doi: DOI HERE
Advance Access Publication Date: Day Month Year
Paper

PAPER

Supplementary Information for Deep reinforcement learning identifies personalized intermittent androgen deprivation therapy for prostate cancer

Yitao Lu,[1] Qian Chu,[2] Zheng Li,[3] Mengdi Wang,[4] Robert Gatenby[5] and Qingpeng Zhang[6, *]

[1]School of Data Science, City University of Hong Kong, Hong Kong SAR, China, [2]Department of Thoracic Oncology, Tongji Hospital, Huazhong University of Science and Technology, 430030, Wuhan, China, [3]Department of Radiology, Tongji Hospital, Huazhong University of Science and Technology, 430030, Wuhan, China, [4]Department of Electrical and Computer Engineering and the Center for Statistics and Machine Learning, Princeton University, 08544, NJ, U.S.A, [5]Department of Integrated Mathematical Oncology and the Cancer Biology and Evolution Program, H. Lee Moffitt Cancer Center and Research Institute, 33612, FL, U.S.A and [6]Musketeers Foundation Institute of Data Science and the Department of Pharmacology and Pharmacy, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China

[*]Corresponding author. qpzhang@hku.hk

Abstract

The evolution of drug resistance leads to treatment failure and tumor progression. Intermittent androgen deprivation therapy (IADT) helps responsive cancer cells compete with resistant cancer cells in intratumoral competition. However, conventional IADT is population-based, ignoring the heterogeneity of patients and cancer. Additionally, existing IADT relies on pre-determined thresholds of prostate-specific antigen to pause and resume treatment, which is not optimized for individual patients. To address these challenges, we framed a data-driven method in two steps. First, we developed a time-varied, mixed-effect, and generative Lotka-Volterra (tM-GLV) model to account for the heterogeneity of the evolution mechanism and the pharmacokinetics of two ADT drugs Cyproterone acetate (CPA) and Leuprolide acetate (LEU) for individual patients. Then, we proposed a reinforcement-learning-enabled individualized IADT framework, namely, I^{2}ADT, to learn the patient-specific tumor dynamics and derive the optimal drug administration policy. Experiments with clinical trial data demonstrated that the proposed I^{2}ADT can significantly prolong the time to progression of prostate cancer patients with reduced cumulative drug dosage. We further validated the efficacy of the proposed methods with a recent pilot clinical trial data. Moreover, the adaptability of I^{2}ADT makes it a promising tool for other cancers with the availability of clinical data, where treatment regimens might need to be individualized based on patient characteristics and disease dynamics. Our research elucidates the application of deep reinforcement learning to identify personalized adaptive cancer therapy.

Key words: Adaptive therapy, Prostate cancer, Personalized medicine, Reinforcement learning.

Supplementary Information

S1. Algorithms for validation

S2. Details of PcaC construction

There are two phenotypes of prostate cancer cells prior to initiating intermittent androgen deprivation therapy (IADT): responsive (hormone-dependent) and resistant (hormone-independent) cells, as described by (3). In the tumor microenvironment, resistant phenotypes can gain advantages through genetic or epigenetic mutations, leading to competition between the two phenotypes. This dynamic is expressed in Formula (1), where the definitions of each variables are described in Table 1 of main text:

$$\frac{d\boldsymbol{x}}{dt} = RX(\mathbf{1} - (K^{-1}\boldsymbol{A}(t)\boldsymbol{x})^{\alpha} - \mathcal{D}). \tag{1}$$

In the context of interacting phenotypes, both internal competition and external pressures influence the two phenotypes, which are regarded as permanently bounded variations of the system represented by Equation 1. Under these conditions, equilibrium cannot be achieved and is instead replaced by ultimate boundedness (a compact set of values in the state space) (4; 5). Consequently, the competitive community matrix is established as follows:

$$A(t) = \left\{ \begin{matrix} 1 & \frac{1}{1+e^{\gamma t}} \\ \frac{1}{1+e^{\gamma t}} & 1 \end{matrix} \right\}, \gamma \in \Re_+, \tag{2}$$

```
Algorithm 1 Online Train tM-GLV
Data: IADT clinical data: $\mathcal{X}(0:\tau)$
Input  : Data: $\mathcal{X}(0:\tau)$
         Initial parameters: $\theta^0_{online}$
Output: online tM-GLV model: $\mathcal{M}(a;\theta_{online})$
Train tM-GLV
    $\theta_{online} \leftarrow \theta^0_{online}$
    Max Iteration $\leftarrow N_{online}$
    while $k < N_{online}$ do
        $P_k \leftarrow \xi - torch$ solver
        $\mathcal{L} \leftarrow MSE(P_k, P)$          ▷ $P$ is ground truth PSA level
        $\theta^k_{online} \leftarrow \theta^{k-1}_{online} + \eta \cdot \nabla\mathcal{L}$
    end
```

```
Algorithm 2 delayed-I$^2$ADT
Data: IADT clinical data: $\mathcal{X}(0:\tau)$
Input  : Predition term: $T$
         $\mathcal{X}(0:\tau)$
         Online Patient: $\mathcal{M}(a;\theta^0_{online})$
         Real Patient: $\mathcal{M}(a;\theta_{offline})$
Output: tM-GLV model: $\mathcal{M}(a;\theta_{online})$
         Dosing Policy: $\mathcal{A}(\theta_A)$
Init PPO
    $\theta_A \leftarrow \theta^0_A$
    Max Iteration $\leftarrow N_A$
    Update steps $\leftarrow s$
    Online updating tM-GLV times: $c \leftarrow 0$
    Converge Signal: $\mathcal{C} \leftarrow$ False
while $k < N_A$ do
    Explore/Exploit Online Patient $(\mathcal{M}(a;\theta^c_{online}))$          ▷
     Decaying-$\epsilon$ greedy search
    Updates agent parameters $\theta^k_A$
    if PPO agent converges then
        $c+=1$
        $\mathcal{C} \leftarrow$ True
    end
    if $\mathcal{C}$ then
        $a_{(\tau+cT):(\tau+(c+1)T)} \leftarrow \mathcal{A}(\theta^k_A)$

        $\mathcal{X}((\tau + cT) : (\tau + (c+1)T)) \leftarrow$
         $\mathcal{M}(a_{(\tau+cT):(\tau+(c+1)T)};\theta_{offline})$          ▷ Offline model
         represents the real patient

        $\mathcal{X}(0 : (\tau + (c+1)T)) \leftarrow$ concatenate($\mathcal{X}(0 :$
         $\tau + cT), \mathcal{X}(\tau + cT : (\tau + (c+1)T)))$

        $\theta^{c+1}_{online} \leftarrow$ Algorithm 1$(\mathcal{X}(0 : (\tau + (c+1)T)), \theta^c_{online})$
         ▷ (Algorithm 1)

        Updates online tM-GLV model: $\mathcal{M}(a;\theta^{c+1}_{online})$          ▷
         Updating the training environment for PPO

        $\mathcal{C} \leftarrow 0$
    end
end
```

where $A_{12} = A_{21} := a(t) = 1/(1 + e^{\gamma t})$, and both are positive. The two phenotypes are in direct competition, with $a(t)$ representing the percentage of resource overlap between them. The overlap is 100% for identical phenotypes, and initially set at 50% for distinct phenotypes. By setting the resistance index $\gamma > 0$, $a(t) = \frac{1}{1+e^{\gamma t}} <= a(0)$, indicating a decreasing trend in resource overlap. This decrease is attributed to competition-induced mutations and epigenetic modifications within the cancer population. As a result, the competition intensity weakens over time due to fewer shared resources, as illustrated in Fig 3.b of main text.

The drug-induced decay term in Equation (1) is assumed to follow a first-order decay process (2) (i.e., a metric of drug exposure). In our study, this term is defined as the linear relationship presented in Equation (3). To ensure meaningful extrapolations for untested dosage regimens per patient, the pharmacokinetics of both drugs included in our model were considered. Cyproterone acetate (CPA) was administered twice daily, and given its half-life of 1.5 days (8), the dynamics of CPA's effect remain constant during therapy. Leuprolide acetate (LEU) was administered intramuscularly at a dosage of 7.5 mg every 4 weeks in a depot suspension format (9; 10).

$$\mathcal{D} = \beta d(t), \quad \beta > 0, \tag{3}$$

where β is a patient-specific parameter, denoting the first-order decaying process, and $d(t)$ represents the normalized drug effects, combining pharmacokinetic knowledge (2). In this case, $d(t)$ is proportional to serum hormone levels. Clinical studies (9; 10) have found that serum testosterone initially increases during the first week, then becomes suppressed to castrate levels. Consequently, $d(t)$ decreases in the first week and reaches a stable level after continuous drug administration.

In our proposed method, we employed a non-compartmental model to estimate drug exposure. Specifically, we assumed a constant drug concentration in the blood for CPA since it is taken twice daily. Therefore, a constant drug effect is assigned to CPA, which is normalized as 1 if taken, otherwise 0. For LEU, we referenced the work of (10) and found that the drug concentration in blood plasma initially increases and then decreases to a steady level over the administration course of a month. The corresponding plasma testosterone level exhibits similar patterns. For illustration of these findings, please refer to Fig 1&5 in work (10). The figures show that with depot-injected LEU, the testosterone level initially increases in the first week and then decreases to the castrate level by week 4. Subsequent injections maintain the testosterone level at the castrate level. It is essential to note that the drug effect for LEU is not defined by the drug concentration in plasma; rather, the corresponding plasma testosterone level reflects the actual drug effects. Based on this information, we have normalized the drug effect by setting it to negative in the first week, decreasing linearly from 0 to -0.5. In the following three weeks, the drug effect gradually increases from -0.5 to 1 linearly with time. Subsequent doses maintain the testosterone at the castrate level with a drug effect of 1. If no maintenance dosage is present, the LEU drug effect resets to 0.

For further information on drug pharmacokinetics, please refer to (9; 10; 11; 12) for LEU, and (8) for CPA.

Regarding the mathematical relationship between prostate cancer cell count and serum PSA levels, it is widely assumed that PSA level dynamics can be simplified as shown in Equation (4):

$$\frac{dP}{dt} = \rho \sum_i \boldsymbol{x} - \phi P, \tag{4}$$

where ρ denotes the rate at which PSA is released from cancer cells, and ϕ represents the decomposition rate of serum PSA. The PSA decay rate is set as a population-wide uniform parameter, with $\phi = 0.25(\text{day}^{-1})$, given that the serum PSA half-life is 2.5 days (13; 14; 15).

By combining Equations (1 ∼ 4), the mathematical model for simulating the prostate cancer cell (PCaC) environment is presented as system (5):

$$\begin{cases} \frac{d\boldsymbol{x}}{dt} &= RX(\mathbf{1} - (K^{-1}\boldsymbol{A}(t)\boldsymbol{x})^{\alpha} - \mathcal{D}), \\ \frac{dP}{dt} &= \rho \sum_i \boldsymbol{x} - \phi P. \end{cases} \tag{5}$$

S3. Learning an adaptive dosing policy by reinforcement learning

S3.1 Details about PPO algorithm

Reinforcement learning (RL) is a continuous process where an agent interacts with an environment at discrete time steps. At each time step, the agent receives the environment's state (s_t) and selects an action (a_t). The environment responds with a new state (s_{t+1}) and a reward (r_{t+1}) associated with the action. After each cycle, the agent updates the value function $V(s)$ or action-value function $Q(s, a)$ based on a certain policy π, where π maps states $s \in S$ to actions $a \in A$, i.e., $\pi : S \to A : a = \pi(s)$ (16; 17).

In RL problems with large state-action spaces, it can be cumbersome to store a separate value function for every possible state. Policy gradient methods were proposed as an alternative, which estimate the policy gradient and plug it into a stochastic gradient ascent algorithm. The gradient estimator has the form:

$$\hat{g} = \hat{\mathbb{E}}_t[\nabla\theta \log \pi_\theta(a_t|s_t)\hat{A}_t] \tag{6}$$

where π_θ is a stochastic policy parameterized by θ, and $\hat{A}_t$ is an estimator of the advantage function at time step t. PPO is a type of policy gradient method that uses a clipped surrogate objective function, which includes an estimator of the advantage function. The clipped surrogate objective function is defined as:

$$L^{CLIP}(\theta) = \hat{\mathbb{E}}_t[\min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon)\hat{A}_t)] \tag{7}$$

where θ represents the parameters of the policy, $r_t(\theta)$ is the ratio of the new policy to the old policy, and $\hat{A}_t$ is the estimated advantage function at time step t, which in our algorithm we use monte carlo method to estimate. The clipping parameter ϵ controls the maximum deviation of the new policy from the old policy, which is set as 0.2 in our algorithm. PPO's clipped surrogate objective function balances the trade-off between exploration and exploitation, and prevents the policy from deviating too far from the previous policy, leading to stable and effective learning in RL problems.

In addition to the surrogate objective, PPO also includes loss values for learning the critic functions and entropy to encourage exploration. The critic functions estimate the value function, which is the expected total discounted reward starting from a given state and following the current policy. The value loss is defined as the mean squared error between the estimated value and the actual value. The entropy loss encourages the policy to explore by adding a term that penalizes policies with low entropy, which measures the randomness of the policy distribution. The total loss function is a linear combination of the clipped surrogate objective, the value loss, and the entropy loss, with hyperparameters that control the relative weight of each term. The objective function is optimized using stochastic gradient descent or a variant of it, such as Adam or RMSprop, to update the policy parameters θ.

S3.2 Details about states, action spaces, and the reward assignment

Learning policies from model-free algorithms need a We introduce the states and action spaces, and the reward assignment in this section. The neural network architecture of the algorithm and the hyperparameter setting are given section 1.7.

In section 1.2, two phenotypes of the prostate cancer cells and the biomarker indicator (serum PSA level) were included in the system (5). Hence, at each time step t, an observation of cell counts ($x_{t,1}$ and $x_{t,2}$) and PSA level (p_t) was made as the current states $s_t = (x_{t,1}, x_{t,2}, p_t)$. Additional feature combinations of s_t can provide more information for model training. In a precise manner, the instant growth/decay rates $\dot{s_t}$ are indicative of the PCaC environment, reflecting the current drug and competition pressures, which can be obtained directly from the current states s_t. Moreover, time t was also included in the states. Hence, the states for PCaC environment was given by $S_t = (s_t, \dot{s_t}, t)$.

Moreover, the action space was discreetly composed by the doses of two drugs in time step t and can be formulated as $A_t = (I_{t,l}, I_{t,c})$.

For the purposes of this work, a successful treatment policy is defined as one in which resistance cancer cells, being suppressed, are co-living with response cancer cells to provide as long as high-quality survival time for patients. Within this measure of success optimality is defined as a trade-off between the highest survival time and the lowest expected cumulative dosing over the course of the treatment.

Using the combination of features S_t in the non-BlackBox model can help provide treatment information, which helps model training. Moreover, an additional feature combination of model-informed learning, which is the related information of states S_t provided by the system (5), can support the reward assignment.

In determining the explicit formulation of the reward function, the key is to describe the drug efficacy and the competition intensity in the PCaC environment. Drugs are assumed to affect the response population solely; thus, the change of the response population concentration $c_{1,t}$ provides a direct indicator of the drug efficacy, as follows:

$$r_{drug,t} = d_1(1 - c_{1,t}). \tag{8}$$

, where d_1 is a constant.

Furthermore, long-term control of prostate cancer involves the survival and development of resistant cancer cells. The only prohibited factor to the resistance population is the competition pressure from the response population. Hence, the competition intensity is included in the reward function, denoted as $r_{comp,t} = d_2(1 - c_{2,t})$, where $c_{2,t}$ denotes the resistance population concentration and d_2 is constant. To guide a low dosing and intermittent administration strategy, a penalty of continuous historic dosage was then assigned to the step reward:

$$p_{drug,t} = \sum_{\alpha} \sum_{i=t-\hat{t}}^{t} \eta^{t-i} w_\alpha \frac{I_{\alpha,i}}{I_{\alpha,max}}, \tag{9}$$

where $\hat{t}$ denotes the continuous administration time, η is the decaying penalty ratio for the historical drug administration, and w_α is the drug-specific penalty parameter. Eventually, the reward function was assigned as follows:

$$r_{step,t} = r_{drug,t} + r_{comp,t} - p_{drug,t}. \tag{10}$$

With limited resources in the tumor microenvironment, if insufficient dosages were administered, the responsive cancer cells proliferated and quickly reached their carrying capacity, making the system stay in an unwanted state with a high tumor level. The reward function (10), based on changes from one decision step to the next, provides an insufficient penalty. On the one hand, the penalty of the drug $p_{drug,t}$ is relatively low because of the small dosage. On the other hand, the $r_{camp,t}$ reaches its maximal value when the response population reaches the carrying capacity. Hence, the reinforcement learning agent may continue to apply a low-dosing strategy rather than increase the dosage to escape the high tumor level. This low-dosing strategy leads to the zero-dosing problem eventually after the model converges. To address this problem, we assign a progression-free time reward to the step reward function and adopt a metastasis probability model as EOS to avoid the high concentration of the response population.

S4. Proof of ultimate boundedness

From Equations (1, 3), we could rewrite Equation (1) as

$$\dot{x} = Xf(t,x), \tag{11}$$

where $\boldsymbol{x}(t) \in \Re^2_+$ denotes the cell counts for two phenotypes (responsive and resistant separately) and $X = \text{diag}\{x_1, x_2\}$. And $f(t,x)$ is a 2-dimensional function: $\Re_+ \times \Re^2_+ \to \Re^2$ is sufficiently smooth for the definition of $A(t)$ in Equation (3). Hence, we claim that the Equation (11) has a unique solution $\boldsymbol{x}(t; t_0, x_0)$ for all initial conditions $(t_0, \boldsymbol{x_0}) \in \Re_+ \times \Re^2_+$.

For the convenience of analysis, we rewrite Equation (11) as follows:

$$\begin{aligned} \dot{\boldsymbol{x}} &= X(R(\mathbf{1} - \mathcal{D}) + (-R(A(t)\boldsymbol{K}^{-1}\boldsymbol{x})^{\alpha})) \\ &= X(a(t,x) + A^{'}(t,x)\boldsymbol{x}), \end{aligned} \tag{12}$$

where $a(t,x) = d(t) = R(\mathbf{1} - \mathcal{D}) : \Re_+ \to \Re^2$, and $A^{'}(t,x) = -RX^{\alpha-1}\boldsymbol{K}^{-\alpha}A(t)^{\alpha} : \Re_+ \times \Re^2_+ \to \Re^{2\times 2}_+$ is a 2×2 functional matrix and $R = \text{diag}\{r_1, r_2\}, \boldsymbol{K} = \text{diag}\{K_1, K_2\}$.

Let us first give the definition of ultimate boundedness (following (18; 5)) as follows:

Definition 1 The solutions $\boldsymbol{x}(t; t_0, x_0)$ of system (12) are said to be ultimately bounded with respect to the region $\Re^2_+$ if there exists a compact region $\Omega \in \Re^2_+$ and a finite time $t_1 = t_1(t_0, x_0)$ such that for any $(t_0, x_0) \in \Re_+ \times \Re^2_+$ we have $x(t; t_0, x_0) \in \Omega$ for all $t \geq t_1$.

Second, from theorem (2.7) in (5) which is stated as,

Theorem 1 *The solutions $\boldsymbol{x}(t; t_0, x_0)$ of system (12) are said to be ultimately bounded with respect to the region $\Re^2_+$ if there exists a diagonal matrix D and a positive number η such that the matrix $B(t,x)$ defined below satisfies the inequality in 13, and $a(t,x) = d(t)$ is bounded from above as sure.*

Hence, to prove the ultimately bounded property for system (12), we first note that since $a(t,x) = d(t)$ in our case, which is bounded for sure. Then we have to find whether there exists a constant positive diagonal matrix $D = diag\{d_1, d_2\}$ and a positive number η such that the 2×2 symmetric matrix $B(t,x)$ satisfies the three conditions in Formula (13). We know that $d(t)$ is bounded since $\mathcal{D}$ denotes the drug effect in our scenario.

$$\begin{cases} B(t,x) = -\dfrac{1}{2}([A^{'}(t,x)]^T D + DA^{'}(t,x)), \\ min\{\lambda_1, \lambda_2\} \geq \eta; \lambda_m, m = \{1,2\} \quad \textit{is eigenvalues of} \quad B(t,x), \\ \forall (t,x) \in \Re_+ \times \Re^2_+. \end{cases} \tag{13}$$

Since $A(t) = \begin{Bmatrix} 1 & \frac{1}{1+e^{\gamma t}} \\ \frac{1}{1+e^{\gamma t}} & 1 \end{Bmatrix}$, with $a(t) = \frac{1}{1+e^{\gamma t}} \in (0, 0.5)$, by eigendecomposition, $A(t) = Q\Lambda(t)Q^{-1}$, where

$$Q = \begin{Bmatrix} 1 & 1 \\ 1 & -1 \end{Bmatrix}, \Lambda(t) = \begin{Bmatrix} 1 + a(t) & 0 \\ 0 & 1 - a(t) \end{Bmatrix}. \tag{14}$$

Then let $\Gamma(x) = RX^{\alpha-1}\boldsymbol{K}^{-\alpha} = \text{diag}\{\frac{r_1 x_1^{\alpha-1}}{K_1^{\alpha}}, \frac{r_2 x_2^{\alpha-1}}{K_2^{\alpha}}\}$,

$$\begin{aligned} A^{'}(t,x) &= -RX^{\alpha-1}\boldsymbol{K}^{-\alpha}A(t)^{\alpha} \\ &= -RX^{\alpha-1}\boldsymbol{K}^{-\alpha}Q\Lambda^{\alpha}Q^{-1} \\ &= -\Gamma(x)Q\Lambda^{\alpha}Q^{-1}. \end{aligned} \tag{15}$$

Let $D = diag\{\frac{K_1^{\alpha}}{r_1 K_1^{\alpha-1}}, \frac{K_2^{\alpha}}{r_2 K_2^{\alpha-1}}\}$, then

$$\begin{aligned} B(t,x) &= -\frac{1}{2}([A^{'}(t,x)]^T D + DA^{'}(t,x)), \\ B(t,c) &= -\frac{1}{2}(Q^{-1}\Lambda^{\alpha}Q diag\{c_1^{\alpha-1}, c_2^{\alpha-1}\} \\ &\quad + diag\{c_1^{\alpha-1}, c_2^{\alpha-1}\}Q\Lambda^{\alpha}Q^{-1}) \\ &= -\frac{1}{2}((1+a(t))^{\alpha} + (1-a(t))^{\alpha}) \begin{Bmatrix} 2c_1^{\alpha-1} & b(t,c) \\ b(t,c) & 2c_2^{\alpha-1} \end{Bmatrix}, \end{aligned} \tag{16}$$

where $b(t,c) = \frac{(1+a(t))^{\alpha} - (1-a(t))^{\alpha}}{(1+a(t))^{\alpha} + (1-a(t))^{\alpha}}(c_1^{\alpha-1} + c_2^{\alpha-1})$, with $c_i = \frac{x_i}{K_i}, i = 1, 2$.

Since $a(t) = \frac{1}{1+e^{\gamma t}}$ with $\gamma > 0$ and $a(t) \to 0$ as $t \to \infty$, we have $b(t,c) \to 0 \quad as \quad t \to \infty$, indicating that we could always find a T, satisfying when $t > T, 4(c_1 c_2)^{\alpha-1} - b(t,c)^2 > 0$. Hence, $|B(t,x)| > 0$ as $t > T$, proving that $B(t,x)$ is positive-definite when $t > T$. This proves the conditions are satisfied.

S5. Initial value setting for System (5)

To address the initial value problem, it is necessary to establish patient-specific and plausible initial values for system (5). However, due to limited patient-specific information, such as pre-treatment/post-treatment prostate volume, determining the true initial cell counts for each patient is challenging. Fortunately, the average prostate volume for all patients included in reference (7) is available. Consequently, we utilize Equation (17) to establish the initial values for system (5).

$$\begin{cases} x_1(0) = c_1 K_1, \\ x_2(0) = c_2 K_2, \\ K_1 = \dfrac{c_3 \bar{V} P_{max}}{\bar{P} V_{cell}}, \\ K_2 = c_4 K_1. \end{cases} \tag{17}$$

Here, $c_i, i \in 1, 2, 3, 4$ are constants, while $\bar{V}$ and $\bar{P}$ represent the average prostate volume and average initial PSA level for all patients, respectively. P_{max} denotes the maximum empirical PSA level for the patients, and V_{cell} refers to the volume of a single cell.

The parameter settings used in our simulation were determined based on a plausible configuration, with the constant c set to $c = (0.8, 10^{-4}, 1.25, 0.25)$. Initially, responsive cancer cells constituted the majority of the tumor microenvironment, while resistant cancer cells accounted for a small fraction ($O(10^{-5})$) of the population. Nevertheless, our simulation revealed that improved results could be obtained by configuring c_2 and c_4 as learnable parameters.

Note that the initial values are set accordingly. In the clinical trial's first treatment cycle, the average prostate volume decreased from a baseline of $24.7cm^3$ to $14.7cm^3$ (1). We assumed that cancer cells accounted for 50% of the prostate volume decline, resulting in an estimated average volume of $\hat{V} = 5cm^3$. The average pre-treatment PSA value was $\hat{P} = 22.1\mu g/L$ (7), and the volume of a single cancer cell was calculated as $V_{cell} = \frac{4}{3}\pi(5 \times 10^{-4}cm)^3 = 5.236 \times 10^{-10}cm^3$.

S6. Patient selection from the clinical trial data

Data of 91 patients in the clinical trial (1) were obtained from `https://www.nicholasbruchovsky.com/clinicalResearch.html` . For simplicity, in our simulation we only consider dual-drug effects. 19 patients in the folder "Shaw_et_al" were excluded because no drug information was available. We excluded 10 patients who were administrated with more than two drugs (patients 014, 022, 026, 028, 039, 041, 055, 064, 081, and 109). Eventually, we have 62 patients for our analysis.

S7. Neural network architecture and hyperparameter setting

Because the state space is continuous while the action space is discrete, we apply PPO Algorithms proposed by (6). The Q-network has four layers of fully-connected linear networks, for both the actor and the critic networks. Weights and bias were initialized randomly. Note that, applying more complicated networks such as the recurrent neural network and the gated recurrent units will not benefit the performance. No batch normalization was used at hidden layers, but features were re-scaled prior to being added to the replay buffer in the following manner, where K_i denotes the capacity for cell kind i and x_i is the corresponding cell counts, $\dot{x_i}$ is the growth rate.

$$
\begin{aligned}
x_i &\rightarrow \frac{log(x_i+1)}{log(K_i)},\\
\dot{x_i} &\rightarrow \begin{cases} \dfrac{log(\dot{x_i})+1}{log(K_i)}, x_i > 1, \\ \dfrac{log(-\dot{x_i})-1}{log(K_i)}, x_i < -1, \\ \dfrac{\dot{x_i}}{log(K_i)}, otherwise. \end{cases}
\end{aligned}
\tag{18}
$$

Adam optimization is applied with an initial learning rate of 3×10^{-5} for the actor network and 1×10^{-5} for the critic network. We apply the clipping gradient, a decaying learning rate, and the normalization of rewards to stabilize the training. Other parameter settings for PPO are adapted from the original paper, please refer to our GitHub page for details.

S8. Avoid the zero-dosing sub-optimal policy

To circumvent the zero-dosing issue, we employ two strategies. First, we propose a straightforward probability model of metastasis. Based on observations and experiments in (19; 20; 21), we adopt a probability model of cell concentration to simulate the metastasis process. Initially, we define sub-metastasis as occurring with a probability when the concentration of cancer cells c exceeds the threshold $\hat{c}$ (sub-metastasis occurs when $m_{sub}(t) = 1$, otherwise 0). The probability is proportional to $c^{2/3}$, where c represents the concentration of cancer cells. We assume the tumor lesion to be spherical, with only cells on the surface capable of detaching from the lesion and transferring to other organs. Given the carrying capacity in Equation (17), $K \sim 1/V_{cell}$, $r_{cell} \sim K^{-1/3}$. Consequently, the surface area is proportional to $c^{2/3}$.

Taking into account the micro-environmental changes, prostate cancer cells transferred during sub-metastasis have a relatively low survival rate in other organs. Therefore, the confirmation of final metastasis (denoted as $m_{final} = 1$) occurs when sub-metastasis takes place n times, and the final metastasis serves as an additional criterion for the end of the study (EOS), defined as follows:

$$
m_{final} := \delta(\sum m_{sub} < n) = \begin{cases} 0, & if \sum m_{sub} < n \\ 1, & otherwise, \end{cases} \tag{19}
$$

Here, $\delta(\cdot)$ represents a binary-valued function, and

$$
m_{sub} \sim \begin{cases} Bernoulli(c^{2/3}), & if\ \hat{c} < c \\ 0, & otherwise \end{cases} \tag{20}
$$

With this metastasis model, if the agent administers an insufficient drug dosage to patients, responsive cancer cells will rapidly reach their capacity, and metastasis will occur swiftly, resulting in a low reward.

Second, to encourage the agent to optimize its performance, we provide a linearly increasing instant reward as follows:

$$
r_{supp,t} = t \times (1 - c_{2,t}), \tag{21}
$$

where $c_{2,t}$ denotes the cell concentration of the resistant population at time t, and t represents time in months, also the number of steps during one episodic sampling.

By employing these two strategies, the agent can effectively avoid the sub-optimal zero-dosing policy.

S9. Leave-pair-out cross validation

To further assess the prediction accuracy of our proposed tM-GLV model, we train the model for each patient 10 times by randomizing the longitudinal data into validation and training sets (20% for validation and 80% for training). The 95% confidence interval (CI) for the patient-specific parameters can be found in Supplementary File 1. The average value, along with the 95% CI for all parameters for all patients, is provided in Supplementary File 2.

In order to better establish the resistance index as the threshold for distinguishing resistant patients from responsive patients, we employed a leave-pair-out cross-validation, separating responsive and resistant groups. The thresholds were determined to maximize sensitivity and specificity in the training set. Using these thresholds, the model classified patients in the testing set as either responsive or resistant. The overall accuracy achieved an average score of 85.7%, with a narrow 95% CI ranging between 83.7% and 87.7%. Specificity ranged between 91.7% and 92.5%, while sensitivity varied between 91.7% and 92.3%.

Supplementary File 1: 95% CI for all the patient-specific parameters.

Supplementary File 2: The population means for all parameters and their 95% CI.

S10. Comparison of the I^{2}ADT-derived adaptive dosing policies of resistant and response patients

Upon analyzing the results of I^{2}ADT, we identified two notable differences between the resistant and responsive populations.

For responsive patients, an ascending drug administration pattern emerged as drug pressure on responsive cancer cells increased throughout the therapy. For the overall responsive patients, the treatment off-to-on ratio, daily CPA dosage, and monthly LEU dosage can be found in Supplementary Fig 1. As therapy progressed, a decreasing pattern in the treatment off-to-on ratio was observed. A significant increase in drug dosing was evident in I^{2}ADT, while a decreasing pattern was apparent in standard IADT.

For resistant patients, a descending policy was observed, with the treatment off-to-on ratio increasing over time. The average off-to-on treatment ratio and daily CPA and monthly LEU dosages across time are presented in Supplementary Fig 2. During the course of treatment, an increasing pattern in the treatment off-to-on ratio was observed, while the CPA/LEU pattern remained relatively flat compared to the standard IADT, which exhibited an increasing pattern. This insignificancy may due to the number of patients in the resistance population is only 11.

S11. Other supplementary files

Supplementary File 3: Justify the choice of γ as resistance index. We use all the parameters calibrated with the clinical data as the classify. The results of AUC and confusion matrix are shown in this supplementary file. Except the parameter alone, we also show the results with r_2/r_1 as classify. From AUC and confusion matrix, we empirically claim that γ has the best power to differ the resistance from response.

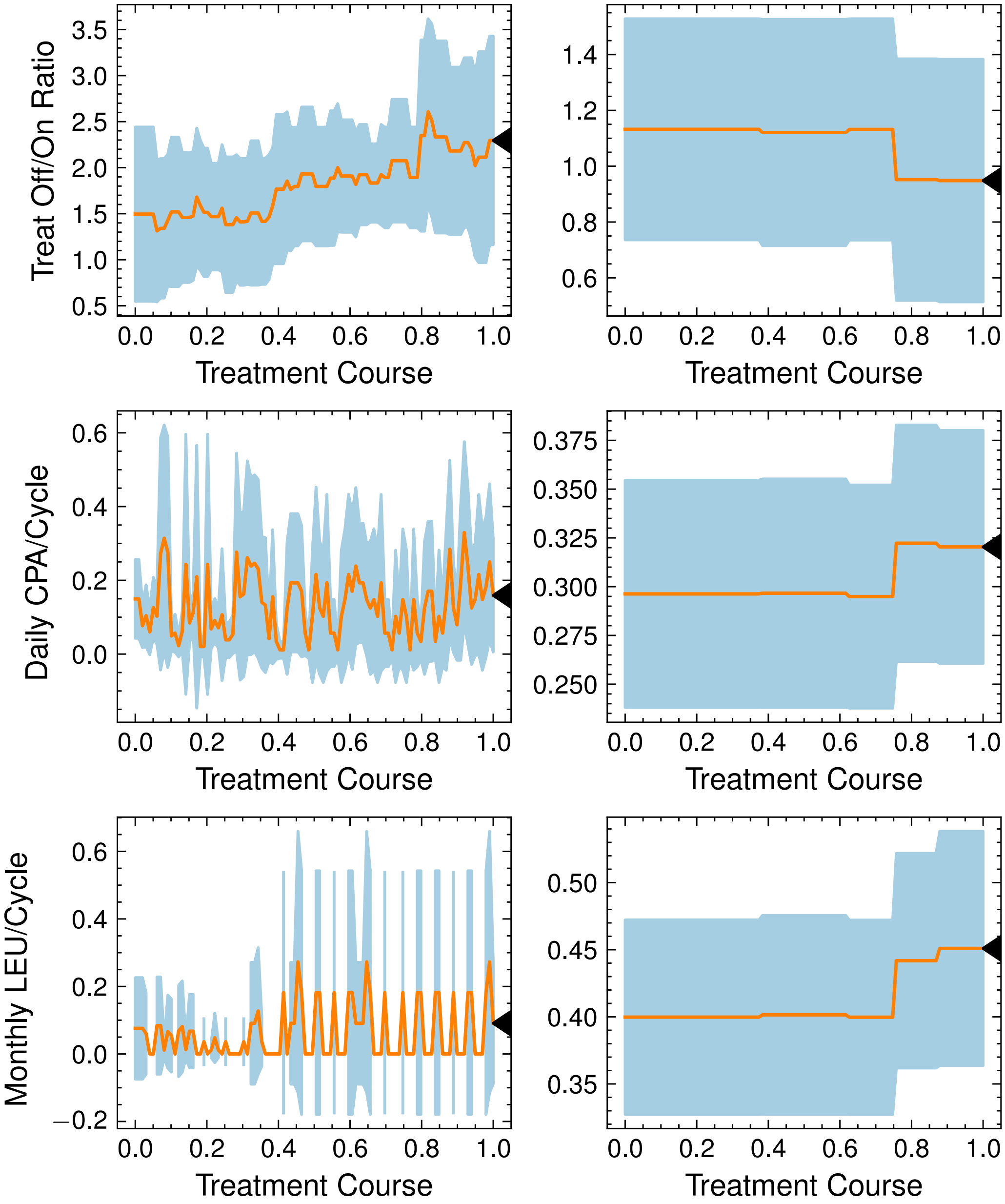


Fig. 1. The treatment off/on ratio, daily CPA dosing, and monthly LEU dosing of the I^2ADT (left) and the standard IADT (right) for responsive patients.

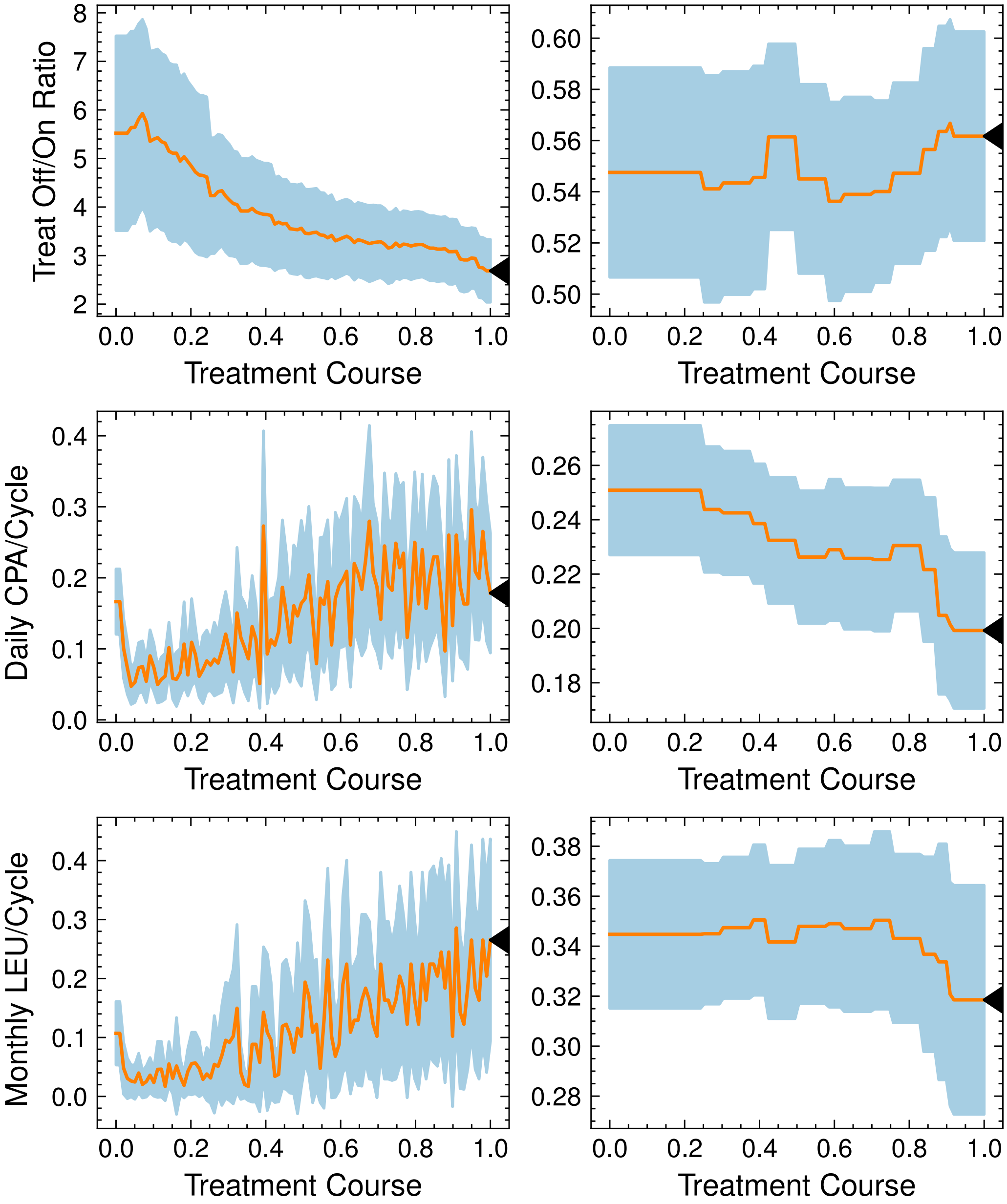


Fig. 2. The treatment off/on ratio, daily CPA dosing, and monthly LEU dosing of the I^2ADT (left) and the standard IADT (right) for resistance patients.

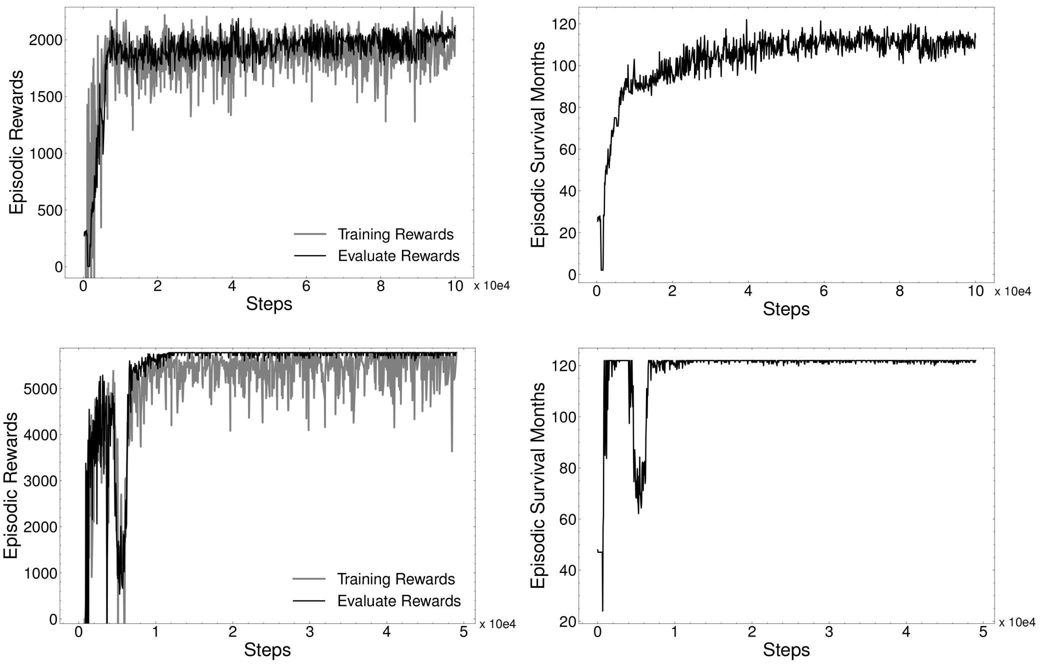


Fig. 3. The episodic reward for training and evaluation reward/survival month with greedy strategy.

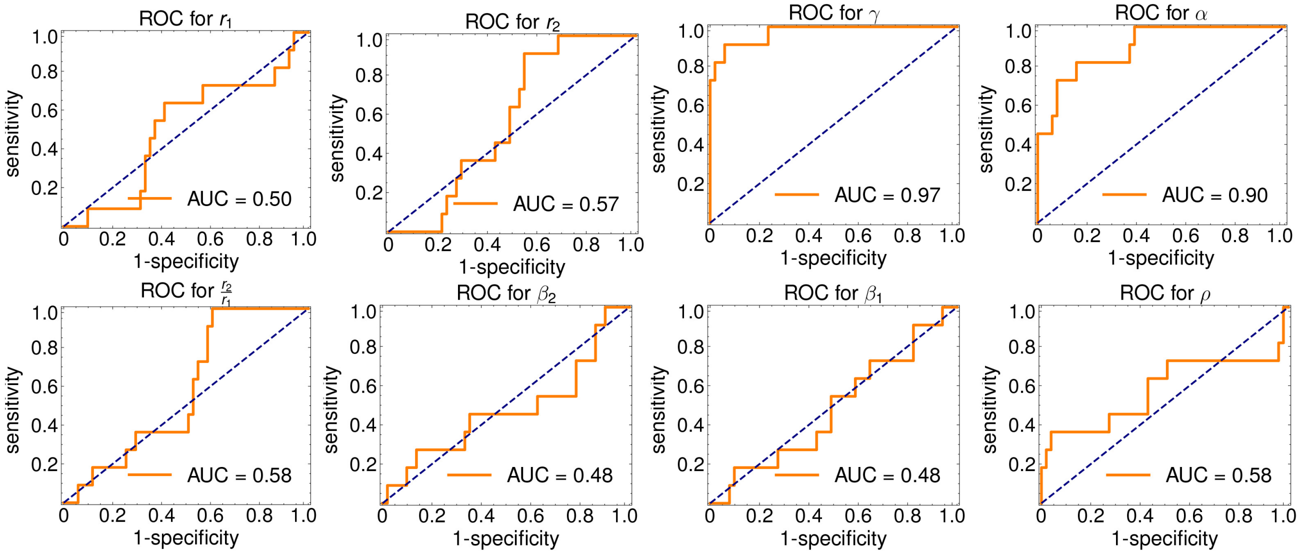
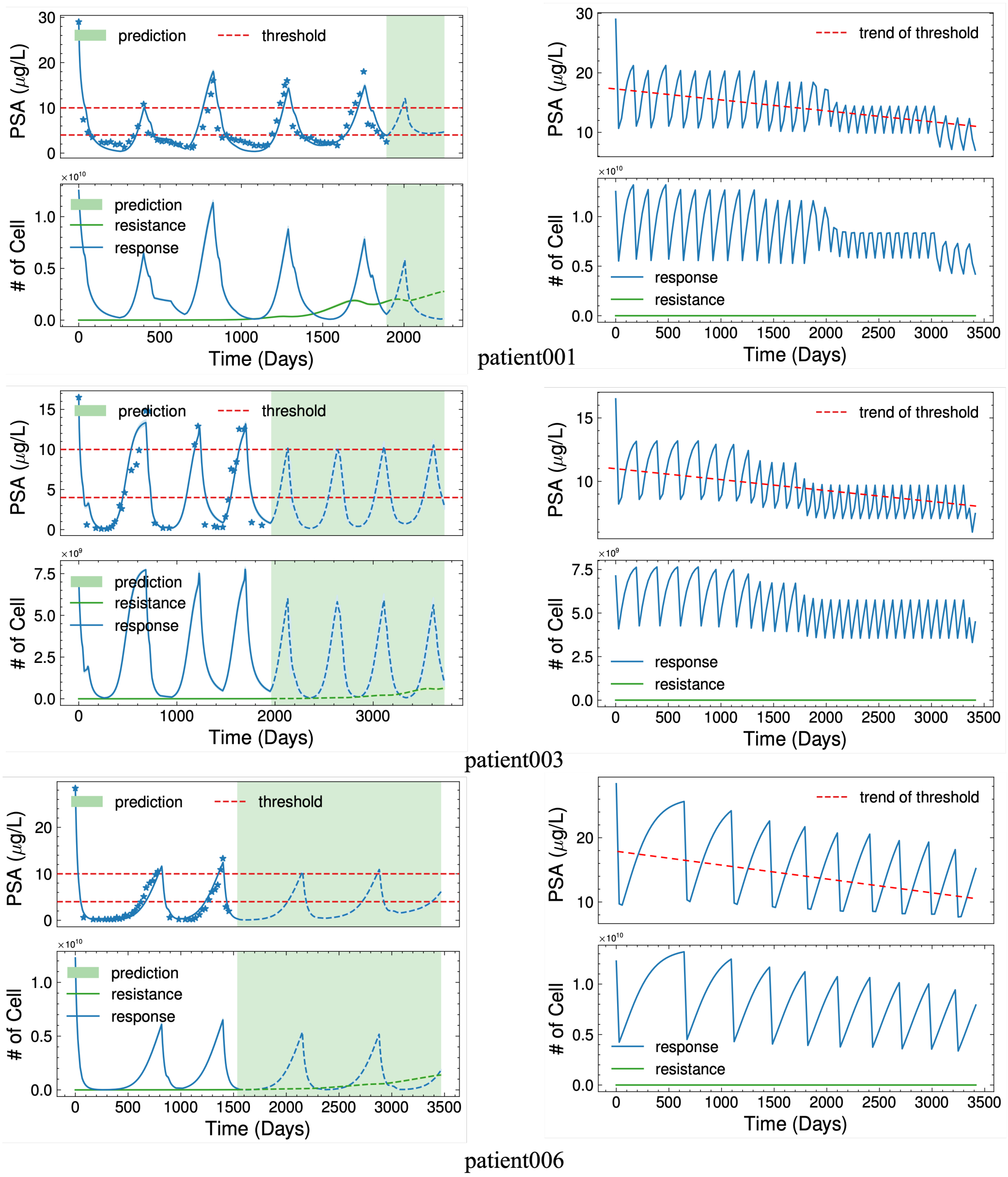


Fig. 4. This figure illustrates the use of all parameters, along with the r_2/r_1 ratio, as classifiers to differentiate between responsive and resistant cases. Apart from γ and α, the remaining parameters demonstrate no predictive power in classification.

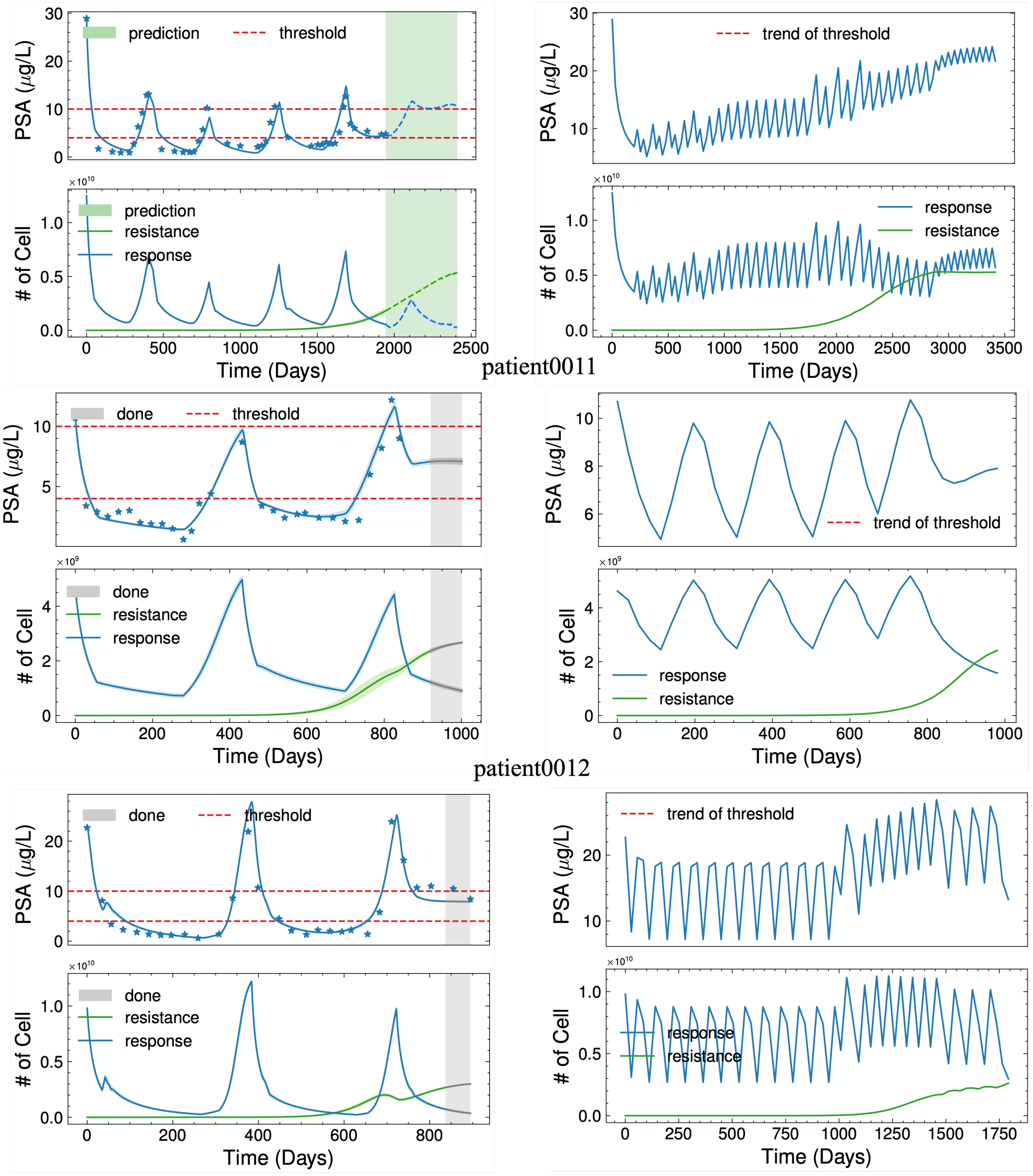


patient001

patient003

patient006

Fig. 5. This figure shows the three patients' evolutionary dynamics for response group with traditional IADT (left) and I^2ADT (right) separately.



patient0011

patient0012

patient0052

Fig. 6. This figure shows the three patients' evolutionary dynamics for resistance group with traditional IADT (left) and I^2ADT (right) separately.

References

1. Bruchovsky N, Klotz L, Crook J, Malone S, Ludgate C, Morris WJ, 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: clinical parameters. Cancer. 2006;107(2):389–395.
2. Ribba B, Holford NH, Magni P, Trocóniz I, Gueorguieva I, Girard P, et al. A review of mixed-effects models of tumor growth and effects of anticancer drug treatment used in population analysis. CPT: pharmacometrics & systems pharmacology. 2014;3(5):1–10.
3. Marusyk A, Almendro V, Polyak K. Intra-tumour heterogeneity: a looking glass for cancer? Nature Reviews Cancer. 2012;12(5):323–334.
4. Ikeda M, Šiljak D. Lotka-Volterra equations: decomposition, stability, and structure. Journal of Mathematical Biology. 1980;9(1):65–83.
5. Ikeda M, Šiljak D. Lotka-Volterra equations: decomposition, stability, and structure part II: nonequilibrium analysis. Nonlinear Analysis: Theory, Methods & Applications. 1982;6(5):487–501.
6. Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. Proximal policy optimization algorithms. arXiv preprint arXiv:170706347. 2017;.
7. Bruchovsky N, Klotz L, Crook J, Phillips N, Abersbach J, Goldenberg SL. Quality of life, morbidity, and mortality results of a prospective phase II study of intermittent androgen suppression for men with evidence of prostate-specific antigen relapse after radiation therapy for locally advanced prostate cancer. Clinical genitourinary cancer. 2008;6(1):46–52.
8. Jentsch D, Schulz V, Wendt H. Pharmacokinetics of cyproterone acetate in normal subjects after im and oral application (authorś transl). Arzneimittel-forschung. 1976;26(5):914–919.
9. Periti P, Mazzei T, Mini E. Clinical pharmacokinetics of depot leuprorelin. Clinical pharmacokinetics. 2002;41(7):485–504.
10. Sharifi R, Soloway M, Correa Jr RJ, Glass AG, Guinan PD, Garnick MB, et al. Clinical study of leuprolide depot formulation in the treatment of advanced prostate cancer. The Journal of urology. 1990;143(1):68–71.
11. Drugs FA. Lupron Depot (leuprolide acetate for depot suspension); 1990.
12. Dlugi AM, Miller JD, Knittle J, Group LS, et al. Lupron depot (leuprolide acetate for depot suspension) in the treatment of endometriosis: a randomized, placebo-controlled, double-blind study. Fertility and sterility. 1990;54(3):419–427.
13. Richardson TD, Wojno KJ, Liang LW, Giacherio DA, England BG, Henricks WH, et al. Half-life determination of serum free prostate-specific antigen following radical retropubic prostatectomy. Urology. 1996;48(6):40–44.
14. Stamey TA, Yang N, Hay AR, McNeal JE, Freiha FS, Redwine E. Prostate-specific antigen as a serum marker for adenocarcinoma of the prostate. New England Journal of Medicine. 1987;317(15):909–916.
15. Oesterling JE, Chan DW, Epstein JI, Kimball Jr AW, Bruzek DJ, Rock RC, et al. Prostate specific antigen in the preoperative and postoperative evaluation of localized prostatic cancer treated with radical prostatectomy. The Journal of urology. 1988;139(4):766–772.
16. Kaelbling LP, Littman ML, Moore AW. Reinforcement learning: A survey. Journal of artificial intelligence research. 1996;4:237–285.
17. Sutton RS, Barto AG. Reinforcement learning: An introduction. MIT press; 2018.
18. Šiljak D, Weissenberger S. Regions of absolute ultimate boundedness for discrete-time systems. at-Automatisierungstechnik. 1973;21(1-12):329–332.
19. Luzzi KJ, MacDonald IC, Schmidt EE, Kerkvliet N, Morris VL, Chambers AF, et al. Multistep nature of metastatic inefficiency: dormancy of solitary cells after successful extravasation and limited survival of early micrometastases. The American journal of pathology. 1998;153(3):865–873.
20. Rejniak KA, Anderson AR. Hybrid models of tumor growth. Wiley Interdisciplinary Reviews: Systems Biology and Medicine. 2011;3(1):115–125.
21. Aceto N, Bardia A, Miyamoto DT, Donaldson MC, Wittner BS, Spencer JA, et al. Circulating tumor cell clusters are oligoclonal precursors of breast cancer metastasis. Cell. 2014;158(5):1110–1122.