Mathematical Model-Driven Deep Learning Enables Personalized Adaptive Therapy

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

Jan 2023

S1 Virtual Patient Parameters

The parameters used in the virtual patient model are taken from Strobl et al. [1] and outlined in Table S1.

Name	Description	Value/Range	Reference
r_S	Sensitive cell proliferation rate	$0.027\,day^{-1}$	Adopted from [2]
r_R	Resistant cell proliferation rate	$0.5r_S - 1.0r_S$	Lower limit [3], upper limit of no cost
d_S, d_R	Natural cell death rate	$0.0r_S - 0.5r_S$	Lower limit given by zero turnover Upper limit adopted from [4]
d_D	Drug-induced cell killing	1.5	Adopted from [5]
N_0	Initial tumor cell density	$0.1 - 0.75$	Values in this range reported by [6]
R_0	Initial resistant cell fraction	$0.001N_0 - 0.1N_0$	Values in this range reported by [7]

Table S1: Parameter values/ranges used for the virtual patient.

For the single patient case considered in Sections 3.1 - 3.3, we consider the case where there is no cost associated with drug resistance, such that $r_R = r_S$. Combined with zero natural death rates d_S, d_R, these parameters replicate a treatment-resistant tumor that is common in late-treatment settings, where conventional treatment protocols typically struggle to contain the disease effectively. Finally we consider a tumor with initial size $N_0 = 0.75$ and initial resistant population $R_0 = 0.01$.

S2 Deep Learning Methods

This section provides further exposition and psuedocode implementations to supplement the explanation of the DRL model in Section 2.3.

We transform cancer treatment into a reinforcement learning problem, in which a computational agent learns to make decisions in an interactive environment by trial and error based on a reward function, allowing the agent to learn from unstructured input data [8]. We construct this as a 'model-free' problem, where information about the dynamics of the environment is not given explicitly, but instead inferred through interactions with the environment. We adopt an Actor-Critic method, which combines the advantages of policy and value-based methods, whereby an 'Actor' updates the policy distribution according to a value function estimated by the 'Critic' [9]. An illustration of this process is given in pseudocode in Algorithm 1.

Note that Operation 5 utilises a baseline comparison for the value of the cumulative reward; this results in a smaller absolute value, which reduces the error in gradient-based updates. This

[1] This uses the standard form for the policy gradient [10].

Algorithm 1 Actor-Critic Method

```
Require: Parameters: $\alpha$ (learning rate), $\gamma$ (discount factor)
Require: Initial values: state $s$, policy parameters $\theta$, value $w$ and action $a$.

for $t \leftarrow 1...T$ do
    Calculate reward $r$ from reward function $R(s, a)$
    Compute next state $s'$ based on previous $P(s'|s, a)$
    Sample next action $a'$ according to policy $\pi_\theta(a'|s')$
    $\theta \leftarrow \theta + \alpha Q_w(s, a) \nabla_\theta \log \pi_\theta(a|s)$    ▷ Update policy parameters[1]
    $\delta = r + \gamma Q_w(s', a') - Q_w(s, a)$    ▷ Compute correction for action value
    $w \leftarrow w + \alpha \delta \nabla_w Q_w(s, a)$    ▷ Update parameters $w$ of value function $Q_w$
    $a \leftarrow a'$; $s \leftarrow s'$
end for
```

choice of baseline is taken from Q Actor-Critic [11], but it is not the only option; a popular alternative is advantage Actor-Critic [12]. Here the baseline $A_w(s, a)\nabla_\theta \log \pi_\theta(a|s)$ is used, where the advantage value $A_w(s, a)$ is given by:

$$A_w(s, a) = Q_w(s, a) - V(s), \tag{1}$$

i.e.,, the added benefit from taking the given action a from state s compared to the expected value V (based on all actions) from state s.

As with all GPU-based deep-learning algorithms, this typically has a high computational cost, and requires specialist architecture to train performant models. We therefore utilize the asynchronous, advantage Actor-Critic (A3C) framework pioneered by Mnih et al. [13] which combines the Advantage Actor-Critic method with a lightweight CPU framework supporting parallel actor-learning asynchronously. The policy and value functions are not updated every timestep and share a convolutional neural network framework with separate softmax and linear outputs for the policy and value respectively. Since the threads are asynchronous, they update from the master policy at different times, and so follow slightly different policies exploring different regions of the environment. A simplistic pseudocode representation of each actor-learner thread is given in Algorithm 2, adapted from [13].

Within this, each worker network is constructed as follows:

1. Input Layers
    - Long Short-Term Memory layer [14] gives 4-dimensional output.
2. Hidden Layers
    - Fully connected layers for each of the sizes: 128, 64, 32, 16, 10.
    - Each layer is multiplied by the previous output to produce a tensor of hidden units.
    - Use a rectified linear activation function.
3. Output Layers
    (a) Policy - Fully connected layer of output size 2, softmax activation function [15].
    (b) Value - Fully connected layer of output size 1, linear activation function.

Note that the output size of the policy is determined by the number of policy options available; this assumes a binary treatment decision (i.e.,, treatment is either given or withheld).

As outlined in Algorithm 2, a reward function is calculated at each step, and is used to weight the update of each thread to the global network. The full reward function applied when training the DRL framework is given in Algorithm 3, based on the parameters in Table 1:

Algorithm 2 Asynchronous Advantage Actor-Critic –adapted from [13]

Require: Parameters: t_{max} (update rate), T_{max} (iteration number), γ (discount factor)
Require: Global shared variables: θ, θ_v (parameter vectors), T (counter)
Require: Thread-specific variables: θ, θ_v (parameter vectors)

```
$t \leftarrow 1$                                                    ▷ Initialise thread step counter
while $T < T_{max}$ do
    Reset gradients $d\theta \leftarrow 0$; $d\theta_v \leftarrow 0$
    Synchronise with global parameters $\theta' \leftarrow \theta$; $\theta'_v \leftarrow \theta_v$
    $t_{start} = t$
    Obtain state $s_t$
    while $t - t_{start} < t_{max}$ do
        Perform action $a_t$ according to policy $\pi(a_t|s_t;\theta')$
        Receive reward $r_t$ and new state $s_{t+1}$
        $t \leftarrow t + 1$
        $T \leftarrow T + 1$
    end while
    $R = V(s_t, \theta'_v)$                                          ▷ Bootstrap from last state
    for $i \in \{t-1, ..., t_{start}\}$ do
        $R \leftarrow r_i + \gamma R$
        $d\theta \leftarrow d\theta + \nabla_{\theta'} \log \pi(a_i|s_i;\theta')(R - V(s_i;\theta'_v))$   ▷ Accumulate gradients wrt $\theta'$
        $d\theta_v \leftarrow d\theta_v + \partial(R - V(s_i;\theta'_v))^2/\partial\theta'_v$
    end for
    Asynchronous update of $\theta$ using $d\theta$ and of $\theta_v$ using $d\theta_v$
end while
```

Algorithm 3 Reward Function used for DRL

Require: Reward Parameters: *base*, *holiday*, *progression*, *survival*
Require: State Variables: n (tumor size), n_0 (initial tumor size), t (time), d (last drug dose)

```
$r_t \leftarrow base$                        ▷ Base reward applied for surviving another timestep
if $n > 1.2n_0$ then                                ▷ If burden is 20% larger than baseline
    $r_t \leftarrow progression$
else if $t > 30 \times 365$ then                       ▷ 30 years denotes indefinite survival
    $r_t \leftarrow r_t + survival$
else if $d = 0$ then                         ▷ If previous treatment period was a drug holiday
    $r_t \leftarrow r_t + holiday$
end if
return $r_t$
```

S3 DRL Variability

In Section 3.1, we present results obtained by training the DRL model on a single patient. These are averaged over 100 evaluations, to account for the stochasticity in patient outcomes. This stochasticity is not a product of the virtual patient itself (whose treatment response is calculated by a deterministic set of differential equations), but rather inherent in the decision making process of the DRL framework, such that each iteration receives a slightly different treatment schedule.

We characterise this stochasticity in Figure S1, showing significant variation between individual evaluations of the patient, with the shortest TTP almost one-third of the longest, for the same patient profile. This variation means that this patient would have a significant probability of performing worse on the DRL strategy compared to AT50, even though the DRL framework has a mean TPP over 200 days greater. In Section 3.2, we subsequently demonstrate how variation in the performance of the DRL model can be reduced by increasing the interval between treatment decisions, enabling the DRL framework to consistently outperform AT50.

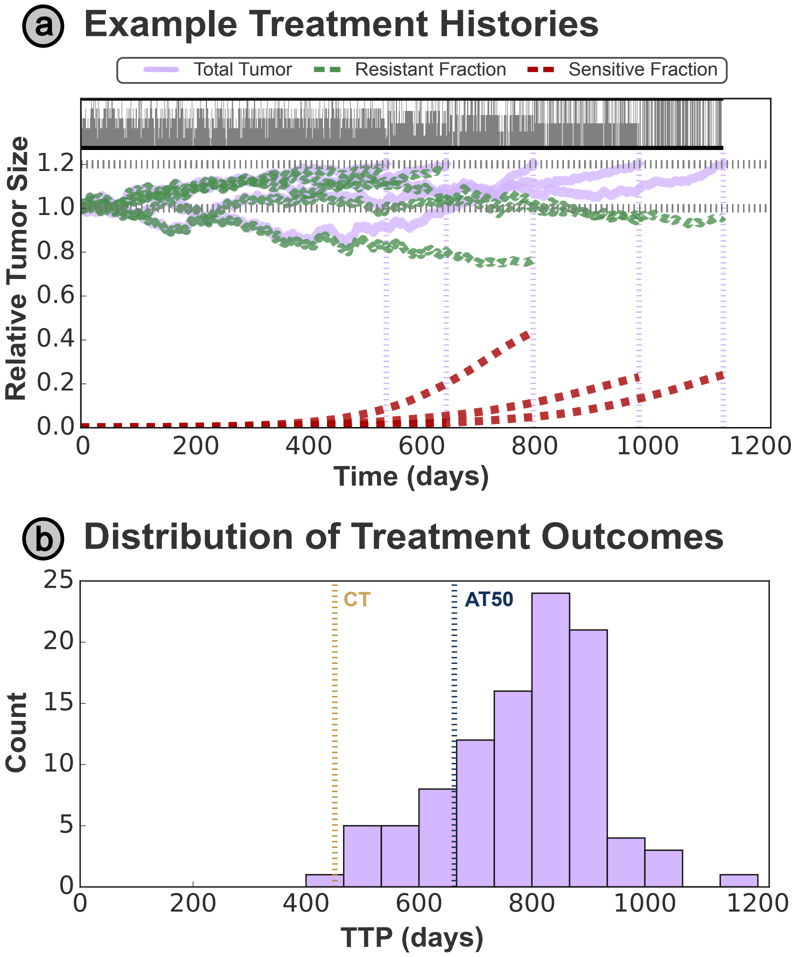

Figure S1: **(a)** Variation in treatment outcomes due to stochasticity in the DRL's decision-making process, evaluated on 5 copies of the virtual patient. **(b)** The distribution of TTP for 100 evaluations, consistently outperforming the predicted TTP for both continuous and adaptive therapy.

S4 Bruchovsky Trial Patients

To replicate the tumor dynamics observed in the clinic, we consider patient data from a prospective Phase II trial of intermittent androgen suppression for locally advanced prostate cancer, conducted by Bruchovsky et al. [16]. This trial used an intermittent strategy that subsequently inspired adaptive therapy to treat biochemical recurrence after radiotherapy, with a continuous lead-in treatment period before allocating treatment holidays according to the patient's PSA dynamics. Enrollment to the study required a rising serum PSA level after the patient received radiotherapy, and no evidence of distant metastasis. Each treatment cycle consisted of 4 weeks of cyproterone acetate given as lead-in therapy, followed by a combination of leuprolide acetate and cyproterone acetate, which ended after a total of 36 weeks. During each cycle, serum PSA and testosterone levels were recorded every 4 weeks. For the full study protocol, see [16].

We formed a cohort of virtual patients from these data by fitting the Lotka–Volterra model introduced in section 2.1 to each patient's treatment history (for details, see Strobl et al. [1]). Patients who developed a metastasis were excluded, to avoid potentially confounding effects from a variation in lesion number. Specifically, we focused on those 12 patients who progressed during the trial, to test whether DRL-informed AT could have improved their TTP. Simulating these parameter sets with the virtual patient model (Equations (1)), however, reveals that only 7 out of the 12 patients are expected to reach progression within 5000 days. This discrepancy can be partly attributed to a differing progression criterion used by Bruchovsky et al. [16], which was based on both PSA and testosterone rises under treatment and thus differed from the PSA-only criterion used here. The second reason is a limitation of the mathematical tumor model used in our study which can not provide a good representation of the treatment dynamics for some patients and instead converges to a spurious steady state solution, where a fully resistant tumor stabilises below the limit for progression (subsequently derived in Section S5.1). Given that the virtual patient model in these cases does not provide a meaningful benchmark we excluded these cases from our analysis. To sum up, we trained the DRL framework on the 7 virtual patients who are expected to reach progression within 5000 days, since they are most significantly impacted by the limitations of the current treatment paradigms.

S5 DRL Framework Sensitivity Analysis

In Section 3.5 we consider the robustness of the DRL framework to variation in patient parameters. There we highlight results from a complete sensitivity analysis, the full results of which are presented in Figure S2 below. The DRL framework was trained on Patient 25 from the Bruchovsky trial ($cost = 0.23, turnover = 0.29, N_0 = 0.42, f_R = 10^{-5}$), and the sensitivity analysis perturbed both the tumor's dynamics and initial composition from these initial values.

S5.1 Progression Analysis under Treatment

Within this sensitivity analysis, we can see there exists a region of parameter space where the Lotka–Volterra model (Section 2.1 - (1)) allows for a steady state which prevents progression. We will first derive this under continuous therapy (CT), where the sensitive population will rapidly be depleted to extinction, leaving a fully resistant population at the steady state. However this final resistant population has no dependence on the drug concentration, and so this steady state may be ultimately achieved under any treatment strategy.

Our original model for the virtual patient is reproduced below:

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

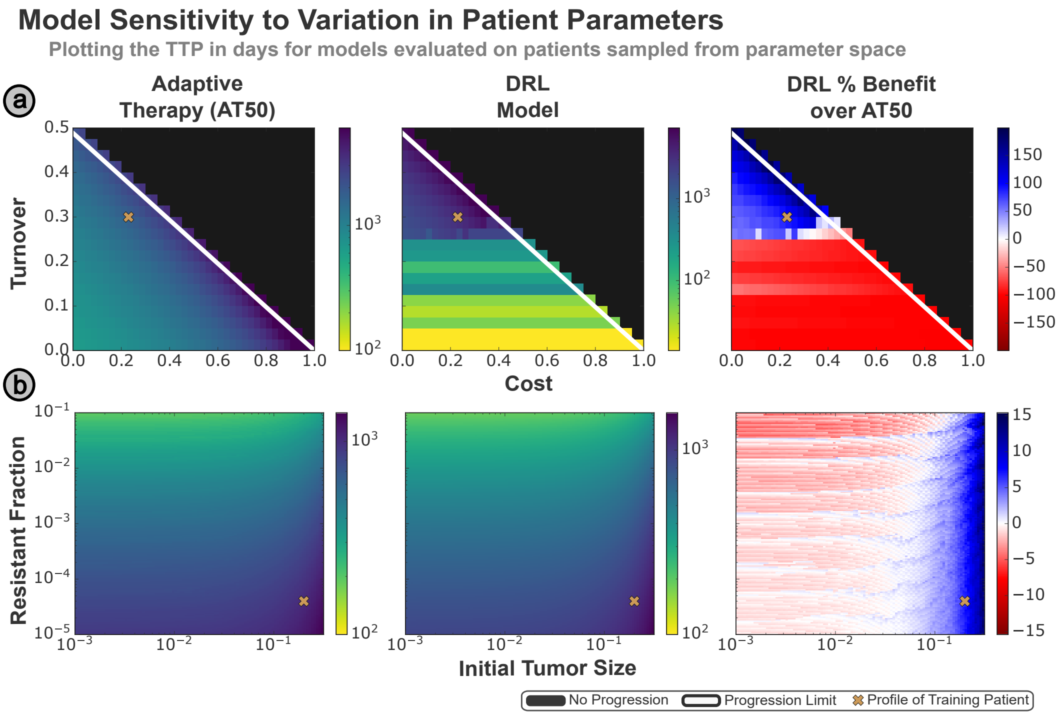

Figure S2: **(a)** The DRL framework is robust to variation in the cost of resistance, but highly sensitive to reduction in the cellular turnover below the value encountered in training, as the tumor grows faster than expected leading to premature progression while the tumor is still sensitive to treatment. **(b)** The DRL framework is less sensitive to changes in the initial tumor parameters, but under-performs AT50 for tumors smaller (i.e., further from carrying capacity) than encountered in training. These tumors grow faster during treatment holidays, and experience reduced competitive suppression.

Considering a fully resistant tumor in steady state $(S(t) = 0; \frac{dR}{dt} = 0)$ defines the resistant cell population R' according to the expression:

$$dR' = r_R R' \left(1 - \frac{R'}{K}\right), \tag{2}$$

which, upon rearrangement, gives the result for R':

$$R' = K \left(1 - \frac{d_R}{r_R}\right). \tag{3}$$

The resistant population reaches a non-zero, constant steady state provided that $r_R > d_r$. This will only result in progression when:

$$1.2(S_0 + R_0) < K \left(1 - \frac{d_R}{r_R}\right), \tag{4}$$

where S_0 and R_0 are the initial populations of sensitive and resistant cells respectively.

Patient profiles which obey this condition will therefore remain stable indefinitely under a CT strategy. Any patient who does not progress under CT will also not progress under the 'rule of thumb' AT strategy, and will either cycle indefinitely, or ultimately experience extinction of the sensitive population and converge to the fully resistant steady state derived above for CT.

S5.2 Progression without Treatment

It is important to note that these stable profiles are only stable under continuous treatment, and may still undergo progression under sub-optimal strategies. However this can only occur through insufficient treatment of the sensitive population, as the resistant population cannot result in progression alone. In this case, under the assumption that $S_0 >> R_0$, we may neglect the resistant population to derive an equivalent condition to (4). When no treatment is given, patients will only reach progression if:

$$1.2S_0 < K\left(1 - \frac{d_S}{r_S}\right). \tag{5}$$

Otherwise, tumors will maintain a non-zero steady state below the progression threshold, provided $r_S > d_S$ (else the tumor will be eliminated). Note that our formulation of the resistance cost requires that $r_S \geq r_R$, while we assume $d_S = d_R$ throughout. This also means that (5) is inherently stricter than (4), i.e., patients who do not progress under continuous treatment may still progress without treatment. It is worth noting that neither requirement depends on the initial resistant fraction, but only compares the initial tumor size to its growth and death rates.

S5.3 Cost-Turnover Space

We may reframe condition (4) for progression under treatment in terms of cost-turnover parameter space. The cost is characterized by the relative proliferation rates for sensitive and resistant cells $(1 - r_R/r_S)$, while the cell turnover represents the natural death rate of cells d/r_s. Rewriting these in terms of the resistant tumor properties:

$$r_R = r_S(1 - \mathit{cost}); \qquad d_r = r_s \times \mathit{turnover}, \tag{6}$$

we may write (4) as:

$$1.2(S_0 + R_0) < K\left(1 - \frac{\mathit{turnover}}{1 - \mathit{cost}}\right). \tag{7}$$

In cost-turnover space, we only observe progression if:

$$\mathit{turnover} < \left(1 - 1.2\frac{S_0 + R_0}{K}\right)(1 - \mathit{cost}). \tag{8}$$

This line is plotted for reference in Figure 5c. Naturally, progression cannot occur for $\mathit{cost} = 1$, as this would correspond to a zero proliferation rate for the resistant cells. More interestingly, it is possible to avoid progression even without a resistance cost, provided the carrying capacity is sufficiently small to restrict the logistic growth of the system.

S6 DRL Robustness to Model Variation

Section 3.5 also explores the robustness of the DRL framework to variation in patient dynamics. Through this, we introduce a modified Lotka– Volterra model introduced by Lu et al. [17], to demonstrate that a pre-trained DRL network can adapt to changes in the underlying tumor dynamics. Explicitly, this model may be written (in non-dimensional form) as:

$$\begin{aligned}
\frac{dS}{dt} &= r_S S\left[1 - \left(\frac{S + \frac{R}{1+e^{\gamma t}}}{K_S}\right)^{\alpha} - d_S D\right], \\
\frac{dR}{dt} &= r_R R\left[1 - \left(\frac{R + \frac{S}{1+e^{\gamma t}}}{K_R}\right)^{\alpha} - d_R D\right],
\end{aligned} \tag{9}$$

where S and R are the sensitive and resistant cell sub-populations respectively, and D is the drug concentration.

For this model, progression was defined as growth in the resistant population alone to $0.1K_R$. The model was parameterized according to Table S2, replicating values used by Lu et al. [17], and chosen to ensure that the profile in question does reach progression.

Name	Description	Value
r_S	Sensitive cell proliferation rate	$0.01365\, day^{-1}$
r_R	Resistant cell proliferation rate	$0.00825\, day^{-1}$
K_S	Carrying capacity for sensitive cells	1.0
K_R	Carrying capacity for resistant cells	0.25
d_S	Drug-induced sensitive cell killing	2.3205
d_R	Drug-induced resistant cell killing	1.3205
S_0	Initial sensitive cell fraction	0.75
R_0	Initial resistant cell fraction	0.01
α	Growth scaling term	1.0
γ	Relative competition	$0.27385\, day^{-1}$

Table S2: Parameter values used for the alternative virtual patient model, taken from Lu et al. [17].

Evaluating a pre-trained DRL framework on this new model, it attained a TTP of 1506 ± 3 days, outperforming the AT50 TTP of 1119 days (Figure S3). We also verify that this benefit is retained across a range of different tumor dynamics, through variation in the growth scaling term (α), with an increase in TTP of 305 days for $\alpha = 0.5$, and 576 days for $\alpha = 2$.

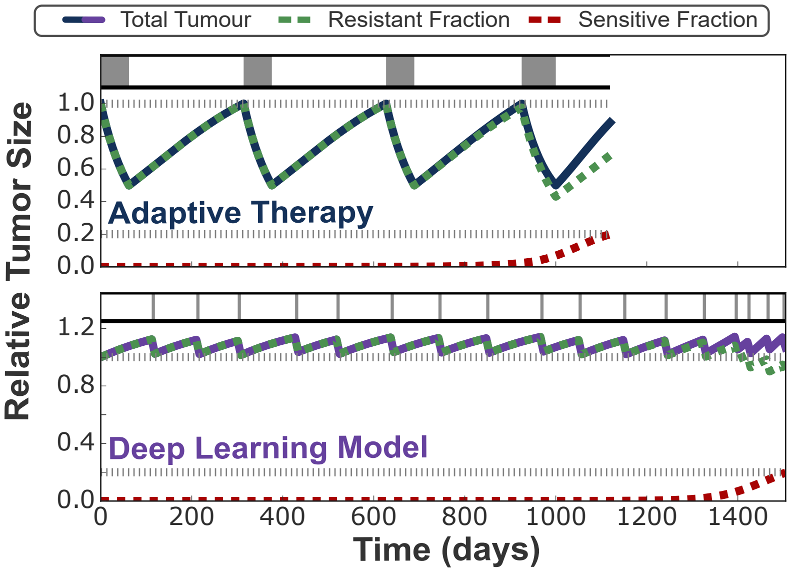

Figure S3: The DRL framework may also adapt to different underlying tumor dynamics, demonstrated by evaluating a pre-trained DRL network on an alternative tumor model (developed by Lu et al. [17]). This model has modified dynamics (achieved through use of exponential growth terms), and the progression criterion is also modified to depend on the resistant subpopulation alone. Despite this, the DRL framework uses similar treatment principles to consistently outperform AT.

In contrast to the previous models, which categorise cells by drug-response, we also considered an alternate model from Brady-Nicholls et al. [18], which differentiated between prostate cancer stem-like (S) and differentiated (D) cells to model tumor response to intermittent androgen deprivation therapy. Stem-like cells divide with per capita rate λ to produce two stem-like cells with probability p_s (symmetric division), with negative feedback $\frac{S}{S+D}$ from differentiated cells, or a stem-like and a non-stem cell (asymmetric division). Unlike androgen-independent stem-like cells, differentiated cells die in response to androgen removal at rate α. Treatment on and off cycles are described with parameter T_x, where $T_x = 1$ when treatment is given, and $T_x = 0$ during holidays.

$$\begin{aligned} \frac{dS}{dt} &= \left(\frac{S}{S+D}\right) p_S \lambda S \\ \frac{dD}{dt} &= \left(1 - \frac{S}{S+D} p_S\right) \lambda S - \alpha T_x D. \end{aligned} \tag{10}$$

The model was parameterized according to Table S3, replicating the values used by Brady-Nicholls et al. [18], with the given patient profile chosen to ensure that the tumor reaches progression. Evaluating the same pre-trained DRL framework on this new model, it attained a TTP of 2841 ± 102 days, outperforming the AT50 TTP of 2123 days.

Name	Description	Value
λ	Stem-like cell proliferation rate	$0.69\, day^{-1}$
p_S	Symmetric division probability	10^{-6}
α	Drug-induced sensitive cell killing	$0.07\, day^{-1}$
S_0	Initial stem-like cell population	10
D_0	Initial differentiated cell population	1000

Table S3: Parameter values used for the stem cell virtual patient model, taken from Brady-Nicholls et al. [18].

S7 Group-Trained DRL Networks

Section 3.5 introduces a DRL model trained on the Bruchovsky patient cohort, demonstrating that such ‘group-trained’ models have increased robustness to variation in tumor dynamics, at the cost of specialisation to a single tumor profile. To supplement this Section, we additionally trained a separate DRL framework on a completely separate set of synthetic profiles, randomly sampled from an enclosed region of parameter space (Figure S4a).

This group acts as an independent validation dataset, demonstrating that such generalized DRL networks are performant without having encountered the test patients during training. For six out of the seven patients, this achieves almost identical TTP as the generalist model trained on the Bruchovsky dataset (Figure S4b), as well as matching or out-performing AT50.

Notably however, this method fails for Patient 12, which in Figure S4a has a significantly reduced turnover compared to the other patients in both the Bruchovsky cohort and the synthetic training space. As demonstrated in Section S5, the DRL framework fails for profiles with a lower turnover than encountered in training. In fact, each DRL model here is only performant for patients with greater or equal turnovers to the lowest value encountered during training - this lowest value for the synthetic group excluded Patient 12 and so the network trained on this group fails to treat that patient.

More specifically, the threshold size in the treatment strategy must be reduced during retraining to account for the lowest turnover, shifting the treatment strategy curve right (Figure S4c) to be more conservative. This accounts for the reduction in TTP for other patients, as the sensitive cell population is maintained at a lower level, reducing resistant suppression. This contributes towards a wider discussion of specialist vs generalist models, and the inherent trade-off between model robustness and optimal performance (i.e. how well a model adapts to new parameter values, against its performance on the training set).

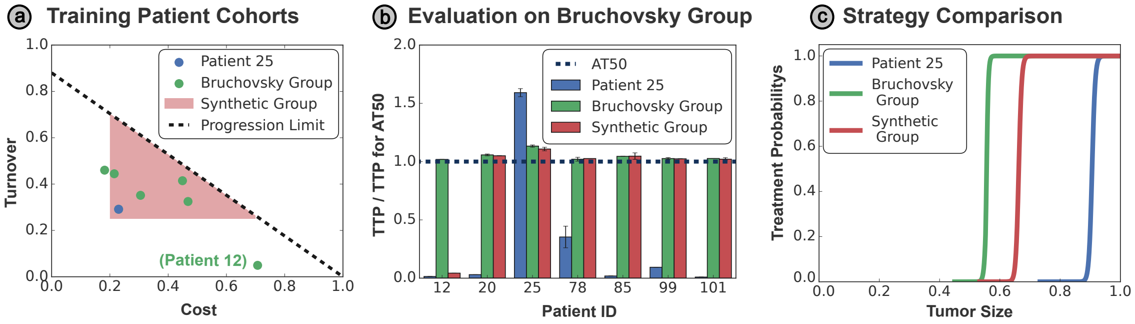

Figure S4: **(a)** A synthetic training cohort was generated by sampling a subset of parameter space, representing patients with the same initial tumor but different underlying dynamics. **(b)** Training the DRL network on this independent cohort achieves comparable performance to the framework trained directly on the Bruchovsky cohort for most patients. The framework fails to treat Patient 12, however, due to their exceptionally low turnover relative to patients encountered in training. **(c)** This difference manifests in the treatment strategies, where the framework trained on the synthetic group has a higher treatment threshold than the completely generalist framework trained on the Bruchovksy group, but below that of the completely specialized framework trained on a single patient.

We can also consider this through the sensitivity of the DRL framework to variation in the decision interval τ. This is particularly pertinent for the clinical translation of this work, as medical appointments and clinical tests are frequently delayed for reasons beyond a clinician's control. This dependence has already been explored somewhat in Section 3.2, where we show that the framework is robust to reductions in τ, but may fail under increases to τ (corresponding to delayed treatment). Exploring this systematically for the DRL framework trained on Patient 25 above (with $\tau = 30$ days), we verify that there is no loss of performance (i.e. a $> 20\%$ reduction in TTP) for $\tau < 30$ days, but there is a loss in performance for $t > 42$ days, as premature progression occurs due to the unanticipated increase in decision interval (Figure S5a). The networks trained on the Synthetic and Bruchovsky patient groups only experienced a loss in performance after 81 and 97 days, respectively. These group-trained networks both demonstrated greater robustness to the variation in τ, and significantly outperformed the network trained on Patient 25 for $\tau > 40$ days ($p < 0.01$), despite having a lower TTP at the training value of τ.

We also consider random variation in decision interval, to replicate the practical realities of clinical scheduling where such delays to treatment are non-uniform. We sample the delay for each decision point from an exponential distribution with a mean of μ days. We find a significant loss of performance on the Patient 25 network for as little as $\mu = 5$ days, reflecting its highly optimized state to the training problem (Figure S5b). This sensitivity to μ may be attributed to the random nature of this sampling - while increasing τ uniformly increases time both on- and off-treatment, such that increased tumor growth off-treatment is compensated by increased suppression on-treatment, no such compensation occurs consistently for every treatment cycle with randomly sampled decision-intervals. This results in premature progression over the course of many treatment cycles. By contrast, group-trained networks based on the Synthetic and Bruchovsky patient groups demonstrated greater robustness to the variation in μ, and with significant losses in performance after only 18 and 28 days, respectively (Figure S5b). In summary, models trained on a wider region of parameter space display greater robustness

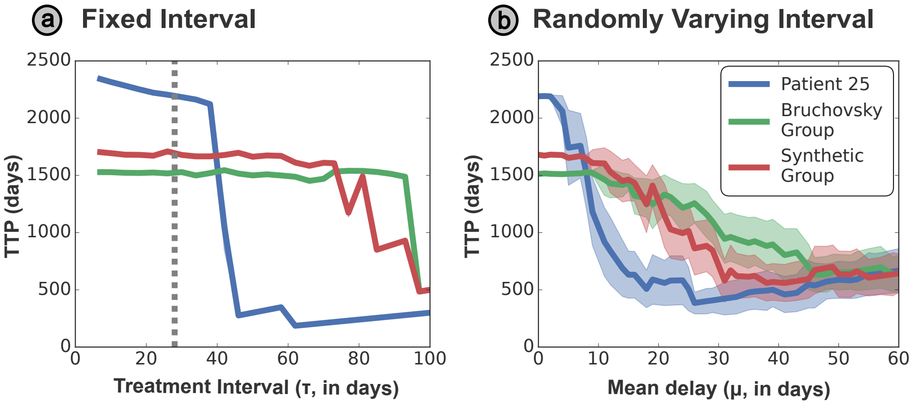

Figure S5: **(a)** TTP for each model, when the treatment interval is varied from the value ($\tau = 30$ days, in gray) used in training. While the model trained on patient 25 outperforms the others at $\tau = 30$, it is much less robust to increases in τ, while the Bruchovsky model demonstrates the greatest robustness along with the lowest TTP at $\tau = 30$. **(b)** To replicate unexpected delays to treatment, a delay is added to each treatment decision, randomly sampled from an exponential distribution with mean μ. Again the Patient 25 model, with the best performance on the training conditions (where $\mu = 0$), is the most susceptible to increases in μ, while the more generalist group-trained models have an increased capacity to cope with this stochasticity in the treatment interval.

(including to variation in parameters that were not varied in training, such as τ), however this comes at a cost of reduced maximal performance on a single patient.

S8 First Cycle Fitting

In Section 3.7, we fit the virtual patient model (1) to the first cycle of adaptive therapy only, following the protocol established by Strobl et al. [1].

However, the dynamics in this first cycle are almost completely determined by the sensitive sub-population alone, with minimal dependence on the resistant sub-population provided its initial size is sufficiently small. For this reason, it is only possible to accurately infer parameters characterized by the sensitive population from the initial cycle data and we decided to only fit for initial tumor size and cellular turnover (hereafter referred to as the sensitive parameters), while fixing the cost of resistance and the initial resistant fraction at average values determined by fitting full patient histories of prior patient cohorts.

While this can limit the accuracy of these fits at later times, as demonstrated in the final two panels of Figure S6 (Patients 99 and 101), this does not impact the ability of the DRL framework to learn effective strategies for a particular patient, provided the estimated values for the sensitive parameters are accurate. As demonstrated in Section 3.5, the effectiveness of the DRL framework is primarily determined by cellular turnover and is robust to significant variation in the cost of resistance.

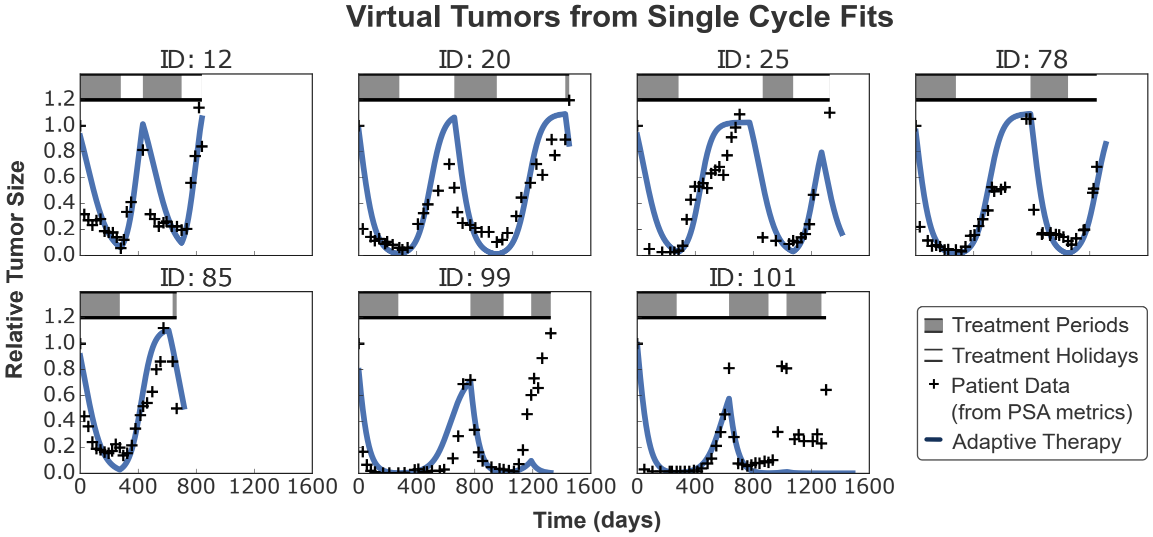

Figure S6: Virtual tumor models based on a fit to the first cycle of patient data only, superimposed over the entire patient history. While the tumor model struggles in some cases to replicate the late time dynamics of the patient, these are dominated by resistant sub-populations within the tumor which do not affect the determination of an optimal treatment strategy for that individual.

References

[1] M. A. Strobl, J. West, Y. Viossat, M. Damaghi, M. Robertson-Tessi, J. S. Brown, R. A. Gatenby, P. K. Maini, and A. R. Anderson, "Turnover modulates the need for a cost of resistance in adaptive therapy," *Cancer Research*, vol. 81, pp. 1135–1147, Feb. 2021.

[2] J. Zhang, J. J. Cunningham, J. S. Brown, and R. A. Gatenby, "Integrating evolutionary dynamics into treatment of metastatic castrate-resistant prostate cancer," *Nature Communications*, vol. 8, Nov. 2017.

[3] J. A. Gallaher, P. M. Enriquez-Navas, K. A. Luddy, R. A. Gatenby, and A. R. Anderson, "Spatial heterogeneity and evolutionary dynamics modulate time to recurrence in continuous and adaptive cancer therapies," *Cancer Research*, vol. 78, pp. 2127–2139, Jan. 2018.

[4] E. Malaise, N. Chavaudra, and M. Tubiana, "The relationship between growth rate, labelling index and histological type of human solid tumours," *European Journal of Cancer (1965)*, vol. 9, pp. 305–312, Apr. 1973.

[5] J. B. West, M. N. Dinh, J. S. Brown, J. Zhang, A. R. Anderson, and R. A. Gatenby, "Multidrug cancer therapy in metastatic castrate-resistant prostate cancer: An evolution-based strategy," *Clinical Cancer Research*, vol. 25, pp. 4413–4421, Apr. 2019.

[6] S. Prokopiou, E. G. Moros, J. Poleszczuk, J. Caudell, J. F. Torres-Roca, K. Latifi, J. K. Lee, R. Myerson, L. B. Harrison, and H. Enderling, "A proliferation saturation index to predict radiation response and personalize radiotherapy fractionation," *Radiation Oncology*, vol. 10, July 2015.

[7] C. Grassberger, D. McClatchy, C. Geng, S. C. Kamran, F. Fintelmann, Y. E. Maruvka, Z. Piotrowska, H. Willers, L. V. Sequist, A. N. Hata, and H. Paganetti, "Patient-specific tumor growth trajectories determine persistent and resistant cancer cell populations during treatment with targeted therapies," *Cancer Research*, vol. 79, pp. 3776–3788, May 2019.

[8] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep reinforcement learning: A brief survey," *IEEE Signal Processing Magazine*, vol. 34, pp. 26–38, Nov. 2017.

[9] J. Peters and S. Schaal, "Natural actor-critic," *Neurocomputing*, vol. 71, pp. 1180–1190, Mar. 2008.

[10] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," in *Proceedings of the 31st International Conference on Machine Learning* (E. P. Xing and T. Jebara, eds.), vol. 32 of *Proceedings of Machine Learning Research*, (Bejing, China), pp. 387–395, PMLR, June 2014.

[11] R. H. Crites and A. G. Barto, "An actor/critic algorithm that is equivalent to q-learning," in *Proceedings of the 7th International Conference on Neural Information Processing Systems*, NIPS'94, (Cambridge, MA, USA), p. 401–408, MIT Press, 1994.

[12] T. Degris, P. M. Pilarski, and R. S. Sutton, "Model-free reinforcement learning with continuous action in practice," in *2012 American Control Conference (ACC)*, IEEE, June 2012.

[13] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in *Proceedings of The 33rd International Conference on Machine Learning* (M. F. Balcan and K. Q. Weinberger, eds.), vol. 48 of *Proceedings of Machine Learning Research*, (New York, New York, USA), pp. 1928–1937, PMLR, June 2016.

[14] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, pp. 1735–1780, Nov. 1997.

[15] J. S. Bridle, “Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition,” in *Neurocomputing*, pp. 227–236, Springer Berlin Heidelberg, 1990.

[16] N. Bruchovsky, L. Klotz, J. Crook, S. Malone, C. Ludgate, W. J. Morris, M. E. Gleave, and S. L. Goldenberg, “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*, vol. 107, no. 2, pp. 389–395, 2006.

[17] Y. Lu, Q. Chu, Z. Li, M. Wang, and Q. Zhang, “Deep reinforcement learning identifies personalized intermittent androgen deprivation therapy for prostate cancer,” May 2022.

[18] R. Brady-Nicholls, J. D. Nagy, T. A. Gerke, T. Zhang, A. Z. Wang, J. Zhang, R. A. Gatenby, and H. Enderling, “Prostate-specific antigen dynamics predict individual responses to intermittent androgen deprivation,” *Nature Communications*, vol. 11, Apr. 2020.