Briefings in Bioinformatics, 2024, **25(2)**, bbae071
https://doi.org/10.1093/bib/bbae071
Problem Solving Protocol

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

Yitao Lu, Qian Chu, Zhen Li, Mengdi Wang, Robert Gatenby and Qingpeng Zhang

Corresponding author: P307J, Graduate House, HKU, HK 00001, China. Tel.: +852-39179024; Fax: +852-28170859; E-mail: 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 and Leuprolide acetate for individual patients. Then, we proposed a reinforcement-learning-enabled individualized IADT framework, namely, I^2ADT, to learn the patient-specific tumor dynamics and derive the optimal drug administration policy. Experiments with clinical trial data demonstrated that the proposed I^2ADT 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^2ADT 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.

Keywords: Adaptive therapy; Prostate cancer; Personalized medicine; Reinforcement learning

INTRODUCTION

Prostate tumor is the second most prevalent cancer and the sixth leading cause of cancer death worldwide [1, 2]. The common treatments of locally advanced prostate cancer are radiotherapy and hormone therapy [3, 4]. Hormone therapy, such as androgen deprivation therapy (ADT), is an effective treatment and is usually applied after the failure of radiotherapy [5, 6]. Similar to other hormone therapies, ADT has side effects, including decreased libido, impotence, hot flashes and sexual effects [7, 8].

The difficulty in treating prostate cancer lies in the development of resistance, which usually leads to treatment failure and tumor progression [8, 9]. There are multi-type cancer cells competing for resources in the resource-limited tumor microenvironment. Such Darwinian dynamics can lead to a rapid proliferation of resistant population. In the conventional drug administration policy, the use of maximum tolerated dose until progression can give the resistant phenotype an advantage over the other competitors, leading to the faster development of tumor resistance [8]. Thus, the intermittent androgen deprivation therapy (IADT) was proposed and validated in several clinical trials especially for patients in phase 1b / phase 2 [6, 10].

Figure 1 illustrates the idea of ADT (a) and IADT (b) separately. If a maximum tolerated dose is adopted, the resistant phenotype has an evolutionary advantage over the responsive phenotypes, leading to quick tumor resistance to ADT (shown in Figure 1a). IADT involves the intermittent administration of drugs, which

Yitao Lu is currently a PhD student with the School of Data Science at the City University of Hong Kong. His research interests lie in the application of reinforcement learning of deep learning in cancer treatments by integrating statistics, computer science, and domain-specific knowledge.
Qian Chu, MD, PhD, is a Professor in Medical Oncology and Radiation Oncology at Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology. Her research focuses on the molecular genetics of lung cancer and glioblastoma, with a particular interest in the Notch pathways and their role in cancer stem cells.
Zhen Li, PhD, is a Professor with the Department of Radiology, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology. His current research focuses on overcoming the critical hurdles that have impeded the development of robust AI models, which include deficiency, isolation, and heterogeneity of CT data generated from diverse institutions.
Mengdi Wang, PhD, is an Associate Professor with the Department of Electrical and Computer Engineering and the Center for Statistics and Machine Learning at Princeton University. Her research interests include data-driven optimization, reinforcement learning, statistical dimension reduction, and system identification.
Robert Gatenby, MD, is the Department Chair of Radiology and co-director of the Cancer Biology and Evolution Program at the H. Lee Moffitt Cancer Center and Research Institute. His research focuses on theoretical and experimental models of evolutionary dynamics in cancer and cancer drug resistance and a tumor's physical microenvironment.
Qingpeng Zhang, PhD, is an Associate Professor at the Musketeers Foundation Institute of Data Science and the Department of Pharmacology and Pharmacy, LKS Faculty of Medicine at The University of Hong Kong. His research interests include medical informatics, AI in drug discovery, and network science.
Received: October 9, 2023. **Revised:** January 11, 2024. **Accepted:** February 3, 2024

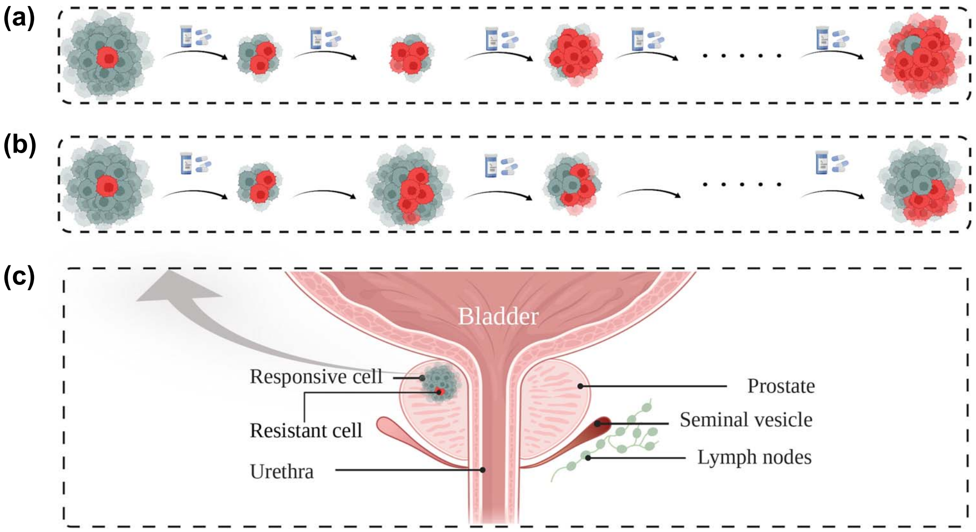

Figure 1. Illustration of ADT and IADT separately (Created with BioRender.com). The three panels, labeled (**A**), (**B**) and (**C**), depict green cells as responsive cancer cells and red cells as resistant cancer cells. In the lowest panel **C**), we illustrate the tumor lesion of prostate cancer. In **A**), continuous dosing kills responsive cancer cells, but the resistant cancer cells quickly dominate the population due to the lack of competition from responsive cells. In contrast, **B**) shows that intermittent dosing allows responsive cancer cells to regrow during therapy-free periods. This enables them to compete more effectively, resulting in the resistant cells being unable to dominate the entire population even in the late stages of treatment.

gives the responsive phenotype the chance to compete with the resistant phenotype (shown in Figure 1b), thus prolonging the time to progression (TTP). In addition, IADT provides quality-of-life benefits by reducing the cumulative drug dosage.

There are two potential design flaws in conventional IADTs. First, many begin with ‘induction treatment” in which ADT is applied at maximum dose continuously for 8 to 9 months and intermittent therapy is then applied only if the prostate-specific antigen (PSA) has been reduced to the normal range. Evolutionarily, this has the effect of strongly selecting for resistance while removing most of the sensitive population. Thus, the critical evolutionary competition that is necessary for suppression of the resistant cells by the expanding sensitive population during treatment cessation is lost. Second, IADTs often impose a rigid treatment schedule which neglects the heterogeneity within the tumor–host interactions which can result in very different proliferation and death rates of prostate cancer cells in different patients. Recent clinical trials and computer simulation studies have suggested an alternative approach with no induction treatment and using pre-define PSA thresholds for suspending and resuming ADT administration [11] may be more successful than the prior trial designs. However, even if such population-based IADTs are effective, they do not fully utilize the patients’ characteristics and clinical information. Thus, they are sub-optimal and the full benefit of IADT in personalized medicine has not been obtained.

Here we hypothesize that optimal evolution-based therapies require detailed integration of patient-specific, treatment-specific and tumor-specific dynamics into the treatment protocol. However, this task is computationally challenging because of the high complexity of models considering the intratumoral dynamics and actual data.

We build upon prior applications and validations of various evolutionary mechanisms for simulating the intratumoral dynamics and IADT responses [10, 12–14]. For example, the classic Lotka–Volterra model was incorporated into an evolutionary game model to simulate the competition mechanism between responsive and resistant tumors [10]. Another recent study explained the resistance occurrence of prostates cancer [14] in IADT by considering stem cells differentiation and evolution.

In addition, artificial intelligence (AI) techniques, particularly reinforcement learning, are promising tools for making optimal treatment decisions that consider different patients’ heterogeneity and tumor’s evolutionary mechanism [15–20]. Recently, model-free reinforcement learning methods have been applied to the dynamic control of cancer. For example, an agent-based model with an associated reinforcement learning framework was proposed to continuously control the drug administration and dynamically regulate the emergence of resistant tumors [21]; this was a theoretical study that was conducted without patient customization or actual clinical data. Another study used reinforcement learning to inform the automated dose adaptation, which achieved human-similar results in non-small cell lung cancer patients [22]. This work did not incorporate the intratumoral evolutionary mechanisms. In both studies, the pharmacokinetics of specific drugs was not considered.

To address the above-mentioned challenges, we propose the reinforcement-learning-enabled individualized IADT (I^2ADT) framework, which learns the patient-specific tumor dynamics from actual patient data and derives the optimal drug administration policy for individual patients using reinforcement learning. Experiments with a multi-center Phase II clinical trial data demonstrate that I^2ADT leads to longer TTP and lower cumulative drug dosage compared with the conventional standard IADT adopted by the trial. Additional validation with external data also demonstrated the efficacy of I^2ADT.

It is important to highlight that our proposed I^2ADT model is fundamentally a data-driven approach, developed based on the clinical data available to us. Consequently, the model’s scope is inherently bound by the extent of these data. As such, certain critical factors, such as age, genetics and lifestyle, were not included in our model due to data accessibility constraints.

In more detail, our model employs a nonlinear, time-varied approach to capture the dynamic interactions between two phenotypes, influenced by key evolutionary processes. This methodology is elaborated in our Methods (2.1) and further in Supplementary S2. We theorize that the model implicitly learns from the clinical data about factors not directly accessible, represented as parameters in Methods (2.1). Significantly, the I^2ADT model is designed to be versatile and adaptable. This flexibility suggests that, with the availability of more comprehensive data, I^2ADT can seamlessly incorporate these additional elements. Such integration would undoubtedly broaden the model's scope, enriching its comprehensiveness and enhancing its practical application.

The contributions of this paper are 4-fold. First, we formulated the patient-specific tumor dynamics by a proposed time-varied, mixed-effect, generalized Lotka–Volterra (tM-GLV) model to describe the dynamic competition between two phenotypes that are sensitive or resistant to treatment. Second, we proposed a deep reinforcement-learning-based framework to define patient-specific tumor evolutionary dynamics and integrate them into the treatment strategy over time. Third, we combined the PSA level and pharmacokinetics to inform the personalized IADT. Fourth, this is a data-driven deep reinforcement learning method for individualized IADT and could be adaptively extended to other cancer types.

MATERIALS AND METHODS

In the Phase II clinical trial, clinicians adopted a population-based policy to treat patients for up to 32–36 weeks in each treatment cycle until progression [6]. 29 out of the 91 patients were excluded from our analysis due to missing data on drug dosages, and some patients took multiple drugs (please refer to the Supplementary S6 for further information). We divided each patient's longitudinal data into training and validation sets through stratified random sampling: 20% of each patient's data in each cycle was randomly selected and removed as the validation set. In the following sections, we introduce the tM-GLV model and essentials of the reinforcement learning for thr proposed I^2ADT.

Ethical approval: This is a retrospective secondary data analysis of open-source database. No ethical approval is required.

Modeling the Prostate Cancer Competition environment

The dynamics of prostate cancer evolution are difficult to characterize with full details, because of the presence of many interacting factors [23]. Following a system control approach, we formulate the ecosystem into a mathematical model capturing the key processes in the population level, namely, selection, competition, mutations, epigenetic modifications and adaptation [24, 25].

Based on the literature [24–29], we developed a time-varied mixed-effect generalized Lotka–Volterra (tM-GLV) model with the aforementioned four processes. Tumors have inherent heterogeneity, and we can usually assume that two phenotypes of prostate cancer cells are present [30] before treatment, namely responsive (hormones-dependent) and resistant (hormones-independent) cells. The resistant cancer cells were minority at first, but they could gain advantage under the androgen suppression conditions. The two phenotypes have fierce competition in the tumor microenvironment because of the high demand for resources [31]. Besides the four processes, we also account for metastasis in the tM-GLV model by adopting a probability model (please refer to Supplementary S8 for details).

Given the context of two competing phenotypes, which are viewed as permanent bounded variations of the system described by system (1), an equilibrium, even if it exists, is never established. Hence, a non-equilibrium model with time-varied disturbances is appropriate to simulate the dynamics with competition-induced mutations. The role of the equilibrium is replaced by the ultimate boundedness, which ensures the populations remain restricted to a certain limiting value in finite time [32, 33]. Hence, the system defined by (1) has no equilibrium but an ultimate boundedness that produces a compact set of values in the states space. When the environment reaches these states, the environment is bounded in this compact set (the proof is given in the Supplementary S4).

Table 1: Definitions of notations in system (1)

Notation	Description	Units
$\mathbf{x}$	(x_1, x_2), vector of the cell counts for two phenotypes	1^1
P	serum PSA level	$\mu g/L$
R	$\text{diag}\{r_1, r_2\}^2$, inherent growth rate for two phenotypes	1/day
X	$\text{diag}\{x_1, x_2\}$, cell counts for two phenotypes	1
$\mathbf{1}$	$(1, 1)$, vector of 1	1
$\mathbf{A}(t)$	time-varied competitive community matrix	1
K	$\text{diag}\{K_1, K_2\}$, carrying capacity for two phenotypes	1
$\mathcal{D}$	drug pressure: drug-induced decay rate for cells	1
α	constant, hyper-parameter of competition-induced decay	1
ρ	the secretion rate of PSA	$\mu g/L$/day
ϕ	the decay rate of PSA	1/day

[1]denotes no unit for this variable. [2] denotes diagonal matrix with the diagonal entries are r_1 and r_2

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

Within the confines of our tM-GLV model, several crucial factors were omitted owing to the limitations in data availability. It is widely acknowledged that variables such as age, weight and genetic factors play significant roles in cancer treatment, as cited in relevant literature. Furthermore, diet and lifestyle are known to substantially influence PSA levels. Unfortunately, due to the unavailability of comprehensive data on these factors, they were not included in our model.

However, we postulate that the clinical outcomes and treatment sequences for each individual are influenced by these factors. And we have learnt the parameters of the model from clinical data. This approach allows us to infer that these parameters, albeit indirectly, encapsulate the effects of the aforementioned factors. Thus, while not explicitly included, the influence of age, weight, genetics, diet and lifestyle is subtly integrated into the model via the parameters learned from clinical outcomes.

To search for optimal patient-specific parameters, we utilized a PyTorch-based solver named ξ-torch [34] to solve the ordinary differentiated equations (refer to Supplementary S5 for initial settings) and Adam optimizer [35] to minimize the least square error between the model-predicted PSA and the ground truth PSA. Considering that the number of cells is tens of millions, a clipping gradient is applied to the optimization process to avoid the gradient explosion.

Model-informed treatment planning with reinforcement learning

While a predictive understanding of the exact evolutionary trajectories toward resistance is essential for effective treatments, it remains a significant challenge. However, controlling the evolution of drug-resistant cancer does not require full predictability or determinism. In a closed-loop system, feedback can help mitigate the reliance on precise trajectory knowledge, as long as the feedback can be obtained at reasonable time intervals; the uncertainty and stochasticity can be approximated in an informed manner, and the controller is robust to changes in the system's behavior and parameter fluctuations [21]. In this work, we apply reinforcement learning as the controller.

Modern RL algorithms can be basically classified into two branches: value-based and policy-based learning algorithms [36]. Deep Deterministic Policy Gradient (DDPG) [37], Trust Region Policy Optimization (TRPO) [38], Proximal Policy Optimization (PPO) [39] and Soft-Actor-Critic (SAC) [40] are powerful algorithms that have been proposed in recent years. However, each algorithm has its strengths and limitations. DDPG is an off-policy actor-critic deterministic algorithm that can only be applied in continuous states and action spaces. TRPO is an on-policy algorithm that uses KL-divergences to control the updating from old policy to new policy. However, its second-order optimization makes it difficult to implement or fine-tune the hyperparameters. Both SAC and PPO are easily implemented and suitable for discrete or continuous action-state spaces, and obtains high data-efficiency and reliable performance. We empirically tested both algorithms and found that applying SAC in our scenario required cautious hyperparameter fine-tuning for each patient; otherwise, it would easily lead to divergence.

In terms of the Prostate Cancer Competition (PCaC) environment constructed in Methods (2.1), though the system provides an environment model (1), obtaining a full-knowledge transition from states to states is challenging. Therefore, an optimal policy for this decision-making process needs to provide dosing guidelines for the next time step based on potential stochastic evolutionary scenarios described by the system (1). The system involves a continuous state space (cell population composition), time-varying and potentially high stochasticity and one or multiple controls (drugs) that can take continuous values when administered intravenously or discrete values when administered orally. Although we use discrete action spaces in our setting, it should be easily generalized into high-dimensional continuous action space in practice. Therefore, a versatile model-free RL algorithm, namely Proximal Policy Optimization (PPO) [39], is preferred due to the complexity of the evolutionary dynamics.

In this work, we use the simulation of the tM-GLV model to train the agent to get the optimal policy. At each time step, the microenvironment information (such as cell counts and PSA levels) is sampled from the non-BlackBox system (1), by which the agent takes actions, and then the PPO trains the agent. To determine the explicit formulation of the reward function, the key is to describe the drug efficacy and the competition intensity in the PCaC environment. In addition, penalty of dosage history is assigned to the reward. Please refer to the Supplementary S3 for detailed description of the states-actions spaces and the rewards assignments.

Note that insufficient dosage can suppress resistant population and reduce cumulative dosage, thus may lead to a sub-optimal policy in which the agent lets responsive population proliferate without control, leading to metastasis and disease progression (denotes as response cancer cells reach high rate of its capacity). Therefore, we will assign a progression-free time reward to the step reward function and apply the aforementioned metastasis probability model as a stopping criteria to avoid the high concentration of the responsive population.

Validation of I^{2}ADT

In this section, we showcase the external validation of our proposed I^{2}ADT approach and address the critical question of how I^{2}ADT can be applied to a new patient.

We used pilot clinical trial data (NCT02415621) from the H. Lee Moffitt Cancer Center, where 17 patients were enrolled in a study group following an IADT dosing policy. This policy required suspending the drug (abiraterone) when PSA levels dropped below 50% of the pretreatment value and resuming treatment when PSA levels returned to baseline. It is essential to note that this suspension criteria is population-based, not patient-specific or optimal. We applied the patients' clinical data to train the patient-specific tM-GLV model, then used PPO to obtain individualized policies, following the same methodology described in the earlier Methods sections.

In addition to external data validation, we introduce a new treatment protocol called *delayed*-I^{2}ADT. When treating a new patient, the clinician develops their treatment policy based on all available clinical and pathological information up to a fixed time stamp, such as a full cycle of standard IADT treatment. To collect more data for model calibration, we implement a weekly PSA test during the first IADT treatment cycle, though the dosing policy is updated monthly. The details are as follows.

We begin by collecting the first τ-month of weekly data from the standard IADT treatment, which includes the initial tumor lesion size, serum PSA level and corresponding dosing history, denoted as $\mathcal{X}(0 : \tau)$. Using the clinical trial data $\mathcal{X}(0 : \tau)$, we propose Algorithm 1 in Supplementary S1 to obtain the tM-GLV model $\mathcal{M}(a; \theta^0_{new})$.

Next, we train the reinforcement learning agent for the tM-GLV model. Once the agent converges, it is said the agent has converged to $\mathcal{A}(\theta^k_A)$. From $\mathcal{A}(\theta^k_A)$, we can predict and obtain the dosing policy for the next T months ahead, which clinicians will review and administer to patients. Clinical surveillance is performed throughout the entire period to gather additional data for tM-GLV model training. The details of the algorithm are shown in Algorithm 2 in Supplementary S1.

RESULTS

Mathematical modeling and simulation of prostate cancer cell evolution

The evolutionary dynamics of prostate cancer cells *in vivo* are complicated. Our model considered two phenotypes, namely responsive (Hormone-Dependent) and resistant (Hormone-Independent). In the beginning, responsive cancer cells dominate the population, and resistant cancer cells account for only a tiny portion due to the inherent heterogeneity and the healthier fitness of responsive phenotype.

According to biological theories [24, 25], two key processes contribute to the evolution of cancer, namely, selection and adaptation, along the way with competition, mutations and epigenetic modifications. Resistant cancer cells are minor, but they exist before the treatment and will take their place under the androgen suppression conditions. A fierce competition likely exists

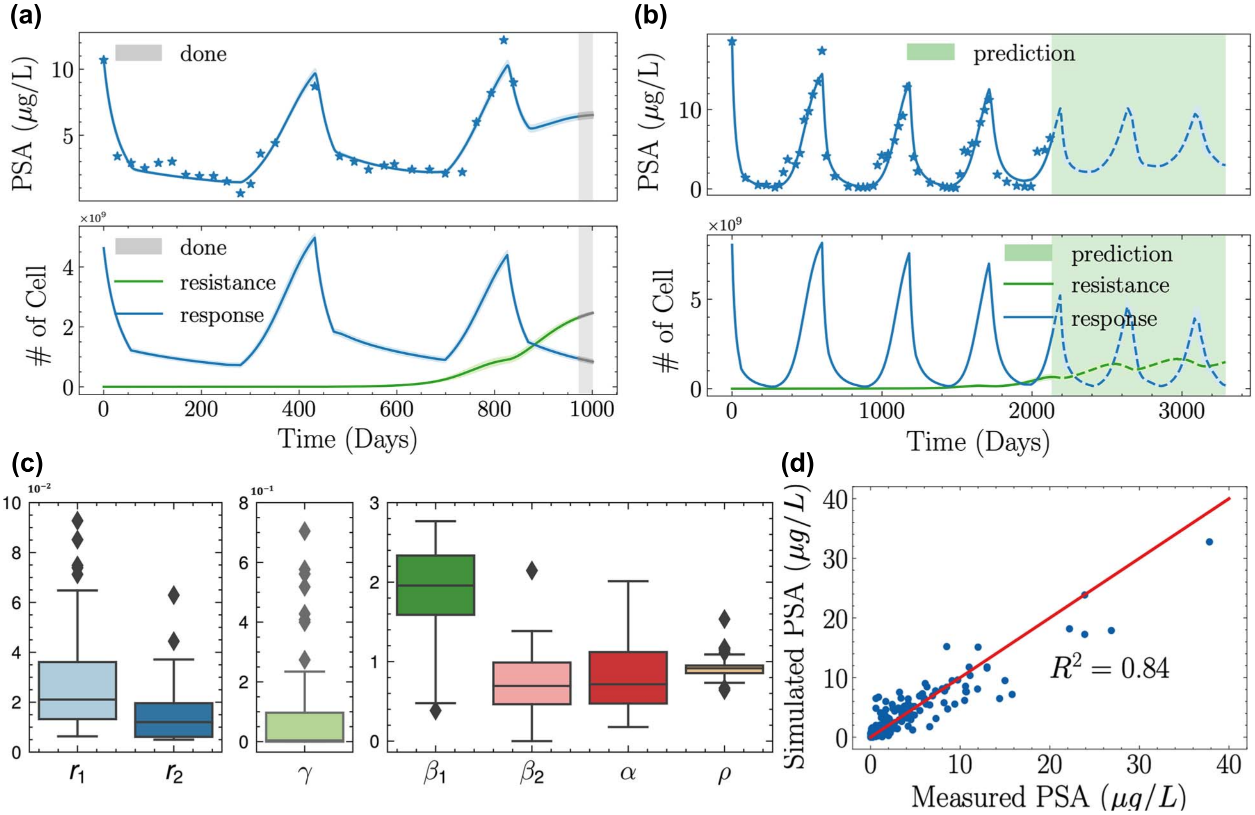

Figure 2. Presenting of the mathematical modeling of prostate cancer dynamics, showing patient-specific treatment responses and resistance development to IADT, alongside parameter distributions and model validation. **A**). The evolutionary dynamics of patient012 who evolved resistance to ADT based on the clinical trial with the application of standard IADT [6]. The upper panel is the fitted curve for the serum PSA level, which is plotted with the ground truth. The lower panel is the corresponding simulation of the dynamics of responsive and resistant phenotypes. We use simulations to predict the PCaC environment with the standard IADT; when the resistant population exceeds 80% of its capacity, we ended the simulation (EOS, a gray background marked as 'done'). The resistant phenotype escapes from the competition pressure in the 3-rd treatment cycle, thereby leading to drug resistance. **B**). The evolutionary dynamics of patient037 who did not evolve resistance to ADT based on the clinical trial involving the application of standard IADT). The upper panel is the fitted curve for the serum PSA level, which is plotted with the ground truth. The lower panel is the corresponding simulation of the dynamics of responsive and resistant phenotypes. The resistant phenotype increases in the later stage of the treatment for both patients. **C**). The distributions of all the patient-specific parameters learned by gradient descent, the mean and the 95% CI are shown in Supplementary File 1. Validation of the mathematical models with the test data.

between the two phenotypes in the tumor microenvironment because of the high demand for resources in this niche [31]. The resistant phenotype can genetically or epigenetically gain advantages through mutations. We calibrated the model with a multicenter Phase II clinical trial by applying the standard IADT to the biochemical recurrence patients after irradiation with localized prostate cancer [6]. The longitudinal data of each patient were utilized for training the patient-specific mathematical model, as described in Methods (2.1).

The PSA dynamics and the model-simulated evolutionary progress of cancer cells are shown in Figures 2a and b. Two representative patients (patient012 and patient037) were presented. Patient037 did not develop resistance to ADT, whereas patient012 developed resistance to ADT with the standard IADT. These observations are based on the criterion that the PSA level exceeded 4 $\mu g/L$ in weeks 24 and 32 in the latest treatment-on period, which was used as the ending criterion of the clinical trial (EOC).

We predicted the evolutionary dynamics for each patient by simulation. The PSA dynamics of patient012 (Figure 2a) showed that the nadir (lowest) PSA level increased gradually with treatment progress, thereby indicating that the patient was gradually developing resistance to ADT. The simulated amount of cancer cells (lower panel, Figure 2a) also showed that the resistant cancer cells were gradually winning the competition against the responsive cancer cells, thereby leading to resistance to ADT in the last clinical cycle, where the concentration of the resistance phenotype exceeded 0.8 (one of the ending criteria of simulation (EOS)). By contrast, Figure 2b shows that patient037 responded to IADT continuously and ended the simulation at the 7*th* cycle (terminal time, set as 120 months).

The interplay between drug dosage and the intratumoral competition showed that for patient012, resistant cancer cells have been suppressed by responsive cancer cells in the competition during the absence of treatment in the first treatment cycle. However, the population size of resistant cancer cells has been increasing. On around day 800, under treatment, the resistant cancer cells finally took the advantage over the responsive cancer cells, leading to drug resistance and treatment failure. In the last treatment cycle, the resistant cancer cells dominated the tumor microenvironment. For patient037, resistant cancer cells were suppressed continuously in the first four clinical cycles and the predicted cycles, where only a slight increase of resistant cancer cells was found in the extrapolated cycles.

To validate the prediction accuracy of our model, we compared the predicted serum PSA level with the ground truth in an out-of-sample experiment (validation set, refer to Methods (2.1)). The results in Figure 2d showed that the model predicted the dynamics of PSA well ($R^2 = 0.84$).

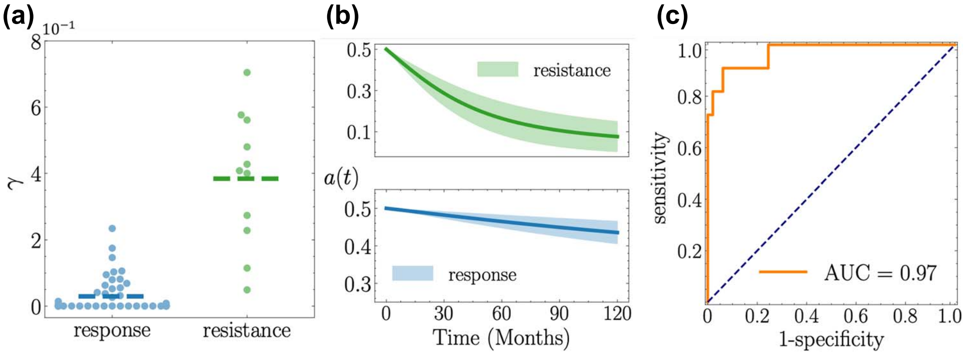

Figure 3. The figure displays a statistical analysis of patient responses to treatment, with panel (**A**) showing the distribution of γ, for response and resistance group, panel (**B**) showing the average change of $A_{21} = \frac{1}{1+e^{\gamma t}}$ with the 95% CI over time for resistant (green) and responsive (blue) patients and panel (**C**) showing the receiver operating characteristic curve for using the value of γ to classify the patients into resistance and response groups. Clinicians' labels in the clinical trial data are used as the ground truth.

These results verified the capability of the proposed tM-GLV model to characterize the resistance development and the individual responses to IADT among the patients. Additionally, the model captured the interplay between drug dosage and intratumoral competition in cancer evolution.

Predict patients' responses using the resistance index

With the model at our disposal, we use the parameters to categorize patients as either resistant or responsive. Our results provide empirical evidence that the best predictive power for differentiating resistance from response comes from the parameter γ (C-statistic=0.97, receiver operating characteristic curve depicted in Figure 2c). Also, given to the imbalanced natural of the positive and the negative samples, we reported Matthews Correlation Coefficient to better illustrate the power of parameter γ, which is around 0.798. For more results in terms of the paramaters analysis, please refer to Supplementary file 3 and Supplementary Figure 4 in Supplementary S11). Consequently, we designate γ as the resistance index.

The distribution of the resistance index for all patients is visualized in the left panel of Figure 3a, where the blue dots signify responsive patients and the green dots denote resistant patients. The average resistance index of resistant patients (0.384, 95% CI: (0.249, 0.519)) is significantly greater than that of responsive patients (0.029, 95% CI: (0.014, 0.044)), with a P-value of 1.28×10^{-6} in the Wilcoxon rank-sum test. Examining the two patients in Figure 2, Patient037, a consistent responder, has a resistance index of 0.0021, while Patient012, a resistant patient, has a resistance index of 0.576.

Using a false positive rate of 10% to establish the threshold, we classify patients into two groups. The resistance index demonstrates a TPR of 0.909 and an FPR of 0.061, with a threshold of 0.115. These findings support our selection of γ as the resistance index in our model. This indicates that the resistance index γ, derived from actual data, captures the key intratumoral evolutionary characteristics of patients and can differentiate responsive patients from resistant ones. We also conducted extensive leave-pair-out cross-validations to further assess the robustness of our results (details can be found in Supplementary S9).

Moreover, we define $a(t) = 1/(1 + e^{\gamma t})$ from the community matrix (see Methods (2) and Supplementary S2 for more information) as the competition coefficient, which represents the degree of resource overlap between the two cancer cell types in the tumor microenvironment. The competition advantage of responsive cancer cells over resistant cancer cells is then calculated as #(responsive cell) × $a(t)$/#(resistance capacity) and vice versa. A higher $a(t)$ value implies greater resource sharing and increased competition between the two sub-populations. In Figure 3b, we plot the coefficient $a(t)$ over time to demonstrate the dynamics of competition intensity. The coefficient decreases rapidly in the resistance group, leading to swift drug resistance, while the decline is more gradual in the response group, resulting in sustained drug responsiveness.

Reinforcement learning informs adaptive drug administration policy for better treatment outcome

The predefined thresholds for suspending and resuming treatment in the standard IADT are not personalized, leading to suboptimal treatment outcomes. The proposed tM-GLV model captures individual patient's intratumoral evolutionary dynamics with personalized parameterizations through the model fitting and validation. In this section, the optimal dosing policies are obtained through Proximal Policy Optimization [39] for 11 resistant patients and 51 responsive patients. For the training convergence and evaluation information, please refer to Figure 3 in Supplementary S11.

I^2ADTon resistance group

The reinforcement learning-derived I^2ADT can significantly postpone resistant patients' TTP. Figure 4a shows the dosing policies, treatment outcome of I^2ADT in left panel, and the corresponding standard IADT on all resistant patients in right panel. Three major differences existed between I^2ADT and standard IADT.

First, the average time of each treatment cycle reduces compared with the standard IADT, with treatment on: 5.0 months versus 13.6 months; treatment off: 8.7 months versus 8.9 months. With such an adaptive dosing policy learned by reinforcement learning, the population of responsive cancer cells oscillated at a relatively high level before the occurrence of resistance, as presented in Figure 4b. The competitive advantage of responsive cancer cells also exhibits such an oscillating pattern, indicating that the proposed I^2ADT could suppress resistant cancer cells by giving competition pressure to responsive cancer cells. The biphasic pattern commonly observed in IADT [25], which also existed in our simulation, as shown in Figure 4c, was prevented in I^2ADT by the shortening of the treatment-on period. The biphasic pattern indicates that after the treatment was turned on for a period of time, the effectiveness of continuous drug treatment

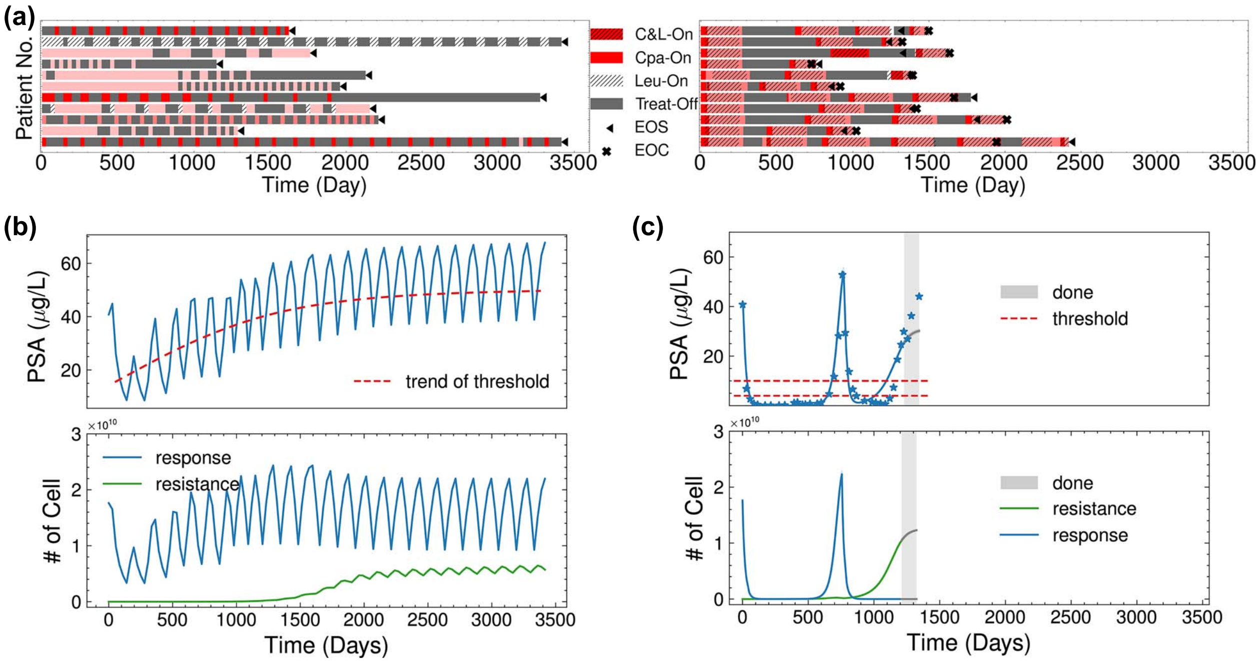

Figure 4. The treating strategies for resistance group, alongside two selected patients' evolutionary dynamics. (**A**) The dosing policies for patients in the resistance group. Each row denotes a resistant patient. The left and right bars present the I^{2}ADT and IADT outcomes, respectively. The black crossing denotes the end of the clinical trial (EOC), and the caret-left icon denotes the end of the simulation (EOS). The color density of Cpa-On and C&L-On is proportional to the CPA dosage. (**B**) and (**C**) present the evolutionary dynamics of PSA and cancer cells under I^{2}ADT and standard IADT in patient099, respectively. In Figure (**B**), there are 24 treatment cycles for 120 months (5 months per cycle in average), while in Figure (**C**), there are two cycles for 43 months (21.5 months per cycle in average).

declines, as reflected by the flattened slope of the patient's PSA curve.

Second, I^{2}ADT resulted in pattern of declining treatment-on-time over time for the resistant population (refer to Supplementary S10). In general, the responsive cancer cells were gradually losing the competition with resistant cancer cells. So, I^{2}ADT helped responsive cancer cells re-grow by reducing the treatment-on time in each cycle. However, due to small number of resistance population, the decline pattern of dosage is not significant. Please refer to the Supplementary S10 for additional analysis of the decline pattern of the dosing policies.

Thirdly, the I^{2}ADT learned through reinforcement learning was dynamic and tailored to each patient's needs. In the initial stages of treatment, I^{2}ADT provided a greater competitive advantage for responsive cancer cells over resistant ones, when compared with both IADT and conventional continuous ADT. This is illustrated in Figure 5a. As treatment progressed and intratumoral competition continued, the competitive advantage of responsive cancer cells gradually decreased to zero in both IADT and ADT. However, in I^{2}ADT, a significant competitive advantage still persisted, allowing responsive cells to compete with the resistant cancer cells and ultimately prolonging the survival time of resistant patients.

To compare the effectiveness of I^{2}ADT with IADT or ADT, we use two surrogates: TTP and progression-free survival (PFS). TTP is defined as the time at which the simulation reaches the end of simulation (EOS) for an individual patient. FPS refers to the time from the initiation of treatment to the occurrence of disease progression (EOS). The EOS is reached when either the resistant cancer cells account for 80% of their capacity or when the simulation reaches its maximum number of steps (120).

Simulation results demonstrate that by maintaining a higher competitive advantage during the early stage, I^{2}ADT significantly prolonged TTP and PFS rates compared with standard IADT or ADT (P-value = 0.0019), as shown in Figures 5b and c. These results indicate that the adaptive dosing can be an effective strategy for delaying the onset of drug resistance and improving patient outcomes.

I^{2}ADT on response group

The 51 responsive patients had a small resistance index γ, which was much smaller than that of resistant patients. Therefore, the resistant cancer cells had been consistently suppressed by the responsive cancer cells, leading to less intense competition. As shown in Figure 3b, the competition advantage of responsive cancer cells declined slowly over time. Similar to the case of the resistance group, the proposed I^{2}ADT learned a dosing policy with shorter treatment cycle (treatment on: 1.3 month versus 13.4 months; treatment off: 3.5 months versus 16.5 months) to fully utilize the competition advantage of responsive cancer cells and to further reduce the risk of developing drug resistance and the cumulative drug dosage.

A distinct difference existed between the I^{2}ADT for responsive patients and that for resistant patients. In general, an ascending pattern of dosage was observed in the I^{2}ADT for most responsive patients by increasing the dosage and/or the treatment-on time in each cycle (as shown in Figure 6b and c for patient106). The resistant cancer cells can be suppressed to a low level. Thus, I^{2}ADT tended to further reduce the overall tumor burden by killing more responsive cancer cells, so that the risk of other disease progression events, such as metastasis and comorbid conditions, could be further reduced. The Supplementary S10 presents details of additional analysis of the ascending dosing policies.

Although the responsive patients did not develop drug resistance in the original clinical trial, I^{2}ADT further reduced the risk of drug resistance in the long run by maximizing the evolutionary competition between the responsive and resistant cancer cells.

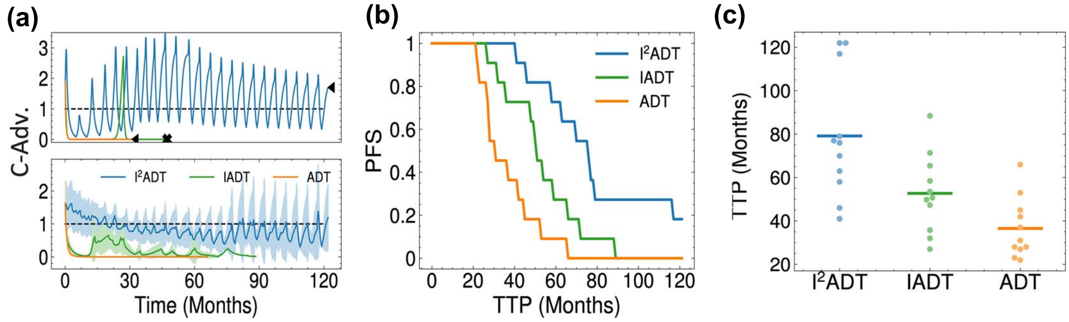

Figure 5. The figure presents the comparison among I^2ADT, IADT and ADT for the resistance group. (**A**) The dynamics of the competition advantage of responsive cancer cells toward the resistant cancer cells in patient036 (upper panel) and all resistant patients with 95% CI (lower panel). (**B**) The PFS rate over time with I^2ADT, IADT and ADT. (**C**) The distribution of TTP with I^2ADT, IADT and ADT.

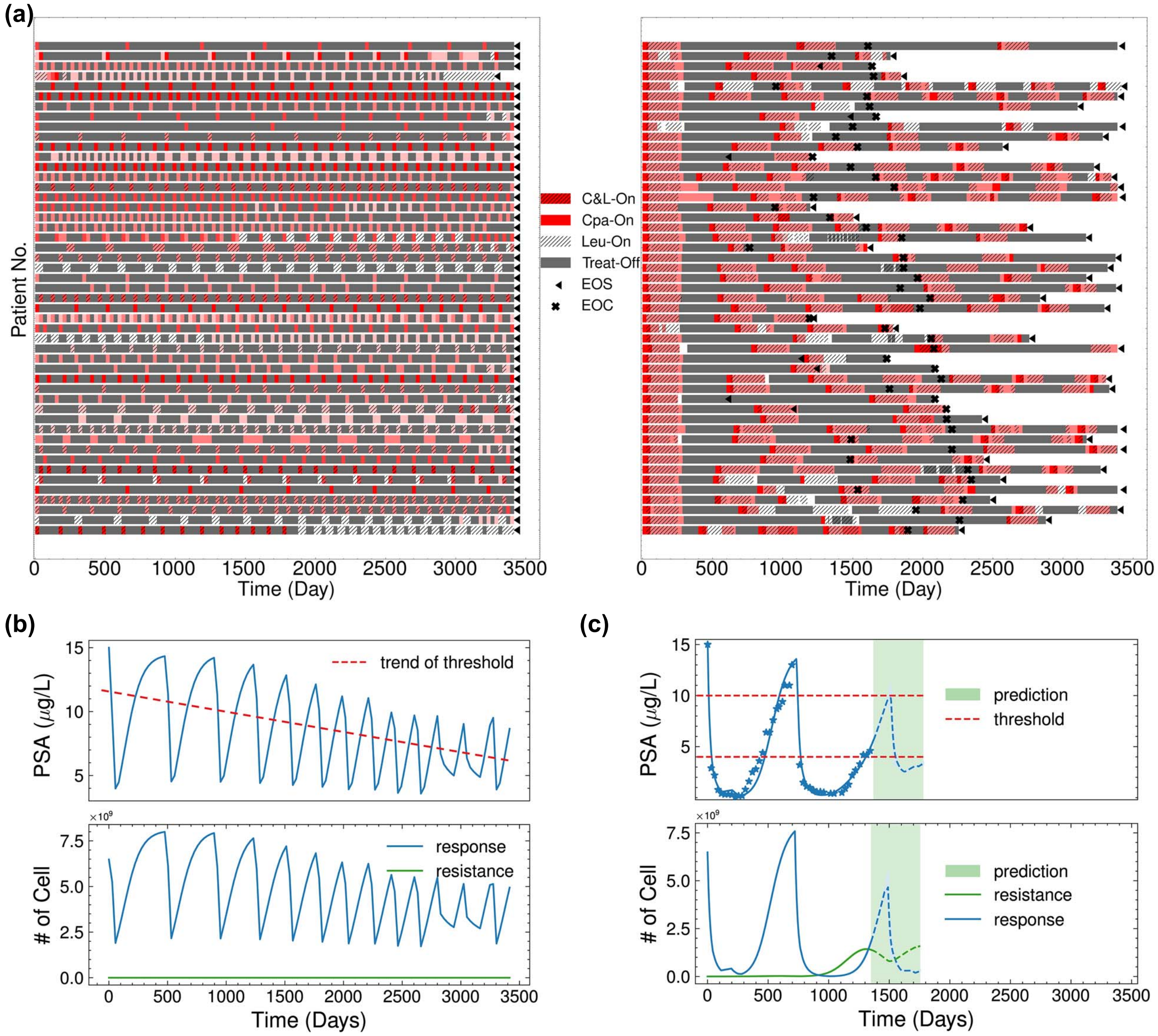

Figure 6. The treating strategies for response group, alongside two selected patients' evolutionary dynamics. (**A**) The dosing policies for patients in the response group. Each row denotes a responsive patient. The left and right bars present the I^2ADT and IADT outcomes, respectively. (**B**) and (**C**) present the evolutionary dynamics of PSA and cancer cells under I^2ADT and standard IADT in patient106, respectively. In Figure (**B**), there are 13 treatment cycles for 120 months (9 months per cycle in average), while in Figure (**C**), there are 2.5 cycles for 62 months (25 months per cycle in average).

Taking Patient106 as an example, even if the resistant cancer cells were under control during the entire period of the clinical trial, and the PSA levels were successfully suppressed under $4\mu g/L$ during the treatment-on period, the population of resistant cancer cells was found to already be increasing. When we used the model to predict the future evolutionary dynamics for

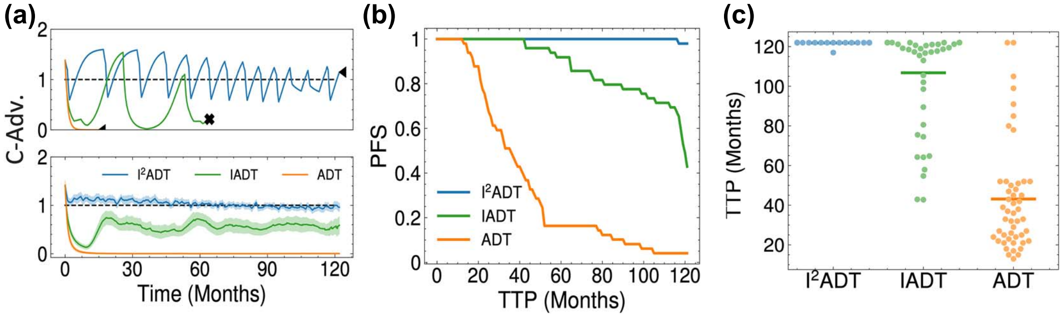

Figure 7. The figure presents the comparison among I^2ADT, IADT and ADT for the response group. (**A**) The dynamics of the competition advantage of responsive cancer cells toward the resistant cancer cells in patient106 (upper panel), and all responsive patients with 95% CI (lower panel). (**B**) The PFS rate over time with I^2ADT, IADT and ADT. (**C**) The distribution of TTP with I^2ADT, IADT and ADT.

Table 2: The percentage of reduction of dosage and treatment-on course of I^2ADT compared with that of the standard IADT. The pP-values of t-test are shown in the brackets.

Items/ deduction rate/ Group	Resistance	Response
Ave. LEU (mg/Month)	87.7% (10^{-5})	71.7% (10^{-16})
Ave. CPA (mg/Day)	60.3% (10^{-6})	43.4%(10^{-13})
Ave. Treat-on Per cycle	27.1% (0.027)	40.3%(10^{-15})

this patient, we found that drug resistance would emerge on day 1750 (EOS). Similar results were observed in most of the responsive patients (Figure 6a). Drug resistance would emerge eventually after a sufficiently long period of time. By reinforcement learning, the proposed I^2ADT prolongs the TTP to 10 years (the maximum period of simulation) and increased the PFS rate significantly.

Similar simulation results demonstrate that by maintaining a higher competitive advantage during the early stage, as shown in Figure 7a., I^2ADT significantly prolonged TTP and PFS rates compared with standard IADT or ADT (P-value = 0.001), as shown in Figures 7. *bandc* but also reduce much more drug dosage, improving life quality. These results indicate that the adaptive dosing can be an effective strategy for delaying the onset of drug resistance and improving patient outcomes.

Dosage reduction

Considering the inevitable adverse events of the treatment-on period of ADT [41], it is preferable to reduce the dosage as long as the disease is under control. We compared the deduction rate of averaged dosage of Cyproterone acetate (CPA), Leuprolide acetate (LEU) by cycle and the overall treatment-on percentage with the standard IADT in (Table 2).

A significant reduction of the dosage of both CPA and LEU and a reduced percentage of the treatment-on period in the treatments of I^2ADT were found, indicating that the proposed I^2ADT can reduce the risk of incidence of adverse events of treatment and enhance the quality of life of prostate cancer patients.

Validation of I^2ADT

To further validate the efficacy of the proposed I^2ADT beyond the simulations, we performed two additional validations.

As shown in the previous sections, the proposed method I^2ADT led a better outcome with less dosage as compared with the standard IADT in the simulations. However, validation with external data is needed to further demonstrate the clinical feasibility of the proposed I^2ADT. Here, we present two additional validations: external data validation (Section (3.4.1)) and individual new patient validation (Section (3.4.2)). The detailed methods and algorithms 1 and 2 are given in the Supplementary S1.

Validation with external clinical trial data

In a recent pilot clinical trial [42], clinicians implemented the IADT approach, which involved suspending treatment when PSA levels decreased to 50% of the pretreatment value and resuming treatment when PSA levels returned to baseline. This strategy is varied from the traditional one. Although this IADT strategy successfully increased TTP and reduced dosage, it remains a population-based approach and is not optimized for individual patients.

We applied the proposed I^2ADT method to patients in this pilot clinical trial by first training the tM-GLV model to obtain the PCaC environment and then using PPO to determine individualized treatment strategies. To facilitate a fair comparison, we extended the IADT dosing policy using the same 50% threshold criteria until the EOS was reached. The resulting dosing policies are illustrated in Figure 8a, and the corresponding PFS/TTP values are presented in the panel b and c of Figure 8. The simulation results revealed that 88.2% (15 out of 17) of patients would experience a longer TTP (P-value= 6.3×10^{-4}) with the I^2ADT approach. With I^2ADT, the average treatment-on and treatment-off durations for each cycle are 1.63 and 8.73 months, respectively. In comparison, the IADT group has average durations of 7.1 and 6.83 months for treatment-on and treatment-off periods, respectively. The average dosage and treatment-on percentage for each cycle were reduced by 55.6% (P-value= 5.7×10^{-6}) when using the I^2ADT method.

To sum up, the external data validation further validated that proposed I^2ADT can achieve better treatment outcome as compared with the conventional IADT.

Validation with individual new patients

In clinical settings, it is common for practitioners to be unaware of a new patient's response to treatment. As such, we designed an experiment in a hypothetical prospective scenario where a new patient undergoes treatment without any prior knowledge of their response to ADT. In this setup, each patient initially receives IADT for a fixed time period (corresponding to the first IADT treatment cycle in our experiments) and then transitions to the personalized I^2ADT approach after gathering data from the IADT treatment. In the clinical trial [6], PSA testing occurs on a monthly basis. To acquire more data and better understand a patient's cancer dynamics, we propose a weekly PSA test during the first IADT treatment cycle. We refer to this strategy as *delayed*-I^2ADT.

In Table 3, we present the results for two representative cases: patient 001 (responsive) and patient 011 (resistant). For both

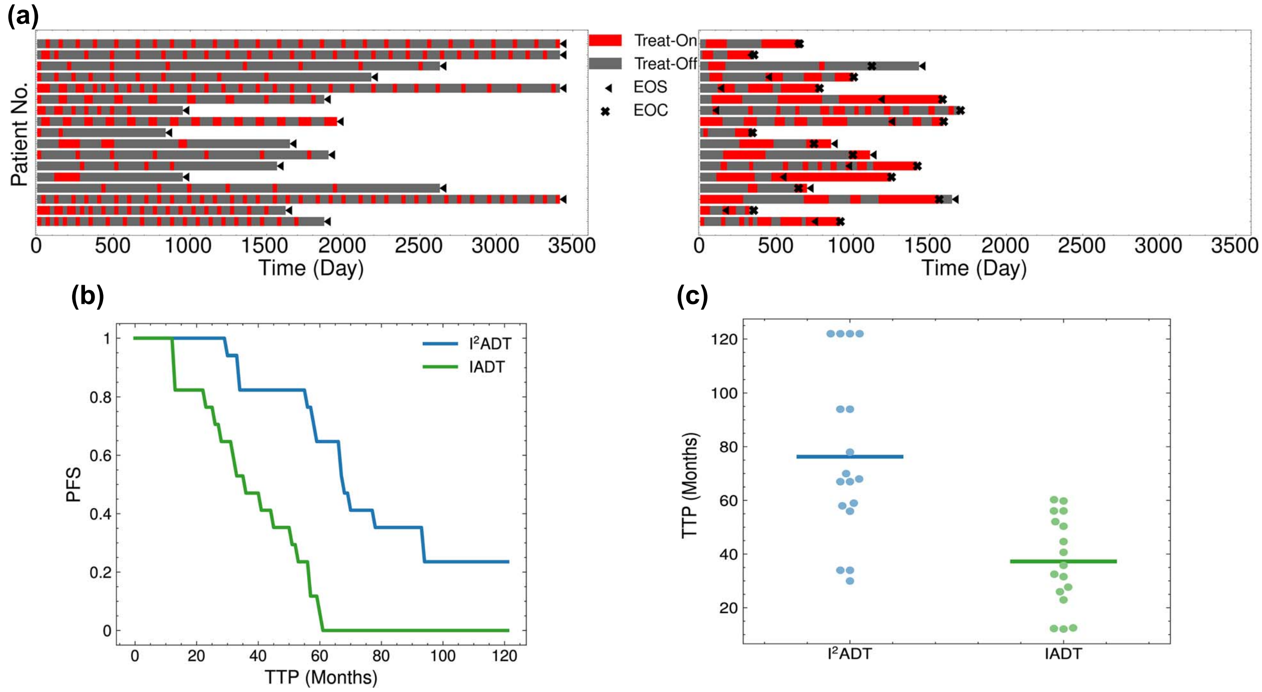

Figure 8. The results of the validation set, including the dosing strategies and the comparison between IADT and I^2ADT. **A**) shows the dosing policies for patients in the validate group from a pilot clinical trial [42]. Each row denotes a patient, and x-axis denotes the time (day). The left and right bars present the outcomes for I^2ADT and IADT described in [42] (with a fixed 50% threshold), respectively. **B**) and **C**) compare the FPS rates over TTP and TTP distributions between IADT and I^2ADT strategies.

Table 3: TTP and average dosage for delayed-I^2ADT, I^2ADT and standard IADT.

	Patient001 (responsive)			**Patient011 (resistant)**		
	delayed-I^2ADT	**I^2ADT**	**IADT**	**delayed-I^2ADT**	**I^2ADT**	**IADT**
TTP (month)	120	120	115	120	120	97
Ave. CPA (mg/day)	56.6	50.4	73	45.8	52.9	81.2
Ave. LEU (mg/month)	0.94	2.52	3.3	0.51	0.06	7.1

patients, *delayed*-I^2ADT achieved similar performance as I^2ADT, exhibiting the same TTP and comparable dosages. Both delayed-I^2ADT and I^2ADT resulted in longer TTP and lower dosages compared with IADT. For patient 001, delayed-I^2ADT led to a higher dosage than I^2ADT, while for patient 011, delayed-I^2ADT led to a lower dosage. The minor difference between delayed-I^2ADT and I^2ADT can be attributed to less intense intratumoral competition during the first IADT treatment cycle. This experiment demonstrates that new patients can benefit from I^2ADT by initially undergoing IADT in the first treatment cycle and subsequently transitioning to a patient-specific, optimized I^2ADT approach for subsequent treatment cycles.

DISCUSSION AND LIMITATIONS

In this work, we propose the I^2ADT dosing strategy in prostate cancer, enhancing the suppression of resistant cells by leveraging the competitive advantage of responsive cells. This approach, I^2ADT, is adaptable for optimizing treatments across various cancer types due to the inherent flexibility of the reinforcement learning. The adaptation needs to do is the mathematical modeling for specific cancer types, the revised reinforcement learning structures and the availability of the clinical data. This adaptability makes it a promising tool for other cancer types, where treatment regimens might need to be individualized based on patient characteristics and disease dynamics. In cancers with well-characterized protocols and rich datasets (e.g. breast cancer, colorectal cancer), the transition is more natural. While the framework of our DRL model could be applied to other cancers, it would require significant customization to incorporate the specific treatment options, progression markers, and patient response criteria relevant to each cancer type. Moreover, rigorous validation through clinical trials or retrospective studies would be essential to establish the model's efficacy and safety in different oncological contexts.

In recognizing the limitations of our study, we aim to provide a comprehensive understanding of the factors that should be considered when interpreting our results and contemplating future research directions. Our AI model demonstrates robust performance within its current application; however, its generalizability is limited by the specificity of the data used for its training and has not been tested across diverse clinical settings. The clinical trial data used in this study focus mainly on drug administration and PSA level measurements, omitting broader physiological factors, genetic factors and lifestyle factors due to data constraints. Although we assume the clinical outcome and treatment reflects these personalized information and then learnt by our model, such exclusions represent the data limitations that future studies should aim to address.

Additionally, while our model integrates the effects of both LEU and CPA, the nuances of their combined effect on the disease's pathway interactions remain to be elucidated. The model's effectiveness could be significantly improved with access to more detailed patient-specific clinical and pathological data, including information on drug combination effects.

The complexity and 'black-box' nature of deep learning models like ours present interpretability challenges that are critical to address for gaining trust in healthcare practices and facilitating the adoption of AI systems. Integration of these systems into existing clinical workflows remains a hurdle, compounded by the need for continuous updates to the model as new data are collected.

In our study, PSA serves as the sole biomarker, despite the knowledge that a more comprehensive biomarker panel, including circulating tumor cells and cell-free DNA, could provide a more nuanced picture of disease progression and improve the calibration of the reinforcement learning algorithm. Moreover, our reinforcement learning model's reward function, based on simulated outcomes, may not fully capture the complexity of clinical scenarios where measuring the direct effects of drugs and the competitive interactions between phenotypes is challenging.

We also acknowledge that the serum hormone levels, maintained at castrate levels during treatment periods in our model, do not reflect the variability observed in patients' recovery of serum testosterone post-treatment [43], a factor not accounted for in our study due to missing information.

This study is retrospective and simulation-based and thus should be interpreted with the caution warranted for such studies. In anticipation of future research, we plan to conduct a pilot clinical trial to validate the I^{2}ADT in prostate cancer patients and explore its applicability to other cancer types. We aim to further refine the model by integrating additional domain knowledge and clinical insights.

CONCLUSION

Here we demonstrate application of the latest reinforcement learning techniques to individualize the intermittent drug administration by characterizing the unique cancer competition environment.

Drug resistance is inevitable in ADT and often leads to treatment failure. By integrating the Darwinian evolution patterns of cancer cells into the drug administration, IADT achieved better outcomes in clinical trials [6, 10, 42]. Existing IADT approaches are population-based and do not consider the heterogeneity of patients. In this paper, we proposed the data-driven and patient-specific I^{2}ADT, which integrates the mathematical models of intratumoral evolutionary dynamics and reinforcement learning to inform personalized intermittent drug administration policies. I^{2}ADT significantly increased the TTP and reduced the cumulative drug dosage, as indicated by the results of the experiments.

Moreover, we propose the *delayed*-I^{2}ADT approach as a practical solution for applying the personalized I^{2}ADT approach to new patients with limited clinical data. Our results show that *delayed*-I^{2}ADT can achieve similar performance to I^{2}ADT, with the similar TTP and dosage. The *delayed*-I^{2}ADT approach provides a new perspective on how to improve the clinical outcome of prostate cancer patients by adopting a personalized treatment plan. Further studies are needed to validate the effectiveness of the *delayed*-I^{2}ADT approach in a clinical setting.

Cancer treatment typically involves multiple lines of therapies and the use of multiple drugs. Every patient has a unique cancer phenotype and tumor microenvironment. In the context of ADT in prostate cancer, our methods yielded a highly personalized dosing policy that maximizes the competition advantage of responsive cancer cells to suppress resistant cancer cells. The I^{2}ADT can be easily extended to optimize the treatment of other cancers.

Though limitations and challenges are presented in this work, as we look to the future, we believe the collaboration of data scientists, pharmacologists and oncologists could further optimize I^{2}ADT and other adaptive therapy strategies. Such interdisciplinary efforts are critical to harnessing the full potential of personalized medicine to enhance cancer treatment outcomes.

Key Points

- We proposed a novel mathematical model to characterize the heterogeneous intertumoral evolutionary dynamics of individual prostate cancer patients.
- We developed a deep reinforcement learning algorithm to derive the personalized optimal drug administration policy.
- Experiments demonstrated that the proposed methods can significantly prolong the time to progression of prostate cancer patients with reduced drug dosage.
- This is the first study harnessing the power of deep reinforcement learning techniques to identify personalized adaptive therapy for metastatic prostate cancer.

SUPPLEMENTARY DATA

Supplementary data are available online at https://academic.oup.com/bib.

REFERENCES

1. Sung H, Ferlay J, Siegel RL, *et al*. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2021;**71**(3):209–49.
2. Litwin MS, Tan H-J. The diagnosis and treatment of prostate cancer: a review. *JAMA* 2017;**317**(24):2532–42.
3. Smith MR, Saad F, Coleman R, *et al*. Denosumab and bone-metastasis-free survival in men with castration-resistant prostate cancer: results of a phase 3, randomised, placebo-controlled trial. *Lancet* 2012;**379**(9810):39–46.
4. Shore ND. Current and future management of locally advanced and metastatic prostate cancer. *Rev Urol* 2020;**22**(3):110.
5. Gillies RJ, Flowers CI, Drukteinis JS, Gatenby RA. A unifying theory of carcinogenesis, and why targeted therapy doesn't work. *Eur J Radiol* 2012;**81**:S48–50.
6. Bruchovsky N, Klotz L, Crook J, *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):389–95.
7. Nguyen PL, Alibhai SMH, Basaria S, *et al*. Adverse effects of androgen deprivation therapy and strategies to mitigate them. *Eur Urol* 2015;**67**(5):825–36.
8. West JB, Dinh MN, Brown JS, *et al*. Multidrug cancer therapy in metastatic castrate-resistant prostate cancer: an evolution-based strategy. *Clin Cancer Res* 2019;**25**(14):4413–21.
9. McGranahan N, Swanton C. Clonal heterogeneity and tumor evolution: past, present, and the future. *Cell* 2017;**168**(4):613–28.
10. Zhang J, Cunningham JJ, Brown JS, Gatenby RA. Integrating evolutionary dynamics into treatment of metastatic castrate-resistant prostate cancer. *Nat Commun* 2017;**8**(1):1816.
11. David Crawford E, Heidenreich A, Lawrentschuk N, *et al*. Androgen-targeted therapy in men with prostate cancer: evolving practice and future considerations. *Prostate Cancer Prostatic Dis* 2019;**22**(1):24–38.

12. Hirata Y, Bruchovsky N, Aihara K. Development of a mathematical model that predicts the outcome of hormone therapy for prostate cancer. *J Theor Biol* 2010;**264**(2):517–27.
13. Baez J, Kuang Y. Mathematical models of androgen resistance in prostate cancer patients under intermittent androgen suppression therapy. *Appl Sci* 2016;**6**(11):352.
14. Brady-Nicholls R, Nagy JD, Gerke TA, *et al*. Prostate-specific antigen dynamics predict individual responses to intermittent androgen deprivation. *Nat Commun* 2020;**11**:1–13.
15. Belkhir S, Thomas F, Roche B. Darwinian approaches for cancer treatment: benefits of mathematical modeling. *Cancer* 2021;**13**(17):4448.
16. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019;**25**(1):44–56.
17. Zhang Z, *et al*. Reinforcement learning in clinical medicine: a method to optimize dynamic treatment regime over time. *Ann Transl Med* 2019;**7**(14):345.
18. Petersen BK, Yang J, Grathwohl WS, *et al*. Deep reinforcement learning and simulation as a path toward precision medicine. *J Comput Biol* 2019;**26**(6):597–604.
19. Gottesman O, Johansson F, Komorowski M, *et al*. Guidelines for reinforcement learning in healthcare. *Nat Med* 2019;**25**(1):16–8.
20. Engelhardt D, Michor F. A quantitative paradigm for decision-making in precision oncology. *Trends Cancer* 2021;**7**(4):293–300.
21. Engelhardt D. Dynamic control of stochastic evolution: a deep reinforcement learning approach to adaptively targeting emergent drug resistance. *J Mach Learn Res* 2020;**21**(1):8392–421.
22. Tseng H-H, Luo Y, Cui S, *et al*. Deep reinforcement learning for automated radiation adaptation in lung cancer. *Med Phys* 2017;**44**(12):6690–705.
23. Basanta D, Scott JG, Fishman MN, *et al*. Investigating prostate cancer tumour–stroma interactions: clinical and biological insights from an evolutionary game. *Br J Cancer* 2012;**106**(1): 174–81.
24. Isaacs JT, Coffey DS. Adaptation versus selection as the mechanism responsible for the relapse of prostatic cancer to androgen ablation therapy as studied in the dunning r-3327-h adenocarcinoma. *Cancer Res* 1981;**41**(12_Part_1):5070–4.
25. Tanaka G, Yoshito Hirata S, Goldenberg L, *et al*. Mathematical modelling of prostate cancer growth and its application to hormone therapy. *Philos Trans A Math Phys Eng Sci* 2010;**368**(1930): 5029–44.
26. Butner JD, Elganainy D, Wang CX, *et al*. Mathematical prediction of clinical outcomes in advanced cancer patients treated with checkpoint inhibitor immunotherapy. *Sci Adv* 2020; **6**(18):eaay6298.
27. Ribba B, Holford NH, Magni P, *et al*. A review of mixed-effects models of tumor growth and effects of anticancer drug treatment used in population analysis. *CPT Pharmacometrics Syst Pharmacol* 2014;**3**(5):1–10.
28. McKane AJ, Newman TJ. Stochastic models in population biology and their deterministic analogs. *Phys Rev E* 2004;**70**(4):041902.
29. Chignola R, Foroni RI. Estimating the growth kinetics of experimental tumors from as few as two determinations of tumor size: implications for clinical oncology. *IEEE Trans Biomed Eng* 2005;**52**(5):808–15.
30. Marusyk A, Almendro V, Polyak K. Intra-tumour heterogeneity: a looking glass for cancer? *Nat Rev Cancer* 2012;**12**(5):323–34.
31. Chang C-H, Qiu J, O'Sullivan D, *et al*. Metabolic competition in the tumor microenvironment is a driver of cancer progression. *Cell* 2015;**162**(6):1229–41.
32. Ikeda M, Šiljak DD. Lotka-volterra equations: decomposition, stability, and structure: part i: equilibrium analysis. *J Math Biol* 1980;**9**(1):65–83.
33. Ikeda M, Šiljak DD. Lotka-volterra equations: decomposition, stability, and structure part ii: nonequilibrium analysis. *Nonlinear Anal Theory Methods Appl* 1982;**6**(5):487–501.
34. Kasim MF, Vinko SM. \xi-torch: differentiable scientific computing library. *arXiv preprint* 2020; arXiv:2010.01921.
35. Kingma DP and Ba J. Adam: A method for stochastic optimization. In: Bengio Y, LeCun Y (eds). *3rd International Conference on Learning Representations, ICLR*, San Diego, CA, USA, 2015.
36. Ivanov S, D'yakonov A. Modern deep reinforcement learning algorithms. *arXiv preprint* 2019; arXiv:1906.10025.
37. Lillicrap TP, Hunt JJ, Pritzel A, *et al*. Continuous control with deep reinforcement learning. In: Bengio Y, LeCun Y (eds). *3rd International Conference on Learning Representations, ICLR*, San Juan, Puerto Rico, 2016.
38. Schulman J, Levine S, Abbeel P, *et al*. Trust region policy optimization. In Bach F and Blei D, editors, *International conference on machine learning*. PMLR, Lille, France, 2015;1889–97.
39. Schulman J, Wolski F, Dhariwal P, *et al*. Proximal policy optimization algorithms. *arXiv preprint* 2017;arXiv:1707.06347.
40. Haarnoja T, Zhou A, Abbeel P, and Levine S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In: Dy J, Krause A (eds). *International conference on machine learning*. PMLR, Stockholmsmässan, Stockholm Sweden, 2018, 1861–70.
41. Bruchovsky N, Klotz L, Crook J, *et al*. 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. *Clin Genitourin Cancer* 2008;**6**(1):46–52.
42. Zhang J, Cunningham J, Brown J, Gatenby R. Evolution-based mathematical models significantly prolong response to abiraterone in metastatic castrate-resistant prostate cancer and identify strategies to further improve outcomes. *Elife* 2022; **11**:e76284.
43. Inoue T, Mizowaki T, Kabata D, *et al*. Recovery of serum testosterone levels and sexual function in patients treated with short-term luteinizing hormone-releasing hormone antagonist as a neoadjuvant therapy before external radiotherapy for intermediate-risk prostate cancer: preliminary prospective study. *Clin Genitourin Cancer* 2018;**16**(2):135–141.e1.