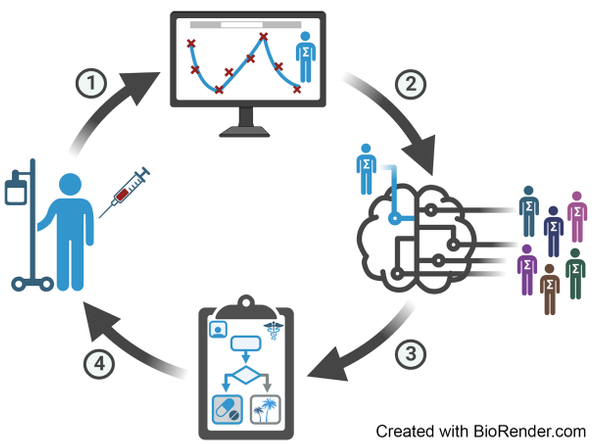
Mathematical Model-Driven Deep Learning Enables Personalized Adaptive Therapy

Kit Gallagher[1,2], Maximilian A.R. Strobl[2], Derek S. Park[2], Fabian C. Spoendlin[1], Robert A. Gatenby[2], Philip K. Maini[1], and Alexander R.A. Anderson[2]

ABSTRACT

Standard-of-care treatment regimens have long been designed for maximal cell killing, yet these strategies often fail when applied to metastatic cancers due to the emergence of drug resistance. Adaptive treatment strategies have been developed as an alternative approach, dynamically adjusting treatment to suppress the growth of treatment-resistant populations and thereby delay, or even prevent, tumor progression. Promising clinical results in prostate cancer indicate the potential to optimize adaptive treatment protocols. Here, we applied deep reinforcement learning (DRL) to guide adaptive drug scheduling and demonstrated that these treatment schedules can outperform the current adaptive protocols in a mathematical model calibrated to prostate cancer dynamics, more than doubling the time to progression. The DRL strategies were robust to patient variability, including both tumor dynamics and clinical monitoring schedules. The DRL framework could produce interpretable, adaptive strategies based on a single tumor burden threshold, replicating and informing optimal treatment strategies. The DRL framework had no knowledge of the underlying mathematical tumor model, demonstrating the capability of DRL to help develop treatment strategies in novel or complex settings. Finally, a proposed five-step pathway, which combined mechanistic modeling with the DRL framework and integrated conventional tools to improve interpretability compared with traditional "black-box" DRL models, could allow translation of this approach to the clinic. Overall, the proposed framework generated personalized treatment schedules that consistently outperformed clinical standard-of-care protocols.

Significance: Generation of interpretable and personalized adaptive treatment schedules using a deep reinforcement framework that interacts with a virtual patient model overcomes the limitations of standardized strategies caused by heterogeneous treatment responses.



Introduction

Drug resistance is responsible for up to 90% of cancer-related deaths (1). It can be present before treatment (intrinsic) or emerge during therapy (acquired) and is driven by a combination of genetic, epigenetic, and environmental processes (2). Much modern cancer research has focused on developing novel therapies to overcome resistance but, especially in the metastatic setting, cure rates remain low and benefits are all too often short-lived (3, 4).

Conventional, standard-of-care treatment schedules for systemic cancer therapies are based on the maximum tolerated dose (MTD) principle. This argues for the administration of treatment at as high a dose and frequency as tolerable, in order to maximize cell kill and thereby the likelihood of cure (5). However, in recent years, it has become increasingly clear that cancers, in particular metastatic cancers, are complex and spatiotemporally heterogeneous and actively evolve under treatment (6). This has prompted a rethinking of the MTD paradigm and has stimulated a growing body of research demonstrating that changes in drug scheduling could delay drug resistance (7, 8, 9, 10). One promising approach is "adaptive therapy" (AT) based on the principle of "competitive control." It suggests that resistant cells, or their precursors, may be present before treatment, but their growth is limited by competition with the drug-sensitive subpopulation (9, 11). However, MTD treatment removes sensitive cells, leading to the competitive release of resistant cells and causing disease progression (12, 13). To address this issue, Gatenby and colleagues (9, 11) proposed AT, which dynamically adjusts treatment to maintain a pool of drug-sensitive cells that compete with emerging resistance, aiming to control, rather than eliminate, the tumor. AT has been shown to extend the time to progression (TTP) *in vivo* for breast

[1]Wolfson Centre for Mathematical Biology, Mathematical Institute, Oxford, United Kingdom. [2]Integrated Mathematical Oncology, Moffitt Cancer Center, Tampa, Florida.

K. Gallagher and M.A.R. Strobl share the first authorship of this article.

P.K. Maini and A.R.A. Anderson share the senior authorship of this article.

Corresponding Authors: Alexander R. Anderson, Moffitt Cancer Center, SRB 4th Floor, 12902 Magnolia Drive, Tampa, FL 33612. E-mail: alexander.anderson@moffitt.org; Kit Gallagher, Mathematical Institute, University of Oxford, Andrew Wiles Building, Radcliffe Observatory Quarter, Woodstock Road, Oxford OX2 6GG, United Kingdom. E-mail: gallagher@maths.ox.ac.uk

Cancer Res 2024;84:1929–41

doi: 10.1158/0008-5472.CAN-23-2040

cancer (14), ovarian cancer (9), lung cancer (15), and melanoma (16) and has, most recently, delivered promising results in a pilot clinical trial in metastatic, castrate-resistant prostate cancer (17, 18).

Castrate-resistant prostate cancer is treated with androgen deprivation therapy, such as abiraterone, which inhibits CYP17A, an enzyme for testosterone auto-production (19). Standard of care is given at MTD via continuous administration until radiographic progression, which occurs after a median of 16.5 months (20). Zhang and colleagues (17, 18) instead applied an adaptive strategy (AT50) where an identical dose to standard of care was given until the tumor burden had reduced by 50% relative to baseline and was subsequently withheld until the tumor burden returned to baseline (NCT02415621). To monitor burden, they combined radiographic imaging with measurements of prostate-specific antigen (PSA) levels, an established serum biomarker (21, 22), which enabled more frequent (monthly) tracking of the disease. In comparison with a matched historical control receiving continuous dosing, the study found that patients undergoing AT had a 19.2-month increase in median progression-free survival, while receiving 46% less drug on average (18).

Although this study is promising, it highlights the need for further research into exactly how we adapt therapy: long-term disease control was achieved for only 4 of 17 patients on the trial (18), and there was significant variation in the adaptive cycling dynamics between patients (illustrated in **Fig. 1A**). Previously we, and others (23, 24, 25, 26), have investigated how the threshold of tumor burden at which treatment is withdrawn in the AT50 protocol affects the outcome. These findings showed that increasing the threshold, so that treatment is withdrawn earlier and at a higher average tumor burden, increases competitive suppression and thereby TTP. Consequently, it has been proposed (24, 25) that the tumor could even be allowed to increase in size beyond its baseline level in what Brady and colleagues (27) have called "range-adaptive" AT. At the same time, these benefits are subject to a trade-off: a higher tumor burden also brings increased risks of phenotype switching, *de novo* mutations, and metastasis (10, 25, 26), indicating that the question of when and how to adapt therapy should ideally be answered on an individual basis. Furthermore, although the AT50 protocol represents an important first step, it is limited in its generalizability. How would we, for example, integrate multiple drugs, unforeseen treatment interruptions (e.g., for delays in data acquisition or toxicity), or patient-specific treatment goals?

From chatbots to self-driving cars, deep learning techniques are revolutionizing the world around us. These "deep" methods use

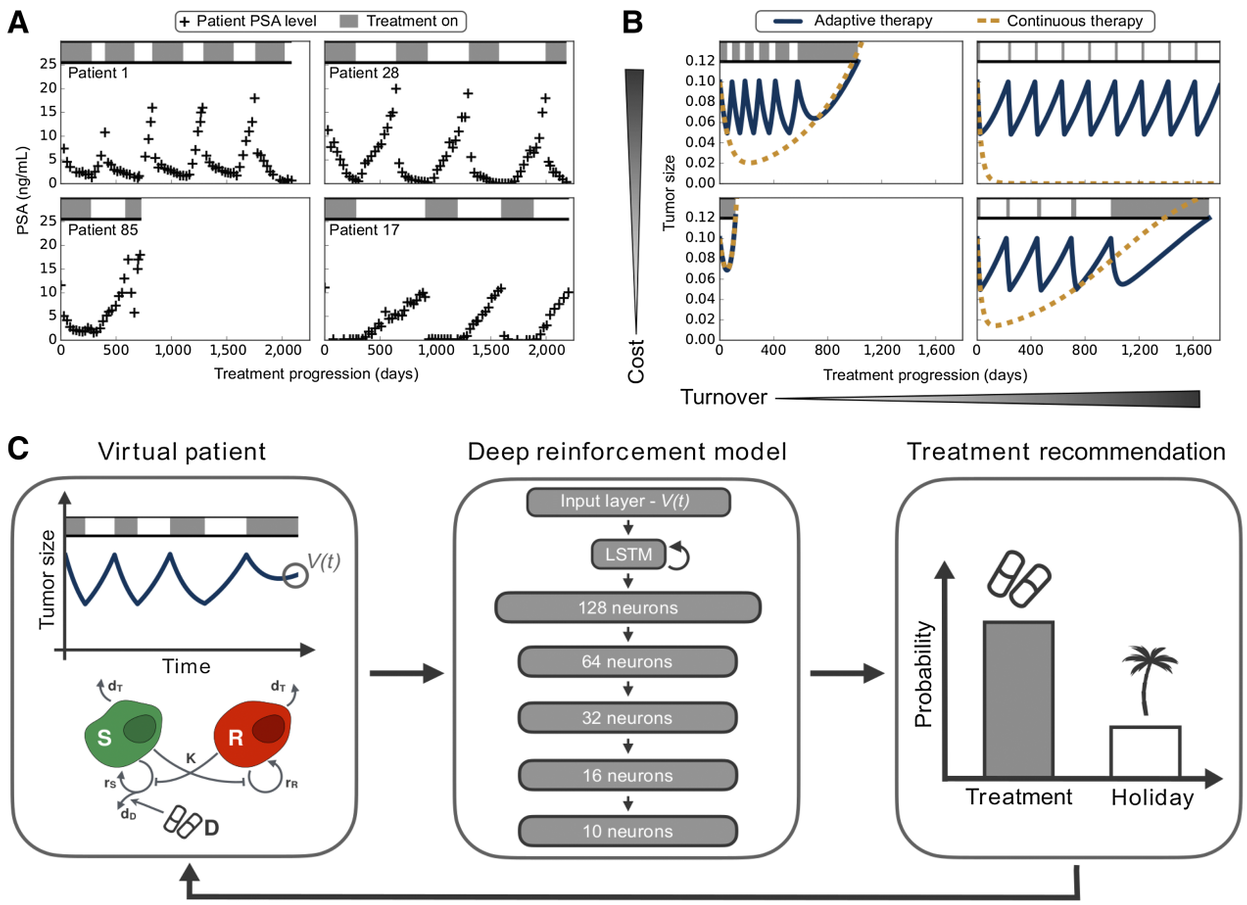


Figure 1.

A, Example treatment records of patients from Zhang and colleagues (17) demonstrating the heterogeneity in tumor dynamics under the AT50-adaptive protocol. PSA levels in the blood were used as a proxy for tumor size. **B,** This diversity in patient response to treatment can be represented by a mathematical tumor model (Eq. A), which replicates variation in tumor growth and drug response rates. **C,** DRL-guided AT. Tumor metrics, such as total size, are fed into a deep reinforcement learning model, which returns a treatment recommendation. In this paper, we test this concept using mathematical tumor models to serve as "virtual patients," and we discuss how to translate these frameworks into clinical practice (model diagram adapted from ref. 23).

artificial neural networks with many intricately connected layers that enable them to learn highly complex relationships between input variables (28). Although initially focused on classification problems, such as cancer diagnosis (29), more recently, so-called deep reinforcement learning (DRL) has extended these methods to decision-making in dynamic and complex environments such as the board game Go (30), autonomous vehicles (31), or problems in healthcare (32).

DRL frameworks have achieved success in a range of drug scheduling problems, from immune response after transplant surgery (33) to controlling drug resistance in bacteria (refs. 34, 35). At each timestep, a deep learning agent is given information on the state of the system (e.g., tumor size), and its output is used to choose from a set of possible actions (e.g., treat vs. not treat; ref. 36). To learn its strategy, the agent is trained through a process of trial and error to maximize a reward function that remunerates success (e.g., tumor shrinkage or cure) and penalizes unfavorable events (e.g., excess drug toxicity; ref. 37). DRL schemes are particularly well suited to this task, as they may account for the long-term effects of actions when maximizing outcomes, even when the relationship between actions and outcomes is not fully known (38). For example, Engelhardt (34) developed a DRL framework within the context of antibiotic resistance to predict precision dosing that adaptively targets harmful bacteria populations with variable drug susceptibility and resistance levels. They introduce a simple DRL framework capable of suppressing proliferation and demonstrate robustness to changes in model parameters, based on discrete-time feedback on the targeted population structure. However, the success of their model relies on the assumption that all strains have some degree of drug response and can ultimately be eliminated, which differs significantly from the context of treatment-nonresponsive tumors in oncology.

In the context of cancer, Maier and colleagues (39) used DRL for adjusting subsequent drug doses, to reduce toxicity in patients with cancer using neutrophil counts as a biomarker for chemotherapy-associated toxicity. They demonstrated that reinforcement learning frameworks have the potential to substantially reduce the incidence of neutropenia, and provide insight into the patient factors that determine treatment recommendations, although they do not interpret the factors underpinning the DRL framework's decision-making process itself.

However, while DRL methods are very promising, their translation into clinical practice faces two key challenges. Unlike a DRL algorithm playing chess, a patient's treatment plan cannot be replayed until the DRL agent has learned its strategy and, secondly, DRL decisions must be interpretable to gain acceptance in the clinic. To address the first issue, studies to date have used mathematical models to serve as "virtual patients," generating the vast quantities of data needed to train a machine-learning algorithm and predicting how the tumor may respond to any hypothetical treatment scenario that could not easily be tested in the clinic. This inherently links the learned strategies to the parameters and assumptions of the specific, underlying model, but how robust are DRL methods when a patient's disease behaves differently than the training model? Can we interpret, and learn from, the treatment strategies suggested by the DRL framework? And how do these strategies perform compared with standard-of-care treatment techniques?

This paper aims to investigate whether deep learning techniques may allow us to integrate evolutionary principles and mathematical models more directly into AT decision-making and uncover novel AT approaches that can be translated into clinical practice. Using a previously characterized and calibrated Lotka–Volterra mathematical model to simulate the intratumor ecological dynamics (see **Fig. 1B**), we test the ability of a DRL algorithm to guide therapy (**Fig. 1C**). We demonstrate that this framework can outperform both standard-of-care and conventional adaptive strategies, and discuss how it can help to uncover interpretable and rational principles for optimal scheduling design. In the second part of this paper, we turn to the key question of how to make DRL-based scheduling clinically feasible, when we cannot be certain about the specific characteristics of a patient's disease, and there is no way to replay or reverse a treatment decision once it is made. We apply our framework to a virtual patient cohort with a range of characteristics, demonstrating its robustness to certain changes in tumor parameters and dynamics. To conclude, we propose a framework in which we integrate mechanistic mathematical models with DRL to deliver dynamic, patient-specific treatment scheduling.

Materials and Methods

Virtual patient model

To benchmark DRL-informed AT rapidly and safely, we use a mathematical model to simulate the treatment response of a "virtual patient." We adopt the two-population Lotka–Volterra model introduced by Strobl and colleagues (see schematic in the leftmost panel of **Fig. 1C**; ref. 23), where $S(t)$ is the number of sensitive cells, and $R(t)$ is the number of resistant cells as a function of time t:

$$\begin{aligned}\frac{dS}{dt} &= r_S S\left(1-\frac{S+R}{K}\right)\times(1-d_D D)-d_S S\\ \frac{dR}{dt} &= r_R R\left(1-\frac{S+R}{K}\right)-d_R R.\end{aligned} \tag{A}$$

Briefly, the model assumes that sensitive and resistant cells proliferate and die at rates r_S and r_R, and d_S and d_R, respectively, and compete for a shared carrying capacity, K. Treatment is assumed to kill sensitive cells at a rate that is proportional to the population's growth rate and the drug concentration, $D(t)$ (with proportionality factor d_D). Although any dosing level could be considered here if the dose-response function is known, we only consider binary dosing ("on/off") in this paper, with the drug concentration on treatment given by unity. Resistant cells are assumed to be fully resistant and can be subject to a cost, so that $r_S \geq r_R$.

To simplify the notation, we define the "cost of resistance" as the proportional difference between resistant and sensitive cell growth rates $(1 - r_R/r_S)$. Similarly, we define cell turnover as the relative death and proliferation rates of sensitive cells d_S/r_S. These values, alongside the initial cell populations (S_0, R_0), define a virtual patient profile. This is used to generate synthetic patient data, by simulating the model equations (Eq. A) when training and testing the deep learning framework. By fitting to clinical data, we can personalize the virtual patient model to individual patient characteristics. All other parameter values, as well as base values and ranges for these four parameters, were adopted from Strobl and colleagues (23) and are given in Supplementary Table S1.

In sections "DRL-guided AT is predicted to improve on intermittent treatment" onward, we use parameter values previously obtained (23) by fitting the Lotka–Volterra model (Eq. A) to publicly available longitudinal response data from patients with prostate cancer undergoing intermittent androgen-deprivation therapy (40). The model was fitted to each patient by minimizing the root mean squared difference between the normalized PSA measurements and simulated tumor volumes. Full details of this process are provided by Strobl and colleagues (23).

Adaptive therapy

We benchmark DRL-informed AT against the following two protocols:

1. **CT** – Continuous therapy (standard of care):

$$D(t) = 1 \forall t \tag{B}$$

2. **AT50** – AT schedule used in the pilot clinical trial by Zhang and colleagues (17, 18). Treatment is given until a 50% decrease from the initial size (N_0) is achieved, then withdrawn until the tumor returns to its initial size:

$$D(t) = \begin{cases} 1, & until\ N(t) < 0.5N_0 \\ 0, & until\ N(t) \geq N_0 \end{cases} \tag{C}$$

We compare the schedules based on their TTP, where we define progression as a 20% growth from initial size, as in prior studies in this area (23, 25, 41).

Deep learning model

To test the feasibility and potential benefits of DRL-driven AT, we develop a prototype in which we use the asynchronous, advantage Actor–Critic (A3C) network pioneered by Mnih and colleagues (42) to drive treatment decision-making. The A3C framework consists of a global network, where many duplicates (workers) update asynchronously during training (**Fig. 2**), which avoids the high computational costs and specialist architecture requirements associated with GPU-based deep learning algorithms (43). The network receives as input the current tumor size and outputs a policy score for each of the two possible actions (treat vs. drug-holiday), reflecting which is predicted to be the more successful. Choosing an input feature set solely comprised of the current tumor size was intended to replicate clinical conditions where further information, such as the proportions of drug-sensitive and -resistant cells, is not practically accessible. To decide whether or not to treat in the next time step, the scores are converted into probabilities, and an action is chosen probabilistically. The network architecture is depicted in **Fig. 1C**, and further details, as well as a pseudocode representation, are given in Supplementary

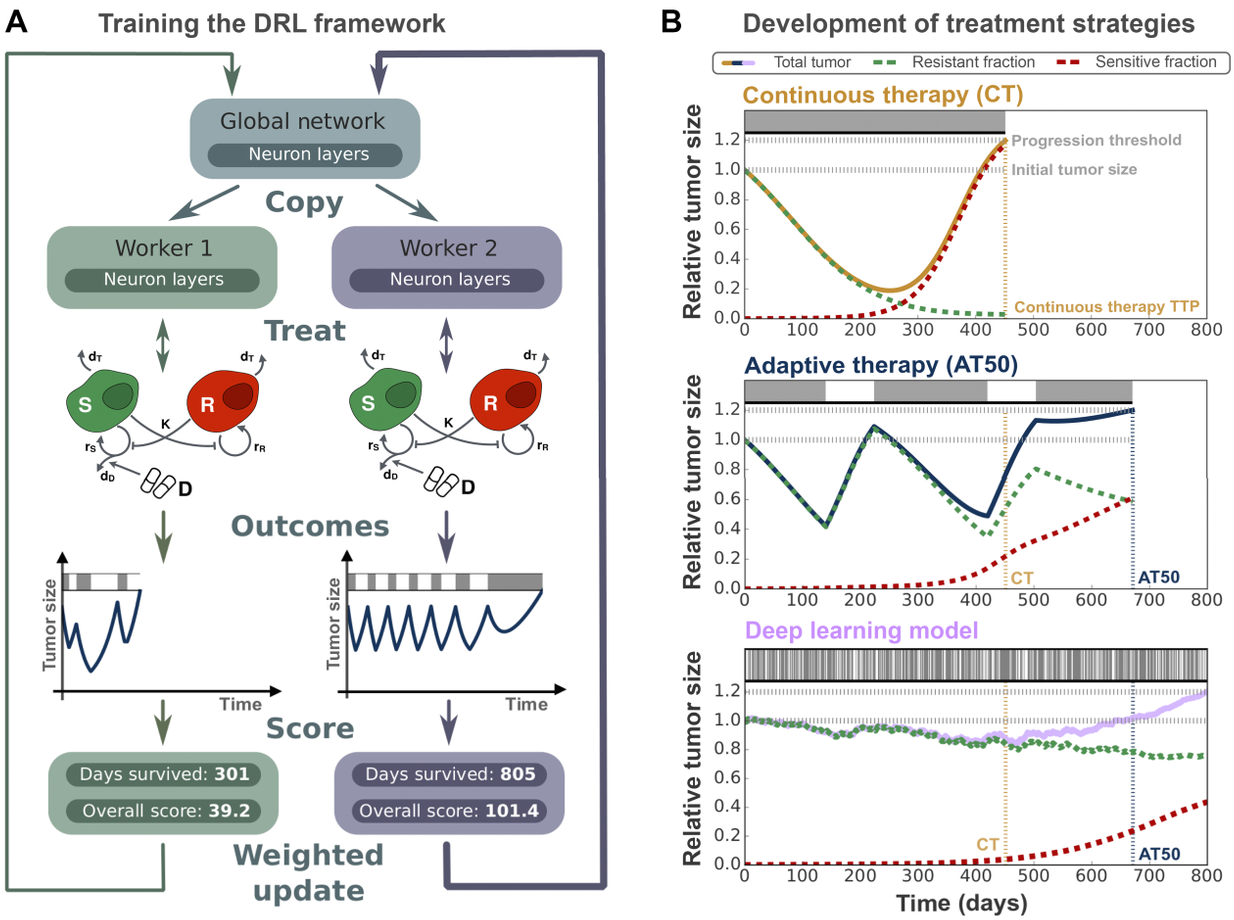


Figure 2.

A, The training process, where treatment strategies are optimized by comparing perturbed copies of the DRL network evaluated on the virtual patient system and by subsequently adjusting the global DRL network based on the relative performance of each copy. **B,** Proof-of-principle application of DRL-guided AT, showing how the deep learning model has learned to carefully adjust treatment to the patient's tumor dynamics. In this way, it can maintain a large, sensitive population and extend TTP significantly, compared with current clinical protocols with monthly treatment decisions, by forcing greater competitive suppression of the resistant cells.

Table 1. Terms, and their default values, in the default reward function for the DRL framework.

Name	Value	Circumstance	Motivation
Base	0.1	Per timestep survived	Maximizes TTP
Holiday	0.05	Per timestep without treatment	Encourages intermittent therapy
Progression	−0.1	Upon progression (20% growth from N_0)	Punishes progression
Survival	5	Upon survival for 30 years	Rewards long-term control

Section S2. A learning rate of 10^{-4} was used throughout, including for transfer learning.

Because the decision-making in the DRL framework is probabilistic, the outcome can vary between iterations (Supplementary Fig. S1). To account for this, we report results as averages over 100 evaluations (unless otherwise stated). This number was determined to be sufficient for the treatment scenarios explored in this paper based on a consistency analysis (not shown).

Reward function

The DRL framework learns treatment strategies through optimizing an objective function, which returns a reward calculated from the most recent state and treatment choice at each timestep. This reward is based on the number of timesteps survived, encouraging the model to maximize TTP. A bonus is given for treatment holidays, to incentivize intermittent treatment strategies. To bound maximal runtime, we add a 30-year survival limit; however, we focus on patient profiles that reach progression well within this period. A list of all terms, alongside their default values and the circumstances under which they are rewarded, is given in **Table 1**. Applicable terms (based on the circumstances under which each term is included) are then summed to give the total reward; a more detailed pseudocode representation of this reward function is provided in Supplementary Section S2.

Discounting is applied to the reward function to determine how important future rewards are to the current state, where a reward that occurs N steps in the future is multiplied by γ^N, for the discounting factor $\gamma \in (0, 1)$. As well as ensuring convergence of the reward sum, this factor is used to tune the prioritization of short and long timescales in the reward function. A value very close to unity ($\gamma = 0.999$) is used throughout to prioritize overall outcomes over shorter-term benefits.

Data availability

All methods, along with the DRL framework, are available at https://github.com/MathOnco/DRL_Personalized_AT and https://codeocean.com/capsule/4619998/tree/v1. Clinical data from Bruchovsky and colleagues (39) were obtained from https://www.nicholasbruchovsky.com/clinicalResearch.html.

Results

This paper aims to investigate whether deep learning can inform cancer therapy scheduling, and whether this can be achieved in a clinically translatable fashion. To this end, we tested DRL-guided AT on cohorts of virtual patients in which the tumor dynamics were driven by a previously established mathematical model (23). This model assumes that the tumor is composed of treatment-sensitive and -resistant cells and that these compete with each other for resources in a Lotka–Volterra fashion (**Fig. 2**).

DRL-guided adaptive therapy can outperform current clinical strategies

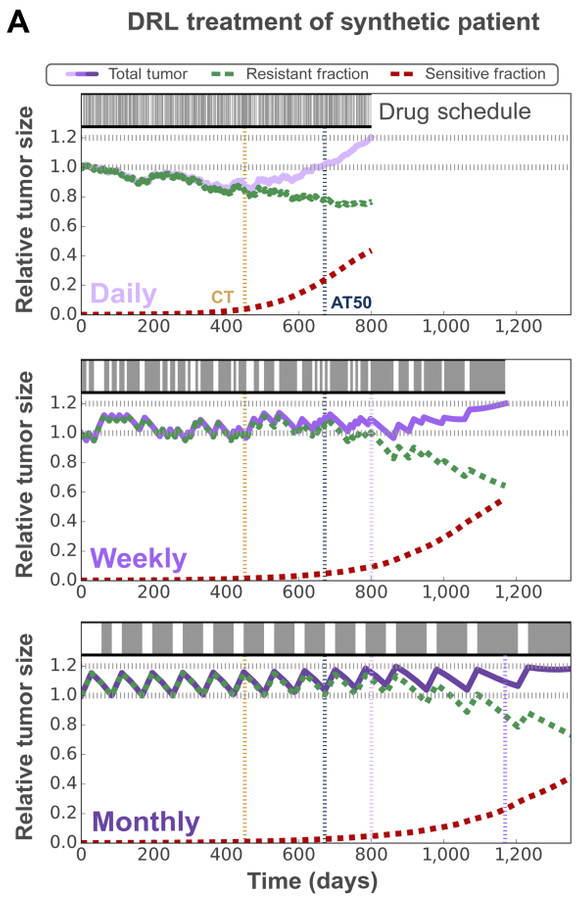
As a first step, we carried out a proof-of-principle case study on a patient assumed to have an initially responsive, but rapidly progressing, disease. This was represented by the virtual patient model, with parameter values taken from Strobl and colleagues (23) and given in Supplementary Section S1. Akin to the AT50 algorithm used by Zhang and colleagues (17), the patient's tumor burden in our DRL-guided protocol is monitored and the decision to continue to treat or not is updated at a fixed "decision frequency." Importantly, however, this decision is not based on a fixed rule of thumb, but on a deep learning model that is carefully trained prior to deployment. During this training process, the DRL framework is applied to a cohort of "training patients"; its performance is scored according to the reward function, and the parameters of the underlying neural network are refined until it converges on a final decision-making policy (**Fig. 2A**).

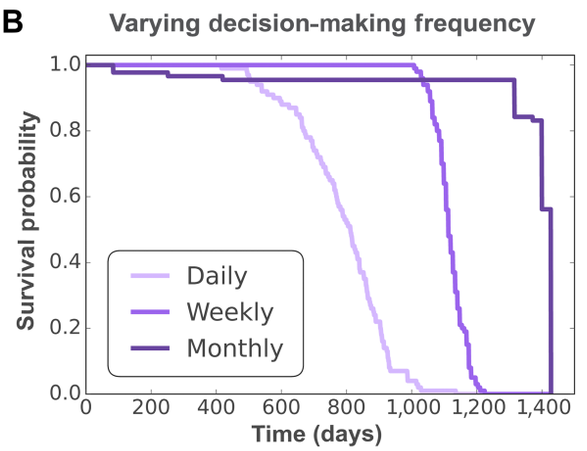
To test whether the DRL could learn an effective treatment strategy, we initially made the idealized assumptions that we could train on a patient identical to the one on which the framework will be deployed and that we could revise treatment every day. Following 2,600 training epochs, we find that DRL-guided treatment is able to control the tumor for longer than conventional treatment schedules (**Fig. 2B**). Through dynamic adjustment of treatment, it achieves an average TTP of 745 days over 100 independent simulations (95% confidence interval [688d, 802d]). The variability in performance is due to the stochastic nature of the DRL decision-making (see Supplementary Section S3 for examples). In comparison, CT and the AT50 therapy progress after 450 and 662 days, respectively (**Fig. 2B**). We conclude that DRL-guided therapy is, in principle, feasible and can improve upon the current AT50 rule-of-thumb approach.

Reducing decision-making frequency can increase performance

To investigate whether, and how, this framework could be used in practice, we next analyzed sensitivity to key parameters in the training and deployment process. In the previous section, we made the somewhat unrealistic assumption that the DRL framework receives tumor size input and reevaluates the treatment strategy on a daily basis. Clearly, this would be both difficult and costly to implement clinically.

Interestingly, and somewhat counterintuitively, we found that reducing the decision frequency of the model increases the expected TTP, despite the reduced information and intervention frequency (**Fig. 3A**). In all cases, this reduction was statistically significant ($P < 0.05$), determined by the probability that a pairwise comparison between model evaluations using different treatment frequencies did not satisfy the aforementioned trend, using 100 evaluations per frequency. In addition, this reduces the computational cost per patient during training. We hypothesize that this improvement in performance is because less frequent decisions are more impactful, enabling the DRL framework to better separate the meaningful trends in the underlying tumor dynamics from random noise in the decision





C
Training vs. evaluation frequency

TTP (days)		Training decision frequency		
		Daily	Weekly	Monthly
Evaluation decision frequency	Daily	745	1,141	1,281
	Weekly	571	1,116	1,341
	Monthly	414	353	1,388

Figure 3.

A, DRL framework performance for different treatment intervals after training on a single-patient profile (Supplementary Section S1). Less frequent decision-making during training allows the DRL to learn more effective treatment schedules. **B,** Kaplan–Meier curves for the same DRL frameworks evaluated across a cohort of 100 virtual patients, where less frequent decisions in training also result in more consistent outcomes, outperforming other strategies in over 90% of cases. **C,** Evaluating the DRL strategy at lower frequencies than used in training results in a stark decrease in TTP (below the diagonal, averaged over 100 patients) as these allow insufficient time to respond to changes in tumor dynamics, whereas there is no significant change at higher frequencies than those used in training. For comparison, the TTP is 672 days under the AT50 scheme and 451 days under conventional CT.

process. This is reflected in the reduced variation in TTP for longer decision frequencies (**Fig. 3B**).

We also considered the role of decision frequency during evaluation. Although there is no increase in mean TTP from reducing the treatment interval, because the model has been optimized for a specific interval during training (shown above the diagonal in **Fig. 3C**), we do observe reduced variance in TTP when treatment decisions are made more frequently in evaluation (not shown). In contrast, if the treatment interval is increased compared with training, TTP is typically reduced (shown in the region below the diagonal in **Fig. 3C**, with $P = 0.2$ for daily training, and $P = 0.035$ for weekly training, again computed via pairwise comparisons between the two data sets). This is because the DRL framework underestimates how quickly the tumor will grow and fails to apply sufficient treatment to keep it from progressing. The tumor thus progresses even though it could have been kept in check with further treatment, a phenomenon we will refer to as "premature progression." To sum up, choosing the frequency at which to consult the DRL framework will require a careful balance between leaving sufficient time to learn from past decisions, while also providing sufficiently frequent decisions to react to the tumor's response or lack thereof.

The DRL framework learns an interpretable treatment policy

Although the performance of the DRL framework is promising, a key challenge to potential clinical translation is its "black-box" nature. To investigate the policy underlying DRL decision-making, we next plot the relationship between the input of the network (current tumor size) and its output (the treatment recommendation) while treating a single patient (**Fig. 4A**). Interestingly, this reveals a clear sigmoidal relationship with a well-defined interpretation: when the current tumor size is above the critical size N^*, the network decides to treat; otherwise, the patient is predicted to benefit more from a treatment holiday (**Fig. 4B**).

In fact, we can dissect the strategy further, showing that the DRL has learned to choose N^* to carefully match the particular patient's treatment dynamics. To do so, we emulate the DRL's decision-making by replacing the neural network in our framework with a simple, binary switch that treats if the tumor size is above some

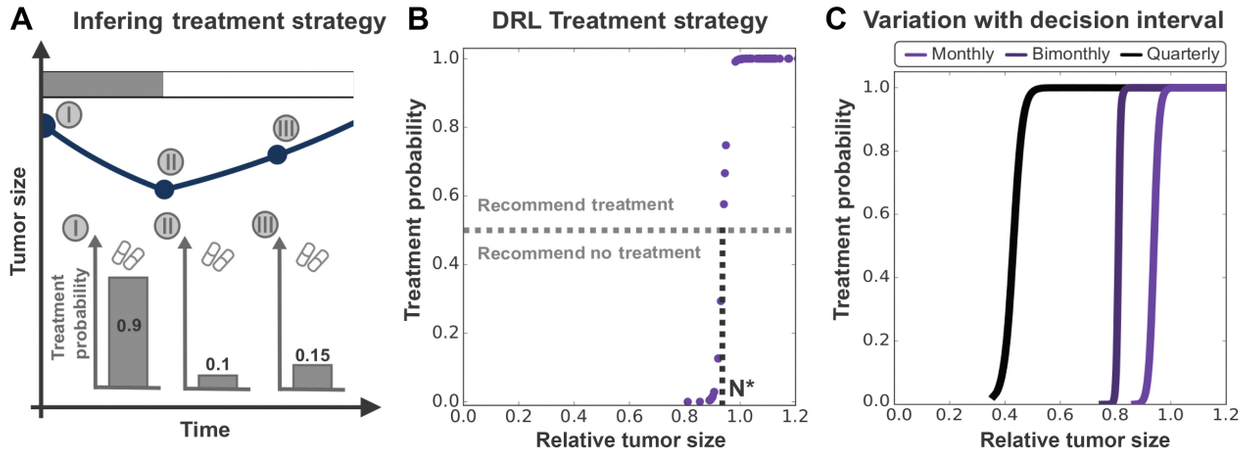


Figure 4.

A, After each decision interval, the DRL framework outputs a treatment probability based on the current tumor size. By recording these over a patient's history, we can infer the underlying treatment strategy learned by the framework. **B,** This strategy is a well-defined sigmoidal relationship between current tumor size and treatment probability, with treatment almost certain above a threshold size, $N^* = 0.83n_0$ (for an initial tumor size n_0). **C,** The value of the threshold size, N^*, varies significantly with the frequency of decision-making in treatment, with more frequent decisions allowing a higher threshold.

threshold, $\tilde{N}$, and otherwise holds. Varying $\tilde{N}$ around the value N^* found by the neural network, we observe that both increasing or decreasing it reduces TTP: if we switch at larger sizes ($\tilde{N} > N^*$), we risk early progression, whereas if we switch at smaller sizes ($\tilde{N} < N^*$) we maintain fewer sensitive cells and can thus exert less competitive suppression on the resistant subpopulation (not shown). We conclude that the DRL has identified a switch point that seeks to optimize the tradeoff between controlling sensitive and resistant cells, which varies with parameters such as the decision interval (**Fig. 4C**). Through this tradeoff, the DRL framework qualitatively replicates strategies previously shown to be analytically optimal in this setting (10, 25).

In this way, DRL methods can help us to derive effective treatment strategies without explicit knowledge of the underlying mathematical system, giving confidence that the DRL may be applied to more complex treatment paradigms (with multiple drugs or dosing levels) where mathematical methods cannot readily derive an optimal treatment schedule. Secondly, we have shown how, by using mechanistic tumor models, we can identify clinically actionable strategies from a "black-box" DRL network and thereby build confidence in the recommended decisions, addressing an important hurdle toward clinical translation.

DRL-guided AT is predicted to improve on intermittent treatment

Although our results so far are promising, they were obtained on a single virtual patient who was assumed, for simplicity, to have neither a cost of resistance nor significant turnover. Although this makes it harder to control resistance, it also results in faster tumor dynamics, which makes it easier for the DRL framework to learn the impact of particular treatment decisions. In the next step, we wanted to test how the DRL framework would perform on more clinically realistic parameter sets. To replicate the tumor dynamics observed in the clinic, we leverage prior work (23) in which we have fitted the Lotka–Volterra model (Eq. A) to patient data from a prospective phase II trial of intermittent androgen suppression for locally advanced prostate cancer conducted by Bruchovsky and colleagues (40). Specifically, we focus on the fits of 7 patients who had progressed on the trial to test whether the DRL-guided AT would have achieved a longer TTP and how this would have compared with AT50 (see Supplementary Section S4 for further details on patient selection). Plotting this cohort in **Fig. 5A**, we observe that these patients differ widely in their dynamics, as reflected in their associated model parameters. Nevertheless, the DRL framework can learn to delay the emergence of drug resistance and increase TTP relative to the clinically trialed intermittent protocol, as exemplified for patient 25 in **Fig. 5B**.

DRL frameworks can be robust to variation in patient parameters

The aim of this paper was to leverage DRL to better tailor AT to the dynamics of an individual patient. Our results so far show great promise but assume that we have perfect knowledge of the governing rules and parameters—which is clearly unrealistic. In the next step, we thus tested how robust a DRL-guided approach is to uncertainty in the underlying dynamics and what strategies we can adopt to better address the individual variations and imperfect knowledge of each patient's parameters in the clinic.

To explore this question, we first carry out a sensitivity analysis in which we systematically perturb the model parameters away from the values the DRL framework encountered in training (for further details see Supplementary Section S5). The DRL framework demonstrates robustness to varying initial resistance fractions and resistance costs associated with the tumor, providing effective treatment relative to AT50 even if these differ significantly from the values of the training patient (Supplementary Fig. S2). In contrast, it is very sensitive to reductions in the initial tumor density and in the rate of turnover (i.e., the ratio between death and growth rates in the tumor), for example, performing worse than AT50 for patients with identical initial tumor composition but a slightly reduced turnover (Supplementary Fig. S2). From the insights in section "The DRL framework learns an interpretable treatment policy" into how the DRL makes its decisions, we can explain this behavior: Changes in turnover and the initial tumor density alter the tumor growth rate relative to that of the training patient and this faster-than-anticipated tumor growth rate leads to premature progression between treatment decisions. This demonstrates how the use of a mechanistic tumor model can dissect the limitations of the DRL framework, differentiating between scenarios in

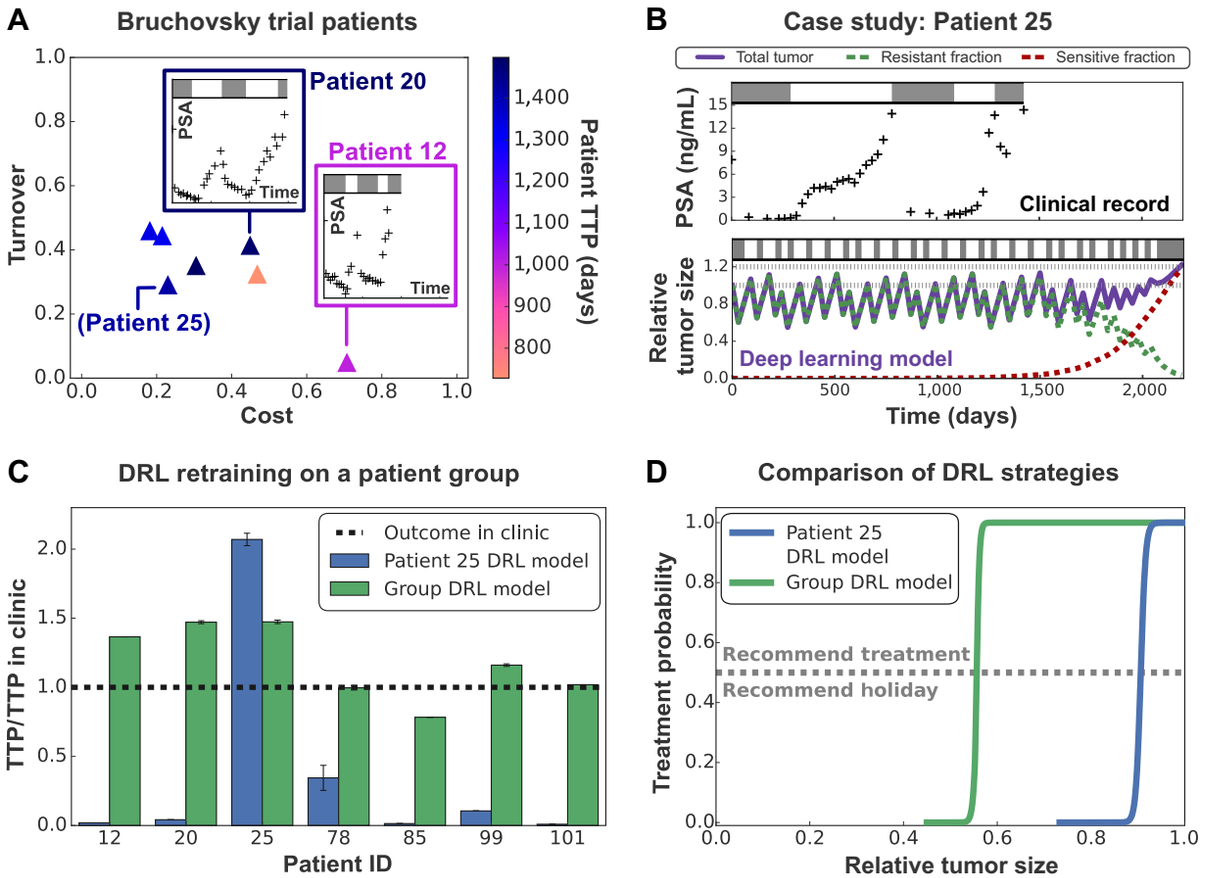


Figure 5.

A, Patient parameter values in cost-turnover parameter space from fits conducted by Strobl and colleagues (23) on patients from the Bruchovsky (41) trial. Selected clinical records are included to illustrate the variation between patients. **B,** Training a DRL framework on a single patient from this cohort generates a specialized model able to significantly outperform outcomes seen in the clinic. **C,** However, individual models struggle when applied to other patient profiles. This can be remedied by training the DRL framework on a virtual patient cohort, at the cost of reduced specialization on each individual. **D,** The generalism manifests in the form of a more conservative strategy that avoids premature progression for all patients in the group but is suboptimal for the least aggressive tumors.

which it is robust (such as to perturbations in the resistant characteristics) and scenarios where it is highly sensitive (such as increases in the tumor growth rate).

In addition to parameter uncertainty, tumor behavior may not align with the virtual patient model (1), as patient tumors may exhibit heterogeneity in the rules governing their dynamics. Hence, we assess the DRL framework's robustness to variations in the underlying tumor dynamics by evaluating our framework on a modified Lotka–Volterra model introduced by Lu and colleagues (44) and reproduced in Supplementary Section S6, with model parameters given in Supplementary Table S2. This model uses a diminishing competition term, such that intratumoral competition decreases over time, with a modified progression criterion based on the growth of the resistant subpopulation alone. Despite these significant differences in dynamics and progression criterion, the (pretrained) DRL framework achieves a TTP of 1,506 $\pm$ 3 days under weekly treatment evaluation, outperforming the AT50 TTP of 1,119 days while reducing the cumulative drug dose by 93% (Supplementary Fig. S3). Additionally in Supplementary Section S6, we benchmark the DRL framework against a stem-cell model published by Brady-Nicholls and colleagues (45) with parameters given in Supplementary Table S3. We demonstrate a benefit of 718 $\pm$ 102 days over AT50, exemplifying the ability of this framework to adapt to models with differing underlying assumptions and mathematical formulations that it was never trained upon.

Generalized treatment strategies through cohort training

As we have just shown, a DRL framework that only "sees" one particular patient profile in training has a limited ability to adapt to different patients or to parameter uncertainty in evaluation. To address this, we tested how well we could enhance the framework's robustness by exposing it not just to one, but to multiple, patient profiles during training. This meant that for each iteration during training, we randomly chose one of the seven patients in the Bruchovsky group as the training profile. Importantly, we did not provide any further information about the patient's tumor dynamics, so the algorithm had to infer that the tumor dynamics might have changed from the previous iteration while providing treatment. In addition, instead of training a new DRL framework from scratch, we used a technique known as "transfer learning," where a preexisting network is retrained for a new task, in order to retain and improve upon the strategy the framework had learned in the single-patient training so far.

Compared with the DRL framework trained on patient 25 alone, this multipatient network is significantly more robust to variation in patient parameters, able to match or outperform the clinical outcomes for all patients ($P < 0.01$ for 100 evaluations of the DRL framework; **Fig. 5C**). However, this comes at the cost of specialization; this cohort-trained DRL framework is significantly less effective on patient 25 (on whom the original framework was trained), although it still outperforms the intermittent strategy by Bruchovsky and colleagues. This is due to a reduced threshold in the treatment strategy so that a lower tumor burden will be maintained throughout treatment (**Fig. 5D**); this prevents fast-growing tumors from reaching progression prematurely but reduces the competitive suppression of the resistant cell population in slower-growing tumors.

The robustness of multipatient networks extends beyond the parameters explicitly varied in the training cohort; in Supplementary Section S7, we vary the time interval at which treatment decisions are made. This was fixed at 30 days in training, but the variation here reflects the realities of treatment delivery in the clinic, where missed and rescheduled appointments often result in delays to planned treatment schedules. Modeling this delay as an exponentially distributed random variable, we found that the patient 25 model experienced a significant decrease in TTP (≥ 20%) for delays averaging less than a week, whereas the group-trained model could tolerate average delays approaching a month (Supplementary Fig. S5). This reinforces the increased robustness of group-trained models to changes in evaluation parameters, although this comes at a cost of reduced performance in the training case.

A further limitation of the multipatient approach is that it is typically only robust to the extent of variation it is exposed to during training. To illustrate this, we generated a cohort of ten "synthetic training patients" by sampling from a defined area of parameter space and used these to train a multipatient network (for details see Supplementary Section S7). We make two key observations: first, DRL-guided AT does not require training on the exact profiles used in evaluation and can match, if not outperform, AT50 if the profile of the patient it is applied to is within, or at least sufficiently close to, the training space (Supplementary Fig. S4). This validation provides confidence in our DRL approach, demonstrating that the DRL framework generates successful treatment schedules for patients it has not "seen" during training, drawing on strategies it has learnt for similar patients in training. However, if a patient falls outside the training space, then even the cohort-trained network can quickly become unreliable (Supplementary Fig. S4). We conclude that training on multiple patient profiles can enhance the robustness of the network, but comes at the price of reduced performance and is still vulnerable if a patient were to fall outside the range of dynamics encountered during training.

Personalized DRL framework

To address these issues, we investigated how we could generate a separate DRL network for each individual patient, fine-tuned to their specific tumor dynamics. However, when a patient first presents, we do not have enough data, for example, on how fast the tumor will respond to treatment, to construct such a "virtual twin." To overcome this problem, we propose to combine old and new methods in a five-step clinical pathway (**Fig. 6A**). By conducting a single, initial AT50 "probing cycle," we collect longitudinal burden data to parameterize the virtual patient model. Next, this model is used to fine-tune a generalist DRL network, trained on the synthetic cohort detailed in Supplementary Section S7 such that no prior knowledge of the patient is assumed; this personalized version is then used to guide subsequent treatment decisions. In this way, we can dynamically tailor the decision-making policy to the dynamics of the individual patient and make an important step toward a clinically feasible implementation of our DRL treatment scheduling framework.

To illustrate this, in **Fig. 6B** we show an example application of this five-step approach for patient 25. When they present in the clinic, they are given an initial cycle of AT50, during which their tumor burden dynamics are recorded and an estimate of their tumor parameters is inferred (see Supplementary Section S8 for numerical details and Supplementary Fig. S6 for the model fits). Using the synthetic, multipatient network from the previous section as the initial network state, we apply transfer learning to retrain the DRL framework to tailor it to this patient. For subsequent treatment cycles, we then switch from AT50 to our DRL-based personalized AT scheme. This successfully maintains a higher stable tumor burden to maximally suppress the drug-resistant subpopulation, increasing TTP by 2.3 years for patient 25 relative to if we had continued on AT50 (see **Fig. 6D** for performance on other patients). In addition, the retraining of our generalized DRL network is feasible in a clinical setting, achieving sufficient personalization to each patient after as little as 2,000 epochs, taking approximately 20 minutes to train on a standard laptop (Intel i5, 4 cores, 1.70 GHz).

A further important challenge in the practical translation of our five-step framework is a reticence to expose a patient to a treatment schedule derived from a "black-box" algorithm. As a workaround, we propose that rather than basing decisions on the raw, numerical output of the network, we could leverage the insights we gained in the section "The DRL framework learns an interpretable treatment policy" to extract and deploy its strategy in an interpretable fashion. To do so, we fit a sigmoid curve to the strategy observed for each personalized network (recall **Fig. 4**) and extract the critical treatment threshold optimized by DRL treatment threshold optimized by DRL for that patient (**Fig 6A**; steps 3 and 4). For example, for patient 12, this would be $0.73n_0$, implying that a clinician should aim to maintain the tumor at 73% of its original size, treating when the tumor is larger and giving treatment holidays when it is smaller. We observe significant variation in these personalized thresholds between patients (**Fig. 6C**), demonstrating the need for tailored approaches to treatment scheduling, with different patients responding best to markedly different strategies ranging from 73% to 100%. Benchmarking these personalized thresholds derived from the DRL frameworks across the cohort, we show that we can match the performance of the "black-box" networks themselves, as well as matching or outperforming the nonpersonalized, intermittent protocol for all patients (**Fig. 6D**). Even for those patients with no significant increase in TTP under the DRL framework, they still benefit from lower drug toxicity, as the cumulative drug dose is significantly reduced compared with an equivalent duration of CT across all patients, with an average reduction of 50% in the cumulative dose over a patient's entire treatment.

This five-step approach heralds a new avenue in personalized medicine, through tailored treatment schedules on a per-patient basis, driven by their initial response to one cycle of treatment. In the clinic, it would also be possible to refit the patient's estimated tumor parameters after data from each subsequent treatment cycle are collected, generating a more accurate prediction of the patient's dynamics that can readjust to changes in the tumor behavior over the timeline of treatment, exemplifying the Adaptive Dosing Adjusted for Personalized Tumorscapes principles we have recently proposed (46). Overall, this approach illustrates a new role for DRL in the development of new scheduling protocols, in which its purpose is not to compute these schedules, but to uncover guiding

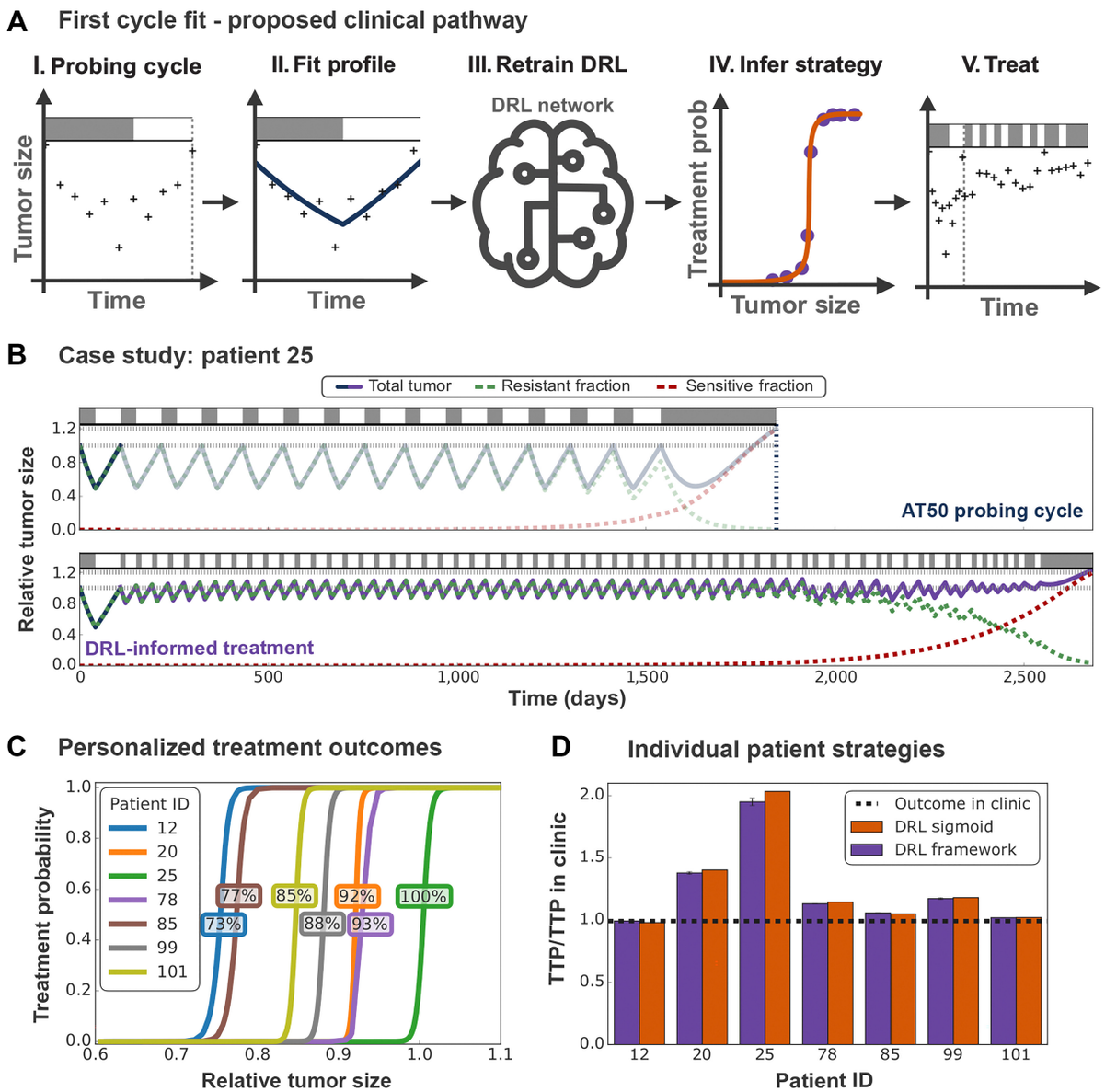


Figure 6.

A, We propose a five-step clinical framework for personalized treatment schedules. Patients undergo an initial "probing" cycle of AT50, to which the virtual patient model is fitted, generating a set of tumor parameters specific to each patient. A copy of the generalized DRL model is then retrained on these personalized parameters, fine-tuning the network to that patient's treatment response. Finally, we extract a sigmoidal treatment strategy fitted to decisions made by the DRL network, providing personalized schedule recommendations throughout the remainder of the treatment schedule. **B,** Example application of the framework to patient 25. Switching to the treatment schedule predicted by the DRL framework after the probing cycle significantly increases the TTP over AT50 (faded line in top panel). **C,** Personalized sigmoidal treatment strategy demonstrated significant interpatient variation in the critical tumor size (labeled for each line) at which treatment is initiated and discontinued. **D,** The personalized DRL framework consistently outperforms the TTP recorded for the patients in the clinic. Moreover, by extracting a personalized, interpretable sigmoidal treatment strategy, we eliminate the "black-box" nature of the framework, while retaining comparable performance.

principles that can subsequently be rationally translated into simpler and more clinically practical protocols.

Discussion

Following the recent clinical implementation of adaptive therapy on castrate-resistant prostate cancer by Zhang and colleagues (18), and multiple ongoing adaptive therapy trials in skin (NCT05651828), prostate (NCT05393791), and ovarian (NCT05080556) cancers, it is of significant interest to better understand whether we can optimize adaptive treatment strategies. Although such optimization may be conducted analytically where binary treatment decisions are applied to simple models, it quickly becomes infeasible in more complex cases. We therefore consider the application of deep learning models,

introducing a reinforcement learning framework wherein the DRL framework interacts with a virtual patient model to develop successful treatment strategies based only on a single clinically accessible tumor metric: total burden.

We demonstrate that this DRL framework can outperform the "rule of thumb" adaptive therapy (AT50) used clinically by Zhang and colleagues (18). This strategy is robust to varying model parameterizations and across multiple underlying tumor models. We uncovered a novel relationship regarding the frequency of treatment decisions (i.e., how often patient data are collected and the current treatment reevaluated), showing that application of treatment should occur at lower frequencies in initial training, to allow the DRL framework to learn meaningful and interpretable treatment strategies. However, this must be balanced by sufficient time to react to tumor growth, with faster-growing tumors requiring more frequent treatment decisions.

Exploring the decision-making process behind the DRL network, we discovered that it had learned to mimic optimal strategies previously derived through optimal control theory (7, 10, 25), maintaining a large sensitive cell population for maximal suppression of the resistant cells while avoiding the progression limit on tumor size. This is enacted through an interpretable "threshold tumor size," below which treatment holidays should be applied to retain competitive suppression of drug-resistant cells through the sensitive population. This threshold is optimized for both the patient profile used in training and the decision-making frequency and provides interpretable and clinically actionable strategies from traditional "black-box" DRL techniques. The fact that the DRL framework uncovered these strategies without knowledge of the underlying mathematical model also illustrates its potential to study more complex models or treatment paradigms (such as multiple treatment drugs or nonbinary dosing levels). The choice of a binary dosing level in this work was made to reflect the single currently approved dosing level of abiraterone; however, the consideration of variable dosing has been shown to impact optimal treatment strategies in other dosing contexts (bioRxiv 2023.03.22.533721; 2023.09.18.558136).

The significant interpatient heterogeneity in the adaptive cycling dynamics observed in the clinic (recall **Fig. 1A**) highlights the importance of decision-making frameworks that are robust to variation between patients and uncertainty in individual patient dynamics. By training the DRL framework on a cohort of patients with differing characteristics, we generate a network that is robust to variation in the values of the underlying tumor parameters. Furthermore, by evaluating a pretrained DRL framework on alternative ordinary differential equation models (Supplementary Section S6), we demonstrated that the framework is sufficiently flexible/model-agnostic to outperform the clinical standard of care across a range of tumor behaviors. It is worth noting, however, that these models were all non-spatial; previous work has demonstrated the important role that space can play in the competitive inhibition of resistant cells (41, 47, 48), which can both help and hinder adaptive therapy as well as change the optimal strategy. An important next step would therefore be to validate our framework in a more realistic, spatial model. Furthermore, we have made the simplifying assumption that PSA exactly reflects the tumor burden dynamics. In reality, the relationship between PSA and tumor burden is more complex as it is influenced by other factors, such as age or body mass index (21, 22), and lesions may differ in how much PSA each releases. Integrating multiple measurement modalities, as well as tackling the management of heterogeneous metastases (49), are important directions for future research.

The increased robustness to parameter and model variation comes at the cost of performance for an individual patient. In short, for the best results, the tumor size threshold at which treatment is withdrawn should be tailored to the tumor characteristics of the individual patient and the frequency of monitoring/decision-making. The "one size fits all" approach adopted, for example, by the current AT50 protocol is suboptimal for the majority of patients: If treatment is given below this threshold, then we suppress the sensitive subpopulation within the tumor unnecessarily, while if it is withheld above this threshold then we risk the patient undergoing premature progression before the next treatment decision.

Historically, AT protocols have focused on a single tumor size threshold (such as in AT50), and others have suggested this threshold should be maximized (24, 25, 27). Given the vastly different tumor dynamics observed between patients in the clinic, we show that a personalized switch criterion for each individual patient is required, based on their dynamics as well as practical constraints imposed by the frequency of treatment decisions in the clinic. To calculate this threshold, we propose a five-step pathway integrating mechanistic modeling with DRL. By using an AT50 probing cycle to characterize the treatment response dynamics of an individual patient, we are able to retrain the DRL network on each patient's dynamics requiring only limited additional computational expenses. Finally, from this personalized DRL framework, we extract a treatment threshold tailored to that patient, which is translated into a simple clinical protocol, to ensure that DRL-informed treatment strategies remain fully interpretable. By prescribing treatment holidays when the PSA is below this threshold value, the clinician can generate a truly personalized treatment schedule tailored to the patient's tumor dynamics and drug response. This schedule is expected to consistently outperform clinical standard-of-care protocols as well as generic AT50, which does not fully account for such interpatient variation. By tailoring our robust, cohort-trained DRL strategies to individual patients, we have demonstrated a clear route for how the results from our *in silico* study could be translated to support clinical decision-making.

In future work, we plan to further leverage uncertainty about patient measurements and dynamics by generating a virtual cohort of patients for each individual patient. We will then use this cohort to retrain our more generalized DRL framework for that individual patient. These virtual cohorts would match the real patient's dynamics within a given error similar to the phase I approach we have previously used (medRxiv 2023.01.18.23284628; ref. 50). Such approaches would retain the robustness benefits from training on a range of patients, while allowing treatment schedules to be tailored to individual patients based on their profiles. In practical terms, this would require minimal retraining of a nonspecialized model using a best estimate of the tumor parameters from fitting to the patient's current clinical history. As more data are collected from a patient, this profile may be continually refined to represent our best understanding of that patient's dynamics at a given point in time. The objective function also provides scope for additional personalization: By modifying how strongly we punish cumulative drug use, treatment schedules could be further deescalated for patients struggling to complete their planned course due to side effects/cytotoxicity.

The approach we have presented here is not the first DRL framework to tackle AT. The work of Lu and colleagues (44) used a different underlying prostate cancer model, but trained the network on both the PSA and the sensitive and resistant populations; although PSA is easily measured in a blood draw, there is no direct way to measure the numbers of sensitive and resistant cells in a real patient. Although they were also able to produce significantly improved outcomes for a single patient, the generalizability of our approach, and the clear path to clinical practice, make our

implementation somewhat more robust. Moreover, our novel discovery of the analytical relationship between tumor dynamics and treatment timing for the treatment threshold sets our work apart in terms of interpretability and also frequency of treatment. With the ongoing revolution in wearables, a future where tumor burden could be measured in real time is not unrealistic. Our results imply that caution is needed in making too frequent treatment decisions in DRL training, because this effectively reduces the impact of each decision and may increase the interference from noise.

The approaches presented here are generalizable to other forms of cancer and illustrate how integrating mathematical modeling and machine learning can provide rational decision-support to tackle the complex and evolving nature of cancer (46). This applies both in situations where total tumor burden or drug load must be managed, either to prevent treatment resistance or to reduce adverse side effects. They may also be extended to multidrug paradigms or to allow variable dosing levels for individual drugs, provided a dose-response function is known. We hypothesize that the application of DRL frameworks may allow optimization in a wide range of clinical settings where clinical standards are currently determined by a "rule of thumb" that may, in many cases, be far from the optimal strategies that a DRL framework could identify.

Authors' Disclosures

No disclosures were reported.

Authors' Contributions

K. Gallagher: Conceptualization, data curation, software, formal analysis, investigation, visualization, methodology, writing–original draft, writing–review and editing. **M.A.R. Strobl:** Conceptualization, data curation, software, supervision, investigation, visualization, methodology, writing–original draft, writing–review and editing. **D.S. Park:** Conceptualization, software, methodology, writing–review and editing. **F.C. Spoendlin:** Investigation, writing–review and editing. **R.A. Gatenby:** Conceptualization, data curation, supervision, funding acquisition, project administration, writing–review and editing. **P.K. Maini:** Conceptualization, resources, supervision, funding acquisition, project administration, writing–review and editing. **A.R.A. Anderson:** Conceptualization, resources, supervision, funding acquisition, visualization, project administration, writing–review and editing.

Acknowledgments

K. Gallagher acknowledges funding from the EPSRC CDT in Sustainable Approaches to Biomedical Science: Responsible and Reproducible Research—SABS:R3 (EP/S024093/1). The authors gratefully acknowledge funding by the NCI via the Cancer Systems Biology Consortium (CSBC; U01CA232382 and U54CA274507 supporting M. Strobl, R. Gatenby, and A. Anderson) and support from the Moffitt Center of Excellence for Evolutionary Therapy.

Note

Supplementary data for this article are available at Cancer Research Online (http://cancerres.aacrjournals.org/).

Received August 16, 2023; revised January 5, 2024; accepted March 21, 2024; published first April 3, 2024.

References

1. Bukowski K, Kciuk M, Kontek R. Mechanisms of multidrug resistance in cancer chemotherapy. Int J Mol Sci 2020;21:3233.
2. Holohan C, Schaeybroeck SV, Longley DB, Johnston PG. Cancer drug resistance: an evolving paradigm. Nat Rev Cancer 2013;13:714–26.
3. Ganesh K, Massagué J. Targeting metastatic cancer. Nat Med 2021;27:3444.
4. Tevaarwerk AJ, Gray RJ, Schneider BP, Smith ML, Wagner LI, Fetting JH, et al. Survival in patients with metastatic recurrent breast cancer after adjuvant chemotherapy: little evidence of improvement over the past 30 years. Cancer 2013;119:1140–8.
5. Perry MC. The chemotherapy source book. vol. 117. Lippincott Williams & Wilkins; 1992.
6. Maley CC, Aktipis A, Graham TA, Sottoriva A, Boddy AM, Janiszewska M, et al. Classifying the evolutionary and ecological features of neoplasms. Nat Rev Cancer 2017;17:605–19.
7. Martin RB, Fisher ME, Minchin RF, Teo KL. Optimal control of tumor size used to maximize survival time when cells are resistant to chemotherapy. Math Biosci 1992;110:201–19.
8. Monro HC, Gaffney EA. Modelling chemotherapy resistance in palliation and failed cure. J Theor Biol 2009;257:292–302.
9. Gatenby RA, Silva AS, Gillies RJ, Frieden BR. Adaptive therapy. Cancer Res 2009; 69:4894–903.
10. Hansen E, Woods RJ, Read AF. How to use a chemotherapeutic agent when resistance to it threatens the patient. PLoS Biol 2017;15:e2001110.
11. Gatenby RA. A change of strategy in the war on cancer. Nature 2009;459: 508–9.
12. Gatenby RA, Brown JS. The evolution and ecology of resistance in cancer therapy. Cold Spring Harbor perspectives in medicine. 2020;10:a040972.
13. Gillies RJ, Verduzco D, Gatenby RA. Evolutionary dynamics of carcinogenesis and why targeted therapy does not work. Nat Rev Cancer 2012;12:487–93.
14. Enriquez-Navas PM, Kam Y, Das T, Hassan S, Silva A, Foroutan P, et al. Exploiting evolutionary principles to prolong tumor control in preclinical models of breast cancer. Sci Transl Med 2016;8:327ra24.
15. Wang J, Zhang Y, Liu X, Liu H. Optimizing adaptive therapy based on the reachability to tumor resistant subpopulation. Cancers 2021;13:5262.
16. Smalley I, Kim E, Li J, Spence P, Wyatt CJ, Eroglu Z, et al. Leveraging transcriptional dynamics to improve BRAF inhibitor responses in melanoma. EBioMedicine 2019;48:17890.
17. Zhang J, Cunningham JJ, Brown JS, Gatenby RA. Integrating evolutionary dynamics into treatment of metastatic castrate-resistant prostate cancer. Nat Commun 2017;8:1816.
18. 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.
19. Fernández-Cancio M, Camats N, Flück C, Zalewski A, Dick B, Frey B, et al. Mechanism of the dual activities of human CYP17A1 and binding to anti-prostate cancer drug abiraterone revealed by a novel V366M mutation causing 17, 20 lyase deficiency. Pharmaceuticals 2018;11:37.
20. Therasse P, Arbuck SG, Eisenhauer EA, Wanders J, Kaplan RS, Rubinstein L, et al. New guidelines to evaluate the response to treatment in solid tumors. J Natl Cancer Inst 2000;92:205–16.
21. Schröder FH, van der Cruijsen-Koeter I, de Koning HJ, Vis AN, Hoedemaeker RF, Kranse R. Prostate cancer detection at low prostate specific antigen. J Urol 2000;163:806–12.
22. Lieberman R. Evidence-based medical perspectives: the evolving role of PSA for early detection, monitoring of treatment response, and as a surrogate end point of efficacy for interventions in men with different clinical risk states for the prevention and progression of prostate cancer. Am J Ther 2004;11:501–6.
23. Strobl MAR, West J, Viossat Y, Damaghi M, Robertson-Tessi M, Brown JS, et al. Turnover modulates the need for a cost of resistance in adaptive therapy. Cancer Res 2021;81:1135–47.
24. Hansen E, Read AF. Modifying adaptive therapy to enhance competitive suppression. Cancers 2020;12:1–13.
25. Viossat Y, Noble R. A theoretical analysis of tumour containment. Nat Ecol Evol 2021;5:826–35.
26. Kim E, Brown JS, Eroglu Z, Anderson ARA. Adaptive therapy for metastatic melanoma: predictions from patient calibrated mathematical models. Cancers 2021;13:823.
27. Brady-Nicholls R, Enderling H. Range-bounded adaptive therapy in metastatic prostate cancer. Cancers 2022;14:5319.
28. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44.
29. Munir K, Elahi H, Ayub A, Frezza F, Rizzi A. Cancer diagnosis using deep learning: a bibliographic review. Cancers 2019;11:1235.

30. Silver D, Huang A, Maddison CJ, Guez A, Sifre L, van den Driessche G, et al. Mastering the game of Go with deep neural networks and tree search. Nature 2016;529:484–9.
31. Bachute MR, Subhedar JM. Autonomous driving architectures: insights of machine learning and deep learning algorithms. Machine Learning with Applications 2021;6:100164.
32. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nat Med 2019;25:24–9.
33. Liu Y, Logan B, Liu N, Xu Z, Tang J, Wang Y. Deep reinforcement learning for dynamic treatment regimes on medical registry data. In: 2017 IEEE International Conference on Healthcare Informatics (ICHI). IEEE; 2017. p. 380–5.
34. Engelhardt D. Dynamic control of stochastic evolution: a deep reinforcement learning approach to adaptively targeting emergent drug resistance. J Mach Learn Res 2020;21:1–30.
35. Weaver DT, King ES, Maltas J, Scott JG. Reinforcement learning informs optimal treatment strategies to limit antibiotic resistance. Proc Natl Acad Sci U S A 2024;121:e2303165121. doi: 10.1073/pnas.2303165121.
36. Eckardt JN, Wendt K, Bornhäuser M, Middeke JM. Reinforcement learning for precision oncology. Cancers 2021;13:4624.
37. Yu C, Liu J, Nemati S, Yin G. Reinforcement learning in healthcare: a survey. ACM Computing Surveys 2023;55:1–36.
38. Zhao Y, Kosorok MR, Zeng D. Reinforcement learning design for cancer clinical trials. Stat Med 2009;28:3294–315.
39. Maier C, Hartung N, Kloft C, Huisinga W, Wiljes J. Reinforcement learning and bayesian data assimilation for model-informed precision dosing in oncology. CPT Pharmacometrics Syst Pharmacol 2021;10:241–54.
40. 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. Cancer 2006;107:389–95.
41. Gallaher JA, Enriquez-Navas PM, Luddy KA, Gatenby RA, Anderson ARA. Spatial heterogeneity and evolutionary dynamics modulate time to recurrence in continuous and adaptive cancer therapies. Cancer Res 2018;78:2127–39.
42. Mnih V, Badia AP, Mirza M, Graves A, Lillicrap T, Harley T, et al. Asynchronous methods for deep reinforcement learning. In: Balcan MF, Weinberger KQ, editors. Proceedings of the 33 International Conference on Machine Learning. vol. 48 of Proceedings of Machine Learning Research. New York, New York, USA: PMLR; 2016. p.1928–37. Available from: https://proceedings.mlr.press/v48/mniha16.html.
43. Gao Y, Liu Y, Zhang H, Li Z, Zhu Y, Lin H, et al. Estimating GPU memory consumption of deep learning models. In: Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering. New York, NY: Association for Computing Machinery; 2020. p. 1342–52.
44. Lu Y, Chu Q, Li Z, Wang M, Gatenby R, Zhang Q. Deep reinforcement learning identifies personalized intermittent androgen deprivation therapy for prostate cancer. Brief Bioinform 2024;25:bbae071. doi: 10.1093/bib/bbae071.
45. Brady-Nicholls R, Nagy JD, Gerke TA, Zhang T, Wang AZ, Zhang J, et al. Prostate-specific antigen dynamics predict individual responses to intermittent androgen deprivation. Nat Commun 2020;11:1750.
46. Strobl MAR, Gallaher J, Robertson-Tessi M, West J, Anderson ARA. Treatment of evolving cancers will require dynamic decision support. Ann Oncol 2023;34:867–84.
47. Strobl MAR, Gallaher J, West J, Robertson-Tessi M, Maini PK, Anderson ARA. Spatial structure impacts adaptive therapy by shaping intra-tumoral competition. Communications Medicine 2022;2:46.
48. Bacevic K, Noble R, Soffar A, Ammar OW, Boszonyik B, Prieto S, et al. Spatial competition constrains resistance to targeted cancer therapy. Nat Commun 2017;8:1995.
49. Gallaher J, Strobl M, West J, Gatenby R, Zhang J, Robertson-Tessi M, et al. Intermetastatic and intrametastatic heterogeneity shapes adaptive therapy cycling dynamics. cancer research. 2023;83:2775–89.
50. Kim E, Rebecca VW, Smalley KSM, Anderson ARA. Phase I trials in melanoma: a framework to translate preclinical findings to the clinic. Eur J Cancer 2016;67:213–22.