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

Yitao Lu (✉ yitao.lu@my.cityu.edu.hk )  
City University of Hong Kong  https://orcid.org/0000-0002-0523-7416

Qian Chu (✉ qianchu@tjh.tjmu.edu.cn )  
Tongji Hospital, Huazhong University of Science and Technology  https://orcid.org/0000-0001-8192-7630

Zhen Li (✉ zhenli@hust.edu.cn )  
Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology  https://orcid.org/0000-0001-8037-4245

Mengdi Wang (✉ mengdiw@princeton.edu )  
Princeton University

Robert Gatenby (✉ Robert.Gatenby@moffitt.org )  
Moffit Cancer Center

Qingpeng Zhang (✉ qingpeng.zhang@cityu.edu.hk )  
City University of Hong Kong  https://orcid.org/0000-0002-6819-0686

Article

Keywords: Prostate cancer, personalized medicine, reinforcement learning, adaptive therapy

DOI: https://doi.org/10.21203/rs.3.rs-1573462/v3

License: © This work is licensed under a Creative Commons Attribution 4.0 International License.  
Read Full License

Additional Declarations: There is NO Competing Interest.
Deep reinforcement learning identifies personalized intermittent androgen deprivation therapy for prostate cancer

Yitao LU¹, Qian CHU², Zheng LI³, Mengdi WANG⁴, Robert Gatenby⁵, Qingpeng ZHANG¹,6*

*For correspondence: qingpeng.zhang@cityu.edu.hk
Q.Z. initialized the idea and supervised the project. Y.L. and Q.Z. collected the data, developed the model, and wrote the initial version of the paper. R.G. collected the data and edited the writing. M.W., Q.C. and Z.L. analyzed the data and edited the writing.

Present address: ¹, 6Hong Kong SAR, China; ², ³Wuhan, China; ⁴Princeton, NJ, U.S.A.; ⁵Tampa, FL, U.S.A.

Abstract
The evolution of drug resistance leads to treatment failure and tumor progression. Intermittent androgen deprivation therapy (IADT) helps responsive cancer cells compete with resistant cancer cells in intratumoral competition. However, conventional IADT is population-based, ignoring the heterogeneity of patients and cancer. Additionally, existing IADT relies on pre-determined thresholds of prostate-specific antigen to pause and resume treatment, which is not optimized for individual patients. To address these challenges, we developed a time-varied, mixed-effect, and generative Lotka-Volterra (tM-GLV) model to account for the heterogeneity of the evolution mechanism and the pharmacokinetics of individual patients. Then, we proposed a reinforcement-learning-enabled individualized IADT framework, namely, I²ADT, to learn the patient-specific tumor dynamics and derive the optimal drug administration policy. Experiments with clinical trial data demonstrated that the proposed I²ADT can significantly prolong the time to progression of prostate cancer patients with reduced cumulative drug dosage. We further validated the efficacy of the proposed methods with a recent pilot clinical trial data. This research elucidates the application of deep reinforcement learning to identify personalized adaptive cancer therapy.

Introduction
Prostate tumor is the second most prevalent cancer and the sixth leading cause of cancer death worldwide Sung et al. (2021); Litwin and Tan (2017); Bite (2005). The common treatments of locally advanced prostate cancer are radiotherapy, and hormone therapy Smith et al. (2012); Shore (2020). Hormone therapy, such as androgen deprivation therapy (ADT), is an effective treatment and is usually applied after the failure of radiotherapy Gillies et al. (2012); Bruchovsky et al. (2006).
Similar to other hormone therapies, ADT has side effects, including decreased libido, impotence, hot flashes, and sexual effects Higano (2003); Sharifi et al. (2005).

The difficulty in treating prostate cancer lies in the development of resistance, which usually leads to treatment failure and tumor progression West et al. (2019); McGranahan and Swanton (2017). There are multi-type cancer cells competing for resources in the resource-limited tumor microenvironment. Such Darwinian dynamics can lead to a rapid proliferation of resistant population. In the conventional drug administration policy, the use of maximum tolerated dose until progression can give the resistant phenotype an advantage over the other competitors, leading to the faster development of tumor resistance West et al. (2019). Thus, the intermittent androgen deprivation therapy (IADT) was proposed and validated in several clinical trials Bruchovsky et al. (2006); Zhang et al. (2017).

Figure 1 illustrates the idea of ADT (a) and IADT (b) separately. If a maximum tolerated dose is adopted, the resistant phenotype has an evolutionary advantage over the responsive phenotypes, leading to quick tumor resistance to ADT (shown in Figure 1.a). IADT involves the intermittent administration of drugs, which gives the responsive phenotype the chance to compete with the resistant phenotype (shown in Figure 1.b), thus prolonging the time to progression (TTP). In addition, IADT provides quality-of-life benefits by reducing the cumulative drug dosage.

Figure 1. Illustration of androgen deprivation therapy (ADT) and intermittent androgen deprivation therapy (IADT) separately (Created with BioRender.com). The two panels, labeled (a) and (b), depict green cells as responsive cancer cells and red cells as resistant cancer cells. In (a), continuous dosing kills responsive cancer cells, but the resistant cancer cells quickly dominate the population due to the lack of competition from responsive cells. In contrast, (b) shows that intermittent dosing allows responsive cancer cells to regrow during therapy-free periods. This enables them to compete more effectively, resulting in the resistant cells being unable to dominate the entire population even in the late stages of treatment.

There are two potential design flaws in conventional IADTs. First, many begin with “induction treatment” in which ADT is applied at maximum dose continuously for 8 to 9 months and intermittent therapy is then applied only if the prostate-specific antigen (PSA) has been reduced to the normal range. Evolutionarily, this has the effect of strongly selecting for resistance while removing most of the sensitive population. Thus, the critical evolutionary competition that is necessary for suppression of the resistant cells by the expanding sensitive population during treatment cessation is lost. Second, IADTs often impose a rigid treatment schedule which neglects the heterogeneity within the tumor-host interactions which can result in very different proliferation and death rates of prostate cancer cells in different patients. Recent clinical trials and computer simulation studies have suggested an alternative approach with no induction treatment and using pre-define PSA thresholds for suspending and resuming ADT administration Crawford et al. (2019) may be more successful than the prior trial designs. However, even if such population-based IADTs are effec-
tive, they do not fully utilize the patients’ characteristics and clinical information. Thus, they are sub-optimal and the full benefit of IADT in personalized medicine has not been obtained.

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

We build upon prior applications and validations of various evolutionary mechanisms for simulating the intratumoral dynamics and IADT responses Zhang et al. (2017); Hirata et al. (2010); Baez and Kuang (2016); Brady-Nicholls et al. (2020). For example, the classic Lotka-Volterra model was incorporated into an evolutionary game model to simulate the competition mechanism between responsive and resistant tumors Zhang et al. (2017). Another recent study explained the resistance occurrence of prostate cancer Brady-Nicholls et al. (2020) in IADT by considering stem cells differentiation and evolution.

In addition, artificial intelligence (AI) techniques, particularly reinforcement learning, are promising tools for making optimal treatment decisions that consider different patients’ heterogeneity and tumor’s evolutionary mechanism Belkhir et al. (2021); Topol (2019); Zhang et al. (2019); Petersen et al. (2019); Gottesman et al. (2019); Engelhardt and Michor (2021). Recently, model-free reinforcement learning methods have been applied to the dynamic control of cancer. For example, an agent-based modeling with an associated reinforcement learning framework was proposed to continuously control the drug administration and dynamically regulate the emergence of resistant tumors Engelhardt (2020); this was a theoretical study that was conducted without patient customization or actual clinical data. Another study used reinforcement learning to inform the automated dose adaptation, which achieved human-similar results in non-small cell lung cancer patients Tseng et al. (2017). This work did not incorporate the intratumoral evolutionary mechanisms. In both studies, the pharmacokinetics of specific drugs was not considered.

To address the abovementioned challenges, we propose the reinforcement-learning-enabled individualized IADT (I²ADT) framework, which learns the patient-specific tumor dynamics from actual patient data and derives the optimal drug administration policy for individual patients using reinforcement learning. Experiments with a multi-center Phase II clinical trial data demonstrate that I²ADT leads to longer TTP and lower cumulative drug dosage compared with the conventional standard IADT adopted by the trial. Additional validation with external data also demonstrated the efficacy of I²ADT.

More specifically, we incorporated a non-linear, time-varied term to describe the dynamic competition advantage between two phenotypes as a result of selection pressure-induced molecular changes, as described in Methods (Modeling the PCaC environment) and Appendix 1.

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

Results

Mathematical modeling and simulation of prostate cancer cell evolution

The evolutionary dynamics of prostate cancer cells in vivo are complicated. Our model considered two phenotypes, namely responsive (Hormone-Dependent) and resistant (Hormone-Independent). In the beginning, responsive cancer cells dominate the population, and resistant cancer cells account for only a tiny portion due to the inherent heterogeneity and the healthier fitness of responsive phenotype.
According to biological theories *Isaacs and Coffey* (1981); *Shimada and Aihara* (2008), four key processes contribute to the evolution of cancer, namely, selection, competition, mutations and epigenetic modifications, and adaptation. Resistant cancer cells are minor, but they exist before the treatment and will take their place under the androgen suppression conditions. A fierce competition likely exists between the two phenotypes in the tumor microenvironment because of the high demand for resources in this niche *Chang et al.* (2015). The resistant phenotype can genetically or epigenetically gain advantages through mutations. We calibrated the model with a multi-center Phase II clinical trial by applying the standard IADT to the biochemical recurrence patients after irradiation with localized prostate cancer *Bruchovsky et al.* (2006). The longitudinal data of each patient was utilized for training the patient-specific mathematical model, as described in Methods (Modeling the PCaC environment).

![Figure 2](image)

**Figure 2.** *a)* The evolutionary dynamics of patient012 who evolved resistance to ADT based on the clinical trial with the application of standard IADT *Bruchovsky et al.* (2006). The upper panel is the fitted curve for the serum PSA level, which is plotted with the ground truth. The lower panel is the corresponding simulation of the dynamics of responsive and resistant phenotypes. We use simulations to predict the Prostate Cancer Competition (PCaC) environment with the standard IADT; when the resistant population exceeds 80% of its capacity, we ended the simulation (EOS, a grey background marked as "done"). The resistant phenotype escapes from the competition pressure in the 3-rd treatment cycle, thereby leading to drug resistance. *b)* The evolutionary dynamics of patient037 who did not evolve resistance to ADT based on the clinical trial involving the application of standard IADT. The upper panel is the fitted curve for the serum PSA level, which is plotted with the ground truth. The lower panel is the corresponding simulation of the dynamics of responsive and resistant phenotypes. The resistant phenotype increases in the later stage of the treatment for both patients. *c)*. The distributions of all the patient-specific parameters learned by gradient descent, the mean and the 95% CI are shown in Supplementary File 1. Validation of the mathematical models with the test data.

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

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

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

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

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

**Predict patients’ responses using the resistance index**

With the model at our disposal, we use the parameters to categorize patients as either resistant or responsive. Our results provide empirical evidence that the best predictive power for differentiating resistance from response comes from the parameter $\gamma$ (C-statistic=0.97, receiver operating characteristic curve depicted in Figure 2.c). For more results, please refer to Supplementary file 3 and Supplementary figure 4 in Appendix 10. Consequently, we designate $\gamma$ as the resistance index.

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

Using a false positive rate of 10% to establish the threshold, we classify patients into two groups. The resistance index demonstrates a TPR of 0.909 and an FPR of 0.061, with a threshold of 0.115. These findings support our selection of $\gamma$ as the resistance index in our model. This indicates that the resistance index $\gamma$, derived from actual data, captures the key intratumoral evolutionary characteristics of patients and can differentiate responsive patients from resistant ones. We also conducted extensive leave-pair-out cross-validations to further assess the robustness of our results (details can be found in Appendix 8).

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

The predefined thresholds for suspending and resuming treatment in the standard IADT are not personalized, leading to sub-optimal treatment outcomes. The proposed tM-GLV model captures individual patient’s intratumoral evolutionary dynamics with personalized parameterizations through the model fitting and validation. In this section, the optimal dosing policies are obtained through Proximal Policy Optimization (Schulman et al. 2017) for 11 resistant patients and 51 responsive patients.

I²ADT on resistance group

The reinforcement learning-derived I²ADT can significantly postpone resistant patients’ time-to-progression (TTP). Figure 4.a shows the dosing policies, treatment outcome of I²ADT in left panel, and the corresponding standard IADT on all resistant patients in right panel. Three major differences existed between I²ADT and standard IADT.

First, the average time of each treatment cycle reduces compared with the standard IADT, with treatment on: 5.0 months v.s. 13.6 months; treatment off: 8.7 months v.s. 8.9 months. With such
Figure 5. (a) The dynamics of the competition advantage of responsive cancer cells towards the resistant cancer cells in patient036 (upper panel) and all resistant patients with 95% CI (lower panel). (b) The PFS rate over time with I2ADT, IADT, and ADT. (c) The distribution of TTP with I2ADT, IADT, and ADT.

an adaptive dosing policy learned by reinforcement learning, the population of responsive cancer cells oscillated at a relatively high level before the occurrence of resistance, as presented in Figure 4b. The competitive advantage of responsive cancer cells also exhibits such an oscillating pattern, indicating that the proposed I2ADT could suppress resistant cancer cells by giving competition pressure to responsive cancer cells. The biphasic pattern commonly observed in IADT Tanaka et al. (2010), which also existed in our simulation, as shown in Figure 4c, was prevented in I2ADT by the shortening of the treatment-on period. The biphasic pattern indicates that after the treatment was turned on for a period of time, the effectiveness of continuous drug treatment declines, as reflected by the flattened slope of the patient's PSA curve.

Second, I2ADT resulted in pattern of declining treatment-on-time over time for the resistant population (refer to Appendix 9 and 10). In general, the responsive cancer cells were gradually losing the competition with resistant cancer cells. So, I2ADT helped responsive cancer cells re-grow by reducing the treatment-on time in each cycle. However, due to small number of resistance population, the decline pattern of dosage is not significant. Please refer to the Appendix 9 and 10 for additional analysis of the decline pattern of the dosing policies.

Thirdly, the I2ADT (intermittent androgen deprivation therapy) learned through reinforcement learning was dynamic and tailored to each patient's needs. In the initial stages of treatment, I2ADT provided a greater competitive advantage for responsive cancer cells over resistant ones, when compared to both IADT and conventional continuous ADT. This is illustrated in Figure 5a. As treatment progressed and intratumoral competition continued, the competitive advantage of responsive cancer cells gradually decreased to zero in both IADT and ADT. However, in I2ADT, a significant competitive advantage still persisted, allowing responsive cells to compete with the resistant cancer cells and ultimately prolonging the survival time of resistant patients.

To compare the effectiveness of I2ADT with IADT or ADT, we use two surrogates: time to progression (TTP) and progression-free survival (PFS). TTP is defined as the time at which the simulation reaches the end of simulation (EOS) for an individual patient. FPS refers to the time from the initiation of treatment to the occurrence of disease progression (EOS). The EOS is reached when either the resistant cancer cells account for 80% of their capacity or when the simulation reaches its maximum number of steps (120).

Simulation results demonstrate that by maintaining a higher competitive advantage during the early stage, I2ADT significantly prolonged TTP and PFS rates compared to standard IADT or ADT (p-value = 0.0019), as shown in Figures 5b&c. These results indicate that the adaptive dosing can be an effective strategy for delaying the onset of drug resistance and improving patient outcomes.

I2ADT on response group

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

Figure 6. (a) The dosing policies for patients in the resistance group. Each row denotes a responsive patient. The left and right bars present the I^2ADT and IADT outcomes, respectively. (b) and (c) present the evolutionary dynamics of PSA and cancer cells under I^2ADT and standard IADT in patient 106, respectively. In Figure (b), there are 13 treatment cycles for 120 months (9 months per cycle in average); while in Figure (c), there are 2.5 cycles for 62 months (25 months per cycle in average).

A distinct difference existed between the I^2ADT for responsive patients and that for resistant patients. In general, an ascending pattern of dosage was observed in the I^2ADT for most responsive patients by increasing the dosage and/or the treatment-on time in each cycle (as shown in Figure 6). The resistant cancer cells can be suppressed to a low level. Thus, I^2ADT tended to further reduce the overall tumor burden by killing more responsive cancer cells, so that the risk of other disease progression events, such as metastasis and comorbid conditions, could be further reduced. The Appendix and 9 and 10 present details of additional analysis of the ascending dosing policies.
Although the responsive patients did not develop drug resistance in the original clinical trial, \textsuperscript{1}I\textsuperscript{2}ADT further reduced the risk of drug resistance in the long run by maximizing the evolutionary competition between the responsive and resistant cancer cells. Taking Patient\textsuperscript{106} as an example, even if the resistant cancer cells were under control during the entire period of the clinical trial, and the PSA levels were successfully suppressed under $4\mu g/L$ during the treatment-on period, the population of resistant cancer cells was found to already be increasing. When we used the model to predict the future evolutionary dynamics for this patient, we found that drug resistance would emerge on day 1750 (EOS). Similar results were observed in most of the responsive patients (Figure 6.a). Drug resistance would emerge eventually after a sufficiently long period of time. By reinforcement learning, the proposed \textsuperscript{1}I\textsuperscript{2}ADT prolongs the TTP to 10 years (the maximum period of simulation) and increased the PFS rate significantly.

Similar simulation results demonstrate that by maintaining a higher competitive advantage during the early stage, \textsuperscript{1}I\textsuperscript{2}ADT significantly prolonged TTP and PFS rates compared to standard IADT or ADT (p-value = 0.001), as shown in Figures 7.b$c$. but also reduce much more drug dosage, improving life quality. These results indicate that the adaptive dosing can be an effective strategy for delaying the onset of drug resistance and improving patient outcomes.

**Dosage reduction**

Considering the inevitable adverse events of the treatment-on period of ADT \textit{Bruchovsky et al.} (2008), it is preferable to reduce the dosage as long as the disease is under control. We compared the deduction rate of averaged dosage of CPA, LEU by cycle and the overall treatment-on percentage with the standard IADT in Table (1).

<table>
<thead>
<tr>
<th>Items</th>
<th>Deduction rate Group resistance</th>
<th>response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ave. LEU (mg/Month)</td>
<td>87.7% ($10^{-15}$)</td>
<td>71.7% ($10^{-16}$)</td>
</tr>
<tr>
<td>Ave. CPA (mg/Day)</td>
<td>60.3% ($10^{-6}$)</td>
<td>43.4% ($10^{-13}$)</td>
</tr>
<tr>
<td>Ave. Treat-on Per cycle</td>
<td>27.1% (0.027)</td>
<td>40.3% ($10^{-15}$)</td>
</tr>
</tbody>
</table>

Table 1. The percentage of reduction of dosage and treatment-on course of \textsuperscript{1}I\textsuperscript{2}ADT compared with that of the standard IADT. The p-values of t-test are shown in the brackets.

A significant reduction of the dosage of both CPA and LEU and a reduced percentage of the treatment-on period in the treatments of \textsuperscript{1}I\textsuperscript{2}ADT were found, indicating that the proposed \textsuperscript{1}I\textsuperscript{2}ADT can reduce the risk of incidence of adverse events of treatment and enhance the quality of life of prostate cancer patients.

**Validation of \textsuperscript{1}I\textsuperscript{2}ADT**

To further validate the efficacy of the proposed \textsuperscript{1}I\textsuperscript{2}ADT beyond the simulations, we performed two additional validations.

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

**Validation with external clinical trial data**

In a recent pilot clinical trial \textit{Zhang et al.} (2022), clinicians implemented the IADT approach, which involved suspending treatment when PSA levels decreased to 50% of the pretreatment value and
resuming treatment when PSA levels returned to baseline. Although this IADT strategy successfully increased TTP and reduced dosage, it remains a population-based approach and is not optimized for individual patients.

We applied the proposed I²ADT method to patients in this pilot clinical trial by first training the tM-GLV model to obtain the PCaC environment and then using PPO to determine individualized treatment strategies. To facilitate a fair comparison, we extended the IADT dosing policy using the same 50% threshold criteria until the EOS was reached. The resulting dosing policies are illustrated in Figure 8, and the corresponding PFS/TTP values are presented in Figure 9. The simulation results revealed that 88.2% (15 out of 17) of patients would experience a longer TTP (p-value = 6.3×10⁻⁴) with the I²ADT approach. With I²ADT, the average treatment-on and treatment-off durations for each cycle are 1.63 and 8.73 months, respectively. In comparison, the IADT group has average durations of 7.1 and 6.83 months for treatment-on and treatment-off periods, respectively. The average dosage and treatment-on percentage for each cycle were reduced by 55.6% (p-value = 5.7 × 10⁻⁶) when using the I²ADT method.

To sum up, the external data validation further validated that proposed I²ADT can achieve better treatment outcome as compared to the conventional IADT.

![Figure 8](image1.png)

**Figure 8.** The dosing policies for patients in the validate group from a pilot clinical trial Zhang et al. (2022). Each row denotes a patient, and x-axis denotes the time (day). The left and right bars present the outcomes for I²ADT and IADT described in Zhang et al. (2022) (with a fixed 50% threshold), respectively.

![Figure 9](image2.png)

**Figure 9.** (a) The PFS rate over time with I²ADT and IADT for validation group. (b) The distribution of TTP with I²ADT and IADT for validation group.

**Validation with individual new patients**

In clinical settings, it is common for practitioners to be unaware of a new patient’s response to treatment. As such, we designed an experiment in a hypothetical prospective scenario where a new patient undergoes treatment without any prior knowledge of their response to ADT. In this setup, each patient initially receives IADT for a fixed time period (corresponding to the first IADT treatment cycle in our experiments) and then transitions to the personalized I²ADT approach after gathering data from the IADT treatment. In the clinical trial Bruchovsky et al. (2006), PSA testing occurs on a monthly basis. To acquire more data and better understand a patient’s cancer dynamics, we propose a weekly PSA test during the first IADT treatment cycle. We refer to this strategy as delayed-I²ADT.
In Table 2, we present the results for two representative cases: patient 001 (responsive) and patient 011 (resistant). For both patients, delayed-I^2 ADT achieved similar performance as I^2 ADT, exhibiting the same TTP and comparable dosages. Both delayed-I^2 ADT and I^2 ADT resulted in longer TTP and lower dosages compared to IADT. For patient 001, delayed-I^2 ADT led to a higher dosage than I^2 ADT, while for patient 011, delayed-I^2 ADT led to a lower dosage. The minor difference between delayed-I^2 ADT and I^2 ADT can be attributed to less intense intratumoral competition during the first IADT treatment cycle. This experiment demonstrates that new patients can benefit from I^2 ADT by initially undergoing IADT in the first treatment cycle and subsequently transitioning to a patient-specific, optimized I^2 ADT approach for subsequent treatment cycles.

<table>
<thead>
<tr>
<th>Patient001 (responsive)</th>
<th>Patient011 (resistant)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>delayed-I^2 ADT</td>
</tr>
<tr>
<td>TTP (month)</td>
<td>120</td>
</tr>
<tr>
<td>Ave. CPA (mg/day)</td>
<td>56.6</td>
</tr>
<tr>
<td>Ave. LEU (mg/month)</td>
<td>0.94</td>
</tr>
</tbody>
</table>

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

Conclusion and Discussion

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

Drug resistance is inevitable in ADT and often leads to treatment failure. By integrating the Darwinian evolution patterns of cancer cells into the drug administration, IADT achieved better outcomes in clinical trials Bruchovsky et al. (2006); Zhang et al. (2017, 2022). Existing IADT approaches are population-based and do not consider the heterogeneity of patients. In this paper, we proposed the data-driven and patient-specific I^2 ADT, which integrates the mathematical models of intratumoral evolutionary dynamics and reinforcement learning to inform personalized intermittent drug administration policies. I^2 ADT significantly increased the time to progression and reduced the cumulative drug dosage, as indicated by the results of the experiments.

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

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

There are a few limitations of this study. First, we acknowledge the limitations of the clinical trial data, which are mainly restricted to drug administration and PSA level measurements during the trial. While our model considers the effects of both Leuprolide acetate and Cyproterone acetate, the quantification of their combination effect remains to be discovered, which could involve the cooperation of pathways for each drug. The effectiveness of I^2 ADT could be further enhanced with the availability of more patient-specific clinical and pathological data, as well as information on the drug-combination effect. Second, PSA is the sole biomarker in the clinical trial data. In clinical practice, more serum biomarkers, such as circulating tumor cells (CTCs) and cell-free DNA (cfDNA),...
needed to be incorporated to better characterize the disease progression and calibrate the reinforcement learning algorithm. Third, in adopting the reinforcement learning to learn individualized policies we formulated a reward function based on the simulation. In order to maximize the total rewards (i.e. drug effects and the competition intensity between two phenotypes), the proposed policy was obtained from the learnt tM-GLV model. However, in real clinical practice, measuring the competition intensity or drug effects are not always feasible. Also, reward function may take various forms given sufficient clinical knowledge. Fourth, the serum hormone was suspended to castrate level during the treatment-on period. However, the recovery of serum testosterone after the suspension of administration would be variable among patients and could take a long time \cite{Inoue2018}, which was not considered in our work due to the missing information. Hence, the model needs to be calibrated according to actual clinical setting while being used in practice.

This is a retrospective and simulation-based study, which should be viewed with caution. In future work, we plan to conduct a pilot clinical trial to validate the I2ADT in prostate cancer patients and patients with other types of cancers. We will also further calibrate the model by incorporating more domain knowledge and clinical experience. We call on the joint effort by data scientists, pharmacologists, and oncologists to create a joint effect to realize the potential of I2ADT and more general adaptive therapies to enhance the treatment outcome of cancer patients.
Methods

In the Phase II clinical trial, clinicians adopted a population-based policy to treat patients for up to 32-36 weeks in each treatment cycle until progression Bruchovsky et al. (2006). 29 out of the 91 patients were excluded from our analysis due to missing data on drug dosages, and some patients took multiple drugs (please refer to the Appendix 5 for further information). We divided each patient’s longitudinal data into training and validation sets through stratified random sampling: 20% of each patient’s data in each cycle was randomly selected and removed as the validation set. In the following sections, we introduce the tM-GLV model and essentials for the reinforcement learning.

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

Modeling the PCaC environment

The dynamics of prostate cancer evolution are difficult to characterize with full details, because of the presence of many interacting factors Basanta et al. (2012). Following a system control approach, we formulate the ecosystem into a mathematical model capturing the key processes in the population level, namely, selection, competition, mutations and epigenetic modifications, and adaptation Isaacs and Coffey (1981); Shimada and Aihara (2008).

Based on the literature Isaacs and Coffey (1981); Shimada and Aihara (2008); Butner et al. (2020); Ribba et al. (2014); Swanson et al. (2001); McKane and Newman (2004); Chignola and Foroni (2005), we developed a time-varied mixed-effect generalized Lotka–Volterra (tM-GLV) model with the aforementioned four processes. Tumors have inherent heterogeneity, and we can usually assume that two phenotypes of prostate cancer cells Marusyk et al. (2012) before treatment, namely responsive (hormones-dependent) and resistant (hormones-independent) cells. The resistant cancer cells were minority at first, but they could gain advantage under the androgen suppression conditions. The two phenotypes have fierce competition in the tumor microenvironment because of the high demand for resources Chang et al. (2015). Besides the four processes, we also account for metastasis in the tM-GLV model by adopting a probability model (please refer to Appendix 7 for details).

Given the context of two competing phenotypes, which are viewed as permanent bounded variations of the system described by system (1), an equilibrium, even if it exists, is never established. Hence, a non-equilibrium model with time-varied disturbances is appropriate to simulate the dynamics with competition-induced mutations. The role of the equilibrium is replaced by the ultimate boundedness, which ensures the populations remain restricted to a certain limiting value in finite time Ikeda and Šiljak (1980, 1982). Hence, the system defined by system (1) has no equilibrium but an ultimate boundedness that produces a compact set of values in the states space. When the environment reaches these states, the environment is bounded in this compact set (the proof is given in the Appendix 3).

To search for optimal patient-specific parameters, we utilized a PyTorch-based solver named \( \xi \)-torch Kasim and Vinko (2020) to solve the ordinary differentiated equations (refer to Appendix 4 for initial settings) and Adam optimizer Kingma and Ba (2014) to minimize the least square error between the model-predicted PSA and the ground truth PSA. Considering that the number of cells is tens of millions, a clipping gradient is applied to the optimization process to avoid the gradient explosion.
Table 3. Definitions of notations in system (1)

<table>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x )</td>
<td>((x_1, x_2)), vector of the cell counts for two phenotypes</td>
<td>1*</td>
</tr>
<tr>
<td>( P )</td>
<td>PSA level</td>
<td>( \mu g/L )</td>
</tr>
<tr>
<td>( R )</td>
<td>( \text{diag}{r_1, r_2}), inherent growth rate for two phenotypes</td>
<td>1/day</td>
</tr>
<tr>
<td>( X )</td>
<td>( \text{diag}{x_1, x_2}), cell counts for two phenotypes</td>
<td>1</td>
</tr>
<tr>
<td>( I )</td>
<td>((1, 1)), vector of 1</td>
<td>1</td>
</tr>
<tr>
<td>( A(t) )</td>
<td>time-varied competitive community matrix</td>
<td>1</td>
</tr>
<tr>
<td>( K )</td>
<td>( \text{diag}{K_1, K_2}), carrying capacity for two phenotypes</td>
<td>1</td>
</tr>
<tr>
<td>( D )</td>
<td>drug pressure: drug-induced decay</td>
<td>1</td>
</tr>
<tr>
<td>( \alpha )</td>
<td>Hyper-parameter in generalizing the competition matrix</td>
<td>1</td>
</tr>
<tr>
<td>( \rho )</td>
<td>the secretion rate of PSA</td>
<td>( \mu g/L ) day</td>
</tr>
<tr>
<td>( \phi )</td>
<td>the decay rate of PSA</td>
<td>1/day</td>
</tr>
</tbody>
</table>

* 1 denotes no unit for this variable.

Model-informed treatment planning with reinforcement learning

While a predictive understanding of the exact evolutionary trajectories towards resistance is essential for effective treatments, it remains a significant challenge. However, controlling the evolution of drug-resistant cancer does not require full predictability or determinism. In a closed-loop system, feedback can help mitigate the reliance on precise trajectory knowledge, as long as the feedback can be obtained at reasonable time intervals, the uncertainty and stochasticity can be approximated in an informed manner, and the controller is robust to changes in the system’s behavior and parameter fluctuations *Engelhardt* (2020). In this work, we apply reinforcement learning as the controller.

Modern RL algorithms can be basically classified into two branches: value-based and policy-based learning algorithms *Ivanov and D’yakonov* (2019). Deep Deterministic Policy Gradient (DDPG) *Lillicrap et al.* (2015), Trust Region Policy Optimization (TRPO) *Schulman et al.* (2015), Proximal Policy Optimization (PPO) *Schulman et al.* (2017), and Soft-Actor-Critic (SAC) *Haarnoja et al.* (2018) are powerful algorithms that have been proposed in recent years. However, each algorithm has its strengths and limitations. DDPG is an off-policy actor-critic deterministic algorithm that can only be applied in continuous states and action spaces. TRPO is an on-policy algorithm that uses KL-divergences to control the updating from old policy to new policy. However, its second-order optimization makes it difficult to implement or fine-tune the hyperparameters. Both SAC and PPO are easily implemented and suitable for discrete or continuous action-state spaces, and obtains high data-efficiency and reliable performance. We empirically tested both algorithms and found that applying SAC in our scenario required cautious hyperparameter fine-tuning for each patient; otherwise, it would easily lead to divergence.

In terms of the PCaC environment constructed in Methods (Modeling the PCaC environment), though the system provides an environment model (6), obtaining a full-knowledge transition from states to states is challenging. Therefore, an optimal policy for this decision-making process needs to provide dosing guidelines for the next time step based on potential stochastic evolutionary scenarios described by the system (6). The system involves a continuous state space (cell population composition), time-varying and potentially high stochasticity, and one or multiple controls (drugs) that can take continuous values when administered intravenously or discrete values when administered orally. Although we use discrete action spaces in our setting, it should be easily generalized into high-dimensional continuous action space in practice. Therefore, a versatile model-free RL algorithm, namely Proximal Policy Optimization (PPO) *Schulman et al.* (2017), is preferred due to the complexity of the evolutionary dynamics.

In this work, we use the simulation of the tM-GLV model to train the PPO to get the optimal
policy. At each time step, the microenvironment information (such as cell counts and PSA levels) is sampled from the non-BlackBox system (1) to train the agent of reinforcement learning by taking the actions given by the agent. To determine the explicit formulation of the reward function, the key is to describe the drug efficacy and the competition intensity in the PCaC environment. In addition, penalty of dosage history is assigned to the reward. Please refer to the Appendix 2 for detailed description of the states-actions spaces and the rewards assignments.

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

**Validation of I²ADT**

In this section, we showcase the external validation of our proposed I²ADT approach and address the critical question of how I²ADT can be applied to a new patient.

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

In addition to external data validation, we introduce a new treatment protocol called delayed-I²ADT. When treating a new patient, the clinician develops their treatment policy based on all available clinical and pathological information up to a fixed time stamp, such as a full cycle of standard IADT treatment. To collect more data for model calibration, we implement a weekly PSA test during the first IADT treatment cycle, though the dosing policy is updated monthly. The details are as follows.

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

Next, we train the reinforcement learning agent for the tM-GLV model. Once the agent converges, it is said to have converged to \( \mathcal{A}(\theta_{\text{f}}^n) \). From \( \mathcal{A}(\theta_{\text{f}}^n) \), we can obtain the \( T \)-month future dosing policy, which clinicians will review and administer to patients. Clinical surveillance is performed throughout the entire period to gather additional data for tM-GLV model training. The details of the algorithm are shown in Algorithm 2.

**Data availability statement:** The initial clinical trial study is reported in Bruchovsky et al. (2006). The additional pilot clinical trial data (NCT02415621) is reported in Zhang et al. (2022). All processed data and the code have been posted on GitHub https://github.com/Michaelrising/PPOProstateCancer.

**Competing interest:** The Authors declare no Competing Financial or Non-Financial Interests.

**Supplementary information:** Please refer to the attached Supplementary files.
Algorithm 1 Online Train tM-GLV

Data: IADT clinical data: $\mathcal{X}(0 : \tau)$
Input : Data: $\mathcal{X}(0 : \tau)$
Initial parameters: $\theta^0_{online}$
Output: online tM-GLV model: $\mathcal{M}(a; \theta_{online})$

Train tM-GLV

\[
\begin{align*}
\theta_{online} &\leftarrow \theta^0_{online} \\
\text{Max Iteration} &\leftarrow N_{online} \\
\text{while } k < N_{online} &\text{ do} \\
\quad P_k &\leftarrow \xi \sim \text{torch solver} \\
\quad \mathcal{L} &\leftarrow MSE(P_k, P) \\
\quad \theta^k_{online} &\leftarrow \theta^{k-1}_{online} + \eta \cdot \nabla \mathcal{L} \\
\text{end}
\end{align*}
\]

$\triangleright$ $P$ is ground truth PSA level

Algorithm 2 delayed-I²ADT

Data: IADT clinical data: $\mathcal{X}(0 : \tau)$
Input : Prediction term: $T$
$\mathcal{X}(0 : \tau)$
Online Patient: $\mathcal{M}(a; \theta^0_{online})$
Real Patient: $\mathcal{M}(a; \theta_{offline})$
Output: tM-GLV model: $\mathcal{M}(a; \theta_{online})$
Dosing Policy: $\mathcal{A}(\theta_A)$

Init PPO

\[
\begin{align*}
\theta_A &\leftarrow \theta^0_A \\
\text{Max Iteration} &\leftarrow N_A \\
\text{Update steps} &\leftarrow s \\
\text{Online updating tM-GLV times: } c &\leftarrow 0 \\
\text{Converge Signal: } C &\leftarrow \text{False} \\
\text{while } k < N_A &\text{ do} \\
\quad \text{Explore/Exploit Online Patient (}\mathcal{M}(a; \theta^c_{online})\text{)} \\
\quad \text{Updates agent parameters } \theta^k_A &\text{ if } PPO \text{ agent converges then} \\
\quad \quad c+ &\leftarrow 1 \\
\quad \quad C &\leftarrow \text{True} \\
\quad \text{end} \\
\quad \text{if } C &\text{ then} \\
\quad \quad a_{(\tau+cT):(\tau+(c+1)T)} &\leftarrow \mathcal{A}(\theta^c_A) \\
\quad \quad \mathcal{X}((\tau + cT) : (\tau + (c + 1)T)) &\leftarrow \mathcal{M}(a_{(\tau+cT):(\tau+(c+1)T)}; \theta_{offline}) \\
\quad \quad \mathcal{X}(0 : (\tau + (c + 1)T)) &\leftarrow \text{concatenate}(\mathcal{X}(0 : \tau + cT), \mathcal{X}(\tau + cT : (\tau + (c + 1)T))) \\
\quad \quad \theta^{c+1}_{online} &\leftarrow \text{Algorithm 1}(\mathcal{X}(0 : (\tau + (c + 1)T)), \theta^c_{online}) \quad \triangleright \text{(Algorithm 1)} \\
\quad \text{Updates online tM-GLV model: } \mathcal{M}(a; \theta^{c+1}_{online}) &\text{ if } PPO \text{ agent converges then} \\
\quad \quad C &\leftarrow 0 \\
\text{end} \\
\text{end}
\end{align*}
\]
References


Drugs FA. Lupron Depot (leuprolide acetate for depot suspension); 1990.


Appendix 1

Details of PcaC construction

There are two phenotypes of prostate cancer cells prior to initiating intermittent androgen deprivation therapy (IADT): responsive (hormone-dependent) and resistant (hormone-independent) cells, as described by Marusyk et al. (2012). In the tumor microenvironment, resistant phenotypes can gain advantage through genetic or epigenetic mutations, leading to competition between the two phenotypes. This dynamic is expressed in Formula (2):

\[
\frac{dx}{dt} = RX(1 - (K^{-1}A(t)x)^\gamma - D).
\]

(2)

In the context of interacting phenotypes, both internal competition and external pressures influence the two phenotypes, which are regarded as permanently bounded variations of the system represented by Equation 2. Under these conditions, equilibrium cannot be achieved and is instead replaced by ultimate boundedness (a compact set of values in the state space) Ikeda and Šiljak (1980, 1982). Consequently, the competitive community matrix is established as follows:

\[
A(t) = \begin{pmatrix}
\frac{1}{1 + e^{-\gamma t}} & \frac{1}{1 + e^{-\gamma t}} \\
\frac{1}{1 + e^{-\gamma t}} & \frac{1}{1 + e^{-\gamma t}}
\end{pmatrix}, \gamma \in \mathbb{R}_+.
\]

(3)

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

The drug-induced decay term in Equation (2) is assumed to follow a first-order decay process Ribba et al. (2014) (i.e., a metric of drug exposure). In our study, this term is defined as the linear relationship presented in Equation (4). To ensure meaningful extrapolations for untested dosage regimens per patient, the pharmacokinetics of both drugs included in our model were considered. Cyproterone acetate (CPA) was administered twice daily, and given its half-life of 1.5 days Jentsch et al. (1976), the dynamics of CPA’s effect remain constant during therapy. Leuprolide acetate (LEU) was administered intramuscularly at a dosage of 7.5 mg every 4 weeks in a depot suspension format Periti et al. (2002); Sharifi et al. (1990).

\[
D = \beta d(t), \quad \beta > 0,
\]

(4)

where \(\beta\) is a patient-specific parameter, and \(d(t)\) represents the normalized drug effects, combining pharmacokinetic knowledge Ribba et al. (2014). In this case, \(d(t)\) is proportional to serum hormone levels. Clinical studies Periti et al. (2002); Sharifi et al. (1990) have found that serum testosterone initially increases during the first week, then becomes suppressed to castrate levels. Consequently, \(d(t)\) decreases in the first week and reaches a stable level after continuous drug administration.

In our proposed method, we employed a non-compartmental model to estimate drug exposure. Specifically, we assumed a constant drug concentration in the blood for CPA since it is taken twice daily. Therefore, a constant drug effect is assigned to CPA, which is normalized as 1 if taken, otherwise 0. For LEU, we referenced the work of Sharifi et al. (1990) and found that the drug concentration in blood plasma initially increases and then decreases...
to a steady level over the administration course of a month. The corresponding plasma
testosterone level exhibits similar patterns. For illustration of these findings, please refer
to Figure 1&5 in work *Sharifi et al. (1990)*. The figures show that with depot-injected LEU,
the testosterone level initially increases in the first week and then decreases to the castrate
level by week 4. Subsequent injections maintain the testosterone level at the castrate level.
It is essential to note that the drug effect for LEU is not defined by the drug concentration in
plasma; rather, the corresponding plasma testosterone level reflects the actual drug effects.
Based on this information, we have normalized the drug effect by setting it to negative in the
first week, decreasing linearly from 0 to -0.5. In the following three weeks, the drug effect
gradually increases from -0.5 to 1 linearly with time. Subsequent doses maintain the testos-
terone at the castrate level with a drug effect of 1. If no maintenance dosage is present, the
LEU drug effect resets to 0.

For further information on drug pharmacokinetics, please refer to Periti et al. (2002);
Sharifi et al. (1990); Drugs (1990); Dlugi et al. (1990) for LEU, and Jentsch et al. (1976) for
CPA.

Regarding the mathematical relationship between prostate cancer cell count and serum
PSA levels, it is widely assumed that PSA level dynamics can be simplified as shown in Equa-
tion (5):

\[
\frac{dP}{dt} = \rho \sum_i x - \phi P.
\]  

(5)

where \( \rho \) denotes the rate at which PSA is released from cancer cells, and \( \phi \) represents the
decomposition rate of serum PSA. The PSA decay rate is set as a population-wide uniform
parameter, with \( \phi = 0.25 \text{ day}^{-1} \), given that the serum PSA half-life is 2.5 days Richardson
et al. (1996); Stamey et al. (1987); Oesterling et al. (1988).

By combining Equations (2 ~ 5), the mathematical model for simulating the prostate
cancer cell (PCaC) environment is presented as system (6):

\[
\begin{align*}
\frac{dx}{dt} &= RX(1 - (K^{-1} A(t)x)^x - D), \\
\frac{dP}{dt} &= \rho \sum_i x - \phi P.
\end{align*}
\]  

(6)
Learning an adaptive dosing policy by reinforcement learning

Details about PPO algorithm

Reinforcement learning (RL) is a continuous process where an agent interacts with an environment at discrete time steps. At each time step, the agent receives the environment’s state \( s_t \) and selects an action \( a_t \). The environment responds with a new state \( s_{t+1} \) and a reward \( r_{t+1} \) associated with the action. After each cycle, the agent updates the value function \( V(s) \) or action-value function \( Q(s, a) \) based on a certain policy \( \pi \), where \( \pi \) maps states \( s \in S \) to actions \( a \in A \), i.e., \( \pi : S \to A : a = \pi(s) \) Kaelbling et al. (1996); Sutton and Barto (2018).

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

\[
\hat{g} = \hat{E}_t [\nabla \theta \log \pi_\theta (a_t | s_t) \hat{A}_t] \tag{7}
\]

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

\[
L_{CLIP}(\theta) = \hat{E}_t [\min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon)\hat{A}_t)] \tag{8}
\]

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

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

Details about states, action spaces, and the reward assignment

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

In section Details of PcaC construction, two phenotypes of the prostate cancer cells and the biomarker indicator (serum PSA level) were included in the system (6). Hence, at each
time step $t$, an observation of cell counts ($x_{t,1}$ and $x_{t,2}$) and PSA level ($p_t$) was made as the current states $s_t = (x_{t,1}, x_{t,2}, p_t)$. Additional feature combinations of $s_t$ can provide more information for model training. In a precise manner, the instant growth/decay rates $\dot{s}_t$ are indicative of the PCaC environment, reflecting the current drug and competition pressures, which can be obtained directly from the current states $s_t$. Moreover, time $t$ was also included in the states. Hence, the states for PCaC environment was given by $S_t = (s_t, \dot{s}_t, t)$.

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

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

Using the combination of features $S_t$ in the non-BlackBox model can help provide treatment information, which helps model training. Moreover, an additional feature combination of model-informed learning, which is the related information of states $S_t$ provided by the system (6), can support the reward assignment.

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

$$ r_{drug,t} = d_1(t)(1 - c_{1,t}). $$  \hspace{1cm} (9)

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

$$ p_{drug,t} = \sum_a \sum_{i=t-1}^{t} \eta^{i-1} w_a \frac{I_{a,i}}{I_{a,max}}, $$  \hspace{1cm} (10)

where $i$ denotes the continuous administration time, $\eta$ is the decaying penalty ratio for the historical drug administration, and $w_a$ is the drug-specific penalty parameter. Eventually, the reward function was assigned as follows:

$$ r_{step,t} = r_{drug,t} + r_{comp,t} - p_{drug,t}. $$  \hspace{1cm} (11)

With limited resources in the tumor microenvironment, if insufficient dosages were administered, the responsive cancer cells proliferated and quickly reached their carrying capacity, making the system stay in an unwanted state with a high tumor level. The reward function (11), based on changes from one decision step to the next, provides an insufficient penalty. On the one hand, the penalty of the drug $p_{drug,t}$ is relatively low because of the small dosage. On the other hand, the $r_{comp,t}$ reaches its maximal value when the response population reaches the carrying capacity. Hence, the reinforcement learning agent may continue to apply a low-dosing strategy rather than increase the dosage to escape the high tumor level. This low-dosing strategy leads to the zero-dosing problem eventually after the model converges. To address this problem, we assign a progression-free time reward to the step reward function and adopt a metastasis probability model as EOS to avoid the high concentration of the response population.
**Appendix 3**

**Proof of ultimate boundedness**

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

\[ \dot{x} = Xf(t, x), \]  

(12)

where \( x(t) \in \mathbb{R}_+^2 \) denotes the cell counts for two phenotypes (responsive and resistant separately) and \( X = \text{diag}(x_1, x_2) \). And \( f(t, x) \) is a 2-dimensional function: \( \mathbb{R}_+ \times \mathbb{R}_+^2 \rightarrow \mathbb{R}^2 \) is sufficiently smooth for the definition of \( A(t) \) in Equation (3). Hence, we claim that the Equation (12) has a unique solution \( x(t; t_0, x_0) \) for all initial conditions \( (t_0, x_0) \in \mathbb{R}_+ \times \mathbb{R}_+^2 \).

For the convenience of analysis, we rewrite Equation (12) as follows:

\[ \dot{x} = X(R(1 - D) + (-R(A(t)K^{-1}x)^m)) = X(a(t, x) + \Lambda'(t, x)x), \]  

(13)

where \( a(t, x) = d(t) = R(1 - D) : \mathbb{R}_+ \rightarrow \mathbb{R}_+^2 \), and \( \Lambda'(t, x) = -RX^{-1}K^{-1}A(t)x : \mathbb{R}_+ \times \mathbb{R}_+^2 \rightarrow \mathbb{R}^{2 \times 2} \) is a 2 \( \times \) 2 functional matrix and \( R = \text{diag}(r_1, r_2) \), \( K = \text{diag}(K_1, K_2) \).

Let us first give the definition of ultimate boundedness (following Šiljak and Weissenberger (1973); Ikeda and Šiljak (1982)) as follows:

**Definition 1** The solutions \( x(t; t_0, x_0) \) of system (13) are said to be ultimately bounded with respect to the region \( \mathbb{R}_+^2 \) if there exists a compact region \( \Omega \in \mathbb{R}_+^2 \) and a finite time \( t_1 \geq t_1(t_0, x_0) \) such that for any \( (t_0, x_0) \in \mathbb{R}_+ \times \mathbb{R}_+^2 \) we have \( x(t; t_0, x_0) \in \Omega \) for all \( t \geq t_1 \).

Second, from theorem (2.7) in Ikeda and Šiljak (1982) which is stated as,

**Theorem 1** The solutions \( x(t; t_0, x_0) \) of system (13) are said to be ultimately bounded with respect to the region \( \mathbb{R}_+^2 \) if there exists a diagonal matrix \( D \) and a positive number \( \eta \) such that the matrix \( B(t, x) \) defined below satisfies the inequality in (14), and \( a(t, x) = d(t) \) is bounded from above as sure.

Hence, to prove the ultimately bounded property for system (13), we first note that since \( a(t, x) = d(t) \) in our case, which is bounded for sure. Then we have to find whether there exists a constant positive diagonal matrix \( D = \text{diag}(d_1, d_2) \) and a positive number \( \eta \) such that the 2 \( \times \) 2 symmetric matrix \( B(t, x) \) satisfies the three conditions in Formula (14). We know that \( d(t) \) is bounded since \( D \) denotes the drug effect in our scenario.

\[
\begin{cases}
B(t, x) = -\frac{1}{2}([A'(t, x)]^T D + DA'(t, x)), \\
\min\{\lambda_1, \lambda_2\} \geq \eta; \lambda_m, m = \{1, 2\} \text{ is eigenvalues of } B(t, x), \\
\forall(t, x) \in \mathbb{R}_+ \times \mathbb{R}_+^2.
\end{cases}
\]  

(14)

Since \( A(t) = \begin{pmatrix} \frac{1}{1+e^{x_1}} & \frac{1}{1+e^{x_2}} \\ 1 & -1 \end{pmatrix} \), with \( a(t) = \frac{1}{1+e^{x_2}} \in (0, 0.5) \), by eigendecomposition, \( A(t) = Q\Lambda(t)Q^{-1} \), where

\[
Q = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \ \Lambda(t) = \begin{pmatrix} 1 + a(t) & 0 \\ 0 & 1 - a(t) \end{pmatrix}.
\]

(15)

Then let \( \Gamma(x) = RX^{-1}K^{-1} = \text{diag}(\frac{r_1x_1^{-1}}{K_1}, \frac{r_2x_2^{-1}}{K_2}) \),

\[
A'(t, x) = -RX^{-1}K^{-1}A(t)x = -RX^{-1}K^{-1}Q\Lambda(t)Q^{-1} = -\Gamma(x)Q\Lambda(t)Q^{-1}.
\]

(16)
Let $D = \text{diag}\left(\frac{K_1 r_1}{K_1 r_1 + 1}, \frac{K_2 r_2}{K_2 r_2 + 1}\right)$, then

$$B(t, x) = -\frac{1}{2}([A'(t, x)]^T D + DA'(t, x)),$$

$$B(t, c) = -\frac{1}{2} (Q^{-1} \Lambda^\alpha Q \text{diag} \{c_1^{\alpha-1}, c_2^{\alpha-1}\} + \text{diag} \{c_1^{\alpha-1}, c_2^{\alpha-1}\} Q \Lambda^\alpha Q^{-1}),$$

$$= -\frac{1}{2} (1 + a(t))^\alpha + (1 - a(t))^\alpha) \begin{pmatrix} 2c_1^{\alpha-1} & b(t, c) \\ b(t, c) & 2c_2^{\alpha-1} \end{pmatrix}, \tag{17}$$

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

Since $a(t) = \frac{1}{1 + e^{\gamma t}}$ with $\gamma > 0$ and $a(t) \to 0$ as $t \to \infty$, we have $b(t, c) \to 0$ as $t \to \infty$, indicating that we could always find a $T$, satisfying when $t > T, 4(c_1 c_2)^{\alpha-1} - b(t, c)^2 > 0$. Hence, $|B(t, x)| > 0$ as $t > T$, proving that $B(t, x)$ is positive-definite when $t > T$. This proves the conditions are satisfied.
To address the initial value problem, it is necessary to establish patient-specific and plausible initial values for system (5). However, due to limited patient-specific information, such as pre-treatment/post-treatment prostate volume, determining the true initial cell counts for each patient is challenging. Fortunately, the average prostate volume for all patients included in reference \textit{Bruchovsky et al.} (2008) is available. Consequently, we utilize Equation (18) to establish the initial values for system (5).

\[
\begin{align*}
    x_1(0) &= c_1 K_1, \\
    x_2(0) &= c_2 K_2, \\
    K_1 &= \frac{c_1 \bar{V} P_{\text{max}}}{P_{\text{cell}}}, \\
    K_2 &= c_4 K_1.
\end{align*}
\]

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

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

In the clinical trial’s first treatment cycle, the average prostate volume decreased from a baseline of 24.7 cm\(^3\) to 14.7 cm\(^3\) \textit{Bruchovsky et al.} (2006). We assumed that cancer cells accounted for 50% of the prostate volume decline, resulting in an estimated average volume of \(\bar{V} = 5\) cm\(^3\). The average pre-treatment PSA value was \(\bar{P} = 22.1\mu g/L\ \textit{Bruchovsky et al.} (2008)\), and the volume of a single cancer cell was calculated as \(V_{\text{cell}} = \frac{4}{3}\pi(5 \times 10^{-4} \text{cm})^3 = 5.236 \times 10^{-10} \text{cm}^3\).
Patient selection from the clinical trial data

Data of 91 patients in the clinical trial *Bruchovsky et al. (2006)* were obtained from https://www.nicholasbruchovsky.com/clinicalResearch.html. For simplicity, in our simulation we only consider dual-drug effects. 19 patients in the folder "Shaw_et_al" were excluded because no drug information was available. We excluded 10 patients who were administrated with more than two drugs (patients 014, 022, 026, 028, 039, 041, 055, 064, 081, and 109). Eventually, we have 62 patients for our analysis.
Appendix 6

Neural network architecture and hyperparameter setting

Because the state space is continuous while the action space is discrete, we apply PPO Algorithms proposed by Schulman et al. (2017). The Q-network has four layers of fully-connected linear networks, for both the actor and the critic networks. Weights and bias were initialized randomly. Note that, applying more complicated networks such as the recurrent neural network and the gated recurrent units will not benefit the performance. No batch normalization was used at hidden layers, but features were re-scaled prior to being added to the replay buffer in the following manner.

\[ x_i \rightarrow \frac{\log(x_i + 1)}{\log(K_i)}, \]

\[ \dot{x}_i \rightarrow \begin{cases} 
\frac{\log(\dot{x}_i) + 1}{\log(K_i)}, & x_i > 1, \\
\frac{\log(-\dot{x}_i) - 1}{\log(K_i)}, & x_i < -1, \\
\frac{\dot{x}_i}{\log(K_i)}, & \text{otherwise.}
\end{cases} \]  

(19)

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

To circumvent the zero-dosing issue, we employ two strategies. First, we propose a straightforward probability model of metastasis. Based on observations and experiments in Luzzi et al. (1998); Rejniak and Anderson (2011); Aceto et al. (2014), we adopt a probability model of cell concentration to simulate the metastasis process. Initially, we define sub-metastasis as occurring with a probability when the concentration of cancer cells $c$ exceeds the threshold $\hat{c}$ (sub-metastasis occurs when $m_{\text{sub}}(t) = 1$, otherwise $0$). The probability is proportional to $c^{2/3}$, where $c$ represents the concentration of cancer cells. We assume the tumor lesion to be spherical, with only cells on the surface capable of detaching from the lesion and transferring to other organs. Given the carrying capacity in Equation (18), $K \sim 1/V_{\text{cell}}$, $r_{\text{cell}} \sim K^{-1/3}$. Consequently, the surface area is proportional to $c^{2/3}$.

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

$$m_{\text{final}} := \delta\left(\sum m_{\text{sub}} < n\right) = \begin{cases} 0, & \text{if } \sum m_{\text{sub}} < n \\ 1, & \text{otherwise} \end{cases} \quad (20)$$

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

$$m_{\text{sub}} \sim \begin{cases} \text{Bernoulli}(c^{2/3}), & \text{if } \hat{c} < c \\ 0, & \text{otherwise} \end{cases} \quad (21)$$

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

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

$$r_{\text{supp}} = n_{\text{steps}} \times (1 - c), \quad (22)$$

where $c$ denotes the cell concentration of the resistant population, and $n_{\text{steps}}$ represents the number of steps in one episode during one episodic sampling.

By employing these two strategies, the agent can effectively avoid the sub-optimal zero-dosing policy.
Leave-pair-out cross validation

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

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

Supplementary File 1: 95% CI for all the patient-specific parameters.
Supplementary File 2: The population means for all parameters and their 95% CI.
Comparison of the $I^2$ADT-derived adaptive dosing policies of resistant and response patients

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

For responsive patients, an ascending drug administration pattern emerged as drug pressure on responsive cancer cells increased throughout the therapy. For the overall responsive patients, the treatment off-to-on ratio, daily CPA dosage, and monthly LEU dosage can be found in Supplementary Figure 2 of Appendix 10. As therapy progressed, a decreasing pattern in the treatment off-to-on ratio was observed. A significant increase in drug dosing was evident in $I^2$ADT, while a decreasing pattern was apparent in standard IADT.

For resistant patients, a descending policy was observed, with the treatment off-to-on ratio increasing over time. The average off-to-on treatment ratio and daily CPA and monthly LEU dosages across time are presented in Supplementary Figure 1 of Appendix 10. During the course of treatment, an increasing pattern in the treatment off-to-on ratio was observed, while the CPA/LEU pattern remained relatively flat compared to the standard IADT, which exhibited an increasing pattern. This insignificance may due to the number of patients in the resistance population is only 11.
Appendix 10

Other supplementary files

**Supplementary Figure 1**: The treatment off/on ratio, daily CPA dosing, and monthly LEU dosing of the $I^2$ADT (left) and the standard IADT (right) for resistance patients.

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

**Supplementary Figure 3**: The episodic reward for training and evaluation reward/survival month with greedy strategy.

**Supplementary Figure 4**: This figure illustrates the use of all parameters, along with the $r_2/r_1$ ratio, as classifiers to differentiate between responsive and resistant cases. Apart from $\gamma$ and $\alpha$, the remaining parameters demonstrate no predictive power in classification.

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

This is a list of supplementary files associated with this preprint. Click to download.

- supplement.pdf