Reinforcement Learning with Action-Derived Rewards for Chemotherapy and Clinical Trial Dosing Regimen Selection

Gregory Yauney GYAUNEY@MEDIA.MIT.EDU
Pratik Shah* PRATIKS@MEDIA.MIT.EDU
Media Lab
Massachusetts Institute of Technology
Cambridge, MA, USA

Abstract

Unstructured learning problems without well-defined rewards are unsuitable for current reinforcement learning (RL) approaches. Action-derived rewards can allow RL agents to fully explore state and action trade-offs in scenarios that require specific outcomes yet are unstructured by external reward. Clinical trial dosing choice is an example of such a problem. We report the successful formulation of clinical trial dosing choice as an RL problem using action-based rewards and learning of dosing regimens to reduce mean tumor diameters (MTD) in patients undergoing simulated temozolomide (TMZ) and procarbazine, 1-(2-chloroethyl)-3-cyclohexyl-l-nitrosourea, and vincristine (PCV) chemo- and radiotherapy clinical trials. The use of action-derived rewards as partial proxies for outcomes is described for the first time. Novel dosing regimens learned by an RL agent in the presence of action-derived rewards achieve significant reduction in MTD for cohorts and individual patients in simulated TMZ and PCV clinical trials while reducing treatment cycle administrations and dosage concentrations compared to human-expert dosing regimens. Our approach can be easily adapted for other learning tasks where outcome-based learning is not practical.

1. Introduction

Reward functions in RL domains are typically derived from a measure external to the chosen representation of the states and actions. Using RL to solve tasks without readily accessible external scalar outcomes is a relatively unexplored field, as many currently studied domains have well-defined outcomes as part of their definitions (Sutton and Barto, 1998). Abstract RL domains often used for theoretical research, such as taxi and gridworld, have predetermined reward states that give reward values independent of the goal state's representation and of the actions taken to get to such states (Sutton and Barto, 1998). More applied domains use similarly abstract notions of reward, such as score or normalized score in the case of the popular Atari domain for deep RL (Mnih et al., 2015). RL has shown significant applicability for a wide variety of tasks such as learning to play Go, winning arcade games, and coordinating the motion of multiple robots (Silver et al., 2016; Mnih et al., 2015; Amato et al., 2016). In medical domains, rewards derived from patient outcomes are especially prevalent in learning problems. For example, recent work applying RL to the treatment of sepsis in intensive care units derives its reward solely from patient survival outcomes rather than from the domain's state and action spaces (Raghu et al., 2017). However, in several

* Corresponding author.

© 2018 G. Yauney & P. Shah.

constrained settings such as clinical trials and stock trading, the relations between state and action representations play a crucial role in influencing outcomes that may be unavailable for learning. Devising suitable non-trivial reward functions to learn contributions from such actions and states to desired outcomes is a promising approach to solve these classes of problems.

Clinical trials to evaluate new drugs, therapies, and vaccines are among the most complex experiments performed in medicine. Nearly half of phase 2 and phase 3 trials fail. For oncology trials, the failure rate rises to two-thirds (Wong et al., 2018). A common theme is the difficulty of predicting clinical results in a wide patient base given limited knowledge of key parameters which need to be considered to test candidate molecules, eliminate adverse events, and identify the drug's half maximal inhibitory concentration. Optimal dosing of chemo- and radiotherapy (CRT) for patients enrolled in oncology clinical trials provides one example of an open-ended problem characterized by complex interactions between different drug properties, dosage and timing of administrations (actions), and effects on tumors (state) and where survival (outcomes) may not be available. Dosing for CRT is currently determined through preclinical experiments and model-based computational methods, such as the optimization of tumor growth inhibition (TGI) models (Cook et al., 2015; Ribba et al., 2012). CRT is often given as a combination of drugs. PCV is a triple drug treatment for glioblastomas (brain tumors) and is often considered toxic. TMZ, which is less toxic and which, in conjunction with RT, has shown greater efficacy than RT alone for glioblastomas (Rinne and Wen, 2015). Thus, there is need for rational CRT experiments in human subjects to optimize for the maximum benefit for the patients while reducing toxicity.

Technical Significance We formulated an RL domain to iteratively explore possibilities for effective dosing to achieve maximum decrease in MTDs of patients. A previously described TGI model developed using PCV and TMZ CRT data from adult diffuse low-grade gliomas (LGG) that captures LGG growth kinetics during and after CRT was used as part of the environment (Ribba et al., 2012). Our modeling and learning approaches are shown in Figure 1. We apply novel state- and action-derived reward functions to learning dosing regimens in TMZ and PCV clinical trials, showing a partition of the trial design space through different trade-offs between actions and outcome proxies. More generally, we

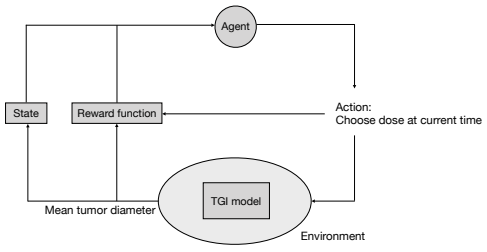


Figure 1: The reinforcement learning agent interacts with an environment containing a tumor growth inhibition (TGI) model. The reward is determined in part by the values used for the reinforcement learning model's state and the agent's most recent action.

present an approach that incorporates action-derived rewards to facilitate the use of RL in domains simultaneously constrained and unstructured by a lack of external rewards.

Clinical Relevance Learned dosing regimens achieve significant reduction in MTD for cohorts and individual patients in TMZ and PCV CRT trials while reducing drug concentrations and administrations compared to human-expert dosing regimens. The system may be useful for exploring possibilities for and ultimately choosing a dosing regimen for clinical trials and other medication dosing scenarios.

2. Background

2.1. Reinforcement learning

RL abstracts a sequential decision-making problem as a Markov decision process (MDP). An MDP is a five-tuple: set of states S, set of actions A, transition function T: $S \times A \times S \rightarrow [0, 1]$, reward function R: $S \times A \times S \rightarrow \mathbb{R}$, and discount rate γ: $[0, 1]$. An RL agent interacts sequentially with the environment with the goal of selecting the action at each timestep t that maximizes expected discounted reward, given by the equation $\sum_{t'=t}^{t_{final}} \gamma^{t'-t} r_{t'}$ (Sutton and Barto, 1998).

Q-learning is an off-policy method for calculating an optimal policy for an RL agent (Kaelbling et al., 1996). The Q-function parameterizes the expected discounted reward by state and action. The optimal Q-value for a state-action pair is given by a rearrangement of the Bellman equation, where s and a represent the state and action at the current timestep and s' and a' represent the state and action at the subsequent timestep:

$$Q^*(s, a) = R(s, a, s') + \gamma \sum_{s' \in S} T(s, a, s') \max_{a'} Q^*(s', a')$$

Actions are sampled according to an epsilon-greedy strategy, after which the optimal action from each state can be found:

$$\pi^*(s) = \text{argmax}_a Q^*(s, a)$$

where $\pi^*(s)$ represents the action a from state s that leads to maximum expected discounted reward (Kaelbling et al., 1996). Deep RL approximates the Q-function with a deep neural network, allowing for learning from rich multidimensional states (Mnih et al., 2015), synchronizing the two periodically (Van Hasselt et al., 2016).

2.2. Reinforcement learning with longitudinal health data

In all reinforcement learning formulations, the current state at each timestep varies across the set of all possible states. Deep RL allows individual states to contain dimensionally richer information, and many domains have focused on high-dimensional visual data (Mnih et al., 2015). In the health context, temporal data in the form of electronic health records and intensive care unit interventions over time have both been represented in an MDP's state (Nemati et al., 2016; Raghu et al., 2017; Prasad et al., 2017). A previous study investigated the use of RL to evaluate discrete combinations of predetermined treatment regimens, as identified and tested in previous trials, at two decision points for multiple

lines of treatment for non-small cell lung cancer (Zhao et al., 2011). Its state space is a tabular representation of patient state variables and previous actions, and a support vector regression is used to approximate the Q-function. The reward function maximizes simulated patient restricted mean survival time, which has been estimated according to a model, and does not use reward derived from actions.

2.3. Dosing and efficacy for chemotherapy clinical trials and TGI models

Phase I clinical trials are conducted to evaluate the toxicity and efficacy of a new drug in humans for improving survival and reducing disease indicators (Cook et al., 2015). Dose-ranging or -finding studies are often conducted at an early preclinical stage for novel therapies to find related values such as the drug's maximum tolerated dose (Scmidt, 1988). A relatively fixed dosing regimen has been decided upon based on preclinical animal models, previous clinical trials, and, in some cases, TGI model optimization before chemotherapy clinical trials are conducted to further evaluate the efficacy and effectiveness of the drug (Cook et al., 2015; Ribba et al., 2012). These dosing regimens often uniformly administer the largest dose concentration and frequency within the realm of predetermined safety calculated from preclinical data. Such dosing regimens often do not generalize to CRT clinical trials (Scmidt, 1988). Individual drug dosages, to the extent that they can vary between patients, can be calculated as a fixed function of patient body mass or body surface area, though there is reason to believe the latter overestimates the drug's efficacy (Gurney, 2002)

2.4. Reward functions in open-ended domains

RL requires the existence or construction of a scalar reward function. Many canonical domains, such as gridworld, have predefined goal states as part of their definition, and others, such as the Atari learning environment, have a measure such as score that can be intuitively interpreted as reward (Sutton and Barto, 1998; Mnih et al., 2015). Previous work has investigated reward shaping, or learning with modified reward functions, but this approach assumes the prior existence of an external reward that can be modified (Ng et al., 1999). The subfield of inverse RL infers reward functions from observed optimal behavior but requires the existence of a corpus of known optimal behavior (Ng et al., 2000). Preference-based RL methods learn reward functions from expert preferences. These methods make it difficult to interrogate and understand policy possibilities in the absence of *a priori* knowledge of expert preferences (Wirth et al., 2017).

3. Approach

We characterized the behavior of an RL agent in environments with under-constrained or ill-defined external rewards using a representative problem of determining dosing regimens for simulated CRT trials for multiple notions of optimality.

3.1. Environment: TMZ and PCV CRT data

TGI models, often systems of linear or differential equations used to determine a tumor's change in size over time in response to the presence of drugs, can be used to simulate clinical trial datasets. We chose a previously-described ordinary differential equation model

of LGGs in response to PCV and TMZ chemotherapy as part of the RL environment to simulate data (Ribba et al., 2012). This TGI model allows for flexibilities in treatment type (PCV or TMZ), dose cycle administration times, dose concentrations, trial design parameters, and patient parameters. Longitudinal MTD trajectories for each patient are its sole outputs.

3.2. RL formulation

We formalize the problem of determining optimal dosing for simulated TMZ or PCV CRT trials as an MDP, where the environment contains the TGI model described above. The agent interacts with the environment each timestep of the simulated trial, iteratively choosing a dosing regimen for each patient. The state space is a function of a patient's recent MTD trajectory or of a group of patients' recent MTD trajectories, and the action space encodes whether or not a PCV or TMZ dose cycle is administered at the current timestep. The agent learns aggregately across patients, choosing actions either individually, one patient at a time, or for all patients as a group. Figure 1 shows the relationship between the agent and the trial environment.

State space: The state space of the agent is the codomain of a continuous function of the patient's recent MTD trajectory. We use double deep Q-learning to encode and learn from such a state space (Van Hasselt et al., 2016). The network has two hidden layers, the first with 32 nodes and the second with 64 nodes, followed by an output layer with the same number of nodes as possible actions. Rectified linear unit activation functions were used after each hidden layer. The state also contains the current month to obey the Markov assumption. The input to the network is a sample of the current patient's recent MTD trajectory or of a group of patients' MTD trajectories. During initial optimization experiments, we empirically determined a ten-month window to be descriptive enough to learn from without incurring a prohibitively long training time. Trials with a more varied simulated patient parameter distribution would likely require a larger temporal window for determining the current state.

Action space: The agent is responsible for choosing a patient's dosing regimen over the course of the simulated trial. We discretize the possible dosing options to allow for a discretized action space, leading to a combinatorially large but not intractable number of possible states for each patient. At each timestep, the agent can choose to initiate a dose cycle of fixed duration or to withhold the dose. Dose cycles can only be initiated with a frequency determined by a corresponding expert trial (Peyre et al., 2010; Ricard et al., 2007). Concentration choices were restricted to a full concentration in fixed concentration experiments but were allowed to be 25%, 50%, 75%, and 100% of a maximum allowable dose for variable concentration experiments.

Transition function: Transitions between states are deterministically calculated by sending the patient parameters and next dose to the TGI model environment. In some experiments, we introduce a cap on the total amount of drug that can be administered per patient. If the agent attempts to give a patient a dose once that cap has been reached, the state transitions as though no dose had been administered.

Reward functions: The only information provided by the clinical trial environment is MTD trajectory, so the reward must take into account MTD if there is to be any proxy

for patient outcomes. In all experiments, overall MTD reduction is incentivized through a reward at the end of each episode. Positive rewards result from a decrease in MTD and negative rewards from an increase. The clinical goal of MTD reduction is additionally encouraged through smaller linear rewards each time the agent selects an action. Different dose penalties were introduced to investigate the effect of additional objectives beyond MTD reduction in the trial design space. Possible penalties could be small or large with respect to the terminal MTD reward. We also performed experiments without any dose penalty to find the optimal dosing regimen that reduces MTD without any other considerations. The form of the reward function at each timestep t and subsequent timestep t' is:

$$R = c(MTD_t - MTD_{t'}) - penalty \cdot concentration$$

where *concentration* is the percent of the unit dose administered at t (0 if no administration). The form of the final episodic reward is similar but takes a longer view by comparing the final observation to the initial observation:

$$R_{final} = c_{final}(MTD_{initial} - MTD_{final})$$

In practice, c was set to 1 and c_{final} was set to 10. Absolute sizes of the relative dose penalties were set after initial optimization experiments at 1 for the small penalty and 5 for the large for TMZ and 1 for the small penalty and 10 for the large for PCV. The values correspond to the MTD reduction required to make a single treatment cycle worthwhile.

3.3. Simulated trial parameters

Trials were conducted with 50 simulated patients whose initial parameters were sampled from previously reported patient parameter log-normal distributions (Ribba et al., 2012). Each learning episode runs a trial for the same fixed period of time, after which the learning process begins a new episode with the same initial patient parameters. Each learning experiment was run for at most 20,000 episodes, empirically determined to provide enough time for convergence. After learning, the learned policy was applied to the same 50 patients. Figure 2 shows representative applications of expert and learned policies to patients under different trial design parameters.

TMZ: Dose cycles could be administered every month for a maximum of 30 cycles with a final observation one year after the last month of the dose administration portion of the trial. Dose cycles had a fixed duration. Dose parameters for the majority of experiments were fixed at 200 $mg/m^2/d$ of TMZ for five days to most closely resemble the fixed dosing in expert TMZ trials found in (Ricard et al., 2007). In variable dose concentration experiments, doses of 50, 100, 150, and 200 $mg/m^2/d$ of TMZ, all for five days, could be administered.

PCV: Dose cycles could be administered every six weeks for a maximum of six cycles with a final observation one year after the last month of the dose administration portion of the trial. Dose cycles had a fixed duration. Dose parameters for the majority of experiments were fixed at 110 mg/m^2 of CCNU on day one of the cycle, 60 mg/m^2 of procarbazine on days eight to 21, and 1.4 mg/m^2 of vincristine on days eight and 29 to most closely resemble the fixed dosing in expert PCV trials found in (Peyre et al., 2010). In experiments with variable dose concentrations, all drugs could be administered at 25%, 50%, 75%, or the full concentrations.

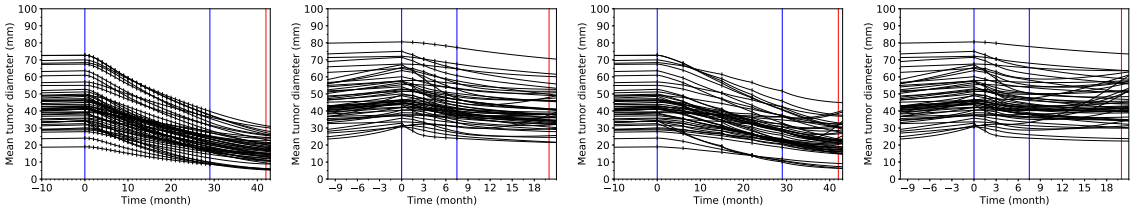


Figure 2: In the presence of large dose penalties, the agent administers fewer TMZ and PCV doses compared to corresponding expert policies. Representative expert and learned policies applied to simulated patients. Each black line represents a mean tumor diameter trajectory for a single patient. The agent could give doses only in the months between the vertical blue lines. The vertical red line represents the time of the final observation. Black markers indicate an administered dose of unit concentration. (a) Expert policy, simulated TMZ trial; (b) Expert policy, simulated PCV trial; (c) Learned TMZ regimen with a large dose penalty applied to the same patients as in panel a; (d) Learned PCV regimen with a large dose penalty applied to the same patients as in panel b.

3.4. Data analysis and evaluation

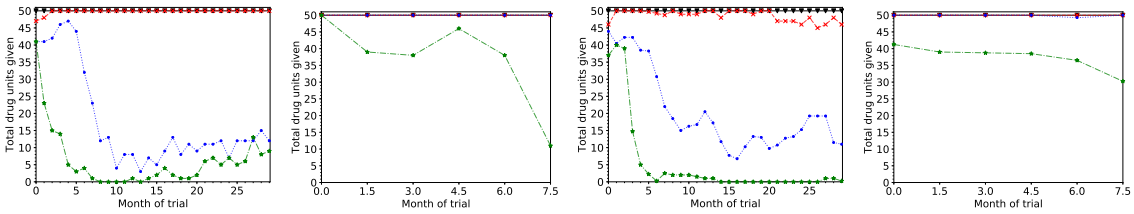
We compare our learned dosing policies to the TMZ and PCV clinical trial dosing regimens to which the underlying TGI model was fit (Ricard et al., 2007; Peyre et al., 2010; Ribba et al., 2012). Comparisons of interest were the reduction in dose cycles administered by the RL agent and the difference in average percentage MTD reduction when compared to that achieved by human-expert-directed TMZ and PCV trials. We also compare our results for the total amount of drug administered per simulated trial and the percentage of MTD reductions across dose penalties and across trial design parameters.

4. Results

We conducted three basic experiments for each combination of parameters that we varied to explore the effect of increasingly constrained reward structures under different agent flexibilities. One set of simulated experiments used the treatment TMZ and the other used PCV. For both treatments, we conducted experiments where a) the agent is able to treat different simulated patients independently (referred to as patient-based experiments), and b) where the agent must treat all simulated patients with the same dose at each timestep (called trial-based experiments). Another set of experiments allowed the agent to administer at each timestep a) a fixed unit dose of TMZ or PCV, and b) doses at 25%, 50%, 75%, and 100% of the unit TMZ or PCV dose's concentration. The agent always had the option of not giving a dose at each timestep for TMZ and PCV experiments. Table 1 shows the combinations of treatment, allowed dose concentrations, and personalization multiplexed with the three action-penalty reward functions. Table 1 also compares the average percentage of MTD reduction between learned policies and expert policies on patients sampled from the same distributions and under similar trial design parameters.

	Trial parameters			Average MTD change			
	Treatment	Conc.	Type	Expert policy	No penalty	Small penalty	Large penalty
A	TMZ	Fixed	Patient	-61.04% ± 11.63%	-60.95% ± 11.64%	-51.89% ± 15.31%	-35.97% ± 17.05%
B	TMZ	Fixed	Trial	-62.18% ± 10.50%	-62.17% ± 10.51%	-54.09% ± 14.05%	-46.27% ± 17.93%
C	TMZ	Variable	Patient	-60.51% ± 10.68%	-60.23% ± 10.80%	-39.35% ± 46.02%	-7.15% ± 40.80%
D	TMZ	Variable	Trial	-62.86% ± 11.41%	-62.72% ± 11.43%	-54.03% ± 15.21%	-45.69% ± 18.06%
E	PCV	Fixed	Patient	-21.37% ± 8.61%	-21.37% ± 8.61%	-21.37% ± 8.61%	-19.10% ± 7.51%
F	PCV	Fixed	Trial	-19.32% ± 10.66%	-19.32% ± 10.66%	-13.30% ± 15.32%	-8.12% ± 19.24%
G	PCV	Variable	Patient	-25.20% ± 9.10%	-25.20% ± 9.10%	-25.20% ± 9.10%	-22.63% ± 9.35%
H	PCV	Variable	Trial	-23.98% ± 9.67%	-23.98% ± 9.67%	-23.32% ± 9.85%	-20.97% ± 11.75%

Table 1: In the absence of a dose penalty, the application of learned policies led to comparable average MTD reductions as expert policies. Average percentage MTD change across all simulated patients when measured at the start of the trial and the final observation. Each row describes three experiments and a corresponding expert policy for comparison for each combination of treatment, personalization, concentration, and dose penalty. Fixed: the agent could give doses with the unit concentration. Variable: the agent could give doses at 25%, 50%, 75%, and 100% of the unit concentration. Conc: concentration.



(a) (b) (c) (d)

Figure 3: In the presence of penalties, the agent learned to administer fewer TMZ and PCV doses. Total amount of drug administered per month by policies learned under different dose penalties in various trial parameter configurations. All scenarios rewarded mean tumor diameter reduction and allowed individualized treatment. Line color corresponds to dose penalty: black, expert; red, none; blue, small; green, large. (a) TMZ, fixed dose; (b) PCV, fixed dose; (c) TMZ, variable dose; (d) PCV, variable dose.

4.1. Expert baselines

For each experiment, we simulate an equivalent clinical trial using the parameters and dosing policies from relevant real-world clinical trials with PCV and TMZ (Ricard et al., 2007; Peyre et al., 2010). These expert policies gave every simulated patient a dose of unit concentration at each timestep for the duration of the trial irrespective of the patients' MTD trajectories (Figure 2a-b).

			Month of trial																											
	Conc.	Penalty	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
A	Fixed	None	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
B		Small	1	1	1	0	1	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0
C		Large	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	1
D	Variable	None	1	1	1	1	.25	.75	1	1	.75	.75	.75	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
E		Small	1	1	.75	.75	0	0	.25	.5	0	.5	0	0	.5	0	.5	0	0	.5	0	0	.5	0	.5	0	0	.25	.5	0	0	.5
F		Large	1	.5	0	0	0	0	0	0	.25	.25	0	.25	.25	0	.25	0	.25	0	.25	.25	0	.25	0	.25	.25	0	.25	.25	0	.25

Table 2: Learned policies for trial-based TMZ experiments where all simulated patients received the same dose each month. Agents with a fixed concentration could give a unit dose, whereas agents with a variable concentration could give doses at 25%, 50%, 75%, and 100% of the unit dose's concentration. Penalty: size of the dose penalty the agent tried to minimize while maximizing mean tumor diameter reduction. Conc: concentration.

			Month of trial					
	Conc.	Penalty	0	1.5	3	4.5	6	7.5
A	Fixed	None	1	1	1	1	1	1
B		Small	1	1	1	1	0	0
C		Large	1	1	1	0	0	0
D	Variable	None	1	1	1	1	1	1
E		Small	.75	1	1	1	1	1
F		Large	1	1	1	.5	.5	.5

Table 3: Learned policies for trial-based PCV experiments where all simulated patients received the same dose each month. Agents with a fixed concentration could give a unit dose, whereas agents with a variable concentration could give doses at 25%, 50%, 75%, and 100% of the unit dose's concentration. Penalty: the size of the dose penalty the agent tried to minimize while maximizing mean tumor diameter reduction. Conc: concentration.

4.2. Optimal MTD reduction with no dose penalty: similar to expert policies

We performed experiments with no dose penalty wherein the only reward was to incentivize MTD reduction.

Fixed dose concentration: In the patient-based TMZ experiment where the agent was able to give or withhold a fixed unit dose each timestep, all patients received treatment at 28 out of 30 possible administrations, and most patients received all possible administrations (Figure 3a, red line). MTD was reduced by the learned policy on average $60.95\% \pm 11.64\%$ compared to the corresponding expert policy's $61.04\% \pm 11.63\%$ reduction (Table 1, row A).

In the trial-based TMZ experiment, all 50 patients received all but one possible administration (Table 2, row A). MTD was reduced on average $62.17\% \pm 10.51\%$ and was very similar to the $62.18\% \pm 10.50\%$ reduction achieved by corresponding expert policy (Table 1, row B).

In patient-based and trial-based PCV experiments with fixed concentrations, the agent learned to give treatment at every possible timestep like the expert policies (Figure 3b, red line; Table 3, row A). The learned policies thus achieved average MTD reductions identical to those of the corresponding expert policies: $21.37\% \pm 8.61\%$ for patient-based and $19.32\% \pm 10.66\%$ for trial-based (Table 1, rows E and F).

Variable dose concentration: When agents had the flexibility to administer a smaller concentration than the fixed unit dose, they similarly administered full or nearly full concentrations at most timesteps in the absence of dose penalties. In the variable concentration patient-based TMZ experiment, full doses were administered to all patients at 12 out of 30 possible administrations, and patients received full or 75% concentration doses at the remaining timesteps (Figure 3c, red line). MTD was reduced by $60.23\% \pm 10.80\%$ on average compared to the expert policy's reduction of $60.51\% \pm 10.68\%$ (Table 1, row C).

In the trial-based variable concentration TMZ experiment, the agent learned to administer full doses at 25 out of 30 decision points (Table 2, row D), achieving an average MTD reduction of $62.72\% \pm 11.43\%$ compared to the expert policy reduction of $62.86\% \pm 11.41\%$ (Table 1, row D).

In variable concentration PCV experiments, the agent again learned to give all possible full concentration doses in patient-based (Figure 3d, red line) and trial-based settings (Table 3, row D), achieving reductions equivalent to those achieved by expert policies at $25.20\% \pm 9.10\%$ and $23.98\% \pm 9.67\%$, respectively (Table 1, rows G and H).

4.3. MTD can be reduced with fewer administrations of PCV and TMZ in the presence of small and large dose penalties

TMZ: The trained agent skipped at least one treatment administration during TMZ patient- and trial-based experiments with fixed dose concentration in the presence of small and large dose penalties. In patient-based experiments, the agent administers more doses in the initial months of the trial, with a small but consistent amount administered later on (Figure 3a, blue and green lines). MTD was reduced by $51.89\% \pm 15.31\%$ under the small dose penalty and $35.97\% \pm 17.05\%$ under the large dose penalty, whereas the expert policy reduced MTD by $61.04\% \pm 11.63\%$ in both cases (Table 1, row A).

In trial-based TMZ experiments, doses were administered to all patients every several months, with higher frequency under the small penalty and lower frequency under the large penalty (Table 2, rows B and C). MTD was reduced by $54.09\% \pm 14.05\%$ under the small dose penalty and $46.27\% \pm 17.93\%$ under the large dose penalty, both compared to the expert reduction of $62.18\% \pm 10.50\%$ (Table 1, row B).

PCV: In patient-based PCV experiments with fixed dose concentration, a small dose penalty did not induce the agent to forgo any dose administrations to any patients, while a large penalty did (Figure 3b, blue and green lines). The small penalty caused a learned policy average MTD reduction of $21.37\% \pm 8.61\%$, and the large penalty reduced MTD by $19.10\% \pm 7.51\%$ on average, both compared to an expert reduction of $21.37\% \pm 8.61\%$ (Table 1, row E).

Trial-based PCV experiments had learned policies with four and three dose administrations out of a possible six for small and large penalties, respectively (Table 3, rows B and C). Average MTD reductions of $13.30\% \pm 15.32\%$ under the small penalty and $8.12\% \pm 19.24\%$ under the large penalty were less than the corresponding expert policy's average reduction of $19.32\% \pm 10.66\%$ (Table 1, row F).

Trial parameters			Average MTD change	
Drug	Conc.	Type	Expert policy	Capped total treatments
TMZ	Fixed	Patient	-61.36% ± 12.63%	-35.30% ± 26.30%
TMZ	Fixed	Trial	-62.13% ± 11.99%	-39.94% ± 24.77%
PCV	Fixed	Patient	-20.47% ± 9.77%	15.13% ± 33.51%
PCV	Fixed	Trial	-22.55% ± 7.49%	26.53% ± 24.58%

Table 4: For experiments with a capped number of total treatments per patient, MTD was reduced much less on average than by equivalent expert dosing regimens with no restriction. Fixed: the agent could give doses with the unit concentration. Conc: concentration.

4.4. Restricted maximum number of dose administrations

We additionally performed experiments with a cap on the total numbers of fixed unit concentration dose administrations an agent could administer during a trial. Total administrations were capped for both TMZ and PCV at one-third the amount given by the expert dosing regimen: 10 unit doses per patient for TMZ and two unit doses per patient for PCV. This was done for the purpose of exploring questions of the form: when total treatment administrations are arbitrarily capped, what is the dosing regimen that leads to the greatest MTD reduction? What is the greatest MTD reduction possible with those limited treatment administrations?

TMZ: In patient-based experiments, MTD was reduced by 35.30% ± 26.30% for the learned policy compared to 61.36% ± 12.63% for the expert policy, and in trial-based, 39.94% ± 24.77% for the learned policy compared to 62.13% ± 11.99% for the expert policy (Table 4, rows A and B). Four patients in the former experiment and one patient in the latter experienced MTD increases of 30.75% ± 21.84% on average and 59%, respectively (Supplementary Table 1).

PCV: In PCV experiments, two treatment administrations were not enough to decrease MTD on average. In the patient-based PCV experiment, MTD increased by 15.13% ± 33.51% compared to an expert reduction of 20.47% ± 9.77% (Table 4, row C). In the trial-based PCV experiment, the learned policy increased MTD by 26.53% ± 24.58% on average whereas the expert policy decreased MTD by 22.55% ± 7.49% on average (Table 4, row D). 27 patients in the former and all 50 patients in the latter experiment experienced MTD increases under the learned policies who would have experienced reductions under the expert policies (Supplementary Table 1). Doses were on the whole administered in the first 10 months in TMZ experiments and in the first two months in PCV experiments; the agent learned that earlier doses generally tend to lead to a larger reduction in MTD over the course of the trial (Supplementary Results).

4.5. Patient-resolution results

Supplementary Results reports dosing regimens and outcomes at the resolution of individual patients. The first section presents MTD trajectories of all patients for both learned and expert policies for each experiment. The second section presents MTD outcomes for each patient under learned and expert policies as well as the learned policies applied to each patient.

5. Discussion

5.1. Action-derived rewards

Canonical RL domains use a reward function derived from a directly observed outcome that is independent of the representation of the state and action. Rewards independent from state and action representation have the benefit of not biasing an agent for or against certain actions. This scenario allows the agent to learn associations between actions and the scenarios where they are beneficial and harmful with respect to the independent reward. We utilized a novel approach in this work that imposes structured rewards derived from states and actions in the absence of a reward derived from outcomes. As a baseline, we use reward derived from the MTD trajectories in the state to incentivize the agent to learn from an outcome proxy. An action is incentivized if a positive reward is given after an action is performed and disincentivized if that reward is negative, partially circumventing the credit assignment problem in RL. The dose penalties we introduce in all TMZ and PCV trial parameter combinations force the learning agent to attempt to maximize the state-derived reward while navigating trade-offs between potentially helpful actions and the immediate harm of those actions. The penalties temper the dose's benefit of MTD reduction with the treatment's potential toxicity. Figure 3 shows that action-derived penalties in the form of small and large dose penalties led to a reduction in the number of penalized actions in patient-based experiments, as expected. Tables 2 and 3 show the same for trial-based experiments. Importantly, the agent is in most cases still able to reduce MTD and gain the state-derived reward. Learned policies are only markedly worse when penalties are imposed beyond reason to allow only a third of the amount of drug experts administered (Table 4). Many RL domains, such as some scenarios for gridworld, often impose a small uniform negative reward to encourage agents to reach a goal (Sutton and Barto, 1998). This is distinct from the action-derived rewards we have pursued because it is not conditioned on the actions chosen by the learning agent. Action penalties essentially rank-order and provide a probabilistic prior on the relative importance of measurements that lead to outcomes. Analogous action-derived rewards are possible in the aforementioned domains; examples include a reward or penalty imposed on some or all button presses in the case of Atari, and rewards or penalties incurred by vasopressor or intravenous fluid administrations in the case of sepsis. In such cases, RL agents would have an incentive to use fewer actions, perhaps even at the expense of score or patient outcomes if the penalty is large enough. In scenarios where actions may be expected to be associated with final outcomes, especially when such outcomes are not available or are intentionally blinded, action-derived rewards provide an additional outcome proxy. Quantifying the relationship between such actions and final outcomes is a promising direction for further investigation.

5.2. Clinical impact of learned dosing regimens

Our work also has significant impact on clinical learning tasks where wholly outcome-based rewards may not be available. We first tested if the RL agent learns to use as many dose administrations as human-expert policies for patient- and trial-based TMZ and PCV experiments by imposing no penalty for dose administration. In the absence of dose penalties, the learned policy often gives as many doses as the corresponding expert policy, even though

it is presented with the additional flexibility of not giving a dose or administering reduced concentrations (Figure 3a-d, red lines; Table 2, rows A and D; Table 3, rows A and D). We found that a solely state-derived reward function does not lead to dosing regimens entirely alien from those utilized in contemporary PCV and TMZ trials, likely because the MTD trajectory in the state is a reasonable proxy for the outcomes for which experts optimize. In designing the expert dosing regimens used in this study, the experts seem to have prioritized MTD reduction rather than a smaller amount of administered treatment. When learning with dose penalties, RL agents were able to successfully reduce simulated patient MTD on average despite administering lower amounts of PCV and TMZ (Table 1). By only initiating those dose cycles that led to MTD reduction rewards outweighing the penalties, the agents effectively learned to balance the trade-off in MTD reduction incentivization and dose penalty.

Policies learned under dose penalties in patient-based experiments tended to use fewer doses and administered those doses more frequently in earlier months of the simulated trials rather than later months, indicating that the early treatments may be most essential for MTD reduction (Figure 3). There was an exception for small dose penalties in patient-based PCV experiments: the learned regimens applied the full number of doses possibly because the smaller number of possible dose cycles makes each more important for MTD reduction (Figures 3b, 3d). The same was found for trial-based experiments for TMZ (Table 2) and PCV (Table 3). Table 1 shows a general trend that the larger the dose penalty, the less MTD reduction can be expected in all conducted experiments. It also shows that all PCV and TMZ experiments, even those with large dose penalties, led to average reductions in MTD across all simulated patients, arguing for more careful dosing regimens for patients undergoing therapy. This trend does not hold for those experiments with capped total doses—lack of, small, and large dose penalties all led to similar dosing regimens (Supplementary Results). The trade-off induced by the dose penalties in such cases does not encourage the agent to forgo any of its already limited doses; the number of doses is so limited that reducing them further would likely incur more penalty than reducing MTD through their administration. This is not the case for all of the non-capped total dose experiments, indicating that there are diminishing returns on MTD reduction as the number of treatment administrations increases. Similarly, the observed increase in MTD under learned policies for capped total treatment PCV experiments may be due to the already small number of treatments performed in the expert policy. It may also be due to the shorter dosing window or differences in treatment efficacy and effectiveness. More research can be done to investigate the effect of changing drug timings and relative concentrations within treatment cycles. Further research can additionally investigate the points at which a decrease in the amount of potentially harmful treatment outweighs the potential benefits of that treatment, as such cutoffs often require clinical judgment and are not currently well-characterized.

We also investigated the impact of the different dose penalties in identical patient cohorts. Though overall average MTD was reduced, policies learned under small and large dose penalties occasionally resulted in an increase in MTD for a few individual patients. In five experiments with small or large dose penalties, between one and 14 patients experienced MTD increases who would not have under the corresponding expert policy (Supplementary Table 2). For example, in the trial-based and fixed dose concentration experiment using

TMZ with a large dose penalty, a single patient experienced an MTD increase of 6.22% rather than the 70% decrease that the expert policy would have achieved (Supplementary Table 2). Further work, through MTD trajectory stratification, can lead to effective dosing regimens for all patients. Four of the five experiments with patients who experienced MTD increases were trial-based, indicating that individual patients tended to fare better when an agent could administer individualized treatment regimens (Supplementary Table 2). When an agent must apply a given dose to all patients, it faced additional trade-offs when helping a few patients by incurring penalties across all patients, possibly causing it to dose more conservatively. The average percentage reduction in MTD between equivalent patient-based and trial-based experiments tends to be comparable for TMZ and greater in the case of PCV experiments, possibly attributable to differences in allowed treatment amounts or drug efficacy and effectiveness (Table 1). Action-derived penalties implicitly create a relationship between the existing reward, whether outcome- or state-derived, and the action. They may therefore be less useful in domains with no known or hypothesized relationship between reward and state representation or actions.

Whereas Zhao et al. (2011) used expert policies as building blocks for at most three decision points in the presence of external outcomes, we use state- and action-derived rewards to learn policies for dosing regimens. The previous work learned optimal selections of expert dosing regimens for first- and second-line chemotherapy treatments, and our approach learns a granular association between individual treatment cycles and MTD reduction that can devise treatment regimens for individuals and cohorts of patients. Our approach also incorporates different notions of optimality in the form of various dose penalties without requiring final patient outcomes.

5.3. Summary

Reinforcement learning can be used to learn dosing policies for clinical trials within given trial parameters and agent flexibility parameters. We report the use of action-derived rewards as a proxy for clinical outcomes to impose additional constraints on a state-derived reward. Specifically, we contribute an RL formulation for learning dosing regimens for TMZ and PCV clinical trials as well as a novel incorporation of action-derived penalties. Our framework learns new TMZ and PCV dosing regimens for various trade-offs between potential action harm and outcome proxies carried by the state. The incorporation of action-derived rewards affords increased learning flexibility for this and other domains in the absence of outcomes. We anticipate that action-derived rewards may be useful even when learning from real-world retrospective clinical trial data that may or may not include patient outcome measures like overall survival by providing priors on harmful and beneficial actions. Action-derived rewards allow agents to more fully explore state and action trade-offs in scenarios unstructured by external rewards that would otherwise be unsuitable for RL approaches.

Acknowledgments

The authors would like to thank Aman Rana for technical assistance.

References

Christopher Amato, George Konidaris, Ariel Anders, Gabriel Cruz, Jonathan P How, and Leslie P Kaelbling. Policy search for multi-robot coordination under uncertainty. *The International Journal of Robotics Research*, 35(14):1760–1778, 2016.

Natalie Cook, Aaron R Hansen, Lillian L Siu, and Albiruni R Abdul Razak. Early phase clinical trials to identify optimal dosing and safety. *Molecular Oncology*, 9(5):997–1007, 2015.

H Gurney. How to calculate the dose of chemotherapy. *British Journal of Cancer*, 86(8): 1297, 2002.

Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4:237–285, 1996.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.

Shamim Nemati, Mohammad M Ghassemi, and Gari D Clifford. Optimal medication dosing from suboptimal clinical examples: A deep reinforcement learning approach. In *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the*, pages 2978–2981. IEEE, 2016.

Andrew Y Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *ICML*, volume 99, pages 278–287, 1999.

Andrew Y Ng, Stuart J Russell, et al. Algorithms for inverse reinforcement learning. In *ICML*, pages 663–670, 2000.

Matthieu Peyre, Stephanie Cartalat-Carel, David Meyronet, Damien Ricard, Anne Jouvet, Johan Pallud, Karima Mokhtari, Jacques Guyotat, Emmanuel Jouanneau, Marie-Pierre Sunyach, et al. Prolonged response without prolonged chemotherapy: a lesson from pcv chemotherapy in low-grade gliomas. *Neuro-Oncology*, 12(10):1078–1082, 2010.

Niranjani Prasad, Li-Fang Cheng, Corey Chivers, Michael Draugelis, and Barbara E Engelhardt. A reinforcement learning approach to weaning of mechanical ventilation in intensive care units. *arXiv preprint arXiv:1704.06300*, 2017.

Aniruddh Raghu, Matthieu Komorowski, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Continuous state-space models for optimal sepsis treatment: a deep reinforcement learning approach. In *Machine Learning for Healthcare Conference*, pages 147–163, 2017.

Benjamin Ribba, Gentian Kaloshi, Mathieu Peyre, Damien Ricard, Vincent Calvez, Michel Tod, Branka Čajavec-Bernard, Ahmed Idbaih, Dimitri Psimaras, Linda Dainese, et al. A tumor growth inhibition model for low-grade glioma treated with chemotherapy or radiotherapy. *Clinical Cancer Research*, 18(18):5071–5080, 2012.

Damien Ricard, Gentian Kaloshi, Alexandra Amiel-Benouaich, Julie Lejeune, Yannick Marie, Emmanuel Mandonnet, Michèle Kujas, Karima Mokhtari, Sophie Taillibert, Florence Laigle-Donadey, et al. Dynamic history of low-grade gliomas before and after temozolomide treatment. *Annals of Neurology*, 61(5):484–490, 2007.

Mikael L Rinne and Patrick Y Wen. Treating anaplastic oligodendrogliomas and who grade 2 gliomas: Pcv or temozlomide? the case for temozolomide. *Oncology*, 29(4):265–265, 2015.

R Scmidt. Dose-finding studies in clinical drug development. *European Journal of Clinical Pharmacology*, 34(1):15–19, 1988.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An introduction*. MIT Press, 1998.

Hado Van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. In *AAAI*, volume 16, pages 2094–2100, 2016.

Christian Wirth, Riad Akrour, Gerhard Neumann, Johannes Fürnkranz, et al. A survey of preference-based reinforcement learning methods. *Journal of Machine Learning Research*, 18(136):1–46, 2017.

Chi Heem Wong, Kien Wei Siah, and Andrew W Lo. Estimation of clinical trial success rates and related parameters. *Biostatistics*, 2018.

Yufan Zhao, Donglin Zeng, Mark A Socinski, and Michael R Kosorok. Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer. *Biometrics*, 67(4):1422–1433, 2011.

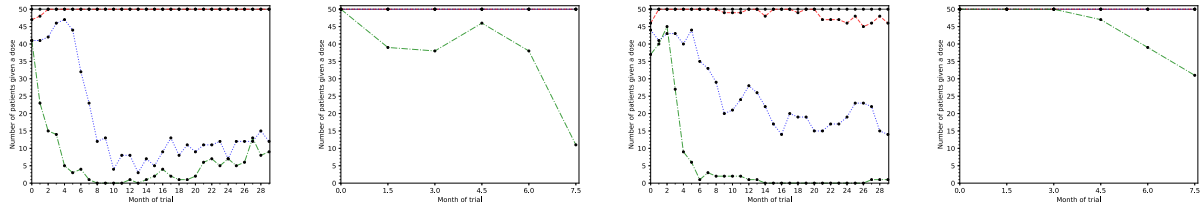
Supplementary Tables and Figures

Trial parameters			Capped total dose			Expert policy	
Drug	Conc.	Type	No. patients	Learned change	Expert change	No. patients	Expert change
TMZ	Fixed	Patient	4	30.75% ± 21.84%	-57.78% ± 9.62%	0	-
TMZ	Fixed	Trial	1	59.00%	-57.08%	0	-
PCV	Fixed	Patient	29	35.66% ± 30.06%	-19.21% ± 11.27%	2	9.39% ± 4.13%
PCV	Fixed	Trial	50	26.56% ± 24.56%	-22.55 ± 7.49	0	-

Supplementary Table 1. Numbers of patients with increases in mean tumor diameter (MTD) and their average percentage increases after application of dosing regimens that had a cap on the total numbers of TMZ (10) or PCV (2). For each penalty, the average change in MTD under the corresponding expert policy for each group of patients is also reported. Fixed: the agent could give doses with the unit concentration. Conc: concentration.

Trial parameters			No penalty			Small penalty			Large penalty			Expert policy	
Drug	Conc.	Type	No. patients	Learned change	Expert change	No. patients	Learned change	Expert change	No. patients	Learned change	Expert change	No. patients	Expert change
TMZ	Fixed	Patient	0	-	-	0	-	-	0	-	-	0	-
TMZ	Fixed	Trial	0	-	-	0	-	-	1	6%	70.5%	0	-
TMZ	Variable	Patient	0	-	-	5	84.4% ± 52.46%	-70.14% ± 15.76%	14	40.79% ± 47.69%	-58.61% ± 11.64%	0	-
TMZ	Variable	Trial	0	-	-	0	-	-	0	-	-	0	-
PCV	Fixed	Patient	0	-	-	0	-	-	0	-	-	0	-
PCV	Fixed	Trial	3	8.67% ± 11.59%	8.82% ± 11.24%	5	18.4% ± 24.75%	0.99% ± 13.49%	12	16.33% ± 23.87	-9.28% ± 13.73	3	8.82% ± 11.24%
PCV	Variable	Patient	0	-	-	0	-	-	0	-	-	0	-
PCV	Variable	Trial	2	4.50% ± 4.95%	4.48% ± 4.46%	2	6.00% ± 5.66%	4.48% ± 4.46%	2	15.50% ± 10.61%	4.48% ± 4.46%	2	4.48% ± 4.46%

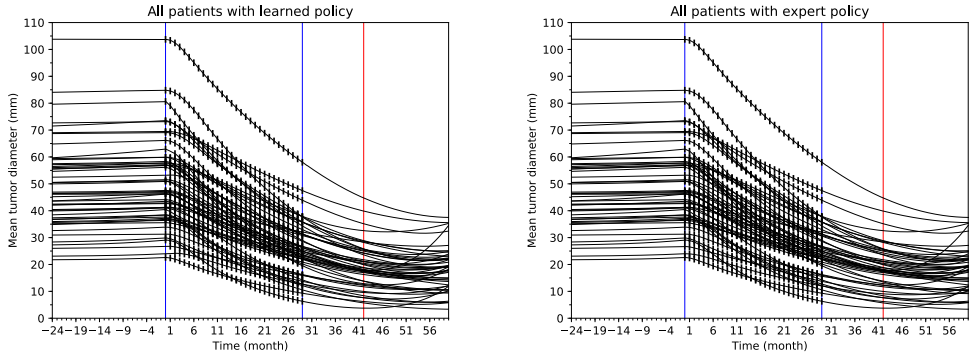
Supplementary Table 2. Numbers of patients with increases in mean tumor diameter (MTD) and their average percentage increases after application of dosing regimens. For each penalty, the average change in MTD under the corresponding expert policy for each group of patients is also reported. Fixed: the agent could give doses with the unit concentration. Variable: the agent could give doses at 25%, 50%, 75%, and 100% of the unit concentration. Conc: concentration.



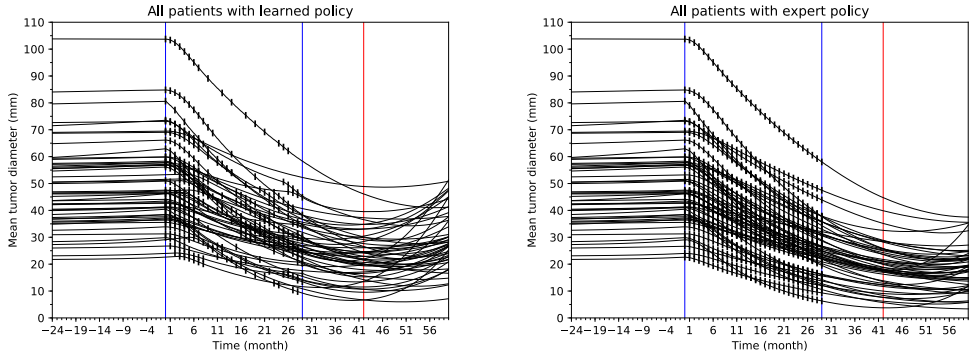
Supplementary Figure 1. Number of patients given a dose per month by policies learned under different dose penalties in various trial parameter configurations. All policies optimized for mean tumor diameter (MTD) reduction and could treat individual simulated patients separately. In the presence of dose penalties, the agent learns in all four trial parameter scenarios that earlier doses tend to lead to larger decreases in MTD. Line color corresponds to dose penalty: black, expert; red, none; blue, small; green, large. Left-to-right: simulated TMZ trial, fixed dose; simulated PCV trial, fixed dose; simulated TMZ trial, variable dose; simulated PCV trial, variable dose.

Supplementary Results: patient-resolution data

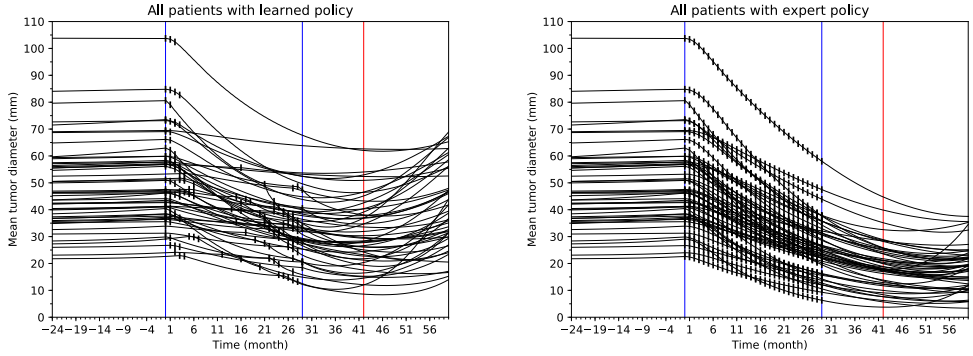
TMZ, patient-based, fixed dose concentration.



TMZ, patient-based, fixed dose concentration, no dose penalty. Left: Application of the learned policy. Right: Application of the expert policy.

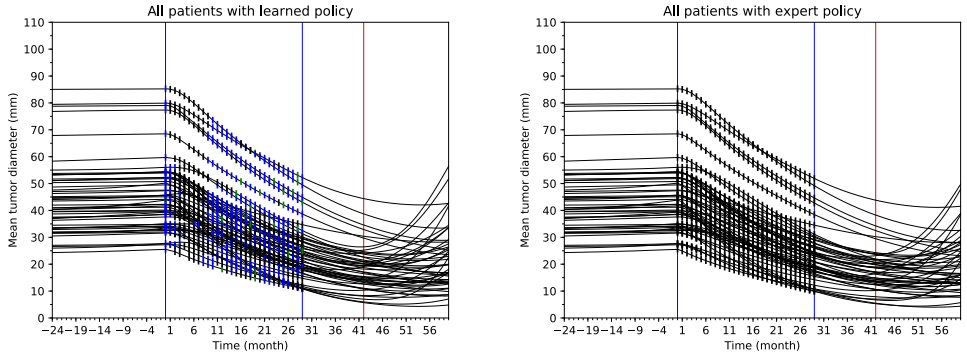


TMZ, patient-based, fixed dose concentration, small dose penalty. Left: Application of the learned policy. Right: Application of the expert policy.

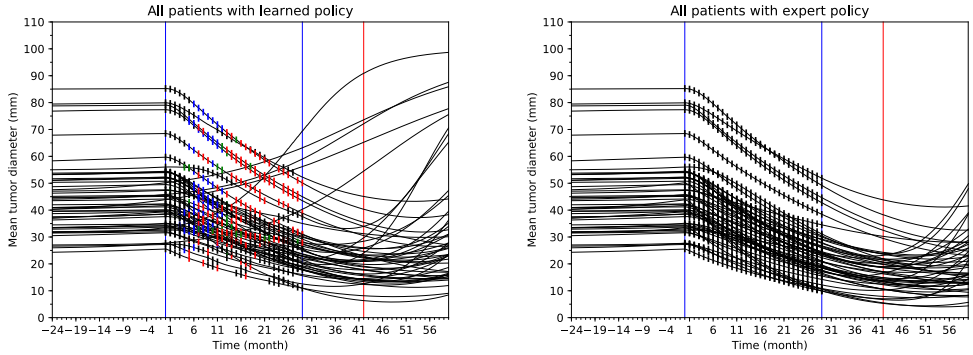


TMZ, patient-based, fixed dose concentration, large dose penalty. Left: Application of the learned policy. Right: Application of the expert policy.

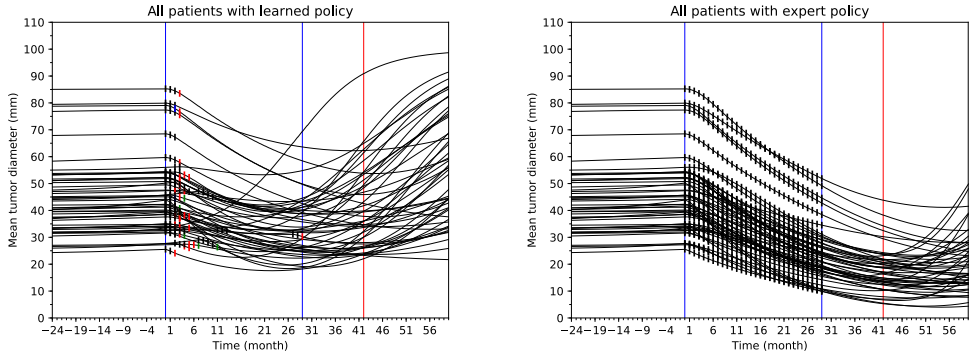
TMZ, patient-based, variable dose concentration.



TMZ, patient-based, variable dose concentration, no dose penalty. Left: Application of the learned policy. Right: Application of the expert policy. The agent could administer doses at 25% (red), 50% (green), 75% (blue), and 100% (black) of the unit concentration.

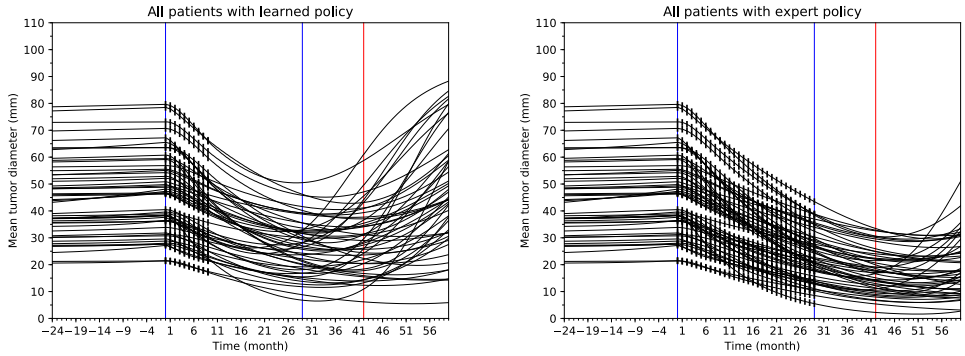


TMZ, patient-based, variable dose concentration, small dose penalty. Left: Application of the learned policy. Right: Application of the expert policy. The agent could administer doses at 25% (red), 50% (green), 75% (blue), and 100% (black) of the unit concentration.

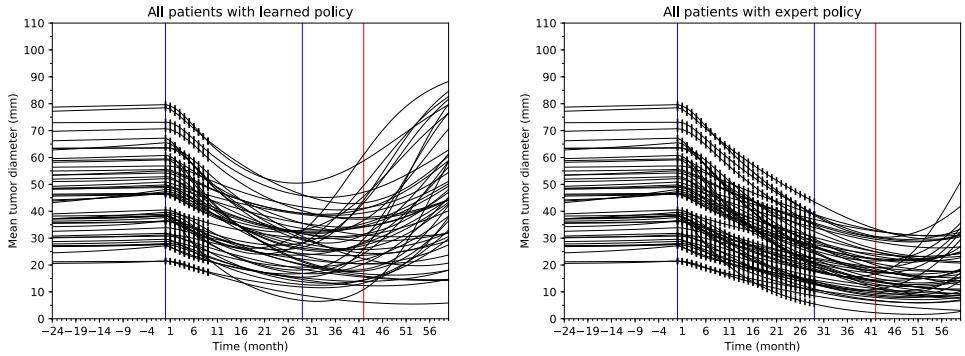


TMZ, patient-based, variable dose concentration, large dose penalty. Left: Application of the learned policy. Right: Application of the expert policy. The agent could administer doses at 25% (red), 50% (green), 75% (blue), and 100% (black) of the unit concentration.

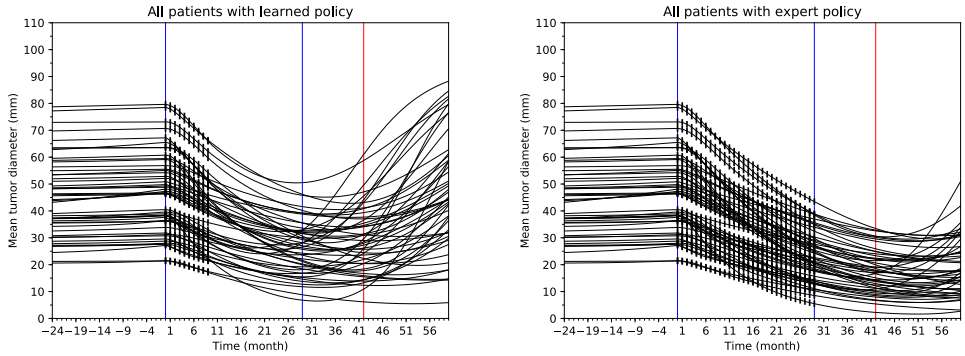
TMZ, patient-based, fixed dose concentration, capped maximum number of treatment administrations.



TMZ, patient-based, fixed dose concentration, no dose penalty, capped total dose of 10 units per patient. Left: Application of the learned policy. Right: Application of the expert policy.

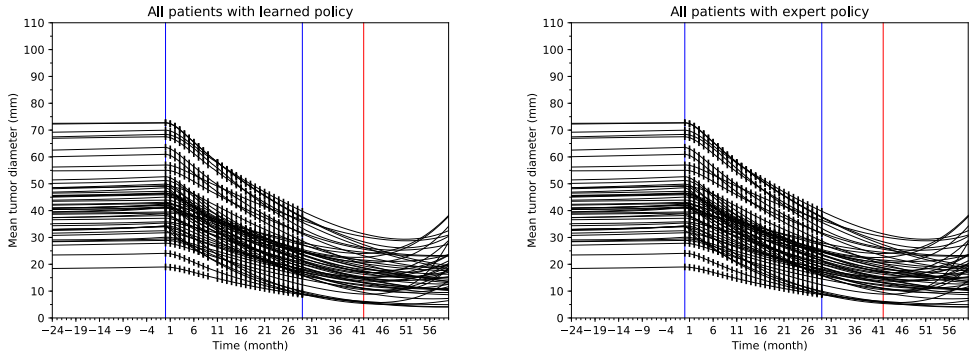


TMZ, patient-based, fixed dose concentration, small dose penalty, capped total dose of 10 units per patient. Left: Application of the learned policy. Right: Application of the expert policy.

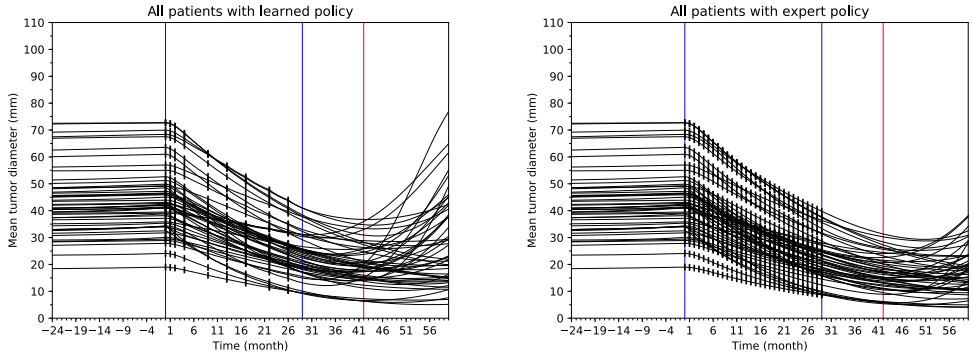


TMZ, patient-based, fixed dose concentration, large dose penalty, capped total dose of 10 units per patient. Left: Application of the learned policy. Right: Application of the expert policy.

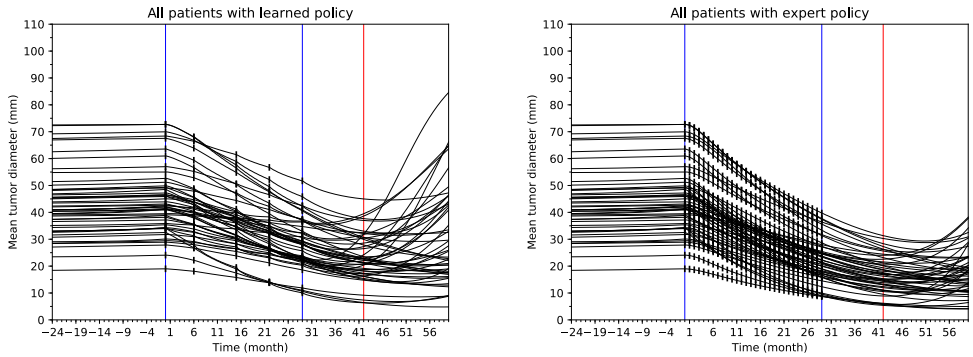
TMZ, trial-based, fixed dose concentration.



TMZ, trial-based, fixed dose concentration, no dose penalty. Left: Application of the learned policy. Right: Application of the expert policy.

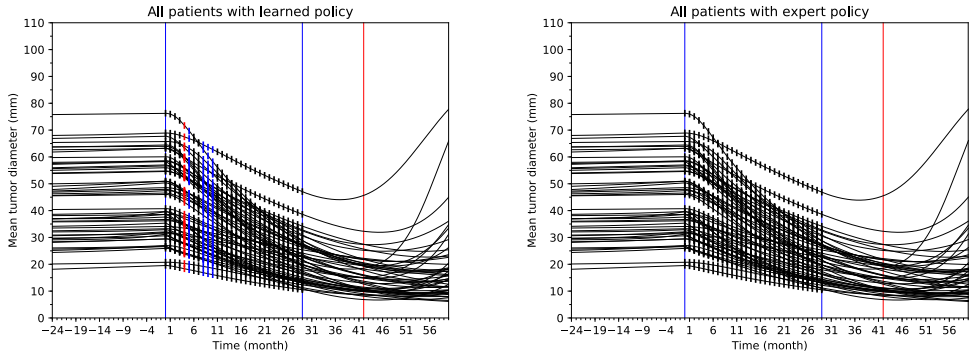


TMZ, trial-based, fixed dose concentration, small dose penalty. Left: Application of the learned policy. Right: Application of the expert policy.

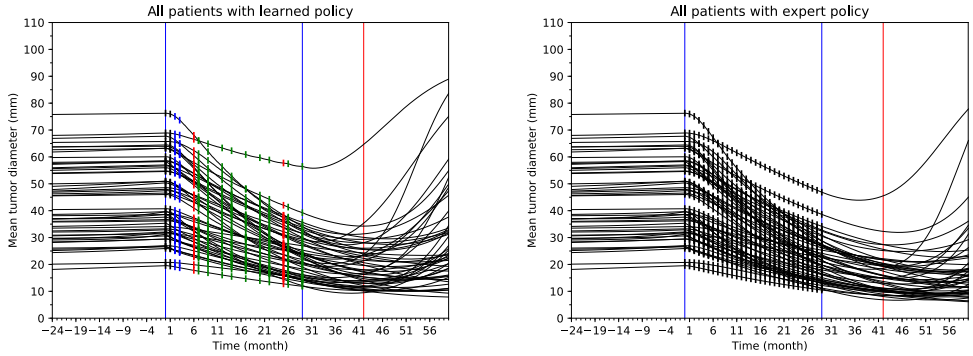


TMZ, trial-based, fixed dose concentration, large dose penalty. Left: Application of the learned policy. Right: Application of the expert policy.

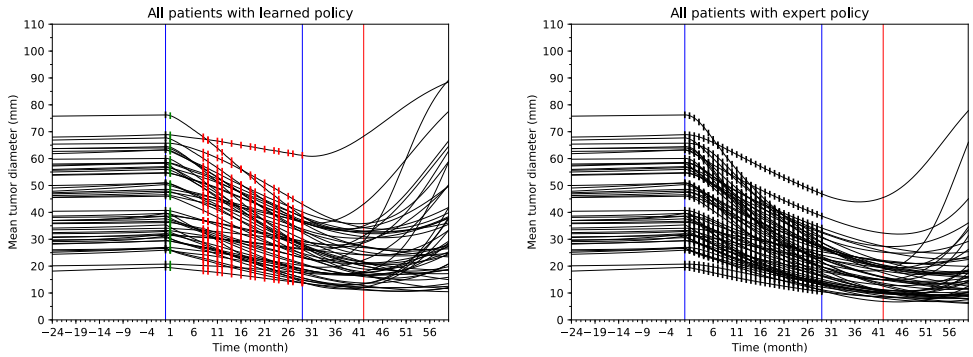
TMZ, trial-based, variable dose concentration.



TMZ, trial-based, variable dose concentration, no dose penalty. Left: Application of the learned policy. Right: Application of the expert policy. The agent could administer doses at 25% (red), 50% (green), 75% (blue), and 100% (black) of the unit concentration.

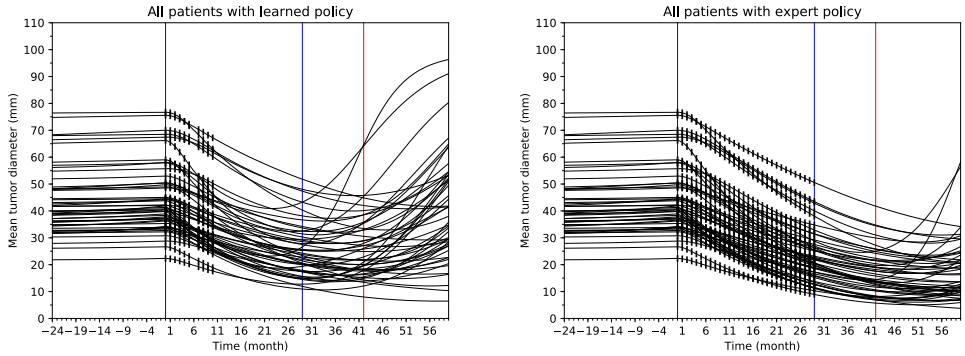


TMZ, trial-based, variable dose concentration, small dose penalty. Left: Application of the learned policy. Right: Application of the expert policy. The agent could administer doses at 25% (red), 50% (green), 75% (blue), and 100% (black) of the unit concentration.

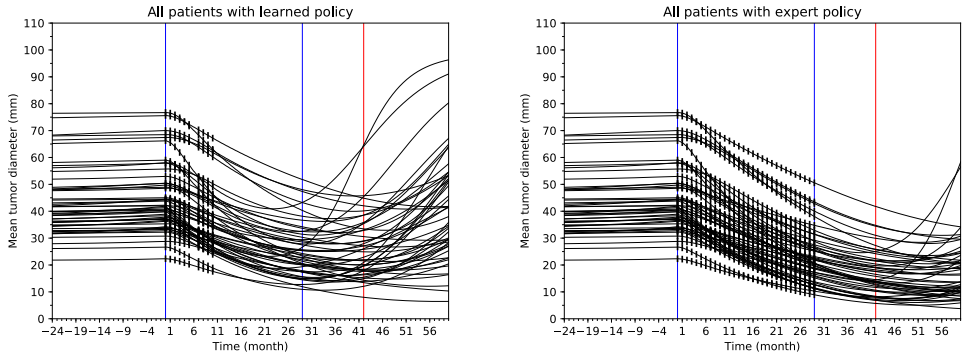


TMZ, trial-based, variable dose concentration, large dose penalty. Left: Application of the learned policy. Right: Application of the expert policy. The agent could administer doses at 25% (red), 50% (green), 75% (blue), and 100% (black) of the unit concentration.

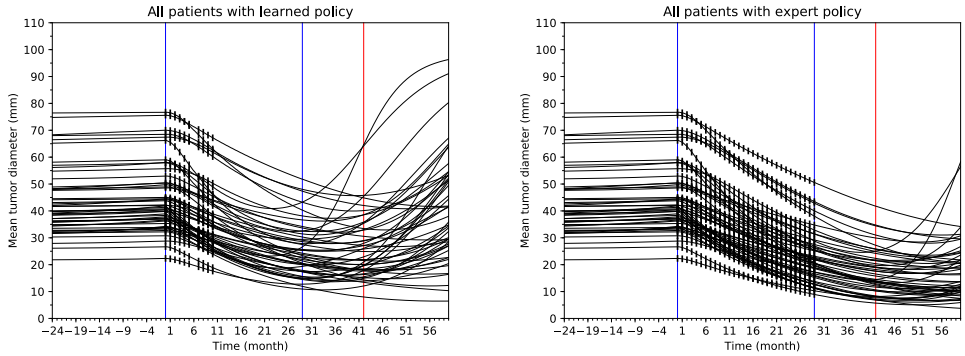
TMZ, trial-based, fixed dose concentration, capped maximum number of treatment administrations.



TMZ, trial-based, fixed dose concentration, no dose penalty, capped total dose of 10 units per patient. Left: Application of the learned policy. Right: Application of the expert policy.

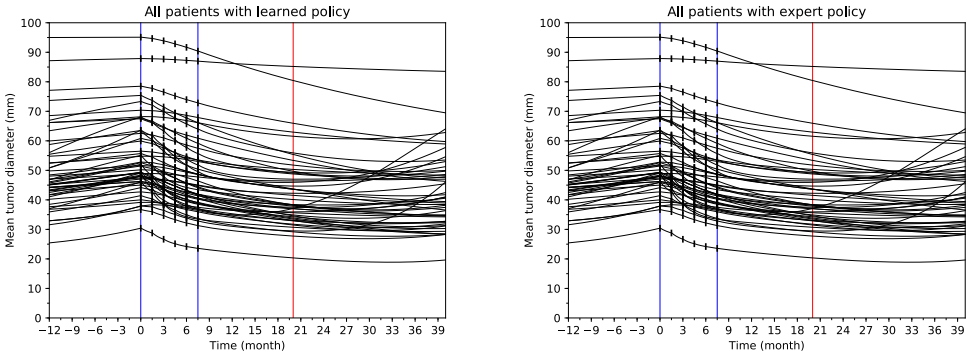


TMZ, trial-based, fixed dose concentration, small dose penalty, capped total dose of 10 units per patient. Left: Application of the learned policy. Right: Application of the expert policy.

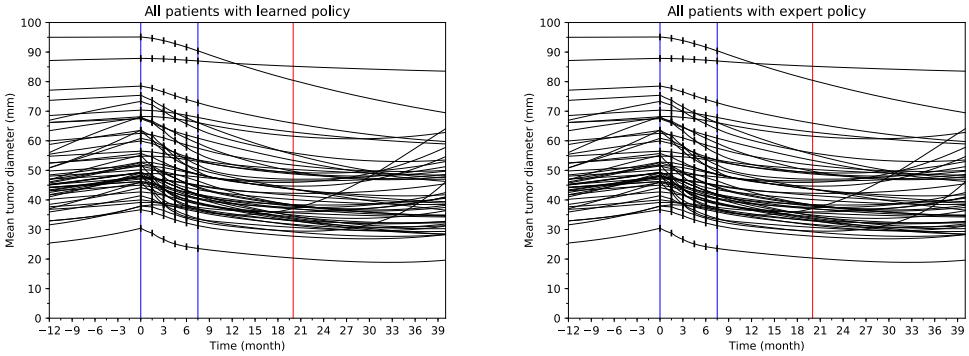


TMZ, trial-based, fixed dose concentration, large dose penalty, capped total dose of 10 units per patient. Left: Application of the learned policy. Right: Application of the expert policy.

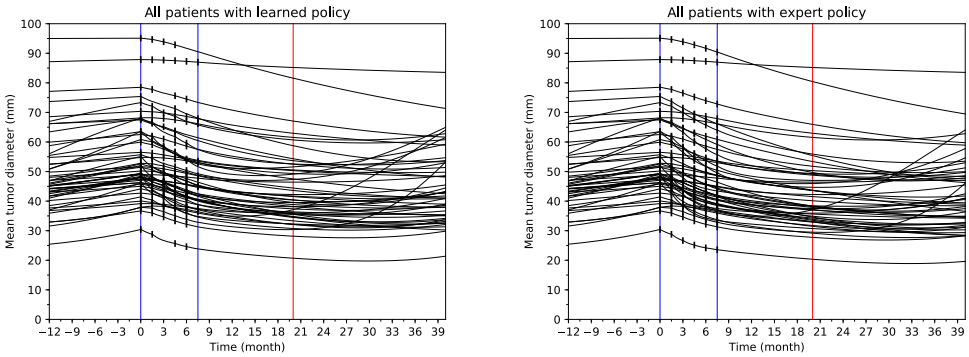
PCV, patient-based, fixed dose concentration.



PCV, patient-based, fixed dose concentration, no dose penalty. Left: Application of the learned policy. Right: Application of the expert policy.

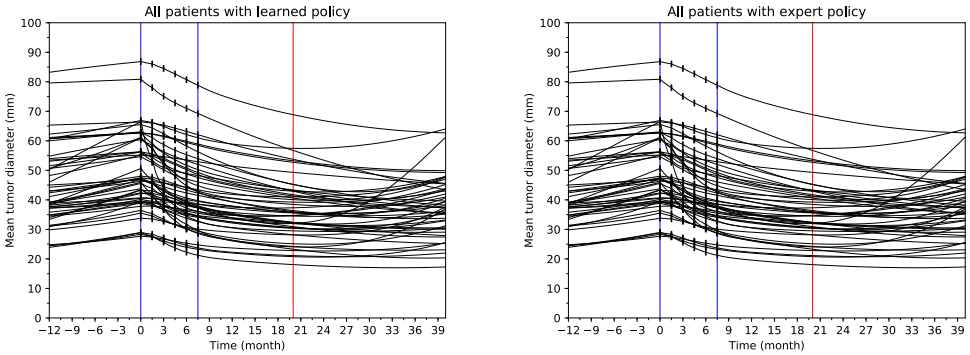


PCV, patient-based, fixed dose concentration, small dose penalty. Left: Application of the learned policy. Right: Application of the expert policy.

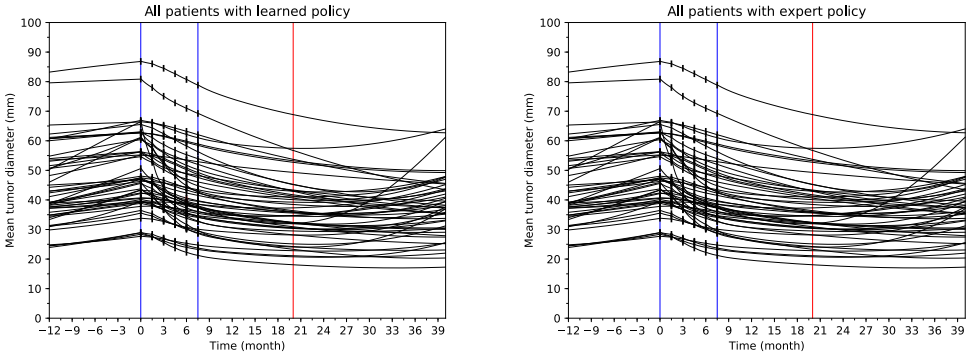


PCV, patient-based, fixed dose concentration, large dose penalty. Left: Application of the learned policy. Right: Application of the expert policy.

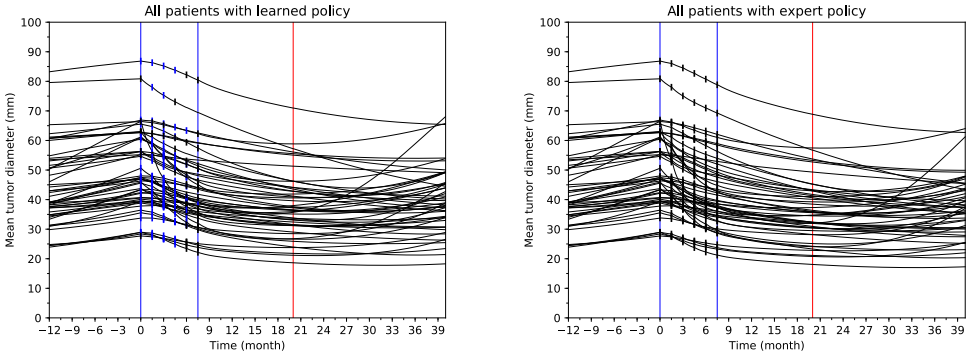
PCV, patient-based, variable dose concentration.



PCV, patient-based, variable dose concentration, no dose penalty. Left: Application of the learned policy. Right: Application of the expert policy. The agent could administer doses at 25% (red), 50% (green), 75% (blue), and 100% (black) of the unit concentration.

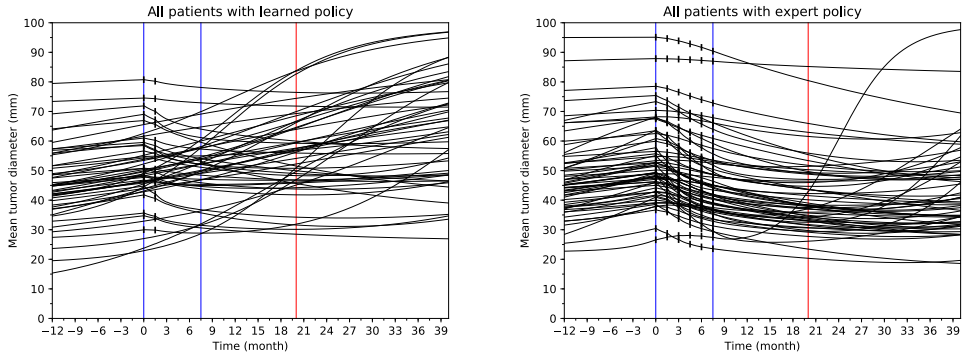


PCV, patient-based, variable dose concentration, small dose penalty. Left: Application of the learned policy. Right: Application of the expert policy. The agent could administer doses at 25% (red), 50% (green), 75% (blue), and 100% (black) of the unit concentration.

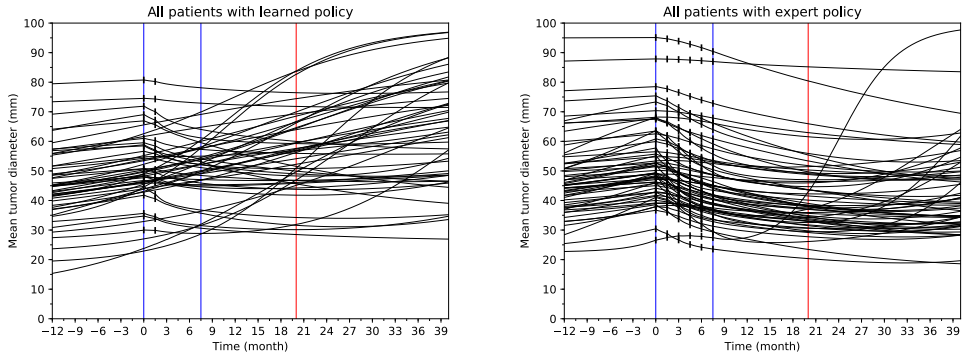


PCV, patient-based, variable dose concentration, large dose penalty. Left: Application of the learned policy. Right: Application of the expert policy. The agent could administer doses at 25% (red), 50% (green), 75% (blue), and 100% (black) of the unit concentration.

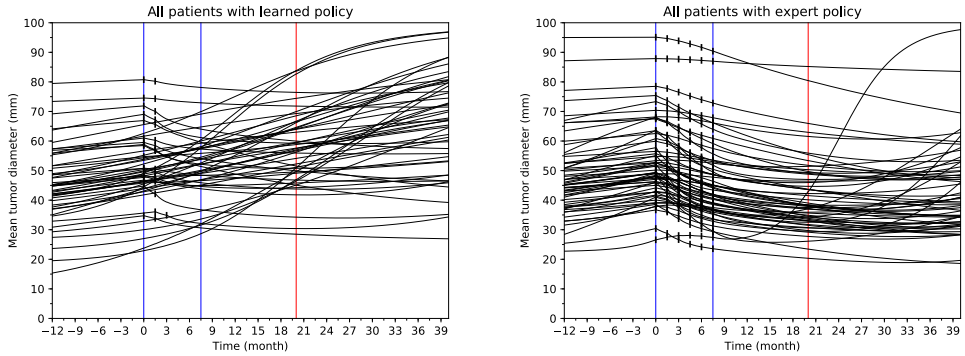
PCV, patient-based, fixed dose concentration, capped maximum number of treatment administrations.



PCV, patient-based, fixed dose concentration, no dose penalty, capped total dose of 2 units per patient. Left: Application of the learned policy. Right: Application of the expert policy.

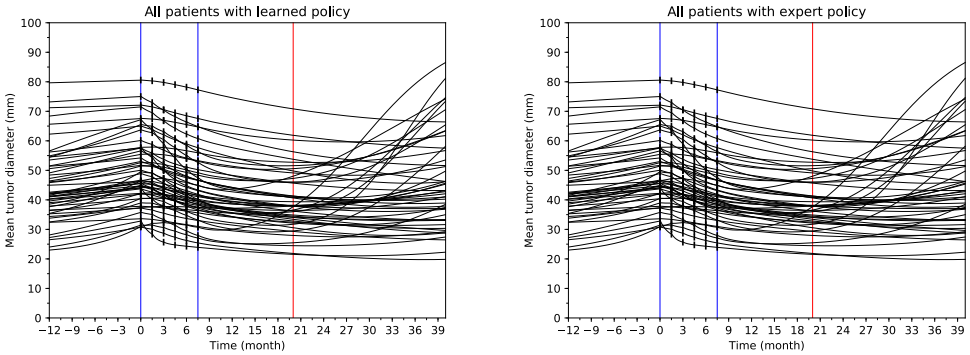


PCV, patient-based, fixed dose concentration, small dose penalty, capped total dose of 2 units per patient. Left: Application of the learned policy. Right: Application of the expert policy.

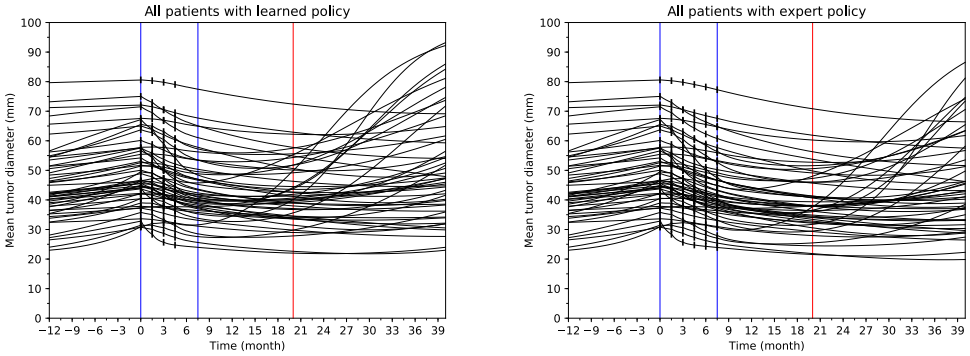


PCV, patient-based, fixed dose concentration, large dose penalty, capped total dose of 2 units per patient. Left: Application of the learned policy. Right: Application of the expert policy.

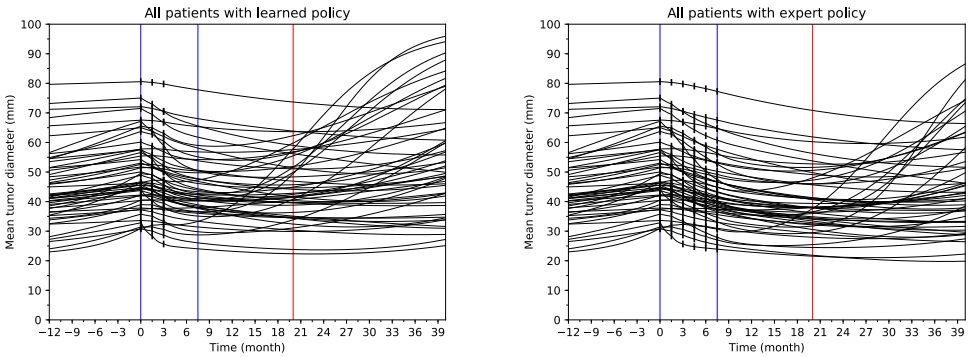
PCV, trial-based, fixed dose concentration.



PCV, trial-based, fixed dose concentration, no dose penalty. Left: Application of the learned policy. Right: Application of the expert policy.

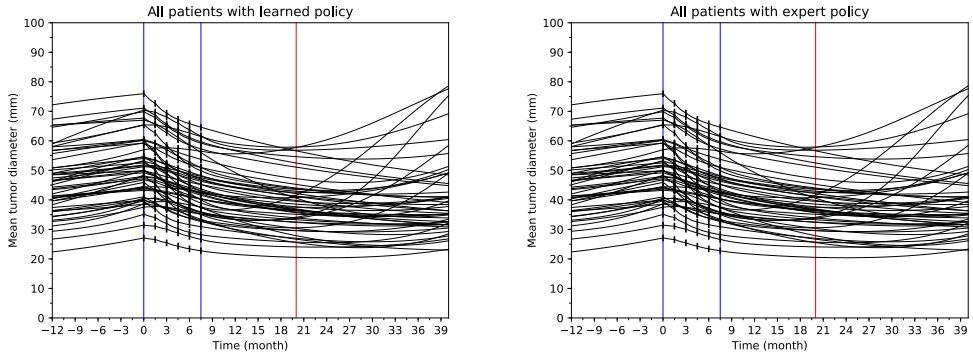


PCV, trial-based, fixed dose concentration, small dose penalty. Left: Application of the learned policy. Right: Application of the expert policy.

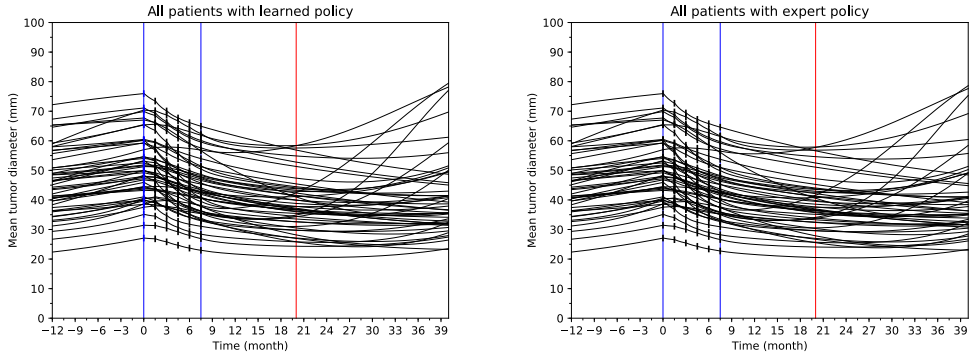


PCV, trial-based, fixed dose concentration, large dose penalty. Left: Application of the learned policy. Right: Application of the expert policy.

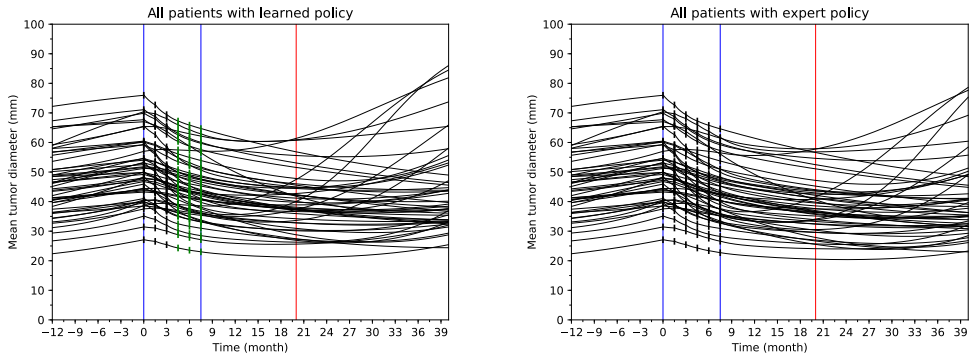
PCV, trial-based, variable dose concentration.



PCV, trial-based, variable dose concentration, no dose penalty. Left: Application of the learned policy. Right: Application of the expert policy.

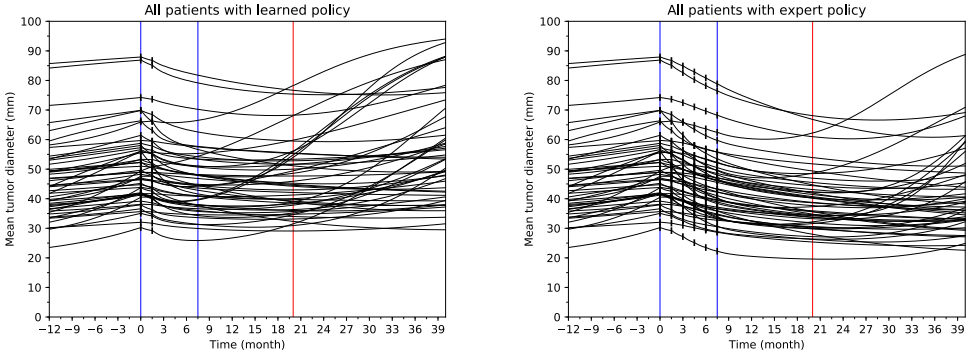


PCV, trial-based, variable dose concentration, small dose penalty. Left: Application of the learned policy. Right: Application of the expert policy.

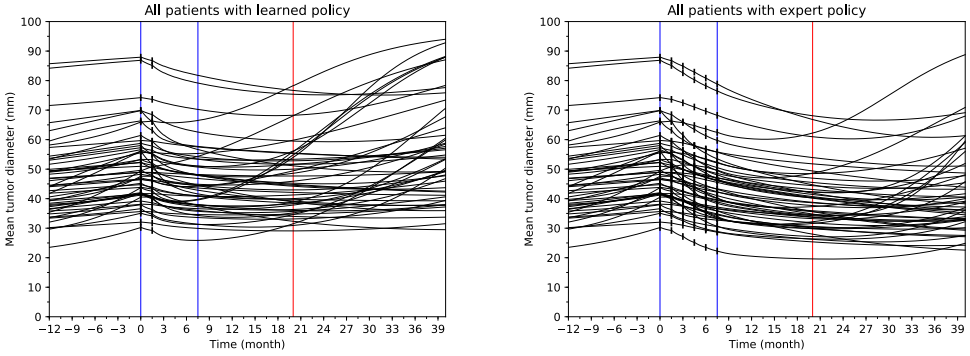


PCV, trial-based, variable dose concentration, large dose penalty. Left: Application of the learned policy. Right: Application of the expert policy.

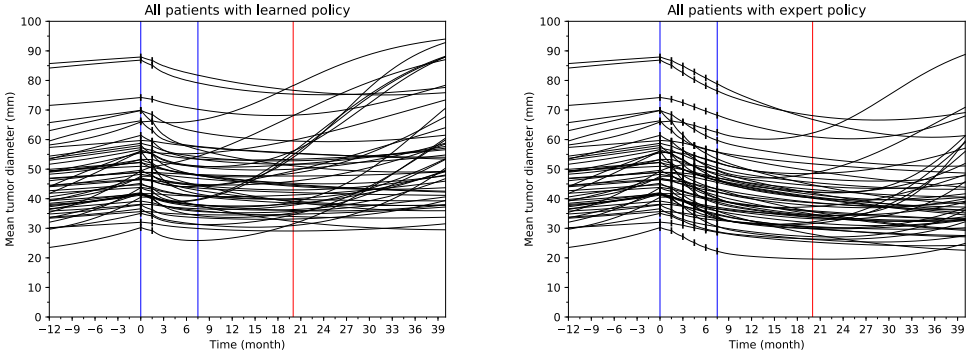
PCV, trial-based, fixed dose concentration, capped maximum number of treatment administrations.



PCV, trial-based, fixed dose concentration, no dose penalty, capped total dose of 2 units per patient. Left: Application of the learned policy. Right: Application of the expert policy.



PCV, trial-based, fixed dose concentration, small dose penalty, capped total dose of 2 units per patient. Left: Application of the learned policy. Right: Application of the expert policy.



PCV, trial-based, fixed dose concentration, large dose penalty, capped total dose of 2 units per patient. Left: Application of the learned policy. Right: Application of the expert policy.

TMZ, patient-based, fixed dose concentration, no dose penalty.

Patient number	MTD at month 0 (mm)	MTD at final observation after learned policy (mm)	MTD at final observation after expert policy (mm)	Dose at month of trial																																
				0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
0	29.10	3.75	3.75	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	26.73	12.55	12.31	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	41.04	18.73	18.73	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
3	41.25	8.37	8.37	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
4	58.33	17.16	17.16	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
5	50.93	20.11	20.11	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
6	56.94	32.57	32.57	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
7	38.45	5.69	5.69	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
8	37.18	7.65	7.65	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
9	31.34	12.22	12.22	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
10	56.15	18.86	18.86	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
11	56.85	12.88	12.88	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
12	84.83	25.81	25.81	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
13	46.42	21.02	21.02	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
14	66.20	19.89	19.89	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
15	36.81	16.25	16.25	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
16	57.76	28.64	28.64	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
17	24.00	9.57	9.07	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
18	46.30	13.81	13.81	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
19	73.19	25.70	25.70	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
20	40.45	17.71	17.71	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
21	57.21	24.27	24.27	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
22	44.15	17.49	17.49	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
23	42.67	28.35	28.35	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
24	80.66	23.55	23.55	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
25	57.87	21.16	21.16	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
26	33.90	11.69	11.69	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
27	59.95	25.64	25.64	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
28	51.58	24.96	24.96	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
29	69.55	39.86	39.86	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
30	22.62	6.21	6.21	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
31	46.23	17.69	17.69	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
32	47.24	25.26	25.26	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
33	73.58	35.15	35.15	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
34	69.12	28.85	28.85	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
35	53.27	24.18	23.66	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
36	35.59	21.15	21.15	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
37	46.81	18.17	18.17	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
38	51.11	18.01	18.01	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
39	37.64	17.44	17.44	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
40	47.53	17.69	17.69	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
41	36.44	18.01	18.01	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
42	103.75	44.73	44.73	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
43	29.93	9.97	9.97	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
44	42.94	16.95	16.95	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
45	43.56	21.74	21.74	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
46	38.44	7.61	7.61	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
47	59.89	28.20	28.20	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
48	62.92	13.78	13.47	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
49	41.33	20.45	20.45	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

TMZ, patient-based, fixed dose concentration, no dose penalty. Mean tumor diameter (MTD) at the beginning of treatment and at the final observation when applying learned and expert policies. Doses applied by the learned policy to each patient.

TMZ, patient-based, fixed dose concentration, small dose penalty.

Patient number	MTD at month 0 (mm)	MTD at final observation after learned policy (mm)	MTD at final observation after expert policy (mm)	Dose at month of trial																																	
				0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
0	29.10	6.68	3.75	0	0	0	0	0	1	1	1	1	1	1	0	0	1	0	1	0	1	0	1	0	0	0	1	1	0	0	1	1	0
1	26.73	17.10	12.31	0	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	41.04	19.95	18.73	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0
3	41.25	10.65	8.37	1	1	1	1	1	1	1	1	0	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	1	0	0	0
4	58.33	19.38	17.16	1	1	1	1	1	1	1	0	0	0	0	0	0	0	1	1	1	0	0	1	1	0	0	0	0	1	0	0	0	0
5	50.93	21.89	20.11	0	0	0	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	0	1	1	1
6	56.94	39.54	32.57	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1
7	38.45	6.68	5.69	1	1	1	1	1	1	1	0	0	0	1	0	1	0	1	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0
8	37.18	9.55	7.65	1	1	1	1	1	1	1	1	0	0	0	1	1	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	1	1
9	31.34	19.70	12.22	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	56.15	20.07	18.86	1	1	1	1	1	1	0	1	0	0	0	1	0	0	0	0	0	0	0	1	1	1	1	0	0	1	0	0	0	0
11	56.85	15.26	12.88	1	1	1	1	1	1	1	1	1	1	0	0	1	1	0	0	1	1	1	0	0	1	0	1	0	1	0	0	0	1
12	84.83	26.55	25.81	1	1	1	1	1	1	1	1	1	1	1	0	0	0	1	0	0	0	0	1	0	0	0	1	1	1	1	1	0	1
13	46.42	21.84	21.02	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	0	1	1	0	0
14	66.20	21.44	19.89	1	1	1	1	1	1	1	1	0	1	0	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	1	1	
15	36.81	24.44	16.25	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	57.76	37.27	28.64	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	24.00	11.50	9.07	0	0	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	0
18	46.30	15.64	13.81	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	1	0	0	0	0	0	0
19	73.19	29.12	25.70	1	1	1	1	1	1	1	1	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0
20	40.45	22.90	17.71	1	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	57.21	26.31	24.27	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	1	1	1	0	0	0	0	0	0	0
22	44.15	19.98	17.49	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	42.67	29.85	28.35	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
24	80.66	24.32	23.55	1	0	1	0	1	0	0	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	1	1	1	0	1	0	1	0
25	57.87	22.93	21.16	1	1	1	1	1	1	1	0	0	0	0	0	1	0	0	0	0	0	1	1	1	1	0	0	0	1	1	0	0	0
26	33.90	16.58	11.69	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
27	59.95	27.64	25.64	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	1	1	1	1	1	0	0	1	1	0	0	0	1	0	0
28	51.58	33.09	24.96	0	0	0	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
29	69.55	48.88	39.86	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30	22.62	13.64	6.21	0	0	0	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
31	46.23	28.41	17.69	0	0	0	1	1	1	1	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
32	47.24	33.03	25.26	0	0	0	0	0	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
33	73.58	36.54	35.15	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	
34	69.12	30.14	28.85	1	1	1	1	1	1	1	1	0	1	0	0	0	0	1	0	0	0	0	0	1	1	1	1	0	0	1	1	0	0
35	53.27	34.81	23.66	0	0	0	0	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
36	35.59	26.10	21.15	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
37	46.81	20.65	18.17	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	1	0	0	0	0	0
38	51.11	24.69	18.01	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
39	37.64	24.28	17.44	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
40	47.53	21.93	17.69	1	1	1	1	1	1	1	1	0	1	0	0	0	0	1	0	0	0	0	1	1	1	1	0	0	1	1	0	0	0
41	36.44	19.76	18.01	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	
42	103.75	46.10	44.73	1	1	1	1	1	1	1	1	0	1	0	0	1	0	1	0	0	1	0	1	0	1	0	0	1	1	1	0	0	0
43	29.93	13.44	9.97	1	1	1	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
44	42.94	17.91	16.95	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	1	1	0	0	0	1	1	1
45	43.56	24.12	21.74	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
46	38.44	10.70	7.61	1	1	1	1	1	1	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0
47	59.89	34.81	28.20	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
48	62.92	14.23	13.47	1	1	1	1	1	1	0	1	1	0	0	1	1	0	0	0	1	1	0	0	0	0	0	1	0	0	0	0	1	0
49	41.33	30.44	20.45	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

TMZ, patient-based, fixed dose concentration, small dose penalty. Mean tumor diameter (MTD) at the beginning of treatment and at the final observation when applying learned and expert policies.

TMZ, patient-based, fixed dose concentration, large dose penalty.

Patient number	MTD at month 0 (mm)	MTD at final observation after learned policy (mm)	MTD at final observation after expert policy (mm)	Dose at month of trial																																
				0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
0	29.10	12.02	3.75	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0
1	26.73	15.01	12.31	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
2	41.04	21.23	18.73	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	1	1	0	0	
3	41.25	29.38	8.37	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	
4	58.33	19.36	17.16	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	0	1	0	0	
5	50.93	39.15	20.11	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
6	56.94	54.03	32.57	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	
7	38.45	14.54	5.69	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
8	37.18	16.47	7.65	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	0	0	0	0	
9	31.34	17.56	12.22	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	
10	56.15	31.91	18.86	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
11	56.85	21.76	12.88	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	0	0	0	1	0	0	0	
12	84.83	46.45	25.81	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
13	46.42	22.75	21.02	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	0	1	1	
14	66.20	46.57	19.89	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
15	36.81	30.32	16.25	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	
16	57.76	52.36	28.64	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
17	24.00	20.36	9.07	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	
18	46.30	28.01	13.81	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
19	73.19	45.14	25.70	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
20	40.45	24.44	17.71	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
21	57.21	42.55	24.27	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
22	44.15	23.22	17.49	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
23	42.67	33.56	28.35	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	
24	80.66	24.22	23.55	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	1	1	1	1	0	
25	57.87	34.72	21.16	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
26	33.90	26.64	11.69	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	
27	59.95	38.09	25.64	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	
28	51.58	40.29	24.96	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
29	69.55	62.62	39.86	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
30	22.62	8.59	6.21	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	
31	46.23	34.70	17.69	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
32	47.24	40.41	25.26	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
33	73.58	52.98	35.15	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
34	69.12	50.92	28.85	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
35	53.27	43.25	23.66	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	
36	35.59	28.40	21.15	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
37	46.81	36.80	18.17	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	
38	51.11	27.75	18.01	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
39	37.64	29.68	17.44	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
40	47.53	36.94	17.69	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	0	0	0	1	0	0	0	0	
41	36.44	26.17	18.01	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	
42	103.75	62.05	44.73	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
43	29.93	12.10	9.97	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	1	0	0	
44	42.94	26.15	16.95	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	1	
45	43.56	25.05	21.74	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	
46	38.44	15.25	7.61	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	
47	59.89	37.83	28.20	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
48	62.92	28.27	13.47	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
49	41.33	37.00	20.45	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	

TMZ, patient-based, fixed dose concentration, large dose penalty. Mean tumor diameter (MTD) at the beginning of treatment and at the final observation when applying learned and expert policies.

TMZ, patient-based, variable dose concentration, no dose penalty.

Patient number	MTD at month 0 (mm)	MTD at final observation after learned policy (mm)	MTD at final observation after expert policy (mm)	Dose at month of trial																												
				0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
0	25.56	8.06	7.89	1	1	1	1	1	1	0.25	0.25	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	0	1
1	34.19	16.35	16.27	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1
2	38.53	13.26	13.26	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
3	47.37	23.72	23.72	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
4	42.32	12.24	12.24	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
5	79.98	43.80	43.80	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
6	34.87	10.61	10.61	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
7	39.64	13.38	13.38	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
8	41.75	20.75	20.33	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1
9	50.80	12.76	12.76	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
10	53.97	20.56	20.56	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
11	40.18	5.50	5.50	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
12	38.67	13.65	13.65	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
13	49.55	19.90	19.90	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
14	39.38	13.68	13.68	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
15	52.18	18.96	18.96	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
16	42.28	18.68	18.64	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1
17	39.86	19.75	19.75	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
18	45.10	16.77	16.77	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
19	68.54	29.46	29.46	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
20	45.60	10.57	10.57	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
21	41.24	17.08	17.06	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1
22	50.18	20.70	20.70	1	1	1	1	1	0.75	1	0.5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
23	59.74	21.84	21.84	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
24	32.02	12.60	12.45	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1
25	27.70	10.27	10.22	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	1	1	1	1	1
26	34.26	14.89	14.85	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1
27	52.33	22.45	21.91	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
28	85.27	38.26	38.26	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
29	31.64	13.84	13.78	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1
30	37.64	17.31	17.14	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1
31	27.24	10.91	10.61	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1
32	47.63	30.26	30.23	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1
33	45.63	24.30	24.30	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
34	43.89	5.38	5.29	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
35	35.11	16.52	16.52	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
36	54.24	23.59	23.59	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
37	33.52	6.92	6.72	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0
38	56.04	23.93	23.93	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
39	33.72	14.49	14.35	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
40	32.59	14.39	14.39	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
41	45.80	24.35	23.17	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
42	77.37	32.47	32.47	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
43	54.18	14.86	14.86	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
44	79.10	34.24	34.24	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
45	33.04	22.44	21.72	1	1	1	1	1	1	1	1	1	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1
46	51.89	18.56	18.56	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
47	27.72	7.50	7.03	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1
48	44.20	21.35	21.35	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
49	54.52	18.83	18.83	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

TMZ, patient-based, variable dose concentration, no dose penalty. Mean tumor diameter (MTD) at the beginning of treatment and at the final observation when applying learned and expert policies.

TMZ, patient-based, variable dose concentration, small dose penalty.

Patient number	MTD at month 0 (mm)	MTD at final observation after learned policy (mm)	MTD at final observation after expert policy (mm)	Dose at month of trial																														
				0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
0	25.56	8.14	7.89	1	1	1	1	0	0.25	0	0	0.25	1	1	0.25	0	0	0	1	1	0.25	0	0	0	0	1	1	1	0	0	0	1	1
1	34.19	17.87	16.27	1	1	0.75	0.75	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0
2	38.53	14.13	13.26	1	1	1	1	1	1	0.75	1	0.75	0	0	0	1	0.25	0.25	0	0	0	0	0	0	1	1	1	1	0	0	0	0	1
3	47.37	25.98	23.72	1	0.5	1	1	1	1	1	1	0	0	0	0	1	1	0.25	0	0	0	0	1	1	1	0	0	0	0	1	1	1	1
4	42.32	12.67	12.24	1	1	1	1	1	0.75	0	0	0.25	0	1	1	1	0.5	0	0	0	1	1	1	0	0	0	0	0	0	1	1	1	0
5	79.98	45.08	43.80	1	1	1	1	1	1	1	0.75	0	0	0	1	1	1	1	0	0	0	1	1	1	1	0	0	0	1	1	1	0	0
6	34.87	13.39	10.61	1	1	1	0.75	1	1	1	1	1	0	0	0	0	0.25	0.25	0	0	0	0	0.25	0	0	0	0	0	0	0	1	1	0
7	39.64	14.27	13.38	1	1	1	1	0.75	0.75	0	0.75	0.25	0	0	0.25	0.25	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0
8	41.75	25.93	20.33	1	1	1	1	1	1	1	0.75	0	0	0	0	1	1	1	0	0	0	0	0	1	1	1	0	0	0	0	0	0	1
9	50.80	13.84	12.76	1	1	1	1	1	1	1	0.75	0.75	0.75	0	0.75	1	0.5	0.5	0.25	0	0.25	1	1	0.25	0	0	0	0	0	1	1	1	0
10	53.97	22.54	20.56	1	1	1	1	1	1	1	0.75	0.75	0	0	0	0.25	0.5	0.25	0.25	0	0	0.25	0.25	0	0	0.25	0.25	0.25	0	0	0	0	1
11	40.18	6.33	5.50	1	1	1	1	1	0.75	0.75	0.75	0.75	0	0.75	1	0.5	0.25	0	0.25	0.25	0.25	0	0	0	0	0	0	1	1	1	0	0	0
12	38.67	24.06	13.65	1	0	0.5	1	1	1	1	1	1	0	0	0	1	0.25	0.25	0.25	0	0	0	0	0	1	1	1	1	0	0	0	0	0
13	49.55	22.15	19.90	1	1	1	1	1	1	1	0.75	0	0	0.25	0.25	0.25	0	0	0.25	0.25	0.25	0	0	0	0	1	1	1	0	0	0	0	0
14	39.38	15.96	13.68	1	1	1	1	1	1	0.75	0.75	0.75	0.75	0.25	0	0	0.25	0	0	0	0.25	0	0	0	0	0	0	0	0	0	0	0	0
15	52.18	21.23	18.96	1	1	1	1	1	1	0.75	0.75	0	0	0	0	1	1	1	0	0	0	0	0	1	1	1	0	0	0	0	0	1	1
16	42.28	19.13	18.64	1	1	1	1	1	1	0	0	0.25	0	1	1	0.5	0	0	0	0	0	1	1	1	0	0	0	0	1	1	1	0	0
17	39.86	23.37	19.75	1	1	1	0.75	0.75	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0
18	45.10	18.85	16.77	1	1	1	1	1	1	1	0.75	0.75	0	1	0.25	0.25	0	0	0.25	0.25	0.25	0	0	0	0.25	0.25	0	0	0	0	0	0	0
19	68.54	30.51	29.46	1	1	1	1	1	0.75	0.75	0.25	0.25	0.25	0.75	0.5	0.5	0	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0	0	0	0.25	0.25	0.25	0	0	0.25
20	45.60	14.37	10.57	1	1	1	1	1	0.75	0.75	0.25	0.25	0.75	0.75	1	0.5	0.5	0.25	0.25	0	0.25	0	0	0	0	0	0	0	0	0	0	0	0
21	41.24	17.73	17.06	1	1	1	1	0	0.25	0	0	0.25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0
22	50.18	73.64	20.70	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	59.74	22.90	21.84	1	1	1	1	1	1	0.75	0.25	0.75	0.75	0.75	0.25	0.75	0.25	0	0.25	0.25	0.25	0.25	0.25	0.25	0	0	0.25	0.25	0	0	0.25	0.25	0.25
24	32.02	14.55	12.45	1	1	1	1	0	0.25	0	0	0	1	1	1	0	0	0	0	1	1	1	0	0	0	1	1	0	0	0	0	0	0
25	27.70	12.64	10.22	1	1	1	1	0.75	0.25	0	0	0	1	1	1	0	0	0	0	1	1	0	0	0	0	0	1	1	1	0	0	0	0
26	34.26	15.56	14.85	1	1	1	1	0	0	0	0	0	1	0	0.25	0	0	0	0	1	1	1	0	0	0	0	1	1	1	0	0	0	0
27	52.33	23.29	21.91	1	1	1	1	1	1	0.75	0	0	0	1	1	1	0	0	0	0	1	1	1	0	0	0	0	0	1	1	1	0	0
28	85.27	40.31	38.26	1	1	1	1	1	1	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.25	0.25	0.5	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0	0.25	0.25	0.25	0	0.25	0.25
29	31.64	14.24	13.78	1	1	1	1	1	1	1	0.25	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	0	0	0	0	1	1	1
30	37.64	20.89	17.14	1	1	1	1	1	1	1	0.25	0	0	0	0	0	1	1	1	0	0	0	0	0	1	1	1	1	0	0	0	0	0
31	27.24	12.34	10.61	0	0	1	1	1	1	0.75	1	1	0.25	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0
32	47.63	31.22	30.23	1	1	1	1	1	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0
33	45.63	25.74	24.30	1	1	1	1	1	0.5	0.75	0.25	0.25	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	1
34	43.89	65.25	5.29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
35	35.11	29.74	16.52	1	1	1	1	1	1	1	0.75	0	0	0	0	0	1	1	1	0	0	0	0	0	1	1	1	0	0	0	0	0	0
36	54.24	24.63	23.59	1	1	1	1	1	1	0.75	0	0	0	1	1	1	0	0	0	0	1	1	1	0	0	0	0	1	1	1	0	0	0
37	33.52	90.91	6.72	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
38	56.04	33.97	23.93	1	0	0	0	0.5	0.5	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0.25	0.25	0	0	0	0	1	1	1	1
39	33.72	17.74	14.35	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0
40	32.59	20.92	14.39	1	1	1	1	0.75	0.75	0.75	0.25	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	1	1	1	0	0
41	45.80	72.19	23.17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
42	77.37	33.94	32.47	1	1	1	1	1	0.75	1	0.75	0.75	0.75	0.75	0.75	0.75	0.5	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0	0.25	0.25	0.25	0
43	54.18	16.26	14.86	1	1	1	1	1	1	0.75	0.75	0.25	0.75	0.75	1	0.75	0.5	0.25	0.25	0	0	1	1	1	0	0	0	0	0	1	1	1	0
44	79.10	35.44	34.24	1	1	1	1	1	0.75	0.75	0.25	0.25	0.75	0.75	0.75	0.5	0.25	0.25	0.25	0.25	0.25	0.5	0.25	0	0.25	0.25	0.25	0	0.25	0.25	0.25	0.25	0
45	33.04	28.43	21.72	1	0	0	0	0	1	0.75	0.5	0.75	0.75	0.5	0.25	0.25	0.25	0.25	0.5	0.5	0.25	0.25	0.25	0.25	0.5	0.5	0.25	0.25	0.25	0.25	0.5	0.5	0.25
46	51.89	19.84	18.56	1	1	1	1	1	1	1	0.75	0.75	0	0	0	1	1	1	0	0	0	0	1	1	1	0	0	0	0	0	1	1	1
47	27.72	54.52	7.03	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
48	44.20	22.44	21.35	1	1	1	1	1	1	0	0	0	0.25	0.25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0
49	54.52	21.62	18.83	1	1	1	1	1	1	0.75	0.25	0.75	0.75	0.75	0.75	0.25	0	0	0	0	1	1	1	0	0	0	0	0	1	1	1	0	0

TMZ, patient-based, variable dose concentration, small dose penalty. Mean tumor diameter (MTD) at the beginning of treatment and at the final observation when applying learned and expert policies.

TMZ, patient-based, variable dose concentration, large dose penalty.

Patient number	MTD at month 0 (mm)	MTD at final observation after learned policy (mm)	MTD at final observation after expert policy (mm)	Dose at month of trial																														
				0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
0	25.56	23.79	7.89	1	1	0.25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	34.19	23.26	16.27	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	38.53	31.92	13.26	1	1	1	0.25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	47.37	46.80	23.72	0	0	1	1	0.25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	42.32	26.11	12.24	1	1	1	0.25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	79.98	62.28	43.80	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	34.87	23.68	10.61	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	39.64	30.07	13.38	1	1	0.25	0.25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	41.75	53.38	20.33	1	1	0.25	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	50.80	49.59	12.76	1	1	1	0.25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	53.97	34.74	20.56	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	40.18	22.04	5.50	1	1	1	0.25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	38.67	65.79	13.65	0	1	1	1	0.25	0.25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	49.55	33.96	19.90	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	39.38	23.67	13.68	1	1	0.25	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	52.18	54.14	18.96	1	1	1	0.25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	42.28	28.03	18.64	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	39.86	33.67	19.75	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	45.10	26.70	16.77	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	68.54	37.81	29.46	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	45.60	24.21	10.57	1	1	0.75	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	41.24	26.15	17.06	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	50.18	43.25	20.70	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	59.74	34.64	21.84	1	1	1	0.25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	32.02	35.95	12.45	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	27.70	54.41	10.22	0	0	1	1	0.25	0.25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
26	34.26	23.75	14.85	1	1	0.25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	52.33	39.43	21.91	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
28	85.27	53.83	38.26	1	1	1	0.25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
29	31.64	25.99	13.78	0	1	1	0.25	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30	37.64	36.08	17.14	1	1	1	0.25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
31	27.24	26.20	10.61	0	0	0	0	1	0.25	0.25	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
32	47.63	39.80	30.23	1	1	0.25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
33	45.63	48.41	24.30	0	1	1	0.25	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
34	43.89	30.15	5.29	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
35	35.11	62.83	16.52	1	1	0.25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
36	54.24	41.77	23.59	1	1	0.75	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
37	33.52	90.91	6.72	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
38	56.04	64.94	23.93	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
39	33.72	27.51	14.35	0	0	1	0.25	0	0.25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0.25	
40	32.59	36.06	14.39	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
41	45.80	61.52	23.17	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
42	77.37	42.54	32.47	1	1	1	0.25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
43	54.18	51.33	14.86	0	0	1	1	0.25	0.25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
44	79.10	45.89	34.24	1	1	0.75	0.25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
45	33.04	39.72	21.72	1	1	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
46	51.89	49.03	18.56	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
47	27.72	32.34	7.03	0	0	0	0	0	0	0	1	1	1	1	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
48	44.20	34.54	21.35	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
49	54.52	58.57	18.83	1	1	1	0.25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

TMZ, patient-based, variable dose concentration, large dose penalty. Mean tumor diameter (MTD) at the beginning of treatment and at the final observation when applying learned and expert policies.

TMZ, patient-based, fixed dose concentration, no dose penalty, capped maximum number of treatment administrations.

Patient number	MTD at month 0 (mm)	MTD at final observation after learned policy (mm)	MTD at final observation after expert policy (mm)	Dose at month of trial																												
				0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
0	60.80	17.78	11.34	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	47.31	26.17	17.93	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	29.62	14.37	6.66	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	38.43	20.43	8.46	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	73.17	47.08	30.64	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	21.54	13.27	7.95	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	39.60	30.92	24.03	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	46.80	45.57	21.49	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	40.50	15.90	12.95	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	46.33	25.73	20.63	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	27.20	13.90	11.19	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	38.34	16.72	14.22	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	27.65	12.35	10.01	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	28.82	21.93	12.43	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	31.86	15.46	13.71	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	48.14	35.43	22.05	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	63.58	22.31	19.92	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	55.54	40.86	29.56	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	46.48	34.62	22.40	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	59.49	31.14	24.90	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	46.73	28.97	24.89	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	63.66	34.08	17.57	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	31.23	20.39	17.55	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	36.51	44.25	17.06	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	33.89	23.59	19.11	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	36.41	29.71	9.48	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
26	49.36	34.34	28.72	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	79.66	58.77	30.49	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
28	21.42	6.16	5.37	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
29	38.93	49.66	10.96	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30	56.88	41.14	32.83	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
31	54.24	27.30	13.95	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
32	55.10	39.37	30.15	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
33	78.51	43.13	33.43	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
34	30.75	14.44	6.87	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
35	65.52	31.23	16.56	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
36	28.02	25.98	8.89	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
37	48.88	20.47	14.06	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
38	59.18	18.39	12.09	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
39	37.64	61.08	16.71	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
40	52.13	36.60	22.11	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
41	70.70	37.06	25.88	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
42	27.69	24.27	10.28	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
43	38.20	10.93	2.23	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
44	50.88	37.69	32.57	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
45	36.19	28.00	14.45	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
46	49.85	31.45	23.32	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
47	38.41	43.20	19.05	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
48	52.92	28.51	15.09	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
49	67.25	24.44	18.54	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

TMZ, patient-based, fixed dose concentration, no dose penalty, capped total dose of 10 units per patient. Mean tumor diameter (MTD) at the beginning of treatment and at the final observation when applying learned and expert policies.

TMZ, patient-based, fixed dose concentration, small dose penalty, capped maximum number of treatment administrations.

Patient number	MTD at month 0 (mm)	MTD at final observation after learned policy (mm)	MTD at final observation after expert policy (mm)	Dose at month of trial																													
				0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
0	60.80	17.78	11.34	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	47.31	26.17	17.93	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	29.62	14.37	6.66	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	38.43	20.43	8.46	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	73.17	47.08	30.64	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	21.54	13.27	7.95	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	39.60	30.92	24.03	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	46.80	45.57	21.49	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	40.50	15.90	12.95	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	46.33	25.73	20.63	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	27.20	13.90	11.19	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	38.34	16.72	14.22	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	27.65	12.35	10.01	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	28.82	21.93	12.43	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	31.86	15.46	13.71	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	48.14	35.43	22.05	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	63.58	22.31	19.92	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	55.54	40.86	29.56	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	46.48	34.62	22.40	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	59.49	31.14	24.90	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	46.73	28.97	24.89	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	63.66	34.08	17.57	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	31.23	20.39	17.55	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	36.51	44.25	17.06	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	33.89	23.59	19.11	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	36.41	29.71	9.48	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
26	49.36	34.34	28.72	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	79.66	58.77	30.49	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
28	21.42	6.16	5.37	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
29	38.93	49.66	10.96	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30	56.88	41.14	32.83	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
31	54.24	27.30	13.95	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
32	55.10	39.37	30.15	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
33	78.51	43.13	33.43	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
34	30.75	14.44	6.87	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
35	65.52	31.23	16.56	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
36	28.02	25.98	8.89	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
37	48.88	20.47	14.06	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
38	59.18	18.39	12.09	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
39	37.64	61.08	16.71	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
40	52.13	36.60	22.11	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
41	70.70	37.06	25.88	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
42	27.69	24.27	10.28	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
43	38.20	10.93	2.23	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
44	50.88	37.69	32.57	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
45	36.19	28.00	14.45	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
46	49.85	31.45	23.32	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
47	38.41	43.20	19.05	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
48	52.92	28.51	15.09	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
49	67.25	24.44	18.54	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

TMZ, patient-based, fixed dose concentration, small dose penalty, capped total dose of 10 units per patient. Mean tumor diameter (MTD) at the beginning of treatment and at the final observation when applying learned and expert policies.

TMZ, patient-based, fixed dose concentration, large dose penalty, capped maximum number of treatment administrations.

Patient number	MTD at month 0 (mm)	MTD at final observation after learned policy (mm)	MTD at final observation after expert policy (mm)	Dose at month of trial																													
				0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
0	60.80	17.78	11.34	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	47.31	26.17	17.93	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	29.62	14.37	6.66	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	38.43	20.43	8.46	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	73.17	47.08	30.64	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	21.54	13.27	7.95	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	39.60	30.92	24.03	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	46.80	45.57	21.49	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	40.50	15.90	12.95	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	46.33	25.73	20.63	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	27.20	13.90	11.19	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	38.34	16.72	14.22	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	27.65	12.35	10.01	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	28.82	21.93	12.43	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	31.86	15.46	13.71	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	48.14	35.43	22.05	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	63.58	22.31	19.92	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	55.54	40.86	29.56	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	46.48	34.62	22.40	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	59.49	31.14	24.90	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	46.73	28.97	24.89	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	63.66	34.08	17.57	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	31.23	20.39	17.55	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	36.51	44.25	17.06	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	33.89	23.59	19.11	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	36.41	29.71	9.48	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
26	49.36	34.34	28.72	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	79.66	58.77	30.49	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
28	21.42	6.16	5.37	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
29	38.93	49.66	10.96	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30	56.88	41.14	32.83	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
31	54.24	27.30	13.95	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
32	55.10	39.37	30.15	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
33	78.51	43.13	33.43	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
34	30.75	14.44	6.87	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
35	65.52	31.23	16.56	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
36	28.02	25.98	8.89	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
37	48.88	20.47	14.06	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
38	59.18	18.39	12.09	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
39	37.64	61.08	16.71	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
40	52.13	36.60	22.11	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
41	70.70	37.06	25.88	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
42	27.69	24.27	10.28	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
43	38.20	10.93	2.23	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
44	50.88	37.69	32.57	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
45	36.19	28.00	14.45	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
46	49.85	31.45	23.32	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
47	38.41	43.20	19.05	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
48	52.92	28.51	15.09	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
49	67.25	24.44	18.54	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

TMZ, patient-based, fixed dose concentration, large dose penalty, capped total dose of 10 units per patient. Mean tumor diameter (MTD) at the beginning of treatment and at the final observation when applying learned and expert policies.

TMZ, trial-based, fixed dose concentration, no dose penalty.

Patient number	MTD at month 0 (mm)	MTD at final observation after learned policy (mm)	MTD at final observation after expert policy (mm)
0	35.90	21.56	21.55
1	45.04	10.99	10.97
2	47.26	10.53	10.53
3	46.60	17.19	17.19
4	41.24	11.11	11.11
5	31.80	13.07	13.06
6	46.19	20.24	20.23
7	32.55	5.97	5.97
8	29.52	13.31	13.30
9	42.15	20.75	20.75
10	52.71	16.78	16.78
11	49.07	21.20	21.18
12	49.90	9.50	9.50
13	19.00	6.18	6.18
14	72.72	27.28	27.27
15	72.78	28.36	28.35
16	42.01	17.28	17.27
17	67.54	26.30	26.29
18	27.89	10.94	10.94
19	61.02	20.12	20.11
20	34.31	5.85	5.83
21	48.57	14.74	14.74
22	51.25	14.89	14.88
23	33.99	14.44	14.44
24	42.46	24.04	24.02
25	57.06	30.28	30.28
26	30.10	8.89	8.88
27	45.94	17.55	17.55
28	37.22	11.98	11.97
29	41.47	13.07	13.07
30	28.94	12.35	12.35
31	38.21	15.29	15.29
32	42.72	18.72	18.71
33	39.57	22.62	22.57
34	43.13	16.92	16.91
35	24.10	5.71	5.70
36	39.11	24.47	24.44
37	49.34	18.53	18.53
38	34.36	15.16	15.16
39	34.13	5.38	5.38
40	55.08	25.61	25.61
41	41.14	15.16	15.14
42	40.95	14.35	14.35
43	43.98	19.37	19.37
44	38.90	16.25	16.25
45	46.47	18.10	18.09
46	70.01	31.33	31.32
47	63.63	23.16	23.15
48	35.39	13.83	13.83
49	68.43	26.17	26.13

TMZ, trial-based, fixed dose concentration, no dose penalty. Mean tumor diameter (MTD) at the beginning of treatment and at the final observation when applying learned and expert policies. Learned dosing regimen shown in Table 2 in the main paper.

TMZ, trial-based, fixed dose concentration, small dose penalty.

Patient number	MTD at month 0 (mm)	MTD at final observation after learned policy (mm)	MTD at final observation after expert policy (mm)
0	35.90	25.84	21.55
1	45.04	18.39	10.97
2	47.26	14.16	10.53
3	46.60	20.57	17.19
4	41.24	14.08	11.11
5	31.80	15.47	13.06
6	46.19	21.51	20.23
7	32.55	6.31	5.97
8	29.52	14.98	13.30
9	42.15	24.58	20.75
10	52.71	18.83	16.78
11	49.07	26.30	21.18
12	49.90	11.21	9.50
13	19.00	7.26	6.18
14	72.72	28.28	27.27
15	72.78	30.93	28.35
16	42.01	19.03	17.27
17	67.54	27.33	26.29
18	27.89	12.29	10.94
19	61.02	20.91	20.11
20	34.31	15.30	5.83
21	48.57	15.23	14.74
22	51.25	19.71	14.88
23	33.99	17.44	14.44
24	42.46	33.02	24.02
25	57.06	33.36	30.28
26	30.10	20.11	8.88
27	45.94	19.09	17.55
28	37.22	14.98	11.97
29	41.47	13.75	13.07
30	28.94	14.29	12.35
31	38.21	16.00	15.29
32	42.72	19.72	18.71
33	39.57	35.98	22.57
34	43.13	20.90	16.91
35	24.10	6.36	5.70
36	39.11	28.84	24.44
37	49.34	19.44	18.53
38	34.36	17.27	15.16
39	34.13	6.61	5.38
40	55.08	28.06	25.61
41	41.14	25.87	15.14
42	40.95	14.85	14.35
43	43.98	20.67	19.37
44	38.90	17.57	16.25
45	46.47	23.77	18.09
46	70.01	34.77	31.32
47	63.63	24.71	23.15
48	35.39	17.78	13.83
49	68.43	36.73	26.13

TMZ, trial-based, fixed dose concentration, small dose penalty. Mean tumor diameter (MTD) at the beginning of treatment and at the final observation when applying learned and expert policies. Learned dosing regimen shown in Table 2 in the main paper.

TMZ, trial-based, fixed dose concentration, large dose penalty.

Patient number	MTD at month 0 (mm)	MTD at final observation after learned policy (mm)	MTD at final observation after expert policy (mm)
0	35.90	30.35	21.55
1	45.04	26.51	10.97
2	47.26	16.36	10.53
3	46.60	23.45	17.19
4	41.24	20.38	11.11
5	31.80	19.15	13.06
6	46.19	22.80	20.23
7	32.55	6.43	5.97
8	29.52	18.24	13.30
9	42.15	27.27	20.75
10	52.71	21.23	16.78
11	49.07	32.12	21.18
12	49.90	15.04	9.50
13	19.00	9.17	6.18
14	72.72	29.91	27.27
15	72.78	33.01	28.35
16	42.01	21.34	17.27
17	67.54	30.60	26.29
18	27.89	14.75	10.94
19	61.02	23.47	20.11
20	34.31	24.27	5.83
21	48.57	16.20	14.74
22	51.25	22.91	14.88
23	33.99	20.63	14.44
24	42.46	38.62	24.02
25	57.06	36.98	30.28
26	30.10	31.98	8.88
27	45.94	22.54	17.55
28	37.22	18.75	11.97
29	41.47	15.40	13.07
30	28.94	17.05	12.35
31	38.21	17.36	15.29
32	42.72	22.92	18.71
33	39.57	39.49	22.57
34	43.13	21.98	16.91
35	24.10	7.36	5.70
36	39.11	32.41	24.44
37	49.34	21.13	18.53
38	34.36	17.24	15.16
39	34.13	6.35	5.38
40	55.08	30.67	25.61
41	41.14	31.77	15.14
42	40.95	16.16	14.35
43	43.98	22.17	19.37
44	38.90	18.15	16.25
45	46.47	28.57	18.09
46	70.01	37.66	31.32
47	63.63	27.83	23.15
48	35.39	20.90	13.83
49	68.43	44.99	26.13

TMZ, trial-based, fixed dose concentration, large dose penalty. Mean tumor diameter (MTD) at the beginning of treatment and at the final observation when applying learned and expert policies. Learned dosing regimen shown in Table 2 in the main paper.

TMZ, trial-based, variable dose concentration, no dose penalty.

Patient number	MTD at month 0 (mm)	MTD at final observation after learned policy (mm)	MTD at final observation after expert policy (mm)
0	32.01	13.05	12.97
1	63.24	10.80	10.67
2	26.99	10.09	10.02
3	47.97	16.05	16.01
4	63.98	20.28	20.25
5	39.63	19.97	19.94
6	51.00	10.40	10.33
7	58.59	19.33	19.18
8	26.61	11.12	11.11
9	35.39	10.59	10.59
10	40.79	18.20	18.19
11	31.43	19.46	19.38
12	46.60	20.54	20.44
13	58.51	27.38	27.36
14	32.90	9.76	9.75
15	68.94	45.68	45.52
16	29.50	9.00	8.95
17	76.29	19.91	19.85
18	48.59	21.84	21.70
19	63.27	32.29	32.29
20	64.52	12.61	12.56
21	47.46	15.11	14.92
22	60.13	13.14	13.11
23	20.78	9.85	9.72
24	38.62	11.46	11.22
25	33.31	15.01	14.97
26	54.83	25.19	25.15
27	38.57	14.85	14.85
28	54.62	25.64	25.60
29	50.95	15.83	15.79
30	36.26	10.86	10.80
31	56.72	27.44	27.39
32	32.10	11.02	10.99
33	36.76	8.09	8.05
34	19.60	8.50	8.46
35	26.84	10.17	10.17
36	39.72	6.85	6.84
37	50.41	21.75	21.69
38	57.06	19.30	19.28
39	30.64	10.20	10.20
40	31.24	14.91	14.91
41	46.08	13.68	13.66
42	67.76	21.83	21.80
43	65.77	17.74	17.49
44	34.10	22.20	22.12
45	58.27	19.57	19.56
46	37.63	18.54	18.54
47	56.03	15.75	15.71
48	29.52	10.27	10.22
49	25.98	8.68	8.60

TMZ, trial-based, variable dose concentration, no dose penalty. Mean tumor diameter (MTD) at the beginning of treatment and at the final observation when applying learned and expert policies. Learned dosing regimen shown in Table 2 in the main paper.

TMZ, trial-based, variable dose concentration, small dose penalty.

Patient number	MTD at month 0 (mm)	MTD at final observation after learned policy (mm)	MTD at final observation after expert policy (mm)
0	32.01	21.39	12.97
1	63.24	16.28	10.67
2	26.99	16.33	10.02
3	47.97	17.38	16.01
4	63.98	22.25	20.25
5	39.63	23.21	19.94
6	51.00	12.13	10.33
7	58.59	25.53	19.18
8	26.61	11.78	11.11
9	35.39	11.03	10.59
10	40.79	19.61	18.19
11	31.43	21.53	19.38
12	46.60	23.12	20.44
13	58.51	31.38	27.36
14	32.90	11.49	9.75
15	68.94	64.66	45.52
16	29.50	10.14	8.95
17	76.29	28.11	19.85
18	48.59	27.97	21.70
19	63.27	34.25	32.29
20	64.52	14.72	12.56
21	47.46	35.37	14.92
22	60.13	14.06	13.11
23	20.78	13.79	9.72
24	38.62	23.48	11.22
25	33.31	17.98	14.97
26	54.83	29.34	25.15
27	38.57	15.38	14.85
28	54.62	28.11	25.60
29	50.95	17.96	15.79
30	36.26	13.02	10.80
31	56.72	30.07	27.39
32	32.10	14.05	10.99
33	36.76	10.35	8.05
34	19.60	10.22	8.46
35	26.84	10.69	10.17
36	39.72	9.48	6.84
37	50.41	26.03	21.69
38	57.06	19.78	19.28
39	30.64	11.88	10.20
40	31.24	16.79	14.91
41	46.08	15.32	13.66
42	67.76	23.89	21.80
43	65.77	25.76	17.49
44	34.10	24.29	22.12
45	58.27	20.65	19.56
46	37.63	23.84	18.54
47	56.03	17.11	15.71
48	29.52	12.99	10.22
49	25.98	13.51	8.60

TMZ, trial-based, variable dose concentration, small dose penalty. Mean tumor diameter (MTD) at the beginning of treatment and at the final observation when applying learned and expert policies. Learned dosing regimen shown in Table 2 in the main paper.

TMZ, trial-based, variable dose concentration, large dose penalty.

Patient number	MTD at month 0 (mm)	MTD at final observation after learned policy (mm)	MTD at final observation after expert policy (mm)
0	32.01	31.30	12.97
1	63.24	24.26	10.67
2	26.99	22.29	10.02
3	47.97	20.06	16.01
4	63.98	25.36	20.25
5	39.63	28.61	19.94
6	51.00	16.69	10.33
7	58.59	32.21	19.18
8	26.61	12.18	11.11
9	35.39	11.43	10.59
10	40.79	20.76	18.19
11	31.43	24.16	19.38
12	46.60	27.38	20.44
13	58.51	34.32	27.36
14	32.90	13.42	9.75
15	68.94	68.32	45.52
16	29.50	12.74	8.95
17	76.29	34.99	19.85
18	48.59	32.69	21.70
19	63.27	35.83	32.29
20	64.52	18.93	12.56
21	47.46	43.13	14.92
22	60.13	16.82	13.11
23	20.78	16.54	9.72
24	38.62	29.16	11.22
25	33.31	20.98	14.97
26	54.83	33.01	25.15
27	38.57	16.08	14.85
28	54.62	30.94	25.60
29	50.95	21.45	15.79
30	36.26	16.65	10.80
31	56.72	32.98	27.39
32	32.10	17.35	10.99
33	36.76	13.71	8.05
34	19.60	12.07	8.46
35	26.84	11.09	10.17
36	39.72	12.15	6.84
37	50.41	30.38	21.69
38	57.06	21.47	19.28
39	30.64	13.40	10.20
40	31.24	17.92	14.91
41	46.08	17.81	13.66
42	67.76	27.22	21.80
43	65.77	34.31	17.49
44	34.10	26.85	22.12
45	58.27	21.45	19.56
46	37.63	29.36	18.54
47	56.03	20.07	15.71
48	29.52	16.69	10.22
49	25.98	17.60	8.60

TMZ, trial-based, variable dose concentration, large dose penalty. Mean tumor diameter (MTD) at the beginning of treatment and at the final observation when applying learned and expert policies. Learned dosing regimen shown in Table 2 in the main paper.

TMZ, trial-based, fixed dose concentration, no dose penalty, capped maximum number of treatment administrations.

Patient number	MTD at month 0 (mm)	MTD at final observation after learned policy (mm)	MTD at final observation after expert policy (mm)	Dose at month of trial																												
				0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
0	50.45	35.86	21.45	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	44.52	27.12	24.17	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	33.75	21.86	18.18	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	33.70	64.74	14.46	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	41.35	13.90	8.36	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	30.78	19.56	10.29	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	34.26	19.79	7.20	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	33.26	14.93	8.11	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	70.03	42.01	34.73	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	42.27	24.90	21.54	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	38.91	20.11	15.76	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	40.85	27.10	17.56	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	37.03	25.82	19.00	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	45.07	28.98	20.36	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	39.87	33.05	21.76	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	37.39	17.62	13.77	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	67.49	45.77	34.43	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	39.81	16.62	10.63	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	26.72	12.09	5.70	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	59.04	20.02	12.91	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	35.39	21.89	18.92	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	32.77	13.42	12.11	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	40.46	16.32	10.21	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	66.24	35.68	13.63	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	48.60	45.46	19.76	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	44.17	28.83	22.33	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
26	40.49	27.96	17.61	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	68.55	40.12	31.01	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
28	33.92	18.16	14.53	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
29	53.02	32.84	24.89	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30	58.06	64.22	25.05	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
31	75.64	44.84	41.76	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
32	28.81	15.18	13.42	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
33	37.10	19.69	8.19	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
34	32.25	7.98	6.91	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
35	36.62	15.49	6.84	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
36	44.53	18.20	13.41	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
37	48.96	14.32	12.57	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
38	37.83	18.43	13.96	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
39	38.34	26.08	8.17	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
40	49.79	38.57	29.92	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
41	57.91	35.37	19.15	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
42	76.75	33.43	27.22	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
43	22.34	17.94	7.68	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
44	50.46	30.54	25.87	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
45	41.89	22.91	11.64	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
46	55.75	38.53	25.71	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
47	42.07	17.88	13.95	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
48	43.85	21.55	16.52	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
49	36.30	21.01	8.07	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

TMZ, trial-based, fixed dose concentration, no dose penalty, capped total dose of 10 units per patient. Mean tumor diameter (MTD) at the beginning of treatment and at the final observation when applying learned and expert policies.

TMZ, trial-based, fixed dose concentration, small dose penalty, capped maximum number of treatment administrations.

Patient number	MTD at month 0 (mm)	MTD at final observation after learned policy (mm)	MTD at final observation after expert policy (mm)	Dose at month of trial																												
				0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
0	50.45	35.86	21.45	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	44.52	27.12	24.17	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	33.75	21.86	18.18	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	33.70	64.74	14.46	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	41.35	13.90	8.36	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	30.78	19.56	10.29	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	34.26	19.79	7.20	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	33.26	14.93	8.11	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	70.03	42.01	34.73	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	42.27	24.90	21.54	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	38.91	20.11	15.76	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	40.85	27.10	17.56	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	37.03	25.82	19.00	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	45.07	28.98	20.36	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	39.87	33.05	21.76	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	37.39	17.62	13.77	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	67.49	45.77	34.43	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	39.81	16.62	10.63	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	26.72	12.09	5.70	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	59.04	20.02	12.91	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	35.39	21.89	18.92	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	32.77	13.42	12.11	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	40.46	16.32	10.21	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	66.24	35.68	13.63	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	48.60	45.46	19.76	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	44.17	28.83	22.33	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
26	40.49	27.96	17.61	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	68.55	40.12	31.01	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
28	33.92	18.16	14.53	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
29	53.02	32.84	24.89	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30	58.06	64.22	25.05	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
31	75.64	44.84	41.76	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
32	28.81	15.18	13.42	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
33	37.10	19.69	8.19	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
34	32.25	7.98	6.91	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
35	36.62	15.49	6.84	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
36	44.53	18.20	13.41	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
37	48.96	14.32	12.57	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
38	37.83	18.43	13.96	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
39	38.34	26.08	8.17	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
40	49.79	38.57	29.92	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
41	57.91	35.37	19.15	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
42	76.75	33.43	27.22	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
43	22.34	17.94	7.68	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
44	50.46	30.54	25.87	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
45	41.89	22.91	11.64	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
46	55.75	38.53	25.71	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
47	42.07	17.88	13.95	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
48	43.85	21.55	16.52	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
49	36.30	21.01	8.07	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

TMZ, trial-based, fixed dose concentration, small dose penalty, capped total dose of 10 units per patient. Mean tumor diameter (MTD) at the beginning of treatment and at the final observation when applying learned and expert policies.

TMZ, trial-based, fixed dose concentration, large dose penalty, capped maximum number of treatment administrations.

Patient number	MTD at month 0 (mm)	MTD at final observation after learned policy (mm)	MTD at final observation after expert policy (mm)	Dose at month of trial																													
				0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
0	50.45	35.86	21.45	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	44.52	27.12	24.17	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	33.75	21.86	18.18	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	33.70	64.74	14.46	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	41.35	13.90	8.36	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	30.78	19.56	10.29	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	34.26	19.79	7.20	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	33.26	14.93	8.11	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	70.03	42.01	34.73	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	42.27	24.90	21.54	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	38.91	20.11	15.76	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	40.85	27.10	17.56	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	37.03	25.82	19.00	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	45.07	28.98	20.36	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	39.87	33.05	21.76	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	37.39	17.62	13.77	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	67.49	45.77	34.43	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	39.81	16.62	10.63	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	26.72	12.09	5.70	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	59.04	20.02	12.91	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	35.39	21.89	18.92	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	32.77	13.42	12.11	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	40.46	16.32	10.21	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	66.24	35.68	13.63	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	48.60	45.46	19.76	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	44.17	28.83	22.33	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
26	40.49	27.96	17.61	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	68.55	40.12	31.01	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
28	33.92	18.16	14.53	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
29	53.02	32.84	24.89	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30	58.06	64.22	25.05	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
31	75.64	44.84	41.76	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
32	28.81	15.18	13.42	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
33	37.10	19.69	8.19	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
34	32.25	7.98	6.91	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
35	36.62	15.49	6.84	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
36	44.53	18.20	13.41	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
37	48.96	14.32	12.57	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
38	37.83	18.43	13.96	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
39	38.34	26.08	8.17	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
40	49.79	38.57	29.92	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
41	57.91	35.37	19.15	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
42	76.75	33.43	27.22	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
43	22.34	17.94	7.68	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
44	50.46	30.54	25.87	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
45	41.89	22.91	11.64	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
46	55.75	38.53	25.71	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
47	42.07	17.88	13.95	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
48	43.85	21.55	16.52	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
49	36.30	21.01	8.07	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

TMZ, trial-based, fixed dose concentration, large dose penalty, capped total dose of 10 units per patient. Mean tumor diameter (MTD) at the beginning of treatment and at the final observation when applying learned and expert policies.

PCV, patient-based, fixed dose concentration, no dose penalty.

Patient number	MTD at month 0 (mm)	MTD at final observation after learned policy (mm)	MTD at final observation after expert policy (mm)	Dose at month of trial					
				0	1.5	3	4.5	6	7.5
0	52.05	34.26	34.26	1	1	1	1	1	1
1	62.53	51.77	51.77	1	1	1	1	1	1
2	41.44	29.16	29.16	1	1	1	1	1	1
3	50.21	38.29	38.29	1	1	1	1	1	1
4	46.11	30.71	30.71	1	1	1	1	1	1
5	49.37	37.08	37.08	1	1	1	1	1	1
6	37.96	29.49	29.49	1	1	1	1	1	1
7	37.94	37.11	37.11	1	1	1	1	1	1
8	95.12	80.42	80.42	1	1	1	1	1	1
9	30.41	20.33	20.33	1	1	1	1	1	1
10	67.48	53.37	53.37	1	1	1	1	1	1
11	60.85	47.00	47.00	1	1	1	1	1	1
12	54.79	49.34	49.34	1	1	1	1	1	1
13	48.06	33.07	33.07	1	1	1	1	1	1
14	49.24	33.51	33.51	1	1	1	1	1	1
15	70.38	62.94	62.94	1	1	1	1	1	1
16	42.73	35.63	35.63	1	1	1	1	1	1
17	36.70	27.74	27.74	1	1	1	1	1	1
18	47.10	42.23	42.23	1	1	1	1	1	1
19	73.34	55.90	55.90	1	1	1	1	1	1
20	48.06	34.89	34.89	1	1	1	1	1	1
21	55.71	43.56	43.56	1	1	1	1	1	1
22	52.81	45.95	45.95	1	1	1	1	1	1
23	45.05	36.96	36.96	1	1	1	1	1	1
24	51.65	43.60	43.60	1	1	1	1	1	1
25	49.30	34.58	34.58	1	1	1	1	1	1
26	56.51	50.31	50.31	1	1	1	1	1	1
27	40.05	31.75	31.75	1	1	1	1	1	1
28	49.17	32.87	32.87	1	1	1	1	1	1
29	55.11	40.11	40.11	1	1	1	1	1	1
30	45.86	37.49	37.49	1	1	1	1	1	1
31	75.39	55.37	55.37	1	1	1	1	1	1
32	63.50	43.68	43.68	1	1	1	1	1	1
33	39.16	35.92	35.92	1	1	1	1	1	1
34	68.07	49.06	49.06	1	1	1	1	1	1
35	78.54	65.93	65.93	1	1	1	1	1	1
36	44.18	31.47	31.47	1	1	1	1	1	1
37	51.47	38.35	38.35	1	1	1	1	1	1
38	63.59	44.67	44.67	1	1	1	1	1	1
39	46.19	37.87	37.87	1	1	1	1	1	1
40	47.98	38.41	38.41	1	1	1	1	1	1
41	46.27	33.84	33.84	1	1	1	1	1	1
42	48.40	33.00	33.00	1	1	1	1	1	1
43	52.63	38.15	38.15	1	1	1	1	1	1
44	87.88	85.19	85.19	1	1	1	1	1	1
45	59.94	48.77	48.77	1	1	1	1	1	1
46	48.94	46.42	46.42	1	1	1	1	1	1
47	47.40	41.53	41.53	1	1	1	1	1	1
48	68.22	61.54	61.54	1	1	1	1	1	1
49	67.80	49.69	49.69	1	1	1	1	1	1

PCV, patient-based, fixed dose concentration, no dose penalty. Mean tumor diameter (MTD) at the beginning of treatment and at the final observation when applying learned and expert policies.

PCV, patient-based, fixed dose concentration, small dose penalty.

Patient number	MTD at month 0 (mm)	MTD at final observation after learned policy (mm)	MTD at final observation after expert policy (mm)	Dose at month of trial					
				0	1.5	3	4.5	6	7.5
0	52.05	34.26	34.26	1	1	1	1	1	1
1	62.53	51.77	51.77	1	1	1	1	1	1
2	41.44	29.16	29.16	1	1	1	1	1	1
3	50.21	38.29	38.29	1	1	1	1	1	1
4	46.11	30.71	30.71	1	1	1	1	1	1
5	49.37	37.08	37.08	1	1	1	1	1	1
6	37.96	29.49	29.49	1	1	1	1	1	1
7	37.94	37.11	37.11	1	1	1	1	1	1
8	95.12	80.42	80.42	1	1	1	1	1	1
9	30.41	20.33	20.33	1	1	1	1	1	1
10	67.48	53.37	53.37	1	1	1	1	1	1
11	60.85	47.00	47.00	1	1	1	1	1	1
12	54.79	49.34	49.34	1	1	1	1	1	1
13	48.06	33.07	33.07	1	1	1	1	1	1
14	49.24	33.51	33.51	1	1	1	1	1	1
15	70.38	62.94	62.94	1	1	1	1	1	1
16	42.73	35.63	35.63	1	1	1	1	1	1
17	36.70	27.74	27.74	1	1	1	1	1	1
18	47.10	42.23	42.23	1	1	1	1	1	1
19	73.34	55.90	55.90	1	1	1	1	1	1
20	48.06	34.89	34.89	1	1	1	1	1	1
21	55.71	43.56	43.56	1	1	1	1	1	1
22	52.81	45.95	45.95	1	1	1	1	1	1
23	45.05	36.96	36.96	1	1	1	1	1	1
24	51.65	43.60	43.60	1	1	1	1	1	1
25	49.30	34.58	34.58	1	1	1	1	1	1
26	56.51	50.31	50.31	1	1	1	1	1	1
27	40.05	31.75	31.75	1	1	1	1	1	1
28	49.17	32.87	32.87	1	1	1	1	1	1
29	55.11	40.11	40.11	1	1	1	1	1	1
30	45.86	37.49	37.49	1	1	1	1	1	1
31	75.39	55.37	55.37	1	1	1	1	1	1
32	63.50	43.68	43.68	1	1	1	1	1	1
33	39.16	35.92	35.92	1	1	1	1	1	1
34	68.07	49.06	49.06	1	1	1	1	1	1
35	78.54	65.93	65.93	1	1	1	1	1	1
36	44.18	31.47	31.47	1	1	1	1	1	1
37	51.47	38.35	38.35	1	1	1	1	1	1
38	63.59	44.67	44.67	1	1	1	1	1	1
39	46.19	37.87	37.87	1	1	1	1	1	1
40	47.98	38.41	38.41	1	1	1	1	1	1
41	46.27	33.84	33.84	1	1	1	1	1	1
42	48.40	33.00	33.00	1	1	1	1	1	1
43	52.63	38.15	38.15	1	1	1	1	1	1
44	87.88	85.19	85.19	1	1	1	1	1	1
45	59.94	48.77	48.77	1	1	1	1	1	1
46	48.94	46.42	46.42	1	1	1	1	1	1
47	47.40	41.53	41.53	1	1	1	1	1	1
48	68.22	61.54	61.54	1	1	1	1	1	1
49	67.80	49.69	49.69	1	1	1	1	1	1

PCV, patient-based, fixed dose concentration, small dose penalty. Mean tumor diameter (MTD) at the beginning of treatment and at the final observation when applying learned and expert policies.

PCV, patient-based, fixed dose concentration, large dose penalty.

Patient number	MTD at month 0 (mm)	MTD at final observation after learned policy (mm)	MTD at final observation after expert policy (mm)	Dose at month of trial					
				0	1.5	3	4.5	6	7.5
0	52.05	35.7	34.26	1	0	1	0	1	0
1	62.53	52.48	51.77	1	1	1	1	1	0
2	41.44	30.74	29.16	1	1	0	1	1	0
3	50.21	39.54	38.29	1	1	1	1	1	0
4	46.11	32.27	30.71	1	0	1	1	0	0
5	49.37	38.84	37.08	1	0	1	1	0	0
6	37.96	30.36	29.49	1	1	1	1	1	0
7	37.94	37.11	37.11	1	1	1	1	1	1
8	95.12	81.56	80.42	1	1	1	1	0	0
9	30.41	20.69	20.33	1	1	0	1	1	0
10	67.48	54.33	53.37	1	0	1	1	0	0
11	60.85	51.58	47	1	1	1	1	0	0
12	54.79	49.34	49.34	1	1	1	1	1	1
13	48.06	35.49	33.07	1	0	1	0	1	0
14	49.24	34.63	33.51	1	0	1	0	1	0
15	70.38	62.94	62.94	1	1	1	1	1	1
16	42.73	36.36	35.63	1	1	1	1	1	0
17	36.7	28.14	27.74	1	1	1	1	1	0
18	47.1	42.23	42.23	1	1	1	1	1	1
19	73.34	60.69	55.9	1	1	0	1	0	0
20	48.06	36.06	34.89	1	1	1	1	1	0
21	55.71	44.4	43.56	1	1	0	1	1	0
22	52.81	47.32	45.95	1	1	1	1	1	0
23	45.05	37.38	36.96	1	1	1	1	1	0
24	51.65	43.6	43.6	1	1	1	1	1	1
25	49.3	35.54	34.58	1	1	0	1	1	0
26	56.51	50.31	50.31	1	1	1	1	1	1
27	40.05	32.14	31.75	1	1	1	1	1	0
28	49.17	33.32	32.87	1	0	1	1	1	0
29	55.11	41.71	40.11	1	0	1	1	0	0
30	45.86	38.1	37.49	1	1	1	1	1	0
31	75.39	58.72	55.37	1	0	0	1	0	0
32	63.5	47.5	43.68	1	0	1	0	0	0
33	39.16	35.92	35.92	1	1	1	1	1	1
34	68.07	52.57	49.06	1	1	0	1	1	0
35	78.54	67.07	65.93	1	1	0	1	1	0
36	44.18	32.09	31.47	1	1	1	1	1	0
37	51.47	40.41	38.35	1	1	0	1	1	0
38	63.59	49.12	44.67	1	1	0	1	1	0
39	46.19	38.18	37.87	1	1	1	1	1	0
40	47.98	39.52	38.41	1	1	1	1	1	0
41	46.27	34.48	33.84	1	1	1	1	1	0
42	48.4	34.8	33	1	0	1	1	0	0
43	52.63	40.36	38.15	1	1	0	1	1	0
44	87.88	85.19	85.19	1	1	1	1	1	1
45	59.94	50.55	48.77	1	1	1	1	0	0
46	48.94	46.42	46.42	1	1	1	1	1	1
47	47.4	41.53	41.53	1	1	1	1	1	1
48	68.22	61.54	61.54	1	1	1	1	1	1
49	67.8	53.67	49.69	1	1	0	1	0	0

PCV, patient-based, fixed dose concentration, large dose penalty. Mean tumor diameter (MTD) at the beginning of treatment and at the final observation when applying learned and expert policies.

PCV, patient-based, variable dose concentration, no dose penalty.

Patient number	MTD at month 0 (mm)	MTD at final observation after learned policy (mm)	MTD at final observation after expert policy (mm)	Dose at month of trial					
				0	1.5	3	4.5	6	7.5
0	65.52	45.16	45.16	1	1	1	1	1	1
1	39.59	32.26	32.26	1	1	1	1	1	1
2	62.92	42.91	42.91	1	1	1	1	1	1
3	47.49	39.24	39.24	1	1	1	1	1	1
4	63.03	45.26	45.26	1	1	1	1	1	1
5	45.90	24.21	24.21	1	1	1	1	1	1
6	40.59	32.24	32.24	1	1	1	1	1	1
7	37.84	34.68	34.68	1	1	1	1	1	1
8	28.40	20.76	20.76	1	1	1	1	1	1
9	50.60	30.34	30.34	1	1	1	1	1	1
10	28.97	21.12	21.12	1	1	1	1	1	1
11	39.18	37.31	37.31	1	1	1	1	1	1
12	46.87	36.37	36.37	1	1	1	1	1	1
13	80.86	56.70	56.70	1	1	1	1	1	1
14	39.40	28.18	28.18	1	1	1	1	1	1
15	38.85	30.61	30.61	1	1	1	1	1	1
16	66.46	35.71	35.71	1	1	1	1	1	1
17	40.71	32.39	32.39	1	1	1	1	1	1
18	55.21	42.96	42.96	1	1	1	1	1	1
19	56.27	41.72	41.72	1	1	1	1	1	1
20	66.91	57.45	57.45	1	1	1	1	1	1
21	48.04	38.45	38.45	1	1	1	1	1	1
22	42.39	34.51	34.51	1	1	1	1	1	1
23	54.64	40.60	40.60	1	1	1	1	1	1
24	86.84	68.83	68.83	1	1	1	1	1	1
25	60.41	43.13	43.13	1	1	1	1	1	1
26	56.26	43.33	43.33	1	1	1	1	1	1
27	47.03	32.33	32.33	1	1	1	1	1	1
28	36.45	23.28	23.28	1	1	1	1	1	1
29	33.81	25.12	25.12	1	1	1	1	1	1
30	62.69	53.24	53.24	1	1	1	1	1	1
31	39.52	24.07	24.07	1	1	1	1	1	1
32	28.92	18.06	18.06	1	1	1	1	1	1
33	46.31	41.01	41.01	1	1	1	1	1	1
34	27.71	22.82	22.82	1	1	1	1	1	1
35	55.22	49.31	49.31	1	1	1	1	1	1
36	43.75	31.94	31.94	1	1	1	1	1	1
37	62.71	52.71	52.71	1	1	1	1	1	1
38	60.96	35.98	35.98	1	1	1	1	1	1
39	56.31	41.30	41.30	1	1	1	1	1	1
40	43.32	34.73	34.73	1	1	1	1	1	1
41	35.56	28.14	28.14	1	1	1	1	1	1
42	47.04	35.56	35.56	1	1	1	1	1	1
43	61.19	40.08	40.08	1	1	1	1	1	1
44	42.18	32.01	32.01	1	1	1	1	1	1
45	45.60	30.92	30.92	1	1	1	1	1	1
46	40.16	30.49	30.49	1	1	1	1	1	1
47	47.48	35.87	35.87	1	1	1	1	1	1
48	43.44	29.49	29.49	1	1	1	1	1	1
49	66.43	53.71	53.71	1	1	1	1	1	1

PCV, patient-based, variable dose concentration, no dose penalty. Mean tumor diameter (MTD) at the beginning of treatment and at the final observation when applying learned and expert policies.

PCV, patient-based, variable dose concentration, small dose penalty.

Patient number	MTD at month 0 (mm)	MTD at final observation after learned policy (mm)	MTD at final observation after expert policy (mm)	Dose at month of trial					
				0	1.5	3	4.5	6	7.5
0	65.52	45.16	45.16	1	1	1	1	1	1
1	39.59	32.26	32.26	1	1	1	1	1	1
2	62.92	42.91	42.91	1	1	1	1	1	1
3	47.49	39.24	39.24	1	1	1	1	1	1
4	63.03	45.26	45.26	1	1	1	1	1	1
5	45.90	24.21	24.21	1	1	1	1	1	1
6	40.59	32.24	32.24	1	1	1	1	1	1
7	37.84	34.68	34.68	1	1	1	1	1	1
8	28.40	20.76	20.76	1	1	1	1	1	1
9	50.60	30.34	30.34	1	1	1	1	1	1
10	28.97	21.12	21.12	1	1	1	1	1	1
11	39.18	37.31	37.31	1	1	1	1	1	1
12	46.87	36.37	36.37	1	1	1	1	1	1
13	80.86	56.70	56.70	1	1	1	1	1	1
14	39.40	28.18	28.18	1	1	1	1	1	1
15	38.85	30.61	30.61	1	1	1	1	1	1
16	66.46	35.83	35.71	1	1	1	1	0.25	1
17	40.71	32.39	32.39	1	1	1	1	1	1
18	55.21	42.96	42.96	1	1	1	1	1	1
19	56.27	41.72	41.72	1	1	1	1	1	1
20	66.91	57.45	57.45	1	1	1	1	1	1
21	48.04	38.45	38.45	1	1	1	1	1	1
22	42.39	34.51	34.51	1	1	1	1	1	1
23	54.64	40.60	40.60	1	1	1	1	1	1
24	86.84	68.83	68.83	1	1	1	1	1	1
25	60.41	43.13	43.13	1	1	1	1	1	1
26	56.26	43.33	43.33	1	1	1	1	1	1
27	47.03	32.33	32.33	1	1	1	1	1	1
28	36.45	23.28	23.28	1	1	1	1	1	1
29	33.81	25.12	25.12	1	1	1	1	1	1
30	62.69	53.24	53.24	1	1	1	1	1	1
31	39.52	24.07	24.07	1	1	1	1	1	1
32	28.92	18.06	18.06	1	1	1	1	1	1
33	46.31	41.01	41.01	1	1	1	1	1	1
34	27.71	22.82	22.82	1	1	1	1	1	1
35	55.22	49.31	49.31	1	1	1	1	1	1
36	43.75	31.94	31.94	1	1	1	1	1	1
37	62.71	52.71	52.71	1	1	1	1	1	1
38	60.96	35.98	35.98	1	1	1	1	1	1
39	56.31	41.30	41.30	1	1	1	1	1	1
40	43.32	34.73	34.73	1	1	1	1	1	1
41	35.56	28.14	28.14	1	1	1	1	1	1
42	47.04	35.56	35.56	1	1	1	1	1	1
43	61.19	40.08	40.08	1	1	1	1	1	1
44	42.18	32.01	32.01	1	1	1	1	1	1
45	45.60	30.92	30.92	1	1	1	1	1	1
46	40.16	30.49	30.49	1	1	1	1	1	1
47	47.48	35.87	35.87	1	1	1	1	1	1
48	43.44	29.49	29.49	1	1	1	1	1	1
49	66.43	53.71	53.71	1	1	1	1	1	1

PCV, patient-based, variable dose concentration, small dose penalty. Mean tumor diameter (MTD) at the beginning of treatment and at the final observation when applying learned and expert policies.

PCV, patient-based, variable dose concentration, large dose penalty.

Patient number	MTD at month 0 (mm)	MTD at final observation after learned policy (mm)	MTD at final observation after expert policy (mm)	Dose at month of trial					
				0	1.5	3	4.5	6	7.5
0	65.52	46.21	45.16	0.75	0.75	0.75	0.75	1	1
1	39.59	33.68	32.26	1	0.75	0.75	1	0	0
2	62.92	43.55	42.91	1	0.75	0.75	1	0	0
3	47.49	40.56	39.24	0.75	0.75	0.75	0.75	0.75	1
4	63.03	46.22	45.26	1	0.75	0.75	1	0	0
5	45.9	25.89	24.21	0.75	0.75	0.75	0.75	1	0
6	40.59	33.08	32.24	1	0.75	0.75	0.75	1	1
7	37.84	35.95	34.68	0.75	0.75	0.75	0.75	1	1
8	28.4	21.17	20.76	0.75	0.75	0.75	0.75	1	1
9	50.6	31.21	30.34	0.75	0.75	0.75	0.75	1	1
10	28.97	21.81	21.12	0.75	0.75	0.75	0.75	1	1
11	39.18	37.31	37.31	1	1	1	1	1	1
12	46.87	37.2	36.37	1	0.75	0.75	1	0	0
13	80.86	57.2	56.7	1	0.75	0.75	1	0	0
14	39.4	29	28.18	0.75	0.75	0.75	0.75	1	1
15	38.85	31.56	30.61	0.75	0.75	0.75	0.75	1	0
16	66.46	36.44	35.71	0.75	0.75	0.75	1	1	0
17	40.71	33.23	32.39	1	1	1	1	1	0
18	55.21	46.19	42.96	0.75	0.75	0.75	0.75	0.75	0.75
19	56.27	42.88	41.72	0.75	0.75	0.75	0.75	1	1
20	66.91	58.84	57.45	0.75	0.75	0.75	0.75	0.75	1
21	48.04	39.6	38.45	0.75	0.75	0.75	0.75	1	1
22	42.39	35.12	34.51	0.75	0.75	0.75	0.75	1	1
23	54.64	41.17	40.6	0.75	0.75	0.75	1	1	0
24	86.84	70.96	68.83	0.75	0.75	0.75	0.75	1	1
25	60.41	44.05	43.13	0.75	0.75	0.75	0.75	1	1
26	56.26	44.13	43.33	0.75	0.75	0.75	1	0	0
27	47.03	33.76	32.33	0.75	0.75	0.75	0.75	1	0
28	36.45	23.96	23.28	0.75	0.75	0.75	0.75	1	1
29	33.81	26.63	25.12	0.75	0.75	0.75	0.75	0.75	1
30	62.69	55.03	53.24	1	1	0.75	1	0	0
31	39.52	26.78	24.07	0.75	0.75	0.75	0.75	0.75	1
32	28.92	18.62	18.06	0.75	0.75	0.75	0.75	1	1
33	46.31	41.88	41.01	1	0.75	0.75	0.75	1	1
34	27.71	23.77	22.82	1	0.75	0.75	0.75	1	0
35	55.22	51.31	49.31	1	1	1	0	0	0
36	43.75	32.41	31.94	0.75	0.75	0.75	1	1	0
37	62.71	55.52	52.71	1	1	1	0	0	0
38	60.96	37.61	35.98	0.75	0.75	0.75	0.75	0.75	1
39	56.31	42.77	41.3	1	0.75	0.75	1	0	0
40	43.32	37.97	34.73	0.75	0.75	0.75	0.75	0.75	0.75
41	35.56	28.48	28.14	0.75	0.75	0.75	0.75	1	1
42	47.04	36.35	35.56	0.75	0.75	0.75	0.75	0.75	1
43	61.19	40.9	40.08	0.75	0.75	0.75	0.75	1	1
44	42.18	35.7	32.01	0.75	0.75	0.75	0.75	0.75	0.75
45	45.6	31.5	30.92	0.75	0.75	0.75	0.75	1	1
46	40.16	31.25	30.49	0.75	0.75	0.75	0.75	1	1
47	47.48	36.51	35.87	0.75	0.75	0.75	0.75	1	1
48	43.44	31.2	29.49	0.75	0.75	0.75	0.75	0.75	1
49	66.43	56.08	53.71	1	1	1	0	0	0

PCV, patient-based, variable dose concentration, large dose penalty. Mean tumor diameter (MTD) at the beginning of treatment and at the final observation when applying learned and expert policies.

PCV, patient-based, fixed dose concentration, no dose penalty, capped maximum number of treatment administrations.

Patient number	MTD at month 0 (mm)	MTD at final observation after learned policy (mm)	MTD at final observation after expert policy (mm)	Dose at month of trial					
				0	1.5	3	4.5	6	7.5
0	54.77	56.87	52.11	1	1	0	0	0	0
1	61.02	55.94	49.73	1	1	0	0	0	0
2	52.25	69.98	35.45	0	0	0	0	0	0
3	53.86	63.97	44.93	0	0	0	0	0	0
4	43.12	66.24	28.61	0	0	0	0	0	0
5	43.95	52.24	32.48	0	0	0	0	0	0
6	69.05	59.24	51.67	1	1	0	0	0	0
7	44.55	82.69	30.33	0	0	0	0	0	0
8	49.59	64.38	38.23	0	0	0	0	0	0
9	66.93	51.54	45.73	1	1	0	0	0	0
10	62.58	83.66	50.28	0	0	0	0	0	0
11	50.70	43.77	38.59	1	1	0	0	0	0
12	32.90	45.12	24.95	0	0	0	0	0	0
13	50.36	59.60	40.98	0	0	0	0	0	0
14	61.65	69.30	45.10	0	0	0	0	0	0
15	27.01	46.98	25.60	0	0	0	0	0	0
16	59.53	49.93	47.19	1	1	0	0	0	0
17	30.01	31.90	25.15	1	1	0	0	0	0
18	65.61	74.45	55.18	0	0	0	0	0	0
19	80.75	76.44	70.75	1	1	0	0	0	0
20	44.77	34.25	31.66	1	1	0	0	0	0
21	42.84	56.25	34.81	0	0	0	0	0	0
22	56.64	48.99	41.96	1	1	0	0	0	0
23	34.73	28.48	25.16	1	1	0	0	0	0
24	45.86	46.10	39.50	1	1	0	0	0	0
25	46.22	56.99	36.77	0	0	0	0	0	0
26	46.78	45.79	40.15	1	1	0	0	0	0
27	43.89	57.41	30.17	0	0	0	0	0	0
28	58.48	45.61	43.35	1	0	0	0	0	1
29	45.51	66.74	32.36	0	0	0	0	0	0
30	48.44	44.22	38.82	1	1	0	0	0	0
31	22.90	51.00	25.72	0	0	0	0	0	0
32	47.76	46.79	42.31	1	1	0	0	0	0
33	74.56	71.81	68.19	1	1	0	0	0	0
34	41.79	31.59	25.57	1	1	0	0	0	0
35	53.46	66.52	43.13	0	0	0	0	0	0
36	49.12	83.85	52.30	0	0	0	0	0	0
37	46.22	54.14	39.76	0	0	0	0	0	0
38	48.21	59.43	45.48	0	0	0	0	0	0
39	50.82	46.80	42.12	1	1	0	0	0	0
40	48.87	57.63	39.62	0	0	0	0	0	0
41	46.50	56.69	34.65	0	0	0	0	0	0
42	58.86	50.77	44.41	1	1	0	0	0	0
43	71.89	59.98	54.73	1	1	0	0	0	0
44	55.19	68.54	41.21	0	0	0	0	0	0
45	23.81	49.72	16.09	0	0	0	0	0	0
46	49.90	64.65	36.55	0	0	0	0	0	0
47	35.65	30.13	27.13	1	1	0	0	0	0
48	48.12	41.80	38.89	1	1	0	0	0	0
49	54.33	48.29	41.35	1	1	0	0	0	0

PCV, patient-based, fixed dose concentration, no dose penalty, capped total dose of 2 units per patient. Mean tumor diameter (MTD) at the beginning of treatment and at the final observation when applying learned and expert policies.

PCV, patient-based, fixed dose concentration, small dose penalty, capped maximum number of treatment administrations.

Patient number	MTD at month 0 (mm)	MTD at final observation after learned policy (mm)	MTD at final observation after expert policy (mm)	Dose at month of trial					
				0	1.5	3	4.5	6	7.5
0	54.77	56.87	52.11	1	1	0	0	0	0
1	61.02	55.94	49.73	1	1	0	0	0	0
2	52.25	69.98	35.45	0	0	0	0	0	0
3	53.86	63.97	44.93	0	0	0	0	0	0
4	43.12	66.24	28.61	0	0	0	0	0	0
5	43.95	52.24	32.48	0	0	0	0	0	0
6	69.05	59.24	51.67	1	1	0	0	0	0
7	44.55	82.69	30.33	0	0	0	0	0	0
8	49.59	64.38	38.23	0	0	0	0	0	0
9	66.93	51.54	45.73	1	1	0	0	0	0
10	62.58	83.66	50.28	0	0	0	0	0	0
11	50.70	43.77	38.59	1	1	0	0	0	0
12	32.90	45.12	24.95	0	0	0	0	0	0
13	50.36	59.60	40.98	0	0	0	0	0	0
14	61.65	69.30	45.10	0	0	0	0	0	0
15	27.01	46.98	25.60	0	0	0	0	0	0
16	59.53	49.93	47.19	1	1	0	0	0	0
17	30.01	31.90	25.15	1	1	0	0	0	0
18	65.61	74.45	55.18	0	0	0	0	0	0
19	80.75	76.44	70.75	1	1	0	0	0	0
20	44.77	34.25	31.66	1	1	0	0	0	0
21	42.84	56.25	34.81	0	0	0	0	0	0
22	56.64	48.99	41.96	1	1	0	0	0	0
23	34.73	28.48	25.16	1	1	0	0	0	0
24	45.86	46.10	39.50	1	1	0	0	0	0
25	46.22	56.99	36.77	0	0	0	0	0	0
26	46.78	45.79	40.15	1	1	0	0	0	0
27	43.89	57.41	30.17	0	0	0	0	0	0
28	58.48	45.61	43.35	1	0	0	0	0	1
29	45.51	66.74	32.36	0	0	0	0	0	0
30	48.44	44.22	38.82	1	1	0	0	0	0
31	22.90	51.00	25.72	0	0	0	0	0	0
32	47.76	46.79	42.31	1	1	0	0	0	0
33	74.56	71.81	68.19	1	1	0	0	0	0
34	41.79	31.59	25.57	1	1	0	0	0	0
35	53.46	66.52	43.13	0	0	0	0	0	0
36	49.12	83.85	52.30	0	0	0	0	0	0
37	46.22	54.14	39.76	0	0	0	0	0	0
38	48.21	59.43	45.48	0	0	0	0	0	0
39	50.82	46.80	42.12	1	1	0	0	0	0
40	48.87	57.63	39.62	0	0	0	0	0	0
41	46.50	56.69	34.65	0	0	0	0	0	0
42	58.86	50.77	44.41	1	1	0	0	0	0
43	71.89	59.98	54.73	1	1	0	0	0	0
44	55.19	68.54	41.21	0	0	0	0	0	0
45	23.81	49.72	16.09	0	0	0	0	0	0
46	49.90	64.65	36.55	0	0	0	0	0	0
47	35.65	30.13	27.13	1	1	0	0	0	0
48	48.12	41.80	38.89	1	1	0	0	0	0
49	54.33	48.29	41.35	1	1	0	0	0	0

PCV, patient-based, fixed dose concentration, small dose penalty, capped total dose of 2 units per patient. Mean tumor diameter (MTD) at the beginning of treatment and at the final observation when applying learned and expert policies.

PCV, patient-based, fixed dose concentration, large dose penalty, capped maximum number of treatment administrations.

Patient number	MTD at month 0 (mm)	MTD at final observation after learned policy (mm)	MTD at final observation after expert policy (mm)	Dose at month of trial					
				0	1.5	3	4.5	6	7.5
0	54.77	56.87	52.11	1	0	0	0	0	0
1	61.02	55.94	49.73	1	1	0	0	0	0
2	52.25	69.98	35.45	0	0	0	0	0	0
3	53.86	63.97	44.93	0	0	0	0	0	0
4	43.12	66.24	28.61	0	0	0	0	0	0
5	43.95	52.24	32.48	0	0	0	0	0	0
6	69.05	59.24	51.67	1	1	0	0	0	0
7	44.55	82.69	30.33	0	0	0	0	0	0
8	49.59	64.38	38.23	0	0	0	0	0	0
9	66.93	51.54	45.73	1	1	0	0	0	0
10	62.58	83.66	50.28	0	0	0	0	0	0
11	50.70	43.77	38.59	1	1	0	0	0	0
12	32.90	45.12	24.95	0	0	0	0	0	0
13	50.36	59.60	40.98	0	0	0	0	0	0
14	61.65	69.30	45.10	0	0	0	0	0	0
15	27.01	46.98	25.60	0	0	0	0	0	0
16	59.53	49.93	47.19	1	1	0	0	0	0
17	30.01	31.90	25.15	0	0	0	0	0	0
18	65.61	74.45	55.18	0	0	0	0	0	0
19	80.75	76.44	70.75	1	1	0	0	0	0
20	44.77	34.25	31.66	1	1	0	0	0	0
21	42.84	56.25	34.81	0	0	0	0	0	0
22	56.64	48.99	41.96	0	0	0	0	0	0
23	34.73	28.48	25.16	1	1	0	0	0	0
24	45.86	46.10	39.50	1	0	0	0	0	0
25	46.22	56.99	36.77	0	0	0	0	0	0
26	46.78	45.79	40.15	1	1	0	0	0	0
27	43.89	57.41	30.17	0	0	0	0	0	0
28	58.48	45.61	43.35	1	1	0	0	0	0
29	45.51	66.74	32.36	0	0	0	0	0	0
30	48.44	44.22	38.82	1	1	0	0	0	0
31	22.90	51.00	25.72	0	0	0	0	0	0
32	47.76	46.79	42.31	0	1	0	0	0	0
33	74.56	71.81	68.19	1	1	0	0	0	0
34	41.79	31.59	25.57	0	0	0	0	0	0
35	53.46	66.52	43.13	0	0	0	0	0	0
36	49.12	83.85	52.30	0	0	0	0	0	0
37	46.22	54.14	39.76	0	0	0	0	0	0
38	48.21	59.43	45.48	0	0	0	0	0	0
39	50.82	46.80	42.12	1	1	0	0	0	0
40	48.87	57.63	39.62	0	0	0	0	0	0
41	46.50	56.69	34.65	0	0	0	0	0	0
42	58.86	50.77	44.41	1	1	0	0	0	0
43	71.89	59.98	54.73	1	1	0	0	0	0
44	55.19	68.54	41.21	0	0	0	0	0	0
45	23.81	49.72	16.09	0	0	0	0	0	0
46	49.90	64.65	36.55	0	0	0	0	0	0
47	35.65	30.13	27.13	0	1	1	0	0	0
48	48.12	41.80	38.89	1	1	0	0	0	0
49	54.33	48.29	41.35	0	0	0	0	0	0

PCV, patient-based, fixed dose concentration, large dose penalty, capped total dose of 2 units per patient. Mean tumor diameter (MTD) at the beginning of treatment and at the final observation when applying learned and expert policies.

PCV, trial-based, fixed dose concentration, no dose penalty.

Patient number	MTD at month 0 (mm)	MTD at final observation after learned policy (mm)	MTD at final observation after expert policy (mm)
0	50.07	37.91	37.91
1	51.74	41.24	41.24
2	46.47	48.04	48.04
3	35.78	24.48	24.48
4	30.96	37.69	37.69
5	30.79	21.84	21.84
6	45.57	36.17	36.17
7	63.75	45.98	45.98
8	60.14	49.12	49.12
9	46.73	34.86	34.86
10	39.00	29.50	29.50
11	43.76	32.83	32.83
12	46.19	37.99	37.99
13	49.52	35.98	35.98
14	44.74	33.41	33.41
15	37.44	37.94	37.94
16	53.02	40.92	40.92
17	32.85	21.54	21.54
18	48.94	34.49	34.49
19	57.58	45.66	45.66
20	54.22	40.65	40.65
21	44.02	34.16	34.16
22	80.59	70.81	70.81
23	37.24	29.31	29.31
24	71.50	53.75	53.75
25	40.54	31.74	31.74
26	72.13	61.87	61.87
27	57.81	52.74	52.74
28	44.68	36.44	36.44
29	37.84	32.44	32.44
30	65.55	51.58	51.58
31	33.65	27.94	27.94
32	56.63	41.32	41.32
33	44.95	32.13	32.13
34	75.01	56.28	56.28
35	42.26	34.04	34.04
36	67.61	60.24	60.24
37	57.59	37.87	37.87
38	31.46	25.46	25.46
39	44.19	39.37	39.37
40	67.15	40.87	40.87
41	46.40	38.11	38.11
42	42.23	36.56	36.56
43	41.93	37.97	37.97
44	55.42	50.58	50.58
45	45.59	30.56	30.56
46	65.02	49.78	49.78
47	40.43	38.16	38.16
48	52.61	46.00	46.00
49	51.46	47.29	47.29

PCV, trial-based, fixed dose concentration, no dose penalty. Mean tumor diameter (MTD) at the beginning of treatment and at the final observation when applying learned and expert policies. Learned dosing regimen shown in Table 3 in the main paper.

PCV, trial-based, fixed dose concentration, small dose penalty.

Patient number	MTD at month 0 (mm)	MTD at final observation after learned policy (mm)	MTD at final observation after expert policy (mm)
0	50.07	39.98	37.91
1	51.74	43.57	41.24
2	46.47	55.32	48.04
3	35.78	27.82	24.48
4	30.96	49.97	37.69
5	30.79	22.62	21.84
6	45.57	37.97	36.17
7	63.75	51.76	45.98
8	60.14	50.09	49.12
9	46.73	36.74	34.86
10	39.00	35.62	29.50
11	43.76	33.57	32.83
12	46.19	38.94	37.99
13	49.52	43.83	35.98
14	44.74	33.84	33.41
15	37.44	39.57	37.94
16	53.02	42.65	40.92
17	32.85	21.99	21.54
18	48.94	37.07	34.49
19	57.58	46.29	45.66
20	54.22	41.17	40.65
21	44.02	34.52	34.16
22	80.59	72.40	70.81
23	37.24	29.77	29.31
24	71.50	54.83	53.75
25	40.54	40.16	31.74
26	72.13	62.90	61.87
27	57.81	55.59	52.74
28	44.68	39.08	36.44
29	37.84	33.58	32.44
30	65.55	57.52	51.58
31	33.65	28.98	27.94
32	56.63	42.54	41.32
33	44.95	34.00	32.13
34	75.01	57.75	56.28
35	42.26	34.49	34.04
36	67.61	62.21	60.24
37	57.59	39.00	37.87
38	31.46	29.79	25.46
39	44.19	41.32	39.37
40	67.15	44.52	40.87
41	46.40	41.89	38.11
42	42.23	44.19	36.56
43	41.93	40.63	37.97
44	55.42	54.35	50.58
45	45.59	31.45	30.56
46	65.02	50.58	49.78
47	40.43	40.20	38.16
48	52.61	49.18	46.00
49	51.46	51.82	47.29

PCV, trial-based, fixed dose concentration, small dose penalty. Mean tumor diameter (MTD) at the beginning of treatment and at the final observation when applying learned and expert policies. Learned dosing regimen shown in Table 3 in the main paper.

PCV, trial-based, fixed dose concentration, large dose penalty.

Patient number	MTD at month 0 (mm)	MTD at final observation after learned policy (mm)	MTD at final observation after expert policy (mm)
0	50.07	41.82	37.91
1	51.74	45.71	41.24
2	46.47	59.85	48.04
3	35.78	31.09	24.48
4	30.96	57.93	37.69
5	30.79	23.81	21.84
6	45.57	39.52	36.17
7	63.75	56.57	45.98
8	60.14	51.08	49.12
9	46.73	38.80	34.86
10	39.00	40.84	29.50
11	43.76	34.40	32.83
12	46.19	39.75	37.99
13	49.52	51.66	35.98
14	44.74	34.38	33.41
15	37.44	40.58	37.94
16	53.02	44.63	40.92
17	32.85	22.35	21.54
18	48.94	39.70	34.49
19	57.58	46.91	45.66
20	54.22	41.64	40.65
21	44.02	34.78	34.16
22	80.59	73.67	70.81
23	37.24	30.34	29.31
24	71.50	56.18	53.75
25	40.54	49.53	31.74
26	72.13	63.84	61.87
27	57.81	57.59	52.74
28	44.68	41.47	36.44
29	37.84	34.53	32.44
30	65.55	62.30	51.58
31	33.65	29.89	27.94
32	56.63	44.15	41.32
33	44.95	36.33	32.13
34	75.01	58.99	56.28
35	42.26	35.09	34.04
36	67.61	63.64	60.24
37	57.59	40.34	37.87
38	31.46	33.84	25.46
39	44.19	42.84	39.37
40	67.15	48.34	40.87
41	46.40	44.92	38.11
42	42.23	49.92	36.56
43	41.93	42.46	37.97
44	55.42	56.83	50.58
45	45.59	32.40	30.56
46	65.02	51.45	49.78
47	40.43	41.56	38.16
48	52.61	51.89	46.00
49	51.46	55.49	47.29

PCV, trial-based, fixed dose concentration, large dose penalty. Mean tumor diameter (MTD) at the beginning of treatment and at the final observation when applying learned and expert policies. Learned dosing regimen shown in Table 3 in the main paper.

PCV, trial-based, variable dose concentration, no dose penalty.

Patient number	MTD at month 0 (mm)	MTD at final observation after learned policy (mm)	MTD at final observation after expert policy (mm)
0	41.06	29.62	29.62
1	60.14	44.02	44.02
2	59.19	45.22	45.22
3	52.52	36.93	36.93
4	67.06	40.54	40.54
5	38.67	41.62	41.62
6	65.51	43.73	43.73
7	43.25	36.42	36.42
8	54.66	33.87	33.87
9	71.10	52.48	52.48
10	60.24	42.11	42.11
11	45.95	26.68	26.68
12	38.55	25.52	25.52
13	51.55	41.27	41.27
14	27.10	20.53	20.53
15	65.39	57.88	57.88
16	48.00	38.55	38.55
17	69.91	54.63	54.63
18	47.77	35.66	35.66
19	35.12	25.69	25.69
20	53.11	39.50	39.50
21	37.55	33.77	33.77
22	44.56	30.31	30.31
23	40.36	34.10	34.10
24	54.03	42.94	42.94
25	60.48	47.17	47.17
26	44.37	40.78	40.78
27	51.83	43.55	43.55
28	48.36	32.37	32.37
29	31.44	24.07	24.07
30	39.90	32.48	32.48
31	75.96	56.58	56.58
32	43.72	37.52	37.52
33	46.78	36.29	36.29
34	57.02	57.77	57.77
35	49.95	37.63	37.63
36	39.98	29.55	29.55
37	44.04	27.85	27.85
38	49.20	36.66	36.66
39	70.38	56.81	56.81
40	49.85	39.80	39.80
41	49.95	35.12	35.12
42	67.71	51.07	51.07
43	47.22	35.90	35.90
44	51.27	42.57	42.57
45	54.57	41.09	41.09
46	59.33	36.51	36.51
47	40.57	30.54	30.54
48	53.30	33.27	33.27
49	39.64	26.98	26.98

PCV, trial-based, variable dose concentration, no dose penalty. Mean tumor diameter (MTD) at the beginning of treatment and at the final observation when applying learned and expert policies. Learned dosing regimen shown in Table 3 in the main paper.

PCV, trial-based, variable dose concentration, small dose penalty.

Patient number	MTD at month 0 (mm)	MTD at final observation after learned policy (mm)	MTD at final observation after expert policy (mm)
0	41.06	29.78	29.62
1	60.14	44.47	44.02
2	59.19	45.41	45.22
3	52.52	37.19	36.93
4	67.06	40.97	40.54
5	38.67	42.46	41.62
6	65.51	44.06	43.73
7	43.25	36.64	36.42
8	54.66	34.40	33.87
9	71.10	52.78	52.48
10	60.24	42.33	42.11
11	45.95	26.96	26.68
12	38.55	25.83	25.52
13	51.55	41.49	41.27
14	27.10	20.71	20.53
15	65.39	58.40	57.88
16	48.00	38.81	38.55
17	69.91	54.87	54.63
18	47.77	35.99	35.66
19	35.12	25.83	25.69
20	53.11	39.90	39.50
21	37.55	34.32	33.77
22	44.56	30.52	30.31
23	40.36	34.89	34.10
24	54.03	43.28	42.94
25	60.48	47.37	47.17
26	44.37	41.12	40.78
27	51.83	43.80	43.55
28	48.36	32.61	32.37
29	31.44	24.33	24.07
30	39.90	32.61	32.48
31	75.96	56.86	56.58
32	43.72	37.78	37.52
33	46.78	36.66	36.29
34	57.02	58.35	57.77
35	49.95	37.80	37.63
36	39.98	29.72	29.55
37	44.04	28.16	27.85
38	49.20	36.90	36.66
39	70.38	57.55	56.81
40	49.85	39.95	39.80
41	49.95	35.44	35.12
42	67.71	51.30	51.07
43	47.22	36.20	35.90
44	51.27	42.91	42.57
45	54.57	41.46	41.09
46	59.33	36.81	36.51
47	40.57	31.03	30.54
48	53.30	33.55	33.27
49	39.64	27.12	26.98

PCV, trial-based, variable dose concentration, small dose penalty. Mean tumor diameter (MTD) at the beginning of treatment and at the final observation when applying learned and expert policies. Learned dosing regimen shown in Table 3 in the main paper.

PCV, trial-based, variable dose concentration, large dose penalty.

Patient number	MTD at month 0 (mm)	MTD at final observation after learned policy (mm)	MTD at final observation after expert policy (mm)
0	41.06	30.13	29.62
1	60.14	46	44.02
2	59.19	45.57	45.22
3	52.52	37.69	36.93
4	67.06	41.79	40.54
5	38.67	47.52	41.62
6	65.51	44.95	43.73
7	43.25	37.28	36.42
8	54.66	35.94	33.87
9	71.1	53.05	52.48
10	60.24	42.32	42.11
11	45.95	27.43	26.68
12	38.55	26.91	25.52
13	51.55	42.05	41.27
14	27.1	21.21	20.53
15	65.39	60.91	57.88
16	48	39.77	38.55
17	69.91	55.58	54.63
18	47.77	37.31	35.66
19	35.12	26.28	25.69
20	53.11	41.16	39.5
21	37.55	37.33	33.77
22	44.56	30.94	30.31
23	40.36	40.11	34.1
24	54.03	44.43	42.94
25	60.48	47.81	47.17
26	44.37	42.74	40.78
27	51.83	44.81	43.55
28	48.36	32.82	32.37
29	31.44	25.62	24.07
30	39.9	32.8	32.48
31	75.96	56.83	56.58
32	43.72	38.88	37.52
33	46.78	37.73	36.29
34	57.02	61.4	57.77
35	49.95	38.11	37.63
36	39.98	30.39	29.55
37	44.04	28.78	27.85
38	49.2	37.24	36.66
39	70.38	60.73	56.81
40	49.85	40.04	39.8
41	49.95	36.63	35.12
42	67.71	51.55	51.07
43	47.22	37.11	35.9
44	51.27	44.56	42.57
45	54.57	42.25	41.09
46	59.33	37.36	36.51
47	40.57	33.63	30.54
48	53.3	34.2	33.27
49	39.64	27.25	26.98

PCV, trial-based, variable dose concentration, large dose penalty. Mean tumor diameter (MTD) at the beginning of treatment and at the final observation when applying learned and expert policies. Learned dosing regimen shown in Table 3 in the main paper.

PCV, trial-based, fixed dose concentration, no dose penalty, capped maximum number of treatment administrations.

Patient number	MTD at month 0 (mm)	MTD at final observation after learned policy (mm)	MTD at final observation after expert policy (mm)	Dose at month of trial					
				0	1.5	3	4.5	6	7.5
0	51.02	54.00	41.41	1	1	0	0	0	0
1	55.56	69.90	48.49	1	1	0	0	0	0
2	38.14	47.61	30.37	1	1	0	0	0	0
3	49.62	56.49	34.79	1	1	0	0	0	0
4	52.35	58.14	35.60	1	1	0	0	0	0
5	43.72	80.73	32.18	1	1	0	0	0	0
6	48.93	53.17	43.10	1	1	0	0	0	0
7	53.31	80.10	42.27	1	1	0	0	0	0
8	41.61	47.03	33.67	1	1	0	0	0	0
9	56.06	68.75	40.27	1	1	0	0	0	0
10	49.46	81.54	48.66	1	1	0	0	0	0
11	36.62	50.93	27.18	1	1	0	0	0	0
12	74.24	80.30	60.98	1	1	0	0	0	0
13	47.04	62.58	37.31	1	1	0	0	0	0
14	86.89	89.98	68.09	1	1	0	0	0	0
15	46.60	49.00	38.60	1	1	0	0	0	0
16	59.72	63.58	49.37	1	1	0	0	0	0
17	61.43	72.94	41.88	1	1	0	0	0	0
18	35.07	57.67	26.37	1	1	0	0	0	0
19	41.21	53.45	31.83	1	1	0	0	0	0
20	44.76	47.60	35.59	1	1	0	0	0	0
21	65.99	88.88	62.25	1	1	0	0	0	0
22	58.75	66.43	46.94	1	1	0	0	0	0
23	48.30	60.36	37.66	1	1	0	0	0	0
24	45.21	50.92	27.97	1	1	0	0	0	0
25	40.56	80.16	31.59	1	1	0	0	0	0
26	41.69	50.56	34.24	1	1	0	0	0	0
27	41.84	65.01	26.23	1	1	0	0	0	0
28	56.60	68.88	44.40	1	1	0	0	0	0
29	39.20	43.59	30.08	1	1	0	0	0	0
30	58.03	65.46	40.70	1	1	0	0	0	0
31	30.24	56.89	19.59	1	1	0	0	0	0
32	56.04	76.28	33.79	1	1	0	0	0	0
33	35.68	40.58	29.80	1	1	0	0	0	0
34	66.35	70.80	51.79	1	1	0	0	0	0
35	57.28	64.74	45.49	1	1	0	0	0	0
36	42.04	54.11	34.43	1	1	0	0	0	0
37	41.85	49.42	29.87	1	1	0	0	0	0
38	41.66	65.34	33.23	1	1	0	0	0	0
39	87.94	90.94	66.81	1	1	0	0	0	0
40	32.09	36.78	25.35	1	1	0	0	0	0
41	53.26	67.80	42.29	1	1	0	0	0	0
42	52.22	55.75	40.39	1	1	0	0	0	0
43	37.94	50.07	30.57	1	1	0	0	0	0
44	69.82	78.95	53.92	1	1	0	0	0	0
45	47.57	90.11	34.62	1	1	0	0	0	0
46	69.89	76.19	44.69	1	1	0	0	0	0
47	46.45	51.55	38.76	1	1	0	0	0	0
48	48.93	51.76	40.16	1	1	0	0	0	0
49	43.86	47.12	35.90	1	1	0	0	0	0

PCV, trial-based, fixed dose concentration, no dose penalty, capped total dose of 2 units per patient. Mean tumor diameter (MTD) at the beginning of treatment and at the final observation when applying learned and expert policies.

PCV, trial-based, fixed dose concentration, small dose penalty, capped maximum number of treatment administrations.

Patient number	MTD at month 0 (mm)	MTD at final observation after learned policy (mm)	MTD at final observation after expert policy (mm)	Dose at month of trial					
				0	1.5	3	4.5	6	7.5
0	51.02	54.00	41.41	1	1	0	0	0	0
1	55.56	69.90	48.49	1	1	0	0	0	0
2	38.14	47.61	30.37	1	1	0	0	0	0
3	49.62	56.49	34.79	1	1	0	0	0	0
4	52.35	58.14	35.60	1	1	0	0	0	0
5	43.72	80.73	32.18	1	1	0	0	0	0
6	48.93	53.17	43.10	1	1	0	0	0	0
7	53.31	80.10	42.27	1	1	0	0	0	0
8	41.61	47.03	33.67	1	1	0	0	0	0
9	56.06	68.75	40.27	1	1	0	0	0	0
10	49.46	81.54	48.66	1	1	0	0	0	0
11	36.62	50.93	27.18	1	1	0	0	0	0
12	74.24	80.30	60.98	1	1	0	0	0	0
13	47.04	62.58	37.31	1	1	0	0	0	0
14	86.89	89.98	68.09	1	1	0	0	0	0
15	46.60	49.00	38.60	1	1	0	0	0	0
16	59.72	63.58	49.37	1	1	0	0	0	0
17	61.43	72.94	41.88	1	1	0	0	0	0
18	35.07	57.67	26.37	1	1	0	0	0	0
19	41.21	53.45	31.83	1	1	0	0	0	0
20	44.76	47.60	35.59	1	1	0	0	0	0
21	65.99	88.88	62.25	1	1	0	0	0	0
22	58.75	66.43	46.94	1	1	0	0	0	0
23	48.30	60.36	37.66	1	1	0	0	0	0
24	45.21	50.92	27.97	1	1	0	0	0	0
25	40.56	80.16	31.59	1	1	0	0	0	0
26	41.69	50.56	34.24	1	1	0	0	0	0
27	41.84	65.01	26.23	1	1	0	0	0	0
28	56.60	68.88	44.40	1	1	0	0	0	0
29	39.20	43.59	30.08	1	1	0	0	0	0
30	58.03	65.46	40.70	1	1	0	0	0	0
31	30.24	56.89	19.59	1	1	0	0	0	0
32	56.04	76.28	33.79	1	1	0	0	0	0
33	35.68	40.58	29.80	1	1	0	0	0	0
34	66.35	70.80	51.79	1	1	0	0	0	0
35	57.28	64.74	45.49	1	1	0	0	0	0
36	42.04	54.11	34.43	1	1	0	0	0	0
37	41.85	49.42	29.87	1	1	0	0	0	0
38	41.66	65.34	33.23	1	1	0	0	0	0
39	87.94	90.94	66.81	1	1	0	0	0	0
40	32.09	36.78	25.35	1	1	0	0	0	0
41	53.26	67.80	42.29	1	1	0	0	0	0
42	52.22	55.75	40.39	1	1	0	0	0	0
43	37.94	50.07	30.57	1	1	0	0	0	0
44	69.82	78.95	53.92	1	1	0	0	0	0
45	47.57	90.11	34.62	1	1	0	0	0	0
46	69.89	76.19	44.69	1	1	0	0	0	0
47	46.45	51.55	38.76	1	1	0	0	0	0
48	48.93	51.76	40.16	1	1	0	0	0	0
49	43.86	47.12	35.90	1	1	0	0	0	0

PCV, trial-based, fixed dose concentration, small dose penalty, capped total dose of 2 units per patient. Mean tumor diameter (MTD) at the beginning of treatment and at the final observation when applying learned and expert policies.

PCV, trial-based, fixed dose concentration, large dose penalty, capped maximum number of treatment administrations.

Patient number	MTD at month 0 (mm)	MTD at final observation after learned policy (mm)	MTD at final observation after expert policy (mm)	Dose at month of trial					
				0	1.5	3	4.5	6	7.5
0	51.02	54.00	41.41	1	1	0	0	0	0
1	55.56	69.90	48.49	1	1	0	0	0	0
2	38.14	47.61	30.37	1	1	0	0	0	0
3	49.62	56.49	34.79	1	1	0	0	0	0
4	52.35	58.14	35.60	1	1	0	0	0	0
5	43.72	80.73	32.18	1	1	0	0	0	0
6	48.93	53.17	43.10	1	1	0	0	0	0
7	53.31	80.10	42.27	1	1	0	0	0	0
8	41.61	47.03	33.67	1	1	0	0	0	0
9	56.06	68.75	40.27	1	1	0	0	0	0
10	49.46	81.54	48.66	1	1	0	0	0	0
11	36.62	50.93	27.18	1	1	0	0	0	0
12	74.24	80.30	60.98	1	1	0	0	0	0
13	47.04	62.58	37.31	1	1	0	0	0	0
14	86.89	89.98	68.09	1	1	0	0	0	0
15	46.60	49.00	38.60	1	1	0	0	0	0
16	59.72	63.58	49.37	1	1	0	0	0	0
17	61.43	72.94	41.88	1	1	0	0	0	0
18	35.07	57.67	26.37	1	1	0	0	0	0
19	41.21	53.45	31.83	1	1	0	0	0	0
20	44.76	47.60	35.59	1	1	0	0	0	0
21	65.99	88.88	62.25	1	1	0	0	0	0
22	58.75	66.43	46.94	1	1	0	0	0	0
23	48.30	60.36	37.66	1	1	0	0	0	0
24	45.21	50.92	27.97	1	1	0	0	0	0
25	40.56	80.16	31.59	1	1	0	0	0	0
26	41.69	50.56	34.24	1	1	0	0	0	0
27	41.84	65.01	26.23	1	1	0	0	0	0
28	56.60	68.88	44.40	1	1	0	0	0	0
29	39.20	43.59	30.08	1	1	0	0	0	0
30	58.03	65.46	40.70	1	1	0	0	0	0
31	30.24	56.89	19.59	1	1	0	0	0	0
32	56.04	76.28	33.79	1	1	0	0	0	0
33	35.68	40.58	29.80	1	1	0	0	0	0
34	66.35	70.80	51.79	1	1	0	0	0	0
35	57.28	64.74	45.49	1	1	0	0	0	0
36	42.04	54.11	34.43	1	1	0	0	0	0
37	41.85	49.42	29.87	1	1	0	0	0	0
38	41.66	65.34	33.23	1	1	0	0	0	0
39	87.94	90.94	66.81	1	1	0	0	0	0
40	32.09	36.78	25.35	1	1	0	0	0	0
41	53.26	67.80	42.29	1	1	0	0	0	0
42	52.22	55.75	40.39	1	1	0	0	0	0
43	37.94	50.07	30.57	1	1	0	0	0	0
44	69.82	78.95	53.92	1	1	0	0	0	0
45	47.57	90.11	34.62	1	1	0	0	0	0
46	69.89	76.19	44.69	1	1	0	0	0	0
47	46.45	51.55	38.76	1	1	0	0	0	0
48	48.93	51.76	40.16	1	1	0	0	0	0
49	43.86	47.12	35.90	1	1	0	0	0	0

PCV, trial-based, fixed dose concentration, large dose penalty, capped total dose of 2 units per patient. Mean tumor diameter (MTD) at the beginning of treatment and at the final observation when applying learned and expert policies.