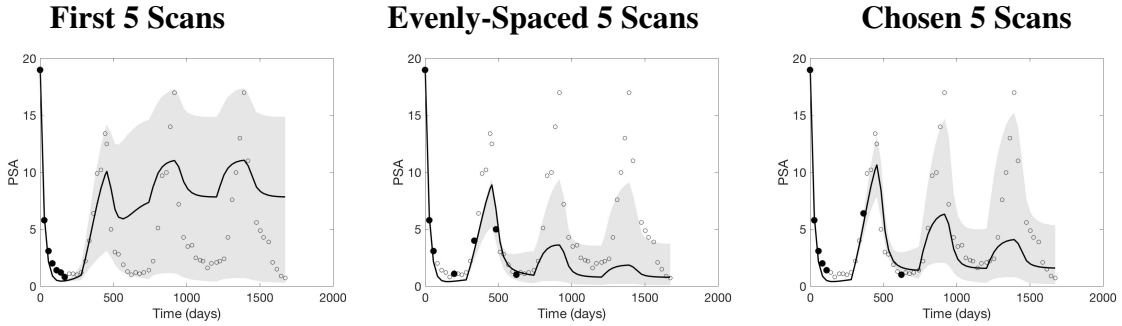

Figure 13. IAS therapy with PSA only. Comparison of design schemes: choosing first five points (left), versus five evenly spaced points (middle), versus algorithm choice of five points (right), using parameters inferred from 1.5 cycles of data to predict 3.5 cycles of data. See Table 3 for corresponding values.

Table 3. IAS therapy with PSA only. Comparing error and uncertainty metrics for three different design schemes: using the first five available points, using five evenly-spaced points, and using the points selected by the algorithm, using parameters inferred from 1.5 cycles of data to predict 3.5 cycles of data. Table corresponds to Figure 13.

	First Scans	Evenly-Spaced Scans	Chosen Scans
Error	17.98	18.41	10.85
Uncertainty	665.15	212.81	297.95

Because none of the three design schemes do an adequate job of predicting future PSA dynamics, we consider whether the predictive capability of our model might be improved by utilizing additional data sources. In particular, the clinical data used in this study also included serum androgen level for each data collection day, corresponding to variable A from the model in Eq (4.5). We now make the day choices for both sets of data (PSA and serum androgen) available to our algorithm, which can select either measurement type to evaluate on a given day. We allow for both metrics to be chosen on a particular day, though the procedure is still sequential (e.g., a mutual information calculation might suggest acquiring PSA data on day 56, and a subsequent calculation performed after recalibrating the parameters using that PSA data point might suggest acquiring androgen data on day 56). Thus, at each step of the sequential design procedure, our algorithm is now selecting the most informative data type/day combination with regards to which (type, day) pair will most reduce the uncertainty in the parameter set $\{q_2, \gamma_1, A_0, \sigma_2\}$. The algorithm is initiated with two points each of PSA and androgen data (both collected on days 0 and 28, the first available data points in the set).

The data acquisition progression and credible interval evolution for this scenario is shown in Figure 14, and the chosen data points of each type are recorded in Table 4. Over the course of 1.5 cycles of treatment, a total of 17 additional points are collected: seven PSA, and 10 androgen data points. It can be seen that the PSA dynamics are very well-fit over the course of the data acquisition procedure; additionally, information about the fluctuation in the serum androgen level is obtained.

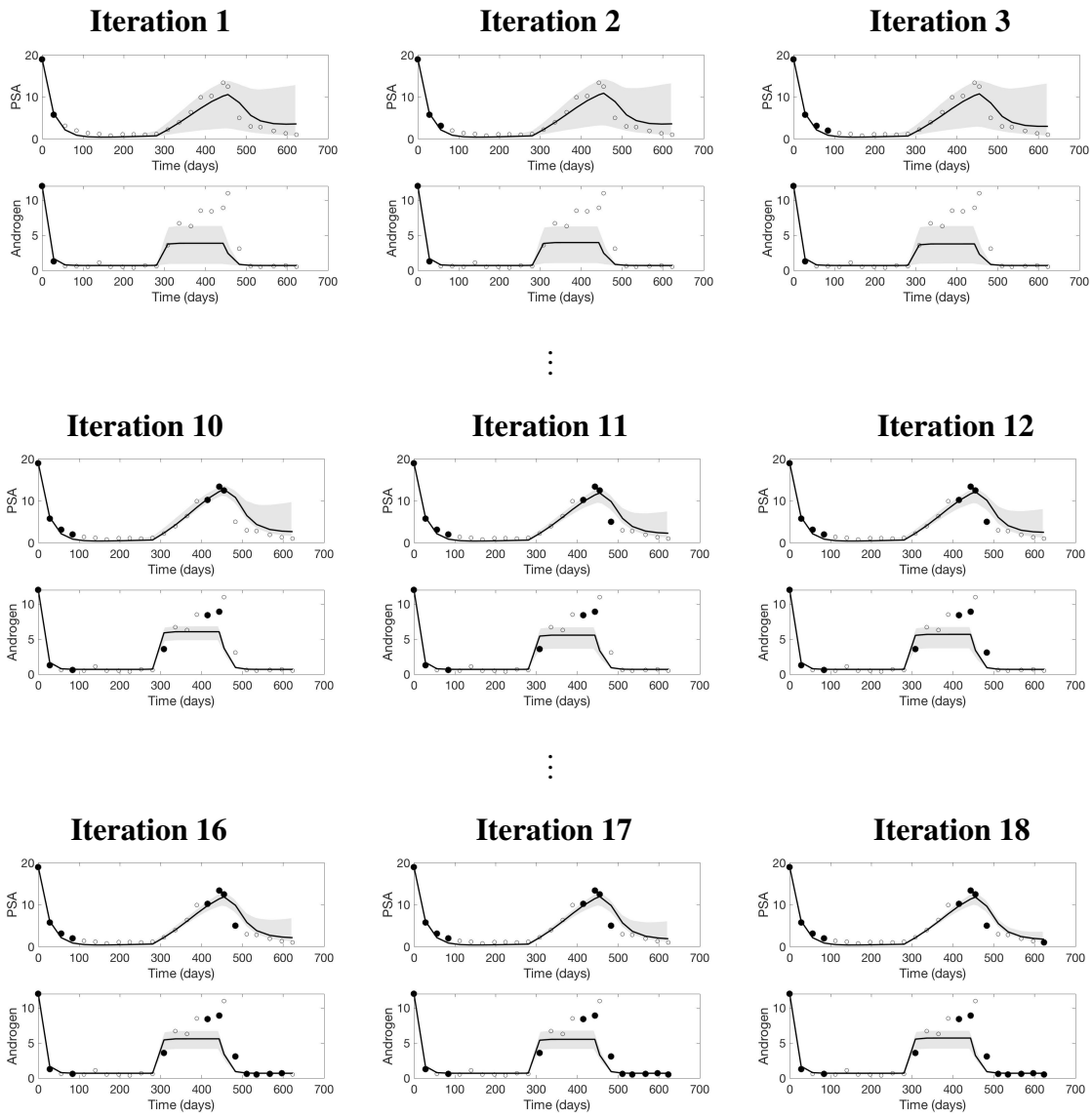

Figure 14. IAS therapy with PSA and serum androgen. Credible interval evolution over scan progression. The first plot shows the initial calibration using the first two points from each of PSA and androgen metrics. The subsequent plots show the progression as the following scans are added.

Table 4. IAS therapy with PSA and serum androgen. Chosen experimental designs, as reflected in Figure 14.

	0	28	56	84	112	140	168	197	224	252	280	308	336	364	388	415	443	455	483	511	535	567	597	623
PSA	x	x	x	x												x	x	x	x					x
Androgen	x	x		x								x				x	x		x	x	x	x	x	x

Though the algorithm collects a total of 17 data points from start to finish, it can be seen in Figure 15 that the error and uncertainty metrics and model trajectories have stabilized by the time that ten additional points have been selected; the additional gain in uncertainty reduction that occurs at point 17 may not be worth the budget required to collect said data or the time spent waiting for the reduction.

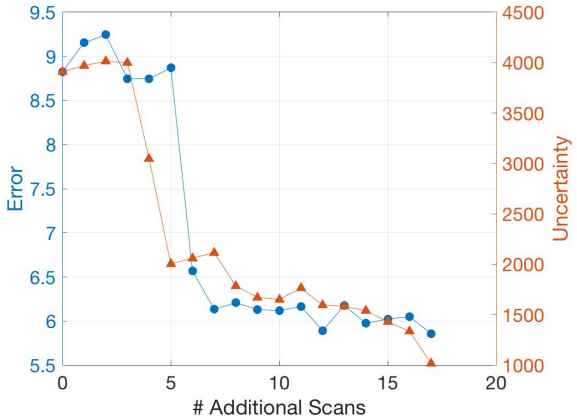

Figure 15. IAS therapy with PSA and serum androgen. Reporting error and uncertainty by additional scan number, fitting to 1.5 cycles of data.

We choose to terminate the algorithm at 10 additional scans (six PSA and four androgen), and once again compare the model output to two potential previously-determined data allocations. For the first scheme, we use the first five PSA and first five androgen data points (corresponding to days 56, 84, 112, 140, and 168); for the second, we use five evenly-spaced points of each type (corresponding to days 84, 224, 364, 483, and 623). The resulting model fits and credible intervals for each of the three design schemes are shown in Figure 16, with numerical values reported in Table 5. The algorithm-chosen design scheme yields the smallest MSE, with uncertainties being comparable between that and the evenly-spaced design scheme.

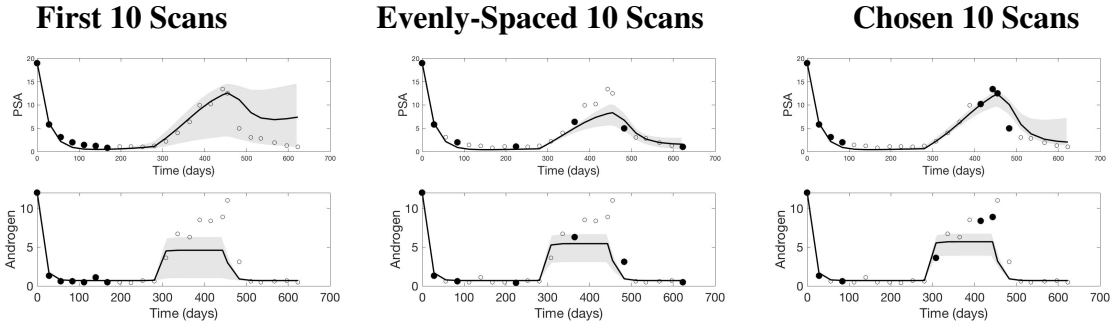

Figure 16. IAS therapy with PSA and serum androgen. Comparison of design schemes: choosing first ten points (left), versus ten evenly-spaced points (middle), versus algorithm choice of ten points (right), fitting to 1.5 cycles of data. (See Table 5 for values.)

Table 5. IAS therapy with PSA and serum androgen. Comparing error and uncertainty metrics for three different design schemes: using the first ten available points, using ten-evenly spaced points, and using the points selected by the algorithm, fitting to 1.5 cycles of data. Table corresponds to Figure 16.

	First Scans	Evenly-Spaced Scans	Chosen Scans
Error	13.29	7.42	6.05
Uncertainty	4231.66	1413.62	1575.57

Once again, we use the parameters inferred during the data acquisition procedure for 1.5 cycles of treatment to project out to 3.5 treatment cycles and compare the predictive capability of all three schemes. The results are reported in Figure 17 and Table 6. The first design scheme once again fails to capture the PSA dynamics. In the evenly-spaced design scheme, while the fit is relatively good within the fitting regime, the oscillations are not maintained into future treatment cycles, yielding a large MSE over the full period. On the contrary, the algorithm-chosen design scheme does the best with regards to capturing future dynamics for both PSA and serum androgen; in particular, this scheme is the only one in which the PSA level continues to fluctuate with each cycle of androgen suppression therapy, and the resulting MSE is far superior to that obtained from the alternate design schemes. In particular, we note that adding information about serum androgen has increased our ability to predict future oscillatory behavior of PSA as compared to using PSA data only (see Figures 13 (right) versus 17 (right) to observe how the peak PSA values are better matched in the latter).

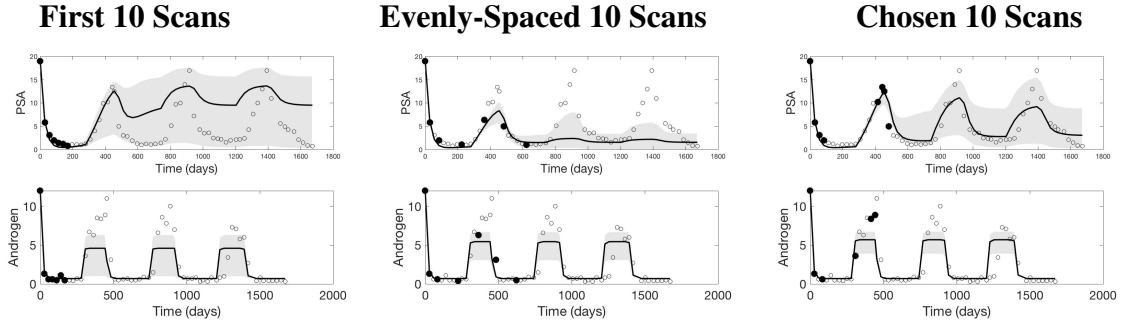

Figure 17. IAS therapy with PSA and serum androgen. Comparing error and uncertainty metrics for three different design schemes: choosing first ten points (left), versus ten evenly-spaced points (middle), versus algorithm choice of ten points (right), using parameters inferred from 1.5 cycles of data to predict 3.5 cycles of data. (See Table 6 for values.)

Table 6. IAS therapy with PSA and serum androgen. Comparing error and uncertainty metrics for three different design schemes: using the first ten available points, using ten evenly-spaced points, and using the points selected by the algorithm, using parameters inferred from 1.5 cycles of data to predict 3.5 cycles of data. Table corresponds to Figure 17.

	First Scans	Evenly-Spaced Scans	Chosen Scans
Error	33.47	22.45	7.62
Uncertainty	22510.54	7223.40	12741.40

As a final analysis, we consider whether the performance of our algorithm is robust across a wide array of patients. From the 66 patients in the study [23], we select the 13 patients for whom data was collected for at least 3.5 cycles of treatment. For each patient, we use our algorithm to select the optimal scanning schedule over the first 1.5 cycles of treatment (using both PSA and androgen collection options), determine the appropriate number of scans to include using our error and uncertainty metrics, and compare the resulting predictive capability of the model for 3.5 cycles of treatment against the evenly-spaced scanning design scheme. The results of these tests are displayed in Figure 18. Analysis of the error comparison reveals that our design scheme outperforms the evenly-spaced scheme in 10 of the 13 patients, with the remaining three patients having nearly comparable scores. With regards to uncertainty, our algorithm outperforms the evenly-spaced design scheme in 11 of the 13 patients. Notably, the largest exception to this trend is Patient 39, which was showcased above. (A closer look at Figure 17 reveals that the "better" performance of the evenly-spaced scheme in this case is actually correlated with the poor performance on the error metric; the oscillatory behavior of the model fades too quickly to capture the trends in the final two cycles, and since the parameter ranges have been chosen to ensure positive model trajectories, the resulting credible interval areas are artifically small.)

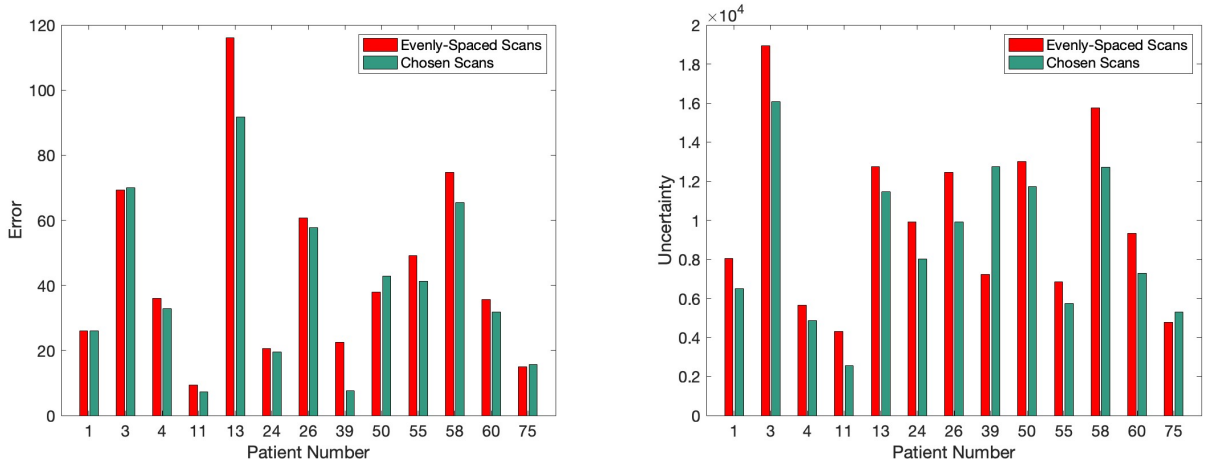

Figure 18. IAS therapy with PSA and serum androgen. Comparing error and uncertainty metrics for the evenly-spaced vs algorithm-chosen design schemes across thirteen patients. Parameters inferred from 1.5 cycles of data are used to predict 3.5 cycles of data.

In summary, our algorithm performs well with regards to inferring parameters using a select few data points, in both the PSA only and PSA/androgen scenarios. For both cases in our showcased Patient 39, the model fit within the data collection regime is preferable to the alternatively considered design schemes. Additionally, the algorithm-chosen design schemes tend to outperform the alternatives when using the inferred parameter values to predict future tumor dynamics across a wide array of patients.

6. Discussion

In this work, we proposed an adaptive-penalized score function to determine an optimal data collection protocol for maximizing the reduction of uncertainty in parameter estimates during model

calibration. This score function was used within a Bayesian sequential design framework, to choose data (type, day) combinations that maximize the information content, while simultaneously incorporating a penalty for data obtained late in the treatment period. The score function presented in this work improves upon a previous score function proposed in [20]. Since the optimal value of the penalization parameter k from the previous work seemed to be closely related to the rate of decay of the data, we hypothesized that this parameter value might be eliminated by gathering information about the expected future changes in dynamics. That is, if we can quantify how close the dynamics are to stabilizing at their final end value, we can use this information to adjust the score function, either increasing the penalty term to encourage acquisition of data in regions where the dynamics are changing rapidly, or by decreasing the penalty term to allow for sparser data collection when information gain from these points would be negligible. The incorporation of this information to the score function not only allowed us to eliminate the penalization parameter k from the previous study, but also aligned well with our goal of reducing unnecessary data collection in non-informative regions.

In addition to updating the score function to reflect information about the expected future change in dynamics, we have also provided a more robust and thorough verification of the algorithm in this investigation. The error-based verification metrics of the previous study are now supported by an uncertainty-based analysis, which relies on the propagation of parameter posterior distributions through the model to produce a 95% credible interval of model trajectories. The forward-looking nature of uncertainty analysis provides a more practical means of deciding when the algorithm might be terminated. By considering the combination of model error, model uncertainty, and the stabilization of model trajectories, we illustrated how these metrics might be used to decide when enough data has been collected to suit the purposes of the investigator.

We have tested this methodology using modeling of prostate cancer as an application with two different sources of high-fidelity data: 1) generating synthetic data from a CA model representing different strengths of response to radiotherapy, and 2) employing clinical data from a study using intermittent androgen suppression therapy. Using both error and uncertainty to assess the predictive power of the corresponding low-fidelity ODE models, we showed that our algorithm chooses data points that can be used to calibrate the low-fidelity models accurately and efficiently. When compared to two alternative pre-determined data collection protocols (choosing the first n available data points and choosing n evenly-spaced data points), the choices made by our algorithm yielded model trajectories that were either comparable or superior in terms of both error (computing the mean-squared error over all possible design conditions) and uncertainty (calculating the area of the credible interval resulting from propagating the estimated parameter posterior densities through the model), often doing so while simultaneously completing the calibration process earlier in the treatment cycle, which could allow for alteration of the treatment protocol in cases where the model predicts a poor outcome. Importantly, in the transition from an idealized setting using synthetic high-fidelity data (the radiotherapy scenario) to the clinical setting using noisy, real-world data (the IAS scenario), these trends were maintained. Additionally, using the clinical data study, our algorithm was shown to produce superior predictive capabilities over the other two alternatives when the inferred parameters were used to make predictions about future tumor dynamics.

Another benefit of our methodology that we expect to see when applied more generally to clinical data is the ability to adapt to settings with noisy data. When measurements are less precise, producing

noisy data, our algorithm senses the variability and will likely choose to sample more points from the noisy range, whereas a predetermined scanning schedule cannot similarly adapt to the data. Additionally, in the future we plan to apply our algorithm to varied clinical schedules, to see how well it can adapt to realistic scheduling issues that may arise due to holidays, staff shortages, etc. Continued testing upon other low-fidelity models that incorporate additional tumor characteristics and allow for different data collection metrics or treatment options will increase the flexibility of our methodology, enabling its application to many different settings in clinical oncology. In future work, we plan to investigate intervention options for patients that are classified early as non-responders, including alternative treatment schedules, combination therapies, and switching to new treatment modalities.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. P. M. Altrock, L. L. Liu, F. Michor, The mathematics of cancer: Integrating quantitative models, *Nat. Rev. Cancer*, **15** (2015), 730–745. https://doi.org/10.1038/nrc4029
2. O. Lavi, M. M. Gottesman, D. Levy, The dynamics of drug resistance: A mathematical perspective, *Drug Resist. Updates*, **15** (2012), 90–97. http://dx.doi.org/10.1016/j.drup.2012.01.003
3. A. Swierniak, M. Kimmel, J. Smieja, Mathematical modeling as a tool for planning anticancer therapy, *Eur. J. Pharmacol.*, **625** (2009), 108–121. https://doi.org/10.1016/j.ejphar.2009.08.041
4. H. M. Byrne, Dissecting cancer through mathematics: From the cell to the animal model, *Nat. Rev. Cancer*, **10** (2010), 221–230. https://doi.org/10.1038/nrc2808
5. R. C. Rockne, A. Hawkins-Daarud, K. R. Swanson, J. P. Sluka, J. A. Glazier, P. Macklin, et al., The 2019 mathematical oncology roadmap, *Phys. Biol.*, **16** (2019), 1–33.
6. D. A. Chambers, E. Amir, R. R. Saleh, D. Rodin, N. L. Keating, T. J. Osterman, et al., The impact of big data research on practice, policy, and cancer care, *Am. Soc. Clin. Oncol. Edu. Book*, **39** (2019), e167–e175.
7. J. E. Bibault, P. Giraud, A. Burgun, Big Data and machine learning in radiation oncology: State of the art and future prospects, *Cancer Lett.*, **382** (2016), 110–117. https://doi.org/10.1016/j.canlet.2016.05.033
8. I. Harshe, H. Enderling, R. Brady-Nicholls, Predicting patient-specific tumor dynamics: How many measurements are necessary?, *Cancers*, **15** (2023), 1368. https://doi.org/10.3390/cancers15051368

9. S. Prokopiou, E. G. Moros, J. Poleszczuk, J. Caudell, J. F. Torres-Roca, K. Latifi, et al., A proliferation saturation index to predict radiation response and personalize radiotherapy fractionation, *Radiat. Oncol.*, **10** (2015), 1–8. https://doi.org/10.1186/s13014-015-0465-x
10. M. Zahid, N. Mohsin, A. Mohamed, J. Caudell, L. Harrison, C. Fuller, et al., Forecasting individual patient response to radiation therapy in head and neck cancer with a dynamic carrying capacity model, *Int. J. Radiat. Oncol. Biol. Phys.*, **111** (2021), 693–704. https://doi.org/10.1016/j.ijrobp.2021.05.132
11. H. Chandarana, H. Wang, R. Tijssen, I. J. Das, Emerging role of mri in radiation therapy, *J. Magn. Reson. Imaging*, **48** (2018), 1468–1478. https://doi.org/10.1002/jmri.26271
12. L. H. Da Cruz, I. Rodriguez, R. Domingues, E. Gasparetto, A. Sorensen, Pseudoprogression and pseudoresponse: imaging challenges in the assessment of posttreatment glioma, *Am. J. Neuroradiol.*, **32** (2011), 1978–1985. https://doi.org/10.3174/ajnr.A2397
13. R. Ljumanovic, J. A. Langendijk, O. S. Hoekstra, D. L. Knol, C. R. Leemans, J. A. Castelijns, Pre-and post-radiotherapy mri results as a predictive model for response in laryngeal carcinoma, *Eur. Radiol.*, **18** (2008), 2231–2240. https://doi.org/10.1007/s00330-008-0986-x
14. F.-G. Wieland, A. L. Hauber, M. Rosenblatt, C. Tönsing, J. Timmer, On structural and practical identifiability, *Current Opin. Syst. Biol.*, **25** (2021), 60–69. https://doi.org/10.1016/j.coisb.2021.03.005
15. S. Gupta, R. Lee, J. Faeder, Parallel tempering with lasso for model reduction in systems biology, *PLoS Comput. Biol.*, **16** (2020), e1007669. https://doi.org/10.1371/journal.pcbi.1007669
16. M. Saccomani, K. Thomaseth, The union between structural and practical identifiability makes strength in reducing oncological model complexity: A case study, *Complexity*, **2018** (2018), 2380650. https://doi.org/10.1155/2018/2380650
17. P. Perdikaris, M. Raissi, A. Damianou, N. D. Lawrence, G. E. Karniadakis, Nonlinear information fusion algorithms for data-efficient multi-fidelity modelling, *Proc. Royal Soc. A Math. Phys. Eng. Sci.*, **473** (2017), 20160751. https://doi.org/10.1098/rspa.2016.0751
18. M. G. Fernández-Godino, C. Park, N. H. Kim, R. T. Haftka, Review of multi-fidelity models, preprint, arXiv:1609.07196.
19. X. Meng, G. E. Karniadakis, A composite neural network that learns from multi-fidelity data: Application to function approximation and inverse pde problems, *J. Comput. Phys.*, **401** (2020), 109020. https://doi.org/10.1016/j.jcp.2019.109020
20. H. Cho, A. L. Lewis, K. M. Storey, Bayesian information-theoretic calibration of radiotherapy sensitivity parameters for informing effective scanning protocols in cancer, *J. Clin. Med.*, **9** (2020), 3208. https://doi.org/10.3390/jcm9103208
21. A. Lewis, R. Smith, B. Williams, V. Figueroa, An information theoretic approach to use high-fidelity codes to calibrate low-fidelity codes, *J. Comput. Phys.*, **324** (2016), 24–43. https://doi.org/10.1016/j.jcp.2016.08.001
22. G. Terejanu, R. R. Upadhyay, K. Miki, Bayesian experimental design for the active nitridation of graphite by atomic nitrogen, *Exp. Thermal Fluid Sci.*, **36** (2012), 178–193. https://doi.org/10.1016/j.expthermflusci.2011.09.012

23. N. Bruchovsky, L. Klotz, J. Crook, S. Malone, C. Ludgate, W. Morris, et al., Final results of the canadian prospective phase II trial of intermittent androgen suppression for men in biochemical recurrence after radiotherapy for locally advanced prostate cancer, *Cancer*, **107** (2006), 389–395. https://doi.org/10.1002/cncr.21989
24. W. Meade, A. Weber, T. Phan, E. Hampston, L. Resa, J. Nagy, et al., High accuracy indicators of androgen suppression therapy failure for prostate cancer—a modeling study, *Cancers*, **14** (2022), 4033. https://doi.org/10.3390/cancers14164033
25. A. Kraskov, H. Stögbauer, P. Grassberger, Estimating mutual information, *Phys. Rev. E*, **69** (2004), 066138. https://doi.org/10.1103/PhysRevE.69.066138
26. H. Cho, A. Lewis, K. Storey, R. Jennings, B. Shtylla, A. Reynolds, et al., A framework for performing data-driven modeling of tumor growth with radiotherapy treatment, in *Springer Special Issue: Using Mathematics to Understand Biological Complexity, Women in Mathematical Biology*, (2021), 179–216. https://doi.org/10.1007/978-3-030-57129-0_8
27. E. J. Hall, A. J. Giaccia, *Radiobiology for the Radiologist*, Philadelphia, 1994.
28. H. Enderling, M. A. Chaplain, P. Hahnfeldt, Quantitative modeling of tumor dynamics and radiotherapy, *Acta Biotheor.*, **58** (2010), 341–353. https://doi.org/10.1007/s10441-010-9111-z
29. V. M. Pérez-García, M. Bogdanska, A. Martínez-González, J. Belmonte-Beitia, P. Schucht, L. A. Pérez-Romasanta, Delay effects in the response of low-grade gliomas to radiotherapy: A mathematical model and its therapeutical implications, *Math. Med. Biol.*, **32** (2015), 307–329. https://doi.org/10.1093/imammb/dqu009
30. D. J. Brenner, E. J. Hall, Fractionation and protraction for radiotherapy of prostate carcinoma, *Int. J. Radiat. Oncol. Biol. Phys.*, **43** (1999), 1095–1101.
31. M. Paczkowski, W. W. Kretzschmar, B. Markelc, S. K. Liu, L. A. Kunz-Schughart, A. L. Harris, et al., Reciprocal interactions between tumour cell populations enhance growth and reduce radiation sensitivity in prostate cancer, *Commun. Biol.*, **4** (2021). https://doi.org/10.1038/s42003-020-01529-5
32. N. Spry, L. Kristjanson, B. Hooton, L. Hayden, G. Neerhut, H. Gurney, et al., Adverse effects to quality of life arising from treatment can recover with intermittent androgen suppression in men with prostate cancer, *Eur. J. Cancer*, **42** (2006), 1083–1092. https://doi.org/10.1016/j.ejca.2006.01.029
33. T. Phan, K. Nguyen, P. Sharma, Y. Kuang, The impact of intermittent androgen suppression therapy in prostate cancer modeling, *Appl. Sci.*, **9** (2019), 36. https://doi.org/10.3390/app9010036

Appendix

Table A1. A summary of the parameters used in the CA model of Section 4.1.2 and their default values. Parameter values are estimated using experimental data from the prostate cancer cell line, PC3, in [31].

Param.	Description	Value	Units
l	Cell size	0.0018	cm
L	Domain length	0.36	cm
$\bar{\tau}_{cycle}$	Mean cell cycle time	Varies	h
c_∞	Background O_2 concentration	2.8×10^{-7}	mol cm^{-3}
D	O_2 diffusion constant	1.8×10^{-5}	cm^{2}s^{-1}
c_Q	O_2 concentration threshold for proliferating cells	1.82×10^{-7}	mol cm^{-3}
c_N	O_2 concentration threshold for quiescent cells	1.68×10^{-7}	mol cm^{-3}
κ_P	O_2 consumption rate of proliferating cells	1.0×10^{-8}	mol cm^{-3}s^{-1}
κ_Q	O_2 consumption rate of quiescent cells	5.0×10^{-9}	mol cm^{-3}s^{-1}
p_{NR}	Rate of lysis of necrotic cells	0.01	hr^{-1}

Table A2. A summary of the parameters used in the low-fidelity ODE model of Section 4.2.1 and their fixed values. Fixed parameter values were estimated using an `fmincon` procedure in Matlab, employing the parameter ranges reported in [24] and initial conditions design from [33].

Param.	Description	Value	Units
μ	max proliferation rate	0.0247	$[\text{day}]^{-1}$
q_1	minimum cell quota for x_1 to proliferate	0.5232	$[\text{nmol}][\text{day}]^{-1}$
q_2	minimum cell quota for x_2 to proliferate	Estimated	$[\text{nmol}][\text{day}]^{-1}$
d	density death rate	0.0294	$[\text{L}]^{-1}[\text{day}]^{-1}$
c	max mutation rate	1.0062e-5	$[\text{day}]^{-1}$
K	half-saturation constant for mutation	1.4667	$[\text{nmol}][\text{day}]^{-1}$
γ_1	androgen production rate by testes	Estimated	$[\text{nmol}][\text{day}]^{-1}$
γ_2	androgen production rate by adrenal gland	0.0602	$[\text{nmol}][\text{day}]^{-1}$
A_0	homeostasis serum androgen level	Estimated	[nmol]
δ	androgen degradation rate	0.0867	$[\text{day}]^{-1}$
m	diffusion rate from A to Q	0.3755	$[\text{day}]^{-1}$
b	baseline PSA production rate	0.0001	$[\text{g}][\text{nmol}]^{-1}[\text{day}]^{-1}$
σ_1	max PSA production rate by x_1	0.9998	$[\text{g}][\text{nmol}]^{-1}[\text{L}]^{-1}[\text{day}]^{-1}$
σ_2	max PSA production rate by x_2	Estimated	$[\text{g}][\text{nmol}]^{-1}[\text{L}]^{-1}[\text{day}]^{-1}$
ϵ	PSA clearance rate	0.0431	$[\text{day}]^{-1}$
$x_1(0)$	initial population of androgen-sensitive prostate cancer cells	0.0199	[L]
$x_2(0)^*$	proportion determining initial population of androgen-resistant cells	0.0169	—
Q_0^*	proportion factor determining initial Q amount	0.5000	—