Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

What a size-only threshold for holding cancer treatment loses in a spatial tumour model

Arjun Grandhe

grandhearjun9@gmail.com

Albuquerque Academy https://orcid.org/0009-0002-8652-3314

Research Article

Keywords: adaptive therapy, discrete monitoring, agent-based model, treatment threshold, spatial structure, premature progression

Posted Date: October 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11542372/v1

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

Additional Declarations: The authors declare no competing interests.

What a size-only threshold for holding cancer treatment loses in a spatial tumour model

Arjun Grandhe

Albuquerque Academy, Albuquerque, NM, USA

grandhearjun9@gmail.com

Abstract

Gallagher et al. derived a size threshold for holding adaptive therapy between clinic visits in non-spatial ordinary differential equation (ODE) models. I tested it in one spatial agent-based model (ABM) against a tolerance of 0.05 for premature progression (progression during a treatment hold). At the derived threshold, premature progression occurred in 0.976, 0.974 and 0.480 of runs at intervals of 7, 14 and 28 days, and at most 0.014 at 42 to 84 days. The ABM's safe threshold differed from the ODE's own by −4.40% to −8.95% at 7 to 56 days, and recalibration to the ABM's drug-free regrowth did not remove this shortfall. Most of it (84.2% to 94.1% at 7 to 28 days) arose in the step from a well-mixed stochastic model to the ABM; that step changes several features at once, so no mechanism is identified. With error-free size readings, measurement history reduced the test log-loss of predicting premature progression at equal size by 48.6%; with 10% measurement error, history was only 2.0% better than exact size. Under a common nominal risk constraint, a hold rule using history and image geometry changed time on drug by −1.4 percentage points and restricted mean time to progression by +4.0 days over 750 days (+0.95 over 365). The gain rested on two of four intervals; non-inferiority sat on the margin at 7 days and was not shown at 14. On new runs its risk was not shown to meet the tolerance at any interval (upper Wilson bounds above 0.05).

Keywords: adaptive therapy, discrete monitoring, agent-based model, treatment threshold, spatial structure, premature progression

1 Introduction

Adaptive therapy treats a tumour only as much as is needed to keep drug-sensitive cells alive, so that they keep competing with resistant cells (Gatenby et al. 2009). A common form holds treatment while the tumour is small and restarts it when the tumour grows. In the clinic the tumour is measured only at visits, so at each visit the rule has to judge whether a tumour left untreated will stay below the progression limit until the next visit. I ask whether a rule that makes that judgement from tumour size alone gets the risk wrong when visits are spaced apart and the tumour has spatial structure.

Gallagher et al. (2025) derived a size threshold for this setting in non-spatial ordinary differential equation (ODE) models. They called progression that occurs because treatment was held for too long between visits premature progression. In the form used here, the threshold $N^*(\tau)$ is the tumour size from which untreated regrowth reaches the progression limit in exactly one monitoring interval τ (Section 2.3). A tumour held below $N^*(\tau)$ at a visit should then not progress before the next visit, provided the model that produced N^* describes regrowth correctly. I use their construction unchanged. What I measure is how much of its promised safety the threshold keeps when the tumour is a spatial agent-based model (ABM).

Switching rules of this kind are already used in patients. In a pilot trial in metastatic castration-resistant prostate cancer, abiraterone was stopped and restarted according to the level of prostate-specific antigen (PSA) relative to its value before treatment (Zhang et al. 2017; H. Lee Moffitt Cancer Center and Research Institute 2015). How that trial was reported and analysed has been debated (Mistry 2021; Zhang et al. 2021).

A threshold can be correct in its own model and still fail in a spatial tumour. An ODE assumes that every cell feels the average crowding of the whole population. On a lattice, only cells next to an empty site can divide, and the drug reaches cells by diffusion. A tumour shrunk by treatment may therefore regrow at a different speed from the one the ODE predicts, and two tumours of the same size may regrow at different speeds. A threshold that is safe in the ODE can then be unsafe for part of a population of spatial tumours.

Spatial structure is known to change how adaptive therapy performs. Strobl et al. (2022) compared an on-lattice ABM with an ODE and found that spatial structure changes competition inside the tumour, and with it the outcome of adaptive therapy. Gallaher et al. (2018) found that spatial heterogeneity changes time to recurrence under continuous and adaptive schedules, and Bacevic et al. (2017) found that spatial competition constrains resistance to targeted therapy. Cell turnover changes whether resistance must carry a fitness cost for adaptive therapy to work (Strobl et al. 2021). West et al. (2023) review the open questions between the mathematics of adaptive therapy and its clinical use.

Some rules already use more than tumour size. In the preclinical study of Enriquez-Navas et al. (2016), the dose was cut by 50% when tumour volume fell by at least 20% between scans and raised by the same amount after a rise of the same size. Gallaher et al. (2018) changed the dose according to the percentage change in population between measurements. Feedback through a model of the tumour has been studied with model predictive control (Hirata et al. 2014) and with dynamic programming (Gluzman et al. 2020). In prostate cancer treated with intermittent androgen deprivation, PSA doubling time before treatment and in the first off-treatment interval was associated with progression (Keizman et al. 2011), and exponential PSA kinetics in the first off-treatment period were associated with shorter cancer-specific survival (Sanchez-Salas et al. 2016). PSA dynamics have also been used to predict individual responses (Brady-Nicholls et al. 2020), and a range-bounded adaptive schedule has been proposed (Brady-Nicholls and Enderling 2022).

My literature search (Appendix A) was not systematic, so I make no claim about the field as a whole. Among the studies I found, none tested a threshold built with the construction of Gallagher et al. (2025) in a spatial model under discrete monitoring, and none asked whether information beyond size changes hold decisions under a fixed risk constraint. I addressed three questions in one spatial ABM. The first is whether the derived threshold keeps premature progression at a stated tolerance in the ABM, how much lower the threshold must be where it does not, and where along a chain of simpler models the shortfall appears. The second is whether information beyond size, such as the history of earlier measurements or the shape of the tumour, predicts premature progression in tumours of the same size. The third is whether a hold rule that uses this information reduces drug exposure or delays progression when it is tuned to the same nominal risk constraint as a size-only rule. The analysis plan for each experiment was written down and dated before its simulations were run.

The derived threshold was unsafe in the ABM at intervals of 7 to 28 days and safe at 42 to 84 days, where the construction's own margin of safety outweighed the shortfall. Recalibrating the threshold to the ABM's drug-free regrowth did not remove the shortfall. At short intervals most of it appeared in the step from a well-mixed stochastic model to the ABM, a step that changes several features at once, so the experiment does not identify which spatial feature is responsible. History and tumour geometry predicted premature progression in tumours of the same size when sizes were read without error, but a hold rule built on that prediction changed drug exposure and time to progression only a little, and only at two of four intervals.

2 Methods

2.1 Agent-based model

The ABM is a spatial tumour simulator built for two drugs, of which only one was used here. Cells occupy the sites of a square lattice with edge $L = 208$ sites, so the lattice has $n_{\max} = 43264$ sites, and each site holds at most one cell. Sensitive cells divide at rate $r_S = 0.040$ per day and are killed by the drug. Resistant cells divide at rate $r_R = 0.035$ per day and are not affected by it. Both die at a background rate $d = 0.005$ per day. A cell can divide only into an empty site among its eight neighbours. Time advances in steps of 1 day, and within a step each cell's death and division are treated as competing risks. A sensitive cell at a site with drug concentration C dies at rate $d + kC$, with $k = 0.15$. All parameter values are listed in Table 1.

The drug concentration $C(x, t)$ follows a reaction-diffusion equation on the lattice (1).

$$\frac{\partial C}{\partial t} = D\,\nabla^2 C - (\kappa + \kappa_{\text{cell}}\,\mathbf{1}_{\text{occ}}(x))C + \kappa\,u(t) \tag{1}$$

Here $D = 0.10$ in lattice units, $\kappa = 0.10$ per day sets the exchange with the circulation, $\kappa_{\text{cell}} = 0.05$ per day is extra uptake at occupied sites, and $u(t)$ is the full dose during treatment and zero during a hold. The simulator's second drug was kept at zero throughout.

Each run starts from a filled disk of radius 31 sites holding $N_0 = 3001$ cells, of which 1% are resistant. The resistant cells are chosen with a random-number stream separate from the simulator's own. Mutation is switched off, so all resistance is present from the start. The tumour has progressed when $N \geq N_{\text{prog}}$, with $N_{\text{prog}} = 1.2N_0 = 3601.2$ cells, and each run ends at progression or at 750 days. The simulator code was not changed at any point in this work.

Treatment follows a discrete-monitoring policy with threshold θ. Decisions are made every τ days, starting on day zero, with τ = 7, 14, 28, 42, 56 or 84 days. At each decision the tumour size N is read without error. If $N \geq \theta$ the drug is given for the next τ days. Otherwise treatment is held for the next τ days.

Table 1**.** Parameters of the agent-based model and of the growth law it shares with the ODE and CTMC. Values are read from the simulator and model code.

parameter	symbol	value	unit
initial tumour size	N_0	3001	cells
initial disk radius		31	sites
lattice edge	L	208	sites
lattice sites	$n_{\max}$	43264	sites
progression limit	N_{prog}	3601.2	cells
pre-existing resistant fraction		1	%
simulation horizon		750	days
time step	Δt	1	days
division rate, sensitive cells	r_S	0.040	per day
division rate, resistant cells	r_R	0.035	per day
background death rate	d	0.005	per day
drug kill coefficient	k	0.15	per day per unit concentration
drug diffusion coefficient	D	0.10	lattice units
drug exchange rate	κ	0.10	per day
drug uptake at occupied sites	κ_{cell}	0.05	per day
growth law, rim coefficient	c	3.534	
growth law, interior turnover	ϕ_{bulk}	0.2712	

2.2 ODE and well-mixed stochastic models

Three simpler models share the ABM's per-cell rates, and each leaves out one feature of it. Comparing their safe thresholds with the ABM's shows where along this chain the threshold stops being safe. The ODE describes the total cell number N. Without treatment it grows as in (2), where $\phi(N)$ is the fraction of cells that have an empty neighbour (3).

$$\frac{dN}{dt} = r_S\,\phi(N)\,N - d\,N \tag{2}$$

$$\phi(N) = (c\,N^{-1/2} + \phi_{\text{bulk}})\,(1 - \frac{N}{n_{\max}}) \tag{3}$$

The first term of ϕ counts cells on the rim of a compact colony and makes growth linear in diameter (Byrne and Drasdo 2009). Cell-based models of this kind have also been used for tumour growth in monolayers and spheroids (Drasdo and Höhme 2005). The second term, ϕ_{bulk}, counts interior cells next to sites emptied by death, and the last factor stops growth when the lattice is full. The constants c = 3.534 and ϕ_{bulk} = 0.2712 were fitted earlier in this project to drug-free growth of a related three-drug lattice model, on a smaller lattice and from a small starting colony. In that same small-colony setting the crowding fraction measured in the two-drug model matched the fitted curve within the scatter of the original fit (Appendix A). They do not describe the configuration used here. On the 208 by 208 lattice, starting from the disk of 3001 cells, the ABM regrew 2.41 times more slowly than this ODE (Section 3.1.3). The derived threshold tested in the main experiment uses these constants because that is the threshold a modeller would take from the ODE. The recalibration below refits the law to the ABM's own drug-free growth on the final configuration. Which difference between the two configurations (lattice size, starting geometry or starting

size) causes the slower regrowth was not determined. Under treatment the ODE kills sensitive cells at a constant effective concentration $C_{\text{eff}} = \kappa/(\kappa + \kappa_{\text{cell}})$ = 0.6667.

The well-mixed continuous-time Markov chain (CTMC) keeps the ODE's crowding factor and per-cell rates but draws each day's divisions and deaths at random. It therefore adds chance variation between tumours (demographic noise) without adding space. In the model called CTMC-nominal the drug acts at C_{eff}, as in the ODE. In the ABM, by contrast, the mean drug concentration over occupied sites rises over the first weeks of treatment and decays after treatment stops, so a short hold is not drug-free. The model called CTMC-pk replaces the constant with a one-compartment pharmacokinetic model fitted to the ABM's mean concentration, with input rate 0.10004 per day and elimination rate 0.13420 per day (time constant 7.45 days).

I arranged the four models as a chain (ODE, CTMC-nominal, CTMC-pk, ABM) in which each model adds one assumption to the one before. The change at each step therefore depends on the order of the chain and is not a share of variance. The chain has no spatial deterministic model and no ODE with the pharmacokinetic model. Its last step, from the CTMC to the ABM, adds contact-limited division, a discrete lattice and drug transport through space at once.

To recalibrate the threshold to the ABM, I measured how long untreated ABM tumours took to grow to N_{prog} from a range of starting sizes around N_0, with 200 runs per starting size. Each run started from a compact disk trimmed to the starting size, so these tumours had never been treated. Inverting the median of these transit times at τ gives $N^*_{\text{recal}}(\tau)$, the size from which the ABM's own median regrowth reaches N_{prog} in one interval. Inverting the fifth percentile gives $N_{q05}(\tau)$, a threshold that allows for the fastest-growing twentieth of tumours. The ODE and CTMC were then rerun under a growth law refitted by least squares to the median drug-free ABM trajectories on the final lattice and starting disk (the recalibrated law), which gave c = 6.32 and ϕ_{bulk} = 0.103. Its relative root-mean-square error against those trajectories was 1.294% on average and 1.784% at worst, above the bar of one percent set in the plan, so the plan used the inverted transit times for N^*_{recal}. The refitted law and the inverted transit times gave thresholds within 1.3–2.4% of each other at every interval. After the first coarse results, the CTMC grid was refined to steps of 0.0025 in the threshold multiplier m defined in Section 2.4.

2.3 Derived threshold

I used the construction of Gallagher et al. (2025) with the ODE of (2) and (3). Let $T(n)$ be the time the untreated ODE takes to grow from size n to N_{prog}. The threshold is the size whose regrowth time equals the monitoring interval (4).

$$T(N^*(\tau)) = \tau \tag{4}$$

With the capacity factor in place I solved (4) numerically, integrating (2) with $n_{\max}$ = 43264. Every value of N^* reported here is this root. Without the capacity factor the threshold has a closed form, given with its dependence on the speed of regrowth in Appendix B.

Three thresholds appear throughout, and I keep them apart. The derived threshold $N^*(\tau)$ is the output of the construction (4). The ODE safe threshold $N^{\text{ODE}}_{\text{safe}}(\tau)$ is the largest threshold at which the deterministic ODE, run under the discrete-monitoring policy, never progresses during a hold; I found it by bisection. The ABM safe threshold $N_{\text{safe}}(\tau)$ is the largest threshold at which premature progression in the ABM stays within the tolerance (Section 2.4). The derived threshold is safe in the ABM when $N^* \leq N_{\text{safe}}$, and I also

report the risk observed at $\theta = N^*$ directly. $N_{\text{safe}}^{\text{ODE}}$ differs from N^* because the derivation assumes the worst case: a tumour just below the threshold at a decision that then grows freely for a full interval. A fixed decision calendar does not produce exactly that case. I split the difference between N^* and N_{safe} into a derivation term and a transfer term (5). The derivation term compares the construction with its own ODE. The transfer term compares the ODE with the ABM. The split was added after the main experiment's data had been read and was planned for the recalibration (Table 2).

$$\text{derivation} = \frac{N_{\text{safe}}^{\text{ODE}}}{N^*} - 1, \qquad \text{transfer} = \frac{N_{\text{safe}}}{N_{\text{safe}}^{\text{ODE}}} - 1, \qquad \text{total} = \frac{N_{\text{safe}}}{N^*} - 1 \tag{5}$$

Under the recalibrated law, the CTMC-to-ABM share is the part of the gap between the ODE and the ABM that lies between CTMC-nominal and the ABM (6). All three safe thresholds are in cells. The numerator is the drop in safe threshold from CTMC-nominal to the ABM and the denominator the drop from the ODE to the ABM, so the share lies between zero and one hundred percent whenever the CTMC's safe threshold lies between the other two. It is an operational decomposition across model stages, not a causal effect of spatial structure: it depends on the order of the chain, and the step it measures changes several model features at once (Section 2.2). Its uncertainty was estimated by resampling the replicate seeds of the ABM and the CTMC (2,000 resamples, the deterministic ODE held fixed), an estimate added after all results had been read; it does not include the error of interpolating between grid points.

$$\text{CTMC-to-ABM share} = \frac{N_{\text{safe}}^{\text{CTMC}} - N_{\text{safe}}}{N_{\text{safe}}^{\text{ODE}} - N_{\text{safe}}} \tag{6}$$

2.4 Premature progression and the safe threshold

All endpoints are defined by the simulator and are not clinical endpoints. A run progresses on the first day on which $N \geq N_{\text{prog}}$, and the run then stops. Replicate s has a premature progression (indicator P_s) if, at some decision, $N < \theta$, treatment is held, and N reaches N_{prog} before the next decision; a progression while the drug is being given is not premature. Over R replicates, p_{prem} is the fraction of runs with a premature progression (7). Time to progression is the day of progression, censored at the 750-day horizon, and its restricted mean (RMST) gives censored runs the full horizon. The model has no death or other competing event, and size is read without error. The safe threshold is the largest threshold that keeps p_{prem} at or below the tolerance q (8). The tolerance q = 0.05 is a reliability level for the model and has no clinical basis. A threshold is called unsafe (anti-conservative) if it lets p_{prem} exceed q.

$$p_{\text{prem}}(\theta) = \frac{1}{R}\sum_{s=1}^{R} P_s(\theta) \tag{7}$$

$$N_{\text{safe}} = \max\{\theta : p_{\text{prem}}(\theta) \leq q\} \tag{8}$$

In the main experiment the threshold was set relative to N^*, as $\theta = mN^*(\tau)$, on a grid of multipliers m, with 500 replicates per grid point. The ODE and the two CTMCs were run on the same grid. The safe threshold was interpolated on that grid and then corrected for the grid's coarseness with two finer measurements (Appendix C). Wherever a corrected value exists, the magnitudes quoted in this paper use it.

2.5 Study design and statistical analysis

Each experiment followed a simulation protocol (Morris et al. 2019) that I wrote and dated before running it. The protocol fixed the hypotheses, the statistics and the rule for reading each result as a gain or not, and the analyses below follow these protocols unless stated. Table 2 lists which analyses were planned and which were added after the data they use had been read. The added analyses are exploratory, and each is marked where it appears. Each experiment drew its random-number seeds from its own range, and no range was used for more than one experiment (Table A1). Each run's seed fixes its starting tumour, so replicate s starts from the same tumour at every threshold of a grid (common random numbers). Confidence intervals (CIs) are 95% percentile intervals from a bootstrap that resamples replicates, keeping together all runs that share a seed. The prediction and policy analyses used 4,000 bootstrap resamples. The checks that validated each simulation program, and the errors they caught during the project, are listed in Appendix F.

Table 2**.** Planned and added analyses. Planned analyses were fixed in a dated plan before the simulations they use were run. Added analyses were made after the data they use had been read and are exploratory.

analysis (section)	status	why it was added
safe thresholds of the ABM, ODE and CTMCs on the main grid; total discrepancy from N^* (Section 3.1.1)	planned	
split of the discrepancy into derivation and transfer terms (Section 2.3)	added for the main experiment, planned for the recalibration	the total changed sign with τ; the split separates the construction from the transfer
share of the gap between the CTMC and the ABM (Section 3.1.4)	planned	
recalibration to the ABM's drug-free regrowth (Section 3.1.3)	planned	
refinement of the CTMC grid (Section 2.2)	added	the first CTMC grid was too coarse to place the crossing
correction of the ABM safe threshold for grid resolution (Section 3.1.2)	added	the main grid was too coarse to place the crossing
prediction comparisons, within-bin ΔAUC, spline-knot refit, training and test sets swapped (Section 3.2)	planned	
absolute scale of the prediction results (Section 3.2.3)	added	to show the size of the relative gains
history gain against the resistant count and against growth since the previous decision (Section 3.2.1)	added (before any policy-test run)	to test what history stands in for
policy test: tuning, scoring, readings, pooled rule, self-calibration (Section 3.3)	planned	
risk at N^*, tolerance, grid and growth-law checks, seed-resampling CIs, policy margin and horizon, replicate spread, measurement noise (Appendix D)	added after all results had been read	in response to review

To study prediction beyond size, I ran the adaptive policy at τ = 7, 14, 28 and 56 days, at 5 thresholds per interval spanning N_{safe}. Each threshold had 500 replicates in a training set and again in a separate test set (20000 runs in all). The unit of analysis was the hold decision. For every decision after the first at which treatment was held, the outcome was whether N reached N_{prog} before the next decision. The first decision was excluded because every run starts at N_0 with no history. The program that logged these decisions repeated the original policy code step for step and only read the simulator's state.

Every prediction model was a logistic regression with a weak ridge penalty (inverse strength 1.0), a cubic B-spline in N with 5 knots at training quantiles, and an indicator for each threshold. The five models differ only in the information they add (Table 3). The size model has only these terms. The history model adds what is known from earlier decisions: the days since treatment was last given (with an indicator for never treated), the growth ratio $\log(N/N_{\mathrm{prev}})$ since the previous decision, the number of earlier treated decisions and the lowest size since treatment was last given. The state model adds the mean drug concentration over

occupied sites, C, and the resistant cell count R. The geometry model adds the shape of the occupied region: rim cells B, resistant rim cells B_R, connected components K, radius of gyration R_g and maximum radius $R_{\max}$. The image-geometry model drops B_R, which cannot be seen without knowing which cells are resistant. All features are those the logging program recorded at the decision, before the decision was applied; C, R, B_R and the exact geometry are simulator quantities that a clinic could not measure in this form. The spline knots were placed at quantiles of N in the training set, with linear extrapolation beyond them, and every column was standardised with the training mean and standard deviation. Each model was fitted by maximum likelihood with the ridge penalty (L-BFGS solver) separately for each interval on the training set, its coefficients were frozen (their hash was recorded), and only then was the test set opened and scored. The feature sets, the penalty, the number of knots and the comparisons were fixed in the plan before any run, so no test outcome informed any choice of features or model. The training and test sets came from separate seed ranges at the same thresholds, and events were rare in both (counts in Section 3.2.1); no reweighting was used.

Table 3**.** Prediction models. Each model contains the terms of the model above it plus the terms listed.

model	information added	terms in the model
size	none (baseline)	spline in N; threshold-cell indicator
history	earlier decisions	size terms; days since treatment was last given; never-treated indicator; $\log(N/N_{\text{prev}})$; number of earlier treated decisions; nadir since treatment
state	drug level and resistant count	history terms; mean drug concentration C; resistant count R
geometry	shape of the tumour	state terms; rim cells B; resistant rim cells B_R; components K; radius of gyration R_g; maximum radius $R_{\max}$
image geometry	shape visible on an image	geometry terms without B_R

I made two primary comparisons. The history model against the size model tests whether history adds information, and the geometry model against the state model tests whether geometry adds information beyond history, drug level and resistant count. Secondary comparisons added the image-geometry features to the state model, to the history model plus C (because a clinic cannot measure R), and to the history model alone. The primary metric was the relative reduction in log-loss (cross-entropy) on the test set (9), where each log-loss is summed over all test hold decisions at an interval, and the pooled value sums over the four intervals before taking the ratio. The history result compares the history model with the size model, and the geometry result compares the geometry model with the state model. Its CI comes from resampling the test seeds, keeping all decisions of a seed together. After the test set had been read, the frozen models were also compared with a base-rate model that predicts the overall event rate, to show the absolute scale. The comparison used log-loss per decision and the Brier score (Brier 1950), the mean squared difference between predicted risk and outcome.

$$\Delta = 1 - \frac{\text{LL}_{\text{test model}}}{\text{LL}_{\text{baseline}}} \tag{9}$$

A comparison counted as a gain if $\Delta \geq 10\%$ and the lower CI bound was above zero, and as no gain if the upper CI bound was below 10%. A pooled result required the same reading, with the same sign, at two of the four intervals. Because the risk of progression rises steeply with N near the limit, a model can improve log-loss simply by ranking sizes better. I therefore also compared the models within narrow bands of size. Test decisions were grouped into bins of 20 cells. A bin was used if it held at least 10 events and 10 non-

events, and at least 3 usable bins were required. The statistic, ΔAUC within size bins, is the event-weighted mean over usable bins of the difference in the area under the receiver operating characteristic curve (AUC). A gain whose within-bin CI includes zero is not described as a gain at matched size.

For the policy test, each rule holds treatment at a decision after the first if N is at most a cap $N_{\mathrm{top}}(\tau)$ and the frozen model's predicted risk $\hat{p}$ is at most a setting p^* (10), and treats otherwise.

$$\text{hold} \Leftrightarrow N \leq N_{\mathrm{top}}(\tau) \text{ and } \hat{p} \leq p^* \tag{10}$$

The cap equals the highest training threshold (3300, 3100, 2950 and 2980 cells at the four intervals), because the models never saw a hold above it. At the first decision every rule copies the tuned size-only decision. The threshold indicator in each model was fixed at one threshold, which adds a constant to the logit that tuning p^* absorbs. Of the five rules compared, the size-only rule holds if $N < \theta$. The primary rule, history-and-image, uses the history model plus the image-geometry features, and so uses only information a clinic could in principle observe. The history rule uses the history model. The full-state rule uses the geometry model, including R and C, and serves only to show what unobservable information adds. The growth-rate rule uses size and threshold plus the change in N since the previous decision.

Each rule was tuned by running it. A deterministic bisection of 8 steps, with 500 replicates per setting, chose the most permissive setting with $p_{\mathrm{prem}} \leq q$ on the tuning replicates. Every rule was therefore tuned to the same nominal risk constraint, which places each tuned risk at or below q but does not make the risks of two rules equal. Whether the rules carried similar risk was judged only from new scoring replicates, through the safety and non-inferiority checks below. Scoring used 1,000 new replicates per rule and interval. The outcomes were the on-drug fraction f (days on drug divided by run length) and the restricted mean time to progression (RMST) over the 750-day horizon, each as a paired difference from the size-only rule on the same replicates. A rule had a safety failure if the Wilson lower bound (Wilson 1927) of its p_{prem} exceeded q. It was non-inferior in risk if the upper CI bound of Δp_{prem} was at most 0.02, a margin chosen without clinical basis. A rule that passed both checks counted as a gain in exposure if its on-drug fraction fell with a CI below zero and as a gain in time if RMST rose with a CI above zero. If neither improved, it showed no gain. A rule that failed either check was not assessable. The pooled value is the equal-weight mean over the four intervals, and a pooled result again required two intervals with the same reading and sign. An interval was flagged as unmatched if the tuned p_{prem} of the history-and-image rule and the size-only rule differed by more than 0.010.

The on-drug fraction was chosen as the primary outcome, in place of total days on drug, before any scoring run. A rule that delays progression keeps a tumour in the study for longer, and so accumulates more days on drug only because the run lasts longer. Total days on drug were recorded as a secondary outcome. Each tuned setting was frozen before any scoring replicate was run, and the size-only rule was tuned first at each interval because every rule copies its first decision.

2.6 Use of AI tools

I used Claude Code (Anthropic) to write the simulation and analysis code and to run the analyses, under my direction and used Claude.ai to assist with task briefs. ChatGPT (OpenAI), DeepSeek and Gemini (Google) critiqued the plans and results over several rounds. Every number in the paper is produced by the scripts described in Appendix A. The error log summarized in Appendix F was generated by Claude Code during the project.

3 Results

3.1 The derived threshold in the spatial model

3.1.1 Derivation and transfer terms

The derived threshold was unsafe in the ABM at 7, 14 and 28 days and safe at 42, 56 and 84 days (Table 4; a direct count added after all results had been read, beside the planned total term). With the threshold set at N^*, premature progression occurred in 488, 487 and 240 of 500 ABM runs at the three short intervals (p_{prem} = 0.976, 0.974 and 0.480), and in 7, 5 and 5 of 500 at the three long ones, with Wilson upper bounds of at most 0.029. The ODE never progressed prematurely at N^*. The direct counts agree with the sign of the total term in Table 5, which is negative exactly at the three short intervals.

The transfer term, which compares the ABM with the ODE's own safe threshold, was negative wherever it was resolved. After the grid correction of Section 3.1.2, it was −7.37%, −8.95%, −8.06% and −4.40% at τ = 7, 14, 28 and 56 days, each with a seed-resampling CI below zero (−5.28% to −3.23% at 56 days, narrower elsewhere; these CIs, and the grid check below, were added after all results had been read). At 42 and 84 days, which were not remeasured on a finer grid, the main-grid values were −4.00% and −0.15%. Without interpolation, the grid points either side of the crossing put the term between −5.47% and −0.74% at 42 days, so it is negative there, and between −5.12% and +2.17% at 84 days, so its sign at 84 days is not resolved. The term was largest at intervals of 7 to 28 days. It did not shrink steadily, since its size at 14 days was larger than at 7 days, both before and after the correction.

The derivation term had the opposite sign. The ODE's own safe threshold was above N^* by +0.84% at 7 days, rising to +37.02% at 84 days. At intervals up to 28 days the ODE's safe threshold is set by runs in which the drug never starts before progression, so there the derivation term is only a lower bound. From 42 days on it is a measurement. The total difference between N^* and the ABM changed sign between 28 and 42 days (−7.28% and +1.55%). The change of sign comes from the growing positive derivation term overtaking the transfer term, which stayed negative at every interval where its sign is resolved.

Table 4. Premature progression at the derived threshold ($\theta = N^*$, multiplier m = one) in the main experiment. Counts are runs with a premature progression out of the replicates at that grid point, with Wilson 95% intervals. The ODE is deterministic. The derived threshold is unsafe where p_{prem} exceeds q = 0.05. Added after all results had been read.

τ (d)	N^* (cells)	ABM runs with premature progression	ABM p_{prem} [Wilson CI]	CTMC-nominal p_{prem}	CTMC-pk p_{prem}	ODE premature progression	derived threshold in the ABM
7	3425.2	488 / 500	0.976 [0.959, 0.986]	0.200	0.146	none	unsafe
14	3255.3	487 / 500	0.974 [0.956, 0.985]	0.050	0.060	none	unsafe
28	2933.5	240 / 500	0.480 [0.437, 0.524]	0.006	0.004	none	unsafe
42	2635.3	7 / 500	0.014 [0.007, 0.029]	0.002	0.000	none	safe
56	2359.7	5 / 500	0.010 [0.004, 0.023]	0.000	0.000	none	safe
84	1872.9	5 / 500	0.010 [0.004, 0.023]	0.000	0.000	none	safe

Figure 1 shows p_{prem} against the threshold for all four models. At 7 days and $\theta = N^*$, p_{prem} was 0.200 in CTMC-nominal and 0.976 in the ABM, while the ODE had no premature progression there. Chance variation alone therefore produces premature progression at short intervals in this model. The ABM's p_{prem} was above CTMC-nominal's at $\theta = N^*$ but not everywhere: it was lower at 21 of 64 of the grid points. From 7 to 84 days the ABM's safe threshold fell by 17.73%, against 45.32% for N^*.

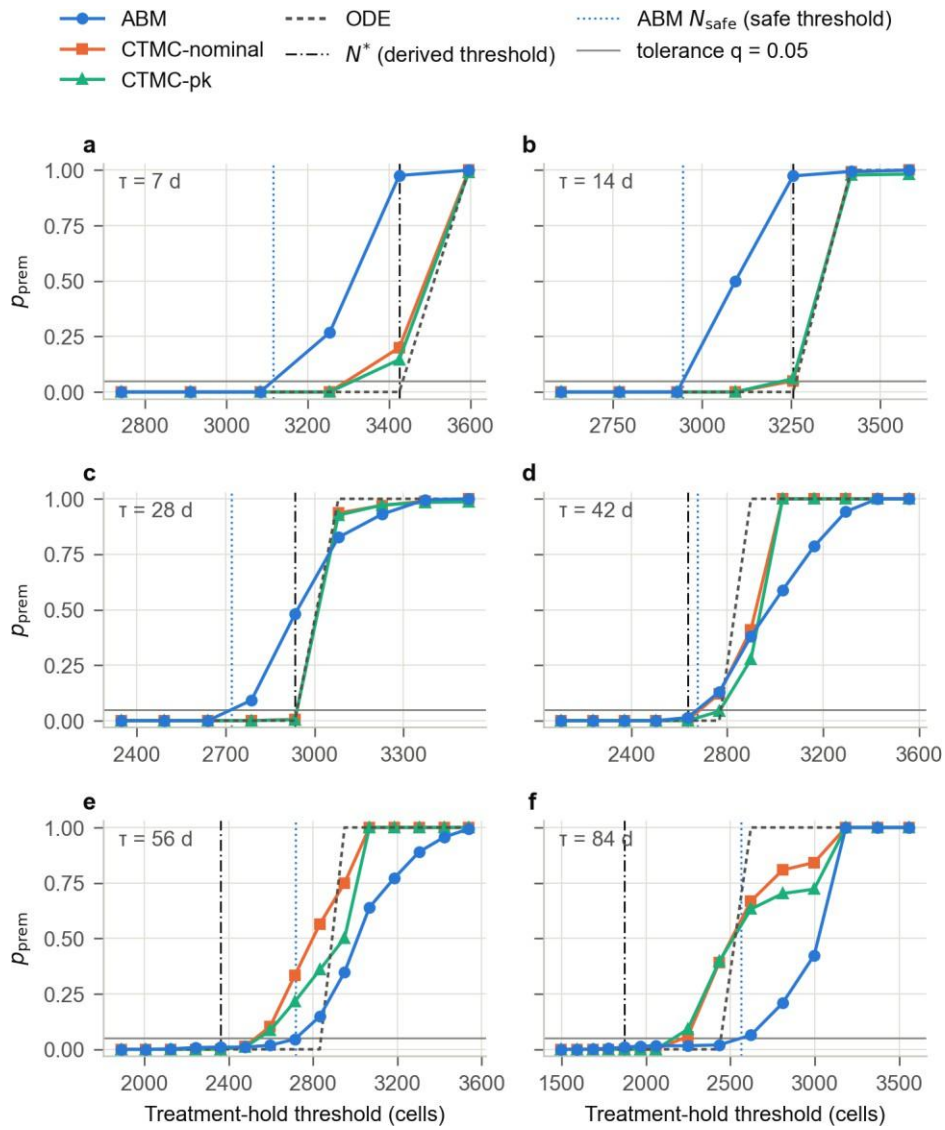

Figure 1. Probability of premature progression p_{prem} against the treatment-hold threshold in the main experiment. ABM, agent-based model; CTMC, continuous-time Markov chain; ODE, ordinary differential equation model. Panels (a)–(f) show the monitoring intervals τ = 7, 14, 28, 42, 56 and 84 days. Each ABM and CTMC point is the mean of 500 replicates. Dash-dot line, derived threshold $N^*(\tau)$; dotted line, ABM safe threshold interpolated on this grid; horizontal line, tolerance q = 0.05.

Table 5. Transfer of the derived threshold to the ABM in the main experiment. N^* is the derived threshold, N_{safe} the ABM safe threshold interpolated on the main grid with its CI, and the ODE safe threshold was found by bisection. The three terms are defined in (5). Every column except the corrected transfer term is on the main grid. The corrected transfer term uses the finer-grid N_{safe} of Table C1 and exists only for the four remeasured intervals.

τ (d)	N^* (cells)	N_{safe} (cells)	CI (cells)	ODE safe threshold (cells)	derivation (%)	transfer, main grid (%)	transfer, corrected (%)	total (%)	grid step (cells)
7	3425.2	3114.6	[3110.5, 3119.9]	3454.0	+0.84	−9.83	−7.37	−9.07	171.3
14	3255.3	2946.1	[2944.8, 2947.6]	3299.4	+1.36	−10.71	−8.95	−9.50	162.8
28	2933.5	2719.9	[2702.3, 2748.0]	3001.0	+2.30	−9.37	−8.06	−7.28	146.7
42	2635.3	2676.2	[2662.4, 2693.2]	2787.7	+5.78	−4.00	--	+1.55	131.8
56	2359.7	2718.3	[2666.4, 2739.8]	2831.6	+20.00	−4.00	−4.40	+15.20	118.0
84	1872.9	2562.4	[2499.5, 2630.8]	2566.2	+37.02	−0.15	--	+36.82	93.6

3.1.2 Grid resolution of the safe threshold

The main grid was too coarse to place the safe threshold accurately, which I found after the main results had been filed. Remeasuring the same policy on the finer grid of the prediction study moved N_{safe} by +84.7, +58.2, +39.2 and −11.2 cells at τ = 7, 14, 28 and 56 days (Table C1). The bias does not always have the same sign (Appendix C).

After the correction the transfer term stayed negative at all four remeasured intervals, the recalibrated threshold stayed unsafe (Section 3.1.3), and the CTMC-to-ABM shares moved by less than two percentage points (Section 3.1.4). What changed is the size of the transfer term at short intervals. At 7 days it went from −9.83% on the main grid to −7.37% (Table C2). The sign of the transfer term also did not depend on the tolerance: with q = 0.025 or 0.10 in place of 0.05, the corrected term stayed negative at 7, 14 and 28 days, and the main-grid term stayed negative from 7 to 56 days. The check was added after all results had been read (Appendix D).

3.1.3 Recalibration to the ABM's regrowth

The ABM regrew more slowly than the ODE built on these constants. Without treatment, the median time to grow from N_0 to N_{prog} was 60.3 days in the ABM and 24.98 days in the ODE, a ratio of 2.41. The ratio compares the two models only and says nothing about tumours in patients. In the closed form of Appendix B, N^* falls as regrowth speeds up, so a model that regrows more slowly calls for a higher threshold. A threshold matched to the ABM's own median regrowth was indeed higher than N^* at every interval, for example 3581.8 against 3425.2 cells at 7 days (Table E1). On regrowth speed alone, then, the ODE-derived threshold should have erred on the safe side in the ABM. At short intervals it erred on the unsafe side. I therefore recalibrated the threshold to the ABM's own regrowth on the final lattice and starting disk, to see whether the gap would close. The refitted growth law reproduced the median drug-free trajectories within 1.784% (Section 2.2), and transit times varied little between runs, with a median coefficient of variation of 4.49%.

The gap did not close. On the recalibration grid, before the grid correction, and compared with the threshold matched to the ABM's median regrowth, N_{safe} differed by −5.94% to −19.10%. Compared with the threshold matched to the fifth percentile of regrowth, it differed by −3.60% to −18.29%, with the extremes at 84 and 28 days. With the corrected ABM values the differences from the fifth-percentile threshold were −9.97%, −14.15%, −17.23% and −9.76% at 7, 14, 28 and 56 days. A threshold calibrated to the ABM's own regrowth, on the median or on the fifth percentile, stayed unsafe at every interval. The ODE run under the refitted law gave the same answer (a comparison added after all results had been read): the corrected ABM safe threshold lay −10.02%, −13.70%, −19.49% and −23.03% from that ODE's safe threshold at the four intervals. The shortfall is therefore not explained by the mismatch in drug-free regrowth speed between the original ODE and the ABM.

3.1.4 Where in the chain of models the gap appears

Under the recalibrated law, most of the gap at short intervals lay between CTMC-nominal and the ABM (Figure 2). With the corrected ABM values the CTMC-to-ABM share (6) was 88.1% (CI 87.2 to 89.0), 94.1% (92.7 to 95.0) and 84.2% (83.6 to 85.9) at 7, 14 and 28 days, against 89.4%, 94.2% and 85.0% on the main grid (Table E2); the CIs come from seed resampling added after all results had been read. For example, at 14 days the numerator was the CTMC safe threshold (3453.2 cells) minus the corrected ABM

value (3004.2 cells), and the denominator the ODE safe threshold (3481.3 cells) minus the same ABM value. When the fine CTMC grids were resampled, the share between the ODE and the CTMC stayed within [10.9,11.4], [6.1,6.9] and [12.9,15.8] percent. At 56 days the share was 45.9%, but the range [47.1,58.3] of the ODE-to-CTMC share includes one half, so the split at 56 days is unresolved. At 42 and 84 days the ODE's safe threshold under the recalibrated law is set by the timing of the decision calendar instead of by regrowth within an interval, so those two intervals have no share.

The split locates the gap in the chain of models and leaves its cause open. Because the chain adds one feature at a time in a fixed order, a different order could assign the gap differently, and the chain contains no spatial deterministic model and no ODE with the pharmacokinetic model with which to test this. The CTMC in the split is CTMC-nominal, so the step to the ABM bundles contact-limited division, lattice discreteness and drug transport through space. The result shows that the gap appears when these features are added together, not that any one of them, or the spatial arrangement of cells alone, causes it.

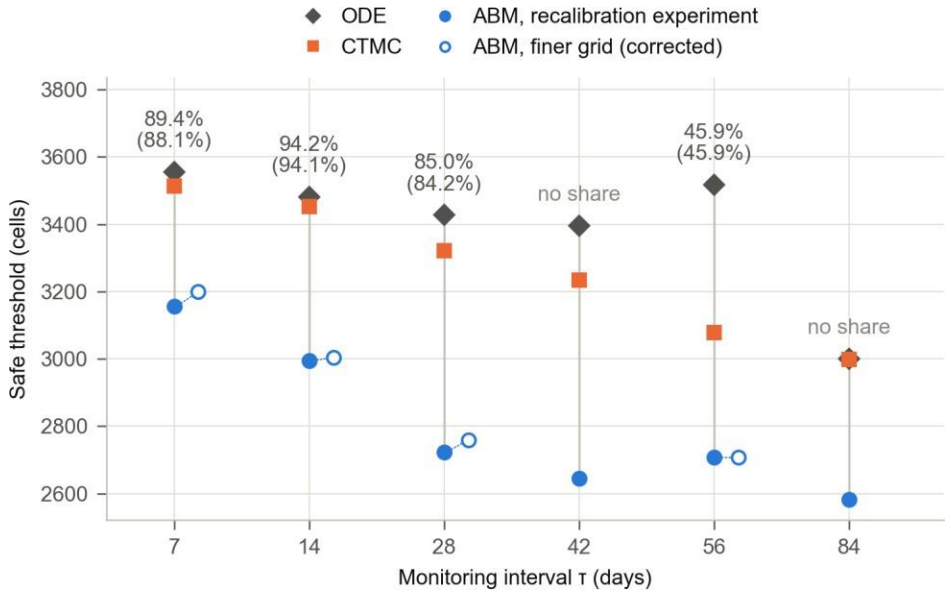

Figure 2. Safe thresholds of the ODE, the CTMC and the ABM under the recalibrated growth law, by monitoring interval. ABM, agent-based model; CTMC, continuous-time Markov chain; ODE, ordinary differential equation model. Filled circles, ABM safe threshold from the recalibration experiment; open circles, the same after the grid correction. The numbers above each interval give the CTMC-to-ABM share (6) on the main grid and, in parentheses, after the correction. At 42 and 84 days the ODE threshold is set by the decision calendar and no share is defined.

3.2 Prediction of premature progression beyond tumour size

3.2.1 Measurement history

The test set held 89069, 39359, 18048 and 8171 hold decisions at τ = 7, 14, 28 and 56 days, of which 431, 403, 479 and 429 ended in premature progression. The training set had 408, 410, 499 and 414 events. Longer intervals give fewer decisions per run and a higher event rate per decision, which rose from 0.484% at 7 days to 5.250% at 56 days.

The history of earlier decisions predicted premature progression better than size and threshold alone at every interval (Figure 3). Adding history reduced test log-loss by +46.2%, +61.5%, +50.7% and +10.4% at τ = 7, 14, 28 and 56 days, and by +48.6% pooled (CI +46.7 to +50.3; Table E3). Within size bins, ΔAUC

was +0.388, +0.369 and +0.377 at the first three intervals, each with a CI above zero. At 56 days it was +0.074 (CI −0.012 to +0.144), so at 56 days the gain is not shown to hold at matched size.

At 7 and 14 days, drug from the last treated interval is still present during a hold, so history could be standing in for leftover drug. With C in both models the gain was +45.5% and +57.1% at those intervals and +46.4% pooled, close to the values without C.

Two further checks were added after the test set had been read. History was not standing in for the resistant count, which a clinic cannot measure: with R in both models the pooled gain from history was +50.3% (CI +48.3 to +52.1), against +48.6% without R, and R alone added +2.1% (CI +1.4 to +2.8) over size and threshold (Table E4). The partial exception was 28 days, where ΔAUC within size bins fell from +0.377 to +0.207 once R was in both models. Growth since the previous decision alone (size and threshold plus the change in N) recovered 35.1% (CI 33.2 to 37.0) of history's pooled gain. The share was lowest at 56 days (−19.1%) and highest at 14 days (52.8%).

3.2.2 Tumour geometry

Lattice geometry added predictive information beyond size, history, leftover drug and resistant count. Adding geometry to the state model reduced test log-loss by +11.5%, +35.4%, +41.8% and +41.2% at the four intervals and by +30.5% pooled (CI +27.3 to +33.5; Table E5). The pooled gain was +30.0% for image geometry without the resistant rim count, +33.3% for image geometry against a model without R, and +33.1% for image geometry added to history alone. The comparison without R is the relevant one in practice, because a clinic cannot measure R, and its gain was larger than the gain against a model that contained R.

At matched size the evidence was thin at short intervals. At 7 and 14 days, ΔAUC within size bins was +0.012 and +0.012, so the gain there comes mainly from sharper probabilities within a size band and little from better ranking. It was +0.035 at 28 days and +0.324 at 56 days, the latter with a wide CI (+0.084 to +0.509). The geometry result concerns prediction and shows neither a mechanism nor better control.

The geometry gain is not a measure of what spatial structure contributes. History is measured without any spatial information, but in this model it arises from spatial dynamics, so the predictive content of the two cannot be divided into spatial and non-spatial parts with these comparisons.

3.2.3 Absolute scale

The gains are reductions on a small number (an analysis added after the test set had been read). Pooled over intervals, 1.126% of test hold decisions ended in premature progression. The log-loss of the base-rate model was 0.0617 nats per decision, the size model was already 58% lower at 0.0258, and the history model reached 0.0132 (Table E6). The pooled Brier score fell from 0.00754 to 0.00395. Refitting with 8 spline knots gave pooled gains of +48.3% and +31.2% for the two primary comparisons. Swapping the training and test sets gave +48.1% and +33.1%, and both checks gave the same reading in every comparison at every interval.

3.2.4 Measurement noise

Every result above uses exact size readings and exact lattice geometry, which no clinic has. In a check added after all results had been read, I multiplied each size reading, and the rim count, radius of gyration and maximum radius, by independent noise with mean one and a coefficient of variation of 5%, 10% or 20%. I then recomputed the history features from the noisy readings, refitted the size, history and history-

and-image models on the noisy training set and scored them on the noisy test set. The outcomes and the decisions were those of the original runs. Against a size model that saw the same noisy readings, history still reduced pooled log-loss by 27.5%, 29.0% and 34.2%, partly because several noisy readings estimate the current size better than one. Against a size model with exact readings, the noisy history model was 11.5% and 2.0% better at the two lower noise levels and −1.0% at the highest. Image geometry over history fell from 33.1% with exact readings to 24.1%, 17.1% and 12.5%. Most of the predictive advantage of history over an exact size reading therefore depends on error-free measurement, and the practical value of these features is untested on real images (Appendix D).

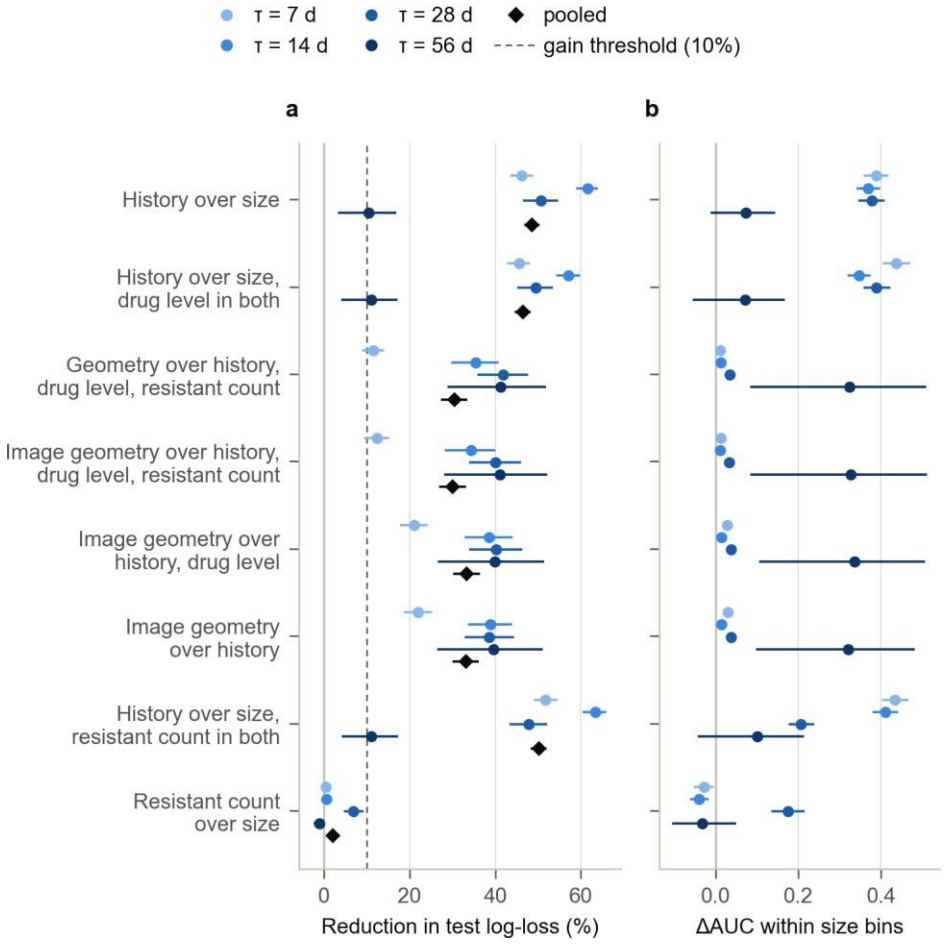

Figure 3**.** Improvement in predicting premature progression at each hold decision, measured on the test set. Each row adds the information named first to a model containing the information named after 'over'. (a) Relative reduction in test log-loss (9), by monitoring interval (blue shades) and pooled (black diamonds). Dashed line, the 10% threshold for a gain. (b) Difference in the area under the ROC curve within bins of similar tumour size. Error bars, 95% bootstrap CIs.

3.3 Hold rules that use more than size

3.3.1 Primary comparison

Under a common nominal risk constraint, the history-and-image rule changed the pooled on-drug fraction by −1.4 percentage points relative to the size-only rule, or −3.7% (CI −4.2 to −3.3). It lengthened RMST by 4.0 days (CI +2.5 to +5.5) within a 750-day horizon (Table 6, Figure 4). The pooled difference in scored

p_{prem} was +0.004 (CI −0.003 to +0.012). By the planned rule the pooled result counted as a gain in both exposure and time, supported at two of the four intervals.

In the scoring runs the size-only rule had a scored p_{prem} of 0.034, 0.028, 0.043 and 0.036 at 7, 14, 28 and 56 days, an on-drug fraction of 0.328, 0.375, 0.393 and 0.432, and an RMST of 422.5, 369.7, 330.9 and 308.2 days. Its tuned p_{prem}, from the tuning runs, was 0.042, 0.046, 0.044 and 0.050 (Table E7); the scoring runs used new seeds, so the two sets of values differ. Against that baseline, the history-and-image rule changed the on-drug fraction by −4.9%, −4.2%, −5.7% and −0.7% and RMST by +7.3, +2.0, +6.8 and −0.1 days.

The achieved risks were similar but not shown to be equal (this and the next paragraph report checks added after all results had been read). In scoring, the size-only rule's p_{prem} had Wilson intervals of 0.024 to 0.047, 0.019 to 0.040, 0.032 to 0.057 and 0.026 to 0.049 at the four intervals. Those of the history-and-image rule were 0.028 to 0.052, 0.028 to 0.052, 0.034 to 0.060 and 0.028 to 0.052. Every upper bound of the history-and-image rule lay above q, so its risk is not shown to meet the tolerance, although no lower bound did either, so no safety failure occurred. The pooled difference in risk from the size-only rule, +0.004 (CI −0.003 to +0.012), also leaves room for a higher risk. Passing the planned non-inferiority checks at two intervals therefore does not show that the rule is safe: its risk on the scoring runs was not statistically established to be within the tolerance.

The pooled means hide large differences between replicates. The paired difference in RMST had a standard deviation of 88.4, 36.3, 28.3 and 7.5 days over replicates at the four intervals. At 7 days the history-and-image rule delayed progression in 60.3% of replicates and advanced it in 38.8%; at 28 days the figures were 53.5% and 13.9%. The paired difference in on-drug fraction had a standard deviation of 4.2 to 5.6 percentage points, several times the mean change. The rule therefore changed individual courses a great deal in both directions, and the pooled gain is a small difference between large changes.

The result is fragile. It rests on the intervals of 7 and 28 days. At 7 days the upper bound of the scored Δp_{prem} was +0.020, exactly the non-inferiority margin of 0.02. With 1,000 replicates, the scored Δp_{prem} moves in steps of one part in 1,000, and a bound one step higher would have failed the margin. The 7-day interval would then have been not assessable, leaving one supporting interval. At 14 days the rule was not assessable because non-inferiority was not shown (upper bound +0.025). At 56 days it showed no gain and was flagged as unmatched, because its tuned p_{prem} (0.038) differed from the tuned p_{prem} of the size-only rule (0.050) by more than 0.010 (Table E7). The image features this rule used are exact lattice geometry, with no measurement noise.

The reading also depends on choices made without clinical basis (checks added after all results had been read; Appendix D). With a non-inferiority margin of 0.01 no interval was assessable and there was no pooled result. With a margin of 0.03 the 14-day interval became a gain in exposure, and the pooled result rested on three intervals. Over a horizon of 365 days in place of 750, the pooled change in RMST was 0.95 days (CI −0.01 to +1.93), so the gain in time was no longer shown; only the 28-day interval kept one (5.1 days, CI 3.4 to 6.8). The gain in time reported above therefore comes mostly from the second year of follow-up, and the result is best described as a small reduction in time on drug at two intervals, with a gain in time that depends on the horizon. The policy conclusion is conditional on the chosen cap, tuning procedure, non-inferiority margin and 750-day horizon, all of them modelling choices; the cap and the tuning procedure could not be varied without new simulations and were not.

Table 6**.** Primary comparison: the history-and-image rule against the size-only rule in the scoring runs (1,000 replicates per rule and interval). The scored p_{prem} is that of the history-and-image rule; the size-only rule's scored values are given in Section 3.3.1 and the tuned values of both rules in Table E7. f is the on-drug fraction and RMST the restricted mean time to progression. Differences are paired by replicate (history-and-image minus size-only), and the pooled row is the equal-weight mean over intervals; in absolute terms the pooled change in on-drug fraction was −1.4 percentage points. Readings follow Section 2.5; not assessable means that non-inferiority in risk was not shown.

τ (d)	scored p_{prem}	relative Δf (%) [CI]	ΔRMST (d) [CI]	scored Δp_{prem} [CI]	reading
7	0.038	−4.9	+7.3 [+1.9, +12.8]	+0.004 [−0.012, +0.020]	gain (exposure, time)
14	0.038	−4.2	+2.0 [−0.2, +4.3]	+0.010 [−0.004, +0.025]	not assessable
28	0.045	−5.7	+6.8 [+5.0, +8.6]	+0.002 [−0.015, +0.019]	gain (exposure, time)
56	0.038	−0.7	−0.1 [−0.6, +0.4]	+0.002 [−0.012, +0.016]	no gain
pooled		−3.7 [−4.2, −3.3]	+4.0 [+2.5, +5.5]	+0.004 [−0.003, +0.012]	gain (exposure, time)

3.3.2 Other rules and safety

The full-state rule, which also sees drug concentration and resistant cells, changed the pooled on-drug fraction by −3.9% (CI −4.3 to −3.5) and met the pooled-result rule on the same two intervals (Table E8). Interval by interval it gave the same readings as the history-and-image rule. The history rule changed the pooled on-drug fraction by −2.4% (CI −2.7 to −2.0) but did not meet the pooled-result rule, because only the 28-day interval showed a gain. At 7 and 14 days it was not assessable, and at 56 days it showed no gain (Table E9).

Running a rule changed the risk it had been tuned to. The growth-rate rule at 7 days had a safety failure, with p_{prem} = 0.069 and a Wilson lower bound of 0.055, above q. The paired Δp_{prem} had a lower CI bound above zero for the growth-rate rule at 7 and 14 days (+0.017 and +0.005) and for the history rule at 14 days (+0.002).

The growth-rate rule, which uses only size and its change since the last decision, changed the pooled on-drug fraction by −2.1% and raised pooled risk, with Δp_{prem} = +0.009 (CI +0.003 to +0.016). Its pooled recovery ratio against the history rule was ρ = 0.89. That value passes the planned bar of 0.80 only arithmetically, because the ratio could be assessed at one interval only, 28 days, where ρ = 0.19. It therefore does not show that size and growth rate carry the information needed for control.

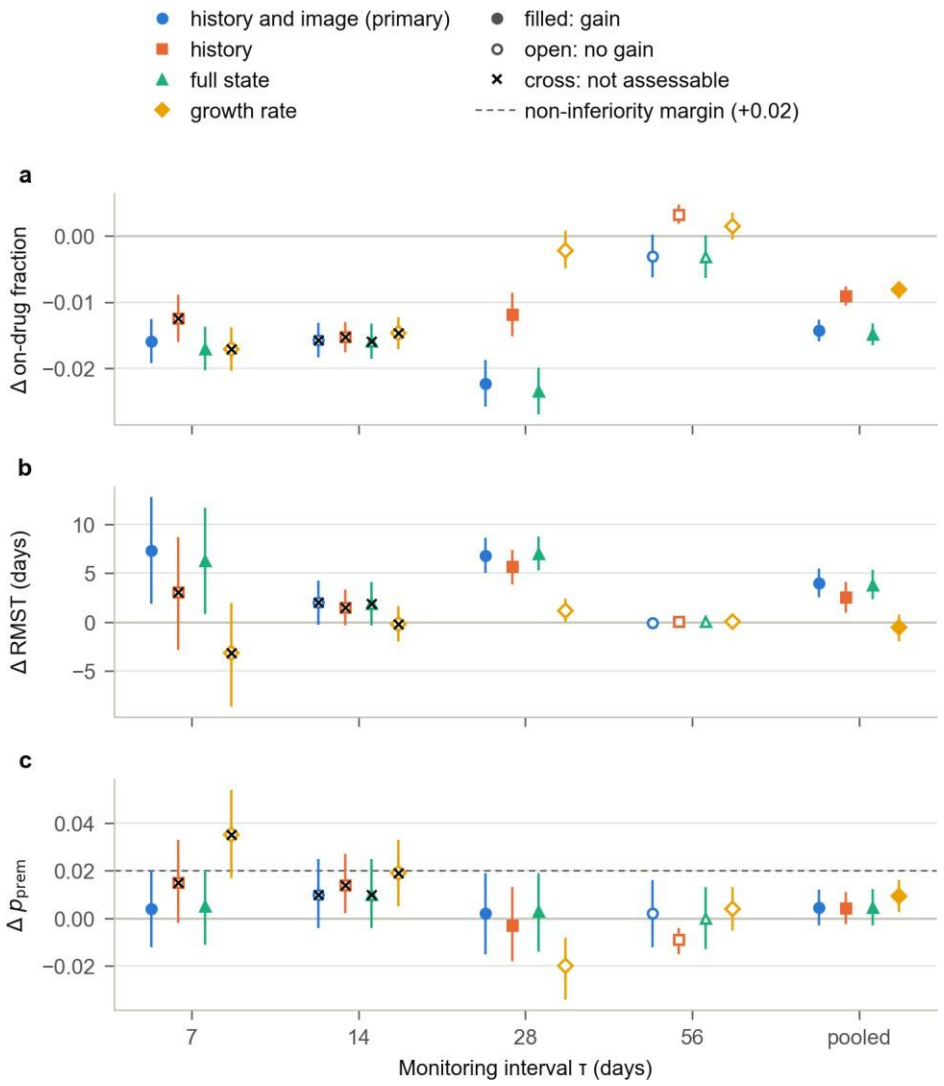

Figure 4. Paired differences between each hold rule and the size-only rule in the scoring runs (1,000 replicates per rule and interval). (a) On-drug fraction. (b) Restricted mean time to progression (RMST) over 750 days. (c) Probability of premature progression p_{prem}; dashed line, non-inferiority margin of 0.02. Filled markers, gain; open markers, no gain; crosses, not assessable. Error bars, 95% bootstrap CIs.

3.3.3 Calibration under the rules

The models' risk estimates did not hold under the rules they drove. For the history-and-image rule, the ratio of observed to expected premature progressions on the hold decisions it made during scoring was 0.28, 0.86, 2.46 and 1.01 at 7, 14, 28 and 56 days. At 7 days the model overstated the risk of the holds it chose, by a factor of more than three, and at 28 days it understated it by a factor of about two and a half. The models were trained on hold decisions made under threshold policies, and a rule that chooses its own holds visits different states. For this reason every rule's p^* was tuned by running the rule, and no model probability was used as a risk estimate. The risk constraint was therefore met, to the extent it was, by tuning on simulated outcomes and not by the model's calibration, and it was checked on new replicates only through the scoring risks above. A different cap, or a tuning procedure that also retrained the models on holds chosen by the rule, was not tested and could change the comparison.

4 Discussion

The first experiment supports two claims and leaves a third open. In this ABM, the threshold built with the construction of Gallagher et al. (2025) did not keep its promised safety at intervals of 7 to 28 days: at the derived threshold almost every run progressed prematurely at 7 and 14 days, and about half did at 28 days. The shortfall relative to the ODE's own safe threshold also survived recalibration of the threshold, and of the ODE, to the ABM's drug-free regrowth on the final configuration. The claim that spatial structure causes the shortfall is not established. The decomposition places most of the short-interval gap in the step from CTMC-nominal to the ABM, but that step changes contact-limited division, lattice discreteness and drug transport at once, and its result depends on the order of the chain. It describes where the gap appears across model stages, not what produces it.

The derivation and transfer terms answer different questions. The derivation term compares the construction with the ODE it was built from, run on a fixed decision calendar, and it erred on the safe side. The transfer term compares that ODE with the ABM on the same calendar, and it erred on the unsafe side wherever its sign was resolved. At short intervals the second outweighed the first. At long intervals the first outweighed the second, and the total changed sign without any change in the ABM. The transfer term was still −4.00% at 42 days (main grid) and −4.40% at 56 days (corrected), so at those intervals the derived threshold was safe in the ABM because the construction's slack exceeded the transfer error, not because the ODE described the ABM well.

Chance variation alone produced premature progression at the shortest interval. At 7 days and $\theta = N^*$, CTMC-nominal progressed prematurely in a fraction 0.200 of runs, while the ODE never did. The construction is a worst case for one deterministic trajectory, built so that premature progression cannot occur in that model. In a population of tumours that regrow at different speeds, the threshold needs a stated tolerance, which the construction does not supply. In this model the problem was sharpest where N^* sits near the limit, as at 7 days, where N^* = 3425.2 cells against a progression limit of 3601.2 cells.

The recalibration result points to a candidate explanation that this study did not test. The recalibration matched the threshold to the drug-free regrowth of compact disks that had never been treated, and every recalibrated threshold lay above N^*, as slower regrowth implies. The ABM's safe threshold nonetheless stayed below even the threshold set at the fifth percentile of those transit times, by −3.60% to −18.29% on the recalibration grid (−9.76% to −17.23% after the grid correction). The holds in the policy follow treatment, and a tumour shrunk by treatment need not regrow like an untreated disk of the same size. The prediction study is consistent with this: at the same size and threshold, the history of earlier decisions separated holds that ended in premature progression from those that did not, with ΔAUC within size bins of +0.388, +0.369 and +0.377 at 7, 14 and 28 days. Read together, the two results suggest that tumour size is not a sufficient state for near-term regrowth in this model, which is the property a single size threshold relies on. Gallagher et al. (2025) met a non-spatial form of the same problem: in a model in which regrowth speeds up as stem cells accumulate, a fixed threshold allowed premature progression and the threshold had to fall over the course of treatment. In the ABM the relevant history differs between tumours, so a threshold would have to follow each tumour's own history, not time alone. Three runs, none done here, would separate the candidate explanations. Recalibrating to regrowth measured from the states the policy holds in should close more of the gap than recalibrating to untreated disks did if treatment history is the cause. A spatial deterministic model would separate lattice structure from chance variation, and an ODE with the pharmacokinetic model would separate drug transport from growth.

History and geometry carried information about premature progression that size did not. The evidence for this is narrower than it first appears. Rules that respond to change between measurements already exist (Enriquez-Navas et al. 2016; Gallaher et al. 2018), and under intermittent androgen deprivation PSA kinetics in the first off-treatment interval are associated with outcome (Keizman et al. 2011; Sanchez-Salas et al. 2016). Here, in tumours of the same size and threshold, on a separate test set, history reduced log-loss by 48.6% and geometry by a further 30.5%, against an event rate of 1.126%. The figures assume exact size readings and exact lattice geometry. With 10% noise on the readings, the history model was only 2.0% better than a size model with exact readings, so most of its advantage depends on error-free measurement. The prediction result is strong at the level of the simulator, and its practical value with real measurements is untested.

The policy test asks a different question from the prediction study, and its answer was weaker. Under the hold rules and risk constraint tested here, features that improved prediction produced a small improvement in control at two of four intervals. The result does not show that history or geometry has little value in general. The study suggests three reasons for the gap without testing any of them. First, information has value only through the decisions it changes (Howard 1966), and the absolute gain in prediction was small: the size model was already 58% below the base-rate model in log-loss, and history lowered the pooled log-loss from 0.0258 to 0.0132 nats per decision. Second, the rules could act on the extra information only in a narrow band of sizes. A rule could hold above the size-only threshold only up to the cap, which was 3300 cells against a tuned size-only threshold of 3215.8 cells at 7 days, and above the cap every rule had to treat. Third, once a model chose the holds, the states it visited changed and its risk estimates no longer matched outcomes (observed-to-expected ratios from 0.28 to 2.46), a case of predictions that change the data they are scored on (Perdomo et al. 2020). Raising the cap, and retraining the models on holds chosen by the rules themselves, would test the second and third reasons.

The comparison under a common risk constraint also sits close to the resolution of the experiment. A shared target in tuning is not evidence that the achieved risks are equal. The size-only rule's tuned and scored p_{prem} differed by up to 0.046 against 0.028 at 14 days, while the non-inferiority margin was 0.02. The growth-rate rule met the tolerance in tuning at 7 days, with a tuned p_{prem} of 0.048, and then had a safety failure in scoring at 0.069. Every gain here is therefore reported with its scored risk, and the primary result changed with the margin and the horizon (Section 3.3.1). Controllers that plan over future states, such as model predictive control (Hirata et al. 2014) and dynamic programming (Gluzman et al. 2020), are a different class from the one-step hold rules tested here, and nothing in this study measures them.

4.1 Limitations

No result here is calibrated to patients, and the results concern one model. There is one ABM, one lattice of 208 by 208 sites, one starting geometry (a single disk of 3001 cells) and one drug, which kills sensitive cells only. Resistance is present from the start at 1%, with no mutation. Untested are other model architectures (off-lattice or three-dimensional models, other agent-based simulators), resistance that emerges during treatment, other treatment schedules and drugs, other tumour geometries and starting sizes, and other threshold constructions. The results do not show a general property of spatial tumour models.

The tolerance $q = 0.05$, the non-inferiority margin of 0.02 and the 750-day horizon are modelling choices without an established clinical basis, and the conclusions are conditional on them. The sign of the transfer term did not change for tolerances of 0.025 and 0.10 at intervals up to 56 days, but the policy reading changed with the margin and lost its gain in time over a 365-day horizon (Appendix D). The rules observe

the number of living cells without error, and a dead cell leaves its site at once, whereas tumour size on an image includes measurement error and can include dead and scar tissue. The image features are exact lattice geometry.

The growth-law constants of the original ODE were fitted, and their transfer to the two-drug model checked, only on a lattice of one hundred by one hundred sites starting from one hundred and twenty cells. On the configuration used here the ABM regrew 2.41 times more slowly than that ODE. The recalibration refitted the law to the final configuration and the shortfall remained, but which difference between the two configurations slows regrowth was not determined.

The intervals of 42 and 84 days were not remeasured on a finer grid. At 42 days the transfer term is negative whichever grid point is used, but at 84 days its sign is not resolved, and the split at 56 days is unresolved. The finer grid used for the correction is itself linearly interpolated. The prediction models are logistic regressions with feature sets fixed in advance, trained on hold decisions under threshold policies, and no other features or model classes were tried. The policy test compared one-step hold rules with one cap and one tuning procedure. The literature search was not systematic, and the analyses listed as added in Table 2 are exploratory.

4.2 Conclusions

Transfer. The threshold construction tested here did not transfer safely to this ABM at monitoring intervals of 7 to 28 days, even after the reported recalibration to the ABM's drug-free regrowth. At 42 to 84 days the derived threshold was safe in the ABM because the construction's own slack outweighed the shortfall. The mechanism of the shortfall remains unresolved.

Prediction. With exact readings, the history of earlier measurements and tumour geometry predicted premature progression in tumours of the same size. With 10% measurement error, most of the advantage of history over an exact size reading disappeared, and the practical value of these features is untested.

Decisions. Under a common nominal risk constraint, a hold rule using history and image geometry changed the time on drug by −1.4 percentage points and RMST by 4.0 days over 750 days. The result rested on two of four intervals, its risk bound sat on the non-inferiority margin at 7 days, and its gain in time was not shown over 365 days. On new runs its risk was not shown to meet the tolerance at any interval. The conclusion is conditional on the chosen cap, tuning procedure, non-inferiority margin and horizon. In this model, better prediction did not produce a robust improvement in control.

Statements and Declarations

Funding. The author received no financial support for this research.

Competing interests. The author declares no competing interests.

Ethics approval and consent to participate. Not applicable. The study used simulated data only.

Data and code availability. All code and the simulation results behind every number are available at https://github.com/arejune/Cancer-Evolution/.

AI declaration. AI tools were used as described in Section 2.6.

References

Bacevic K, Noble R, Soffar A, Wael Ammar O, Boszonyik B, Prieto S, Vincent C, Hochberg ME, Krasinska L, Fisher D (2017) Spatial competition constrains resistance to targeted cancer therapy. Nature Communications 8(1):1995. https://doi.org/10.1038/s41467-017-01516-1

Brady-Nicholls R, Nagy JD, Gerke TA, Zhang T, Wang AZ, Zhang J, Gatenby RA, Enderling H (2020) Prostate-specific antigen dynamics predict individual responses to intermittent androgen deprivation. Nature Communications 11(1):1750. https://doi.org/10.1038/s41467-020-15424-4

Brady-Nicholls R, Enderling H (2022) Range-Bounded Adaptive Therapy in Metastatic Prostate Cancer. Cancers 14(21). https://doi.org/10.3390/cancers14215319

Brier GW (1950) Verification of forecasts expressed in terms of probability. Monthly Weather Review 78(1):1-3. https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

Byrne H, Drasdo D (2009) Individual-based and continuum models of growing cell populations: a comparison. Journal of Mathematical Biology 58(4-5):657-687. https://doi.org/10.1007/s00285-008-0212-0

Drasdo D, Höhme S (2005) A single-cell-based model of tumor growth in vitro: monolayers and spheroids. Physical Biology 2(3):133-147. https://doi.org/10.1088/1478-3975/2/3/001

Enriquez-Navas PM, Kam Y, Das T, Hassan S, Silva A, Foroutan P, Ruiz E, Martinez G, Minton S, Gillies RJ, Gatenby RA (2016) Exploiting evolutionary principles to prolong tumor control in preclinical models of breast cancer. Science Translational Medicine 8(327):327ra24. https://doi.org/10.1126/scitranslmed.aad7842

Gallagher K, Strobl MAR, Anderson ARA, Maini PK (2025) Deriving Optimal Treatment Timing for Adaptive Therapy: Matching the Model to the Tumor Dynamics. Bulletin of Mathematical Biology 87(10):146. https://doi.org/10.1007/s11538-025-01525-y

Gallaher JA, Enriquez-Navas PM, Luddy KA, Gatenby RA, Anderson ARA (2018) Spatial Heterogeneity and Evolutionary Dynamics Modulate Time to Recurrence in Continuous and Adaptive Cancer Therapies. Cancer Research 78(8):2127-2139. https://doi.org/10.1158/0008-5472.CAN-17-2649

Gatenby RA, Silva AS, Gillies RJ, Frieden BR (2009) Adaptive therapy. Cancer Research 69(11):4894-4903. https://doi.org/10.1158/0008-5472.CAN-08-3658

Gluzman M, Scott JG, Vladimirsky A (2020) Optimizing adaptive cancer therapy: dynamic programming and evolutionary game theory. Proceedings of the Royal Society B: Biological Sciences 287(1925):20192454. https://doi.org/10.1098/rspb.2019.2454

H. Lee Moffitt Cancer Center and Research Institute (2015) A Pilot Study of Adaptive Abiraterone Therapy for Metastatic Castration Resistant Prostate Cancer. ClinicalTrials.gov identifier NCT02415621. https://clinicaltrials.gov/study/NCT02415621

Hirata Y, Azuma S, Aihara K (2014) Model predictive control for optimally scheduling intermittent androgen suppression of prostate cancer. Methods 67(3):278-281. https://doi.org/10.1016/j.ymeth.2014.03.018

Howard RA (1966) Information value theory. IEEE Transactions on Systems Science and Cybernetics 2(1):22-26. https://doi.org/10.1109/TSSC.1966.300074

Keizman D, Huang P, Antonarakis ES, Sinibaldi V, Carducci MA, Denmeade S, Kim JJ, Walczak J, Eisenberger MA (2011) The change of PSA doubling time and its association with disease progression in patients with biochemically relapsed prostate cancer treated with intermittent androgen deprivation. The Prostate 71(15):1608-1615. https://doi.org/10.1002/pros.21377

Mistry HB (2021) On the reporting and analysis of a cancer evolutionary adaptive dosing trial. Nature Communications 12(1):316. https://doi.org/10.1038/s41467-020-20174-4

Morris TP, White IR, Crowther MJ (2019) Using simulation studies to evaluate statistical methods. Statistics in Medicine 38(11):2074-2102. https://doi.org/10.1002/sim.8086

Perdomo J, Zrnic T, Mendler-Dünner C, Hardt M (2020) Performative prediction. In: Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research 119:7599-7609. https://proceedings.mlr.press/v119/perdomo20a.html

Sanchez-Salas R, Olivier F, Prapotnich D, Dancausa J, Fhima M, David S, Secin FP, Ingels A, Barret E, Galiano M, Rozet F, Cathelineau X (2016) First off-time treatment prostate-specific antigen kinetics predicts survival in intermittent androgen deprivation for prostate cancer. The Prostate 76(1):13-21. https://doi.org/10.1002/pros.23098

Strobl MAR, West J, Viossat Y, Damaghi M, Robertson-Tessi M, Brown JS, Gatenby RA, Maini PK, Anderson ARA (2021) Turnover Modulates the Need for a Cost of Resistance in Adaptive Therapy. Cancer Research 81(4):1135-1147. https://doi.org/10.1158/0008-5472.CAN-20-0806

Strobl MAR, Gallaher J, West J, Robertson-Tessi M, Maini PK, Anderson ARA (2022) Spatial structure impacts adaptive therapy by shaping intra-tumoral competition. Communications Medicine 2:46. https://doi.org/10.1038/s43856-022-00110-x

West J, Adler F, Gallaher J, Strobl M, Brady-Nicholls R, Brown J, Roberson-Tessi M, Kim E, Noble R, Viossat Y, Basanta D, Anderson ARA (2023) A survey of open questions in adaptive therapy: Bridging mathematics and clinical translation. eLife 12. https://doi.org/10.7554/eLife.84263

Wilson EB (1927) Probable Inference, the Law of Succession, and Statistical Inference. Journal of the American Statistical Association 22(158):209-212. https://doi.org/10.1080/01621459.1927.10502953

Zhang J, Cunningham JJ, Brown JS, Gatenby RA (2017) Integrating evolutionary dynamics into treatment of metastatic castrate-resistant prostate cancer. Nature Communications 8(1):1816. https://doi.org/10.1038/s41467-017-01968-5

Zhang J, Cunningham JJ, Brown JS, Gatenby RA (2021) Response to Mistry. Nature Communications 12(1):329. https://doi.org/10.1038/s41467-020-20175-3

Appendix A Reproducibility and random seeds

Every result can be regenerated from the stored simulation output without rerunning the simulations, and every simulation can be rerun from its seeds. The code repository contains the simulation code, the stored output with checksums, the dated analysis plans, the package versions, a description of every file, and one command line for every step. The command python run.py ledger recomputes every number in this paper, python run.py figures and python run.py tables redraw the figures and tables, python run.py review reruns the checks added after review (Appendix D), and tests/regression.py reruns part of every experiment and compares it with the stored output. The repository's README gives the order of the simulation steps, their seeds, their run times and the commands for a full rerun from scratch. Each figure's sources and the code version that drew it are recorded with the figures.

Each experiment used its own range of seeds (Table A1). Replicate s of an experiment uses the first seed of its range plus s, so the same replicate starts from the same tumour at every threshold. The main experiment had 32500 runs, the recalibration sweep 25500, the prediction study 20000, and the policy test 80000 tuning runs and 20000 scoring runs. The prediction study took 86.7 minutes of wall time and the policy test 45.6 and 21.4 minutes for tuning and scoring, on a machine with variable load.

Table A1. Random-number seeds used by each experiment. Replicate s uses the first seed plus s. A dash marks an experiment whose run count is not reported in this paper.

experiment	first seed	runs
main experiment, ABM and the ODE and CTMC arms	1200000	32500
recalibration: drug-free transit times and threshold sweep	1800000	25500
CTMC under the recalibrated law	1900000	--
CTMC under the recalibrated law, finer grids	2000000	--
prediction study, check of the logging code	2300000	--
prediction study, training set and test set	2100000, 2200000	20000
policy test, check of the rule code	2600000	--
policy test, tuning	2400000	80000
policy test, scoring	2500000	20000
reserved and never used	1700000	--

Every number in this paper is produced by a script that recomputes it from the stored output or checks it against the dated record it came from; no number was typed by hand. The script checks 1015 entries: 905 recomputed from the raw output, 67 design constants (48 read from the code that ran and 19 copied from an analysis plan), and 43 checked word for word against a dated record. It reported 0 failures. Table A2 maps each headline number to the code that computes it, its input, its denominator and its uncertainty.

Table A2. Results audit for the headline numbers. Code paths are those of the code repository; every number is an entry of the ledger written by python run.py ledger. Percentile CIs come from resampling replicate seeds.

number	value	computed by	input	denominator	uncertainty
risk at N^*, τ = 7 d	0.976	analysis/review.py	p_t_main_raw.npz	500 runs	Wilson interval
corrected transfer term, τ = 14 d	−8.95 %	report/ledger.py	p_s_cells, ODE safe thresholds	ODE safe threshold	seed resampling (analysis/review.py)
CTMC-to-ABM share, τ = 14 d	94.1 %	report/ledger.py	p_s_cells, p_t_recal_ctmc_fine, p_t_recal_odeb	ODE minus ABM safe threshold	seed resampling
history log-loss reduction, pooled	+48.6 %	analysis/statesuff.py	p_s_cells (training and test)	size-model log-loss, test set	seed resampling, 4,000
geometry log-loss reduction, pooled	+30.5 %	analysis/statesuff.py	p_s_cells (training and test)	state-model log-loss, test set	seed resampling, 4,000
change in on-drug fraction, pooled	−1.4 points	experiments/policy.py	p_s_policy/score	1,000 paired replicates per interval	seed resampling, 4,000
change in RMST, pooled	+4.0 d	experiments/policy.py	p_s_policy/score	1,000 paired replicates per interval	seed resampling, 4,000
observed-to-expected ratio, τ = 28 d	2.46	experiments/policy.py	p_s_policy/score	expected events on the rule's holds	none

I fitted the growth-law constants of (3) on a three-drug version of the lattice model in an earlier stage of the project, on a smaller lattice and from a small starting colony. Before the two-drug model used them, I ran it without drug and measured $\phi(N)$ directly, as the fraction of cells with an empty neighbour. The measured values lay as close to the three-drug curve as to the two-drug model's own best fit, and the two fits gave nearly the same constants. That check predates the code repository and is not reproduced there. The literature search of Section 1 queried PubMed, PubMed Central and Crossref directly, and every reference cited here was checked field by field against its primary record.

Appendix B Closed form of the threshold

Without the capacity factor, (2) becomes $dN/dt = A\sqrt{N} + BN$ with $A = r_S c$ and $B = r_S \phi_{\text{bulk}} - d$. Substituting $u = \sqrt{N}$ makes the equation linear in u, and setting the regrowth time from N^* to N_{prog} equal to τ gives (11). It is used only as an analytic reference. The construction is that of Gallagher et al. (2025); only this closed form for the unbounded growth law is new here.

$$N^*(\tau) = \left[\frac{(A + B\sqrt{N_{\text{prog}}})e^{-B\tau/2} - A}{B}\right]^2 \tag{11}$$

Dividing A and B by a common factor slows regrowth by that factor and has the same effect as shortening τ by it. Because N^* falls as τ grows, a model that regrows more slowly has a higher N^* at every interval. The same holds with the capacity factor, since slower growth lengthens the regrowth time $T(n)$ from every size n and so moves the root of (4) upward.

Appendix C Grid resolution and the correction

In the main experiment the multiplier m ran from 0.80 in steps of 0.05, with extra upper values at long intervals so that every grid spanned the crossing. Grid points with $\theta \geq N_{\text{prog}}$ were excluded because the drug can never start there. The experiment had 32500 runs at 65 grid points, one of them a continuous-therapy control. I estimated N_{safe} by linear interpolation in p_{prem} between the two grid points on either side of q, after replacing p_{prem} by its running maximum over increasing θ.

Linear interpolation across a coarse grid is biased when p_{prem} rises from zero along a convex curve. The main-grid step was 93.6 to 171.3 cells. The prediction study remeasured the same policy at 5 thresholds per interval, with steps of 50 to 80 cells and 500 replicates per threshold in each of two sets, and the bisection that tuned the size-only rule narrowed it further (Figure C1, Table C1, Table C2).

The bias does not always have the same sign, since the value at 56 days moved down. The final interval of the tuning bisection contains the finer-grid value at 14 and 56 days and lies above it at 7 and 28 days. The finer grid is itself linearly interpolated, and the bisection interval, from 500 replicates in a different seed range, is the finest resolution available.

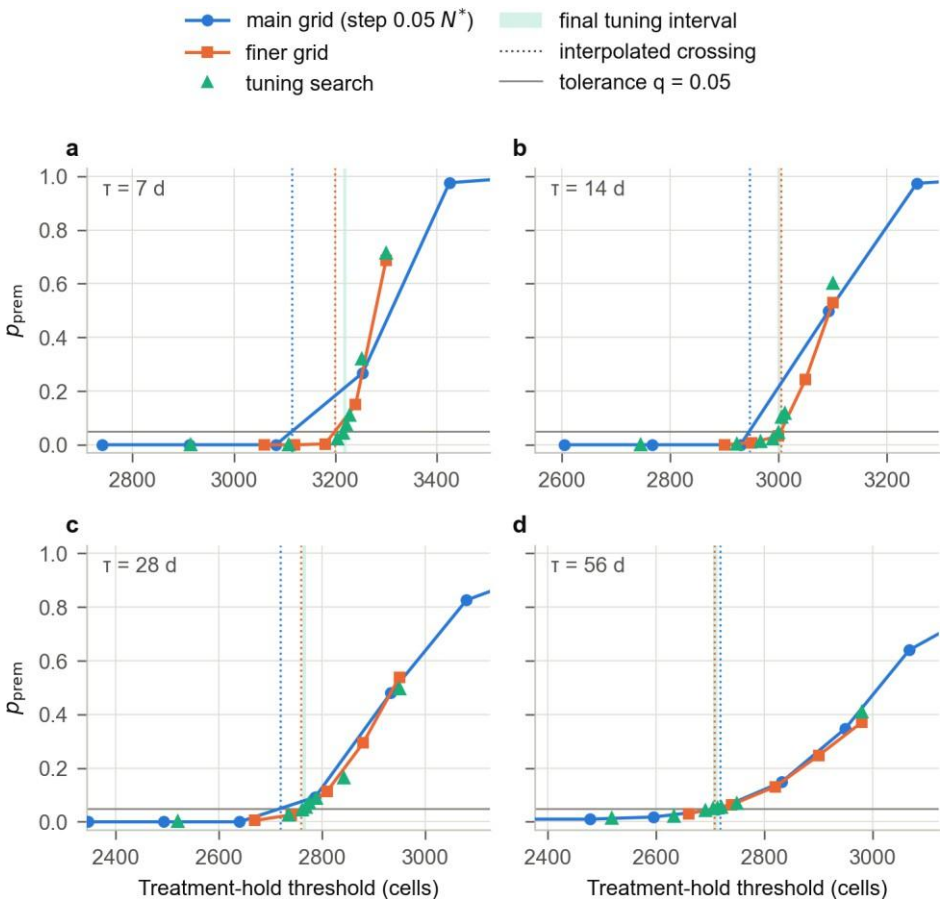

Figure C1. Resolution of the ABM safe threshold. Panels (a)–(d) show τ = 7, 14, 28 and 56 days. p_{prem} against the treatment-hold threshold on the main grid (multiplier step 0.05), on the finer grid (5 thresholds per interval, two sets of 500 replicates) and at the steps of the bisection that tuned the size-only rule (500 replicates per step). Dotted lines, interpolated crossings of q = 0.05, coloured by grid; shaded band, final tuning interval.

Table C1. Correction of the ABM safe threshold for grid resolution. The finer-grid value is interpolated on the grid of the prediction study. The final tuning interval is the last bisection interval of the size-only rule. The intervals of 42 and 84 days were not remeasured.

τ (d)	main-grid N_{safe} (cells)	finer-grid N_{safe} (cells)	finer-grid step (cells)	final tuning interval (cells)	shift (cells)
7	3114.6	3199.3	60	[3215.8, 3221.8]	+84.7
14	2946.1	3004.2	50	[3000.4, 3006.0]	+58.2
28	2719.9	2759.1	70	[2761.9, 2768.6]	+39.2
56	2718.3	2707.1	80	[2705.7, 2712.9]	−11.2

Table C2. Quantities in the main text before (main grid) and after (corrected) the correction of Table C1.

τ (d)	transfer, main grid (%)	transfer, corrected (%)	CTMC-to-ABM share, main grid (%)	CTMC-to-ABM share, corrected (%)	vs N_{q05}, main grid (%)	vs N_{q05}, corrected (%)
7	−9.83	−7.37	89.4	88.1	−11.18	−9.97
14	−10.71	−8.95	94.2	94.1	−14.44	−14.15
28	−9.37	−8.06	85.0	84.2	−18.29	−17.23
56	−4.00	−4.40	45.9	45.9	−9.77	−9.76

Appendix D Sensitivity and robustness checks

The checks in this appendix were added after all results had been read, in response to review, and are exploratory. None runs a new simulation: each recomputes from the stored output (python run.py review in the code repository). Table D1 summarises what each changed and whether the conclusion it bears on still follows.

Tolerance. The ABM safe threshold was recomputed at q = 0.025 and 0.10. The ODE safe threshold does not depend on q, because the ODE either progresses prematurely or does not. On the finer grid the transfer term was −7.67%, −9.52% and −8.81% at the lower tolerance and −6.78%, −8.59% and −6.74% at the higher one, at 7, 14 and 28 days. On the main grid it stayed negative from 7 to 56 days at both tolerances, and at 84 days it was −4.29% and +4.00%.

Grid and interpolation. Without interpolation, the transfer term at the main-grid points either side of the crossing of q was −10.75% and −5.79% at 7 days, −11.20% and −6.27% at 14 days, −12.02% and −7.14% at 28 days, −5.47% and −0.74% at 42 days and −5.12% and +2.17% at 84 days. Seed resampling (2,000 resamples) gave the corrected transfer term CIs of −7.46 to −7.28, −9.02 to −8.87, −8.34 to −7.76 and −5.28 to −3.23 percent.

Growth law. The main experiment's ODE uses the constants fitted on the small configuration, and the recalibration uses constants refitted to the final one. The ABM's corrected safe threshold lay below the ODE's safe threshold under both laws (Table D1), so the direction of the shortfall does not depend on which law the ODE uses. What causes the slower regrowth of the final configuration was not investigated, because that would need new simulations.

Policy test. The readings of the history-and-image rule were recomputed with non-inferiority margins of 0.01 and 0.03, and the on-drug fraction and RMST were recomputed over the first 365 days of each run from the stored decision logs. The cap N_{top} and the tuning procedure could not be varied without new simulations and were not.

Measurement noise. Size readings and the rim count, radius of gyration and maximum radius were multiplied by independent lognormal noise with mean one. The history features were then recomputed from the noisy readings, and the models were refitted on the training set and scored on the test set (Section 3.2.4). The number of connected components and the resistant count were left exact.

Table D1. Sensitivity and robustness checks, all added after the results had been read. Transfer terms are in percent; a negative value means the ABM safe threshold lies below the ODE safe threshold.

check	what changed	result	conclusion still follows
tolerance q = 0.025 and 0.10	ABM safe threshold	corrected transfer term −8.81 and −6.74 at τ = 28 d; main grid negative from τ = 7 to 56 d; at τ = 84 d −4.29 and +4.00	yes, up to τ = 56 d
no interpolation (main grid)	grid points either side of the crossing	transfer term at τ = 42 d between −5.47 and −0.74; at τ = 84 d between −5.12 and +2.17	yes at τ = 42 d; sign unresolved at τ = 84 d
seed resampling	corrected transfer term and CTMC-to-ABM share	transfer CIs below zero at all four intervals; share CIs 83.6 to 95.0 %	yes
growth law of the ODE	constants refitted to the final configuration	corrected transfer term −10.02 to −23.03 %	yes
non-inferiority margin 0.01	policy reading	no interval assessable	no pooled result
non-inferiority margin 0.03	policy reading	τ = 14 d becomes a gain in exposure	yes, on three intervals
horizon 365 d	on-drug fraction and RMST	pooled ΔRMST +0.95 d [−0.01, +1.93]	gain in exposure only
size and image noise, CV 10 %	prediction models refitted	history over exact size +2.0 %; image geometry over history +17.1 %	history gain largely lost
policy cap, tuning procedure	not tested (needs new simulations)		unknown
cause of the slower regrowth	not tested (needs new simulations)		unknown

Appendix E Supplementary tables

Table E1. Recalibration control. $N^*_{\mathrm{recal}}$and N_{q05} invert the median and fifth-percentile drug-free transit times of the ABM. N_{safe} is the ABM safe threshold on the grid around N^*_{recal}, before the grid correction.

τ (d)	N^*_{recal} (cells)	N_{safe} (cells)	vs median (%)	N_{q05} (cells)	vs N_{q05} (%)
7	3581.8	3156.5	−11.88	3553.7	−11.18
14	3518.2	2994.1	−14.90	3499.4	−14.44
28	3366.7	2723.5	−19.10	3333.3	−18.29
42	3211.1	2645.8	−17.60	3166.5	−16.44
56	3044.4	2706.9	−11.09	3000.0	−9.77
84	2744.4	2581.5	−5.94	2677.8	−3.60

Table E2. Safe thresholds under the recalibrated law and the CTMC-to-ABM share (6), on the main grid. The CTMC grid step was 0.0025 in m. The range is that of the ODE-to-CTMC share when the fine CTMC grids are resampled. Corrected shares are in Table C2.

τ (d)	ODE safe threshold (cells)	CTMC safe threshold (cells)	ABM N_{safe} (cells)	CTMC-to-ABM share (%)	ODE-to-CTMC share, range (%)	ODE threshold set by
7	3555.5	3513.2	3156.5	89.4	[10.9,11.4]	regrowth
14	3481.3	3453.2	2994.1	94.2	[6.1,6.9]	regrowth
28	3426.8	3321.6	2723.5	85.0	[12.9,15.8]	regrowth
42	3395.9	3233.9	2645.8	78.4	--	decision calendar
56	3517.2	3078.7	2706.9	45.9	[47.1,58.3]	regrowth
84	3001.0	2998.6	2581.5	99.4	--	decision calendar

Table E3. Measurement history: the history model against the size model on the test set. Δ is the relative reduction in test log-loss (9). ΔAUC within size bins is the event-weighted mean over usable bins of 20 cells, with the number of bins in parentheses. The last column repeats Δ with drug concentration C in both models.

τ (d)	hold decisions, test set	events, test / training	Δ log-loss (%) [CI]	ΔAUC within size bins [CI] (bins)	Δ log-loss, C in both models (%)
7	89069	431 / 408	+46.2 [+43.4, +48.9]	+0.388 [+0.358, +0.418] (6)	+45.5
14	39359	403 / 410	+61.5 [+58.9, +63.9]	+0.369 [+0.341, +0.398] (10)	+57.1
28	18048	479 / 499	+50.7 [+46.6, +54.7]	+0.377 [+0.346, +0.409] (16)	+49.4
56	8171	429 / 414	+10.4 [+3.3, +16.8]	+0.074 [−0.012, +0.144] (7)	+11.1
pooled			+48.6 [+46.7, +50.3]		+46.4

Table E4. Checks on the history gain. History over size with the resistant count R in both models; R alone over size; and the share of the history model's log-loss gain reached by a model of size and threshold plus growth since the previous decision. All were frozen before any policy-test run.

τ (d)	history over size, R in both (%)	R over size (%)	growth-rate model, share of history gain (%)
7	+51.8	+0.5	41.7
14	+63.4	+0.5	52.8
28	+47.9	+6.9	8.4
56	+11.0	−1.1	−19.1
pooled	+50.3 [+48.3, +52.1]	+2.1 [+1.4, +2.8]	35.1 [33.2, 37.0]

Table E5. Tumour geometry: the geometry model against the state model on the test set, and the image-geometry comparisons (image geometry added to the state model; to the history model plus C, without R; and to the history model alone).

τ (d)	geometry, Δ log-loss (%) [CI]	geometry, ΔAUC within size bins [CI] (bins)	image geometry, R in both (%)	image geometry, no R (%)	image geometry over history (%)
7	+11.5 [+8.9, +14.1]	+0.012 [+0.009, +0.016] (6)	+12.3	+21.1	+22.0
14	+35.4 [+29.7, +40.8]	+0.012 [+0.009, +0.017] (10)	+34.3	+38.6	+38.8
28	+41.8 [+35.9, +47.6]	+0.035 [+0.024, +0.047] (16)	+40.1	+40.3	+38.6
56	+41.2 [+28.8, +51.9]	+0.324 [+0.084, +0.509] (7)	+41.1	+39.9	+39.6
pooled	+30.5 [+27.3, +33.5]		+30.0	+33.3	+33.1

Table E6. Absolute scale of the prediction results on the test set. Log-loss is in nats per hold decision. The base-rate model predicts the overall event rate.

τ (d)	event base rate (%)	base-rate model (nats)	size model (nats)	history model (nats)	state model (nats)	geometry model (nats)
7	0.484	0.0306	0.0146	0.0078	0.00703	0.00622
14	1.024	0.0571	0.0306	0.0118	0.01103	0.00712
28	2.654	0.1225	0.0611	0.0301	0.02960	0.01722
56	5.250	0.2058	0.0470	0.0421	0.04231	0.02488
pooled	1.126	0.0617	0.0258	0.0132	0.01254	0.00872

Table E7. Settings chosen in the tuning runs. In parentheses, the tuned p_{prem} (from the 500 tuning replicates at the chosen setting) and the scored p_{prem} (from the 1,000 scoring replicates). The size-only rule tunes a threshold in cells; each other rule tunes $\log_{10} p^*$.

rule	τ = 7 d	τ = 14 d	τ = 28 d	τ = 56 d
size-only	3215.8 (tuned 0.042, scored 0.034)	3000.4 (tuned 0.046, scored 0.028)	2761.9 (tuned 0.044, scored 0.043)	2705.7 (tuned 0.050, scored 0.036)
history and image	−0.703 (tuned 0.046, scored 0.038)	−0.547 (tuned 0.046, scored 0.038)	−0.859 (tuned 0.046, scored 0.045)	−0.234 (tuned 0.038, scored 0.038)
history	−1.484 (tuned 0.048, scored 0.049)	−0.781 (tuned 0.044, scored 0.042)	−1.172 (tuned 0.038, scored 0.040)	−0.312 (tuned 0.038, scored 0.027)
full state	−0.391 (tuned 0.042, scored 0.039)	−0.469 (tuned 0.050, scored 0.038)	−0.703 (tuned 0.046, scored 0.046)	−0.234 (tuned 0.040, scored 0.036)
growth rate	−1.562 (tuned 0.048, scored 0.069)	−1.328 (tuned 0.036, scored 0.047)	−1.328 (tuned 0.026, scored 0.023)	−0.312 (tuned 0.048, scored 0.040)

Table E8. Pooled paired differences of every rule against the size-only rule. The last column states whether the rule for a pooled result (the same reading and sign at two intervals) was met.

rule	relative Δ on-drug fraction (%) [CI]	ΔRMST (d) [CI]	scored Δp_{prem} [CI]	reading	rule for a pooled result
history and image	−3.7 [−4.2, −3.3]	+4.0 [+2.5, +5.5]	+0.004 [−0.003, +0.012]	gain (exposure, time)	met
history	−2.4 [−2.7, −2.0]	+2.6 [+1.0, +4.1]	+0.004 [−0.003, +0.011]	gain (exposure, time)	not met
full state	−3.9 [−4.3, −3.5]	+3.8 [+2.3, +5.4]	+0.004 [−0.003, +0.012]	gain (exposure, time)	met
growth rate	−2.1 [−2.5, −1.8]	−0.5 [−1.9, +0.8]	+0.009 [+0.003, +0.016]	gain (exposure)	not met

Table E9. Secondary rules against the size-only rule in the scoring runs, by interval. Readings follow Section 2.5. Not assessable means that non-inferiority in risk was not shown; a safety failure means that the Wilson lower bound of p_{prem} exceeded q.

rule	τ (d)	scored p_{prem}	relative Δ on-drug fraction (%)	ΔRMST (d)	scored Δp_{prem} [CI]	reading
history	7	0.049	−3.8	+3.0	+0.015 [−0.002, +0.033]	not assessable
history	14	0.042	−4.1	+1.5	+0.014 [+0.002, +0.027]	not assessable
history	28	0.040	−3.0	+5.7	−0.003 [−0.018, +0.013]	gain (exposure, time)
history	56	0.027	+0.7	+0.1	−0.009 [−0.015, −0.004]	no gain (worse)
full state	7	0.039	−5.2	+6.3	+0.005 [−0.011, +0.020]	gain (exposure, time)
full state	14	0.038	−4.2	+1.9	+0.010 [−0.004, +0.025]	not assessable
full state	28	0.046	−6.0	+7.0	+0.003 [−0.014, +0.019]	gain (exposure, time)
full state	56	0.036	−0.7	+0.1	+0.000 [−0.013, +0.013]	no gain
growth rate	7	0.069	−5.2	−3.1	+0.035 [+0.017, +0.054]	safety failure
growth rate	14	0.047	−3.9	−0.2	+0.019 [+0.005, +0.033]	not assessable
growth rate	28	0.023	−0.6	+1.1	−0.020 [−0.034, −0.008]	no gain
growth rate	56	0.040	+0.3	+0.0	+0.004 [−0.005, +0.013]	no gain

Appendix F Validation and errors found during the project

Each program that logged extra state was checked against the code that produced the original results before its runs were used. On its own seed range, the program of the prediction study reproduced the original runs exactly, left the simulator's random-number stream as an unmodified copy left it, and matched its decision outcomes to the run outcomes. On its own seed range, the policy-test program's size-only rule reproduced the original policy, replayed decisions identically, and reproduced every model feature on 2935 decisions.

The errors and withdrawn claims below were found during the project. Each is given with how it was found and the safeguard adopted afterwards.

Six were errors in code. In an earlier two-drug model every resistant escape cell came from a founder cell, so the mutation chain had no effect. Counters in a copy of the simulator, checked to give identical runs, showed this, and since then intermediate state is logged and every logging copy is checked against the original. Founder cells had also been placed on free lattice away from the tumour, and one colony was clipped by the lattice edge. Replaying the starting conditions and grouping runs by distance to the edge found both, and geometry is now checked before outcomes are computed. A wrapper meant to override parameters applied none of them, and its output looked like a confirmation. A check with four arms before launch caught it, and every wrapper is now tested with several arms. A filter in the specification was never implemented, so the raw value was used. Reading the code against the specification found it, and the two are now compared line by line. A script printed a wrong list of certified results and now asserts the rule of the analysis plan that it had broken. A process-start error, from constants set inside the main guard, stopped one launch before any seed was used, and every program now sets its constants at module level.

Eight were errors in analysis or claims. The premise of a dosing controller based on tumour composition had no effect once its observer was corrected: none of 400 stored runs changed when rescored, and stored runs are now rescored before anything is rerun. Two arithmetic errors by Claude Code, on which decisions had been argued, were found by the author's review and a staged rescore. Decompositions are now computed in both orders. A claim that the threshold erred in both directions, with a sign that changed with τ, went to four outside reviews and was withdrawn after a bisection of the stored ODE results. Every discrepancy is now compared with a reference point, the derivation and transfer terms of (5). A prediction of which grid points would be degenerate used the ODE transit time and was wrong in both directions, as the time-to-progression pattern of the runs showed; degenerate points are now found empirically. The first CTMC-to-ABM shares used a CTMC grid whose step was 91–125% of the quantity it measured and were withdrawn, and every bracket is now refined. The main-grid safe threshold was biased by interpolation across a coarse grid, by +84.7 cells at 7 days and not always in the same direction. The finer grids showed this, and Section 3.1.2 corrects it. Wording about sharper probabilities at short intervals was nearly applied to the history comparison, where the within-bin ΔAUC shows it does not hold, so each wording rule is now limited to its own comparison. A recorded statement that the transfer term shrinks steadily with τ is contradicted by its own values. It came to light when every number was regenerated from the stored output, and numbers are now produced by script and never retyped.

Four concerned planning. Two recalibration forecasts by Claude Code were recorded before the runs and failed, and the failure was kept on record. Per-decision data were assumed to exist for the main experiment and did not, so stored output is now listed before any analysis, and no used seed range is rerun for new data. A cost estimate of 179.9 ms per run overstated the cost on an idle machine, where a pilot measured 17.3 ms per run, so costs are now estimated from a pilot under the same conditions. The first proposed safety rule would have fired about half the time from sampling noise alone. Design review caught it before any runs, and the Wilson lower-bound rule with a non-inferiority margin replaced it.

Two concerned references. A generated draft contained an invented PubMed identifier, and a correct identifier carried a wrong first author; every reference is now checked field by field against its primary record. A web tool that summarises pages returned an invented reference, which was discarded, and no summarising tool is used for citations.