LEARNING THE MAXIMUM TOLERATED DOSE FOR CONTINUOUS TOXICITY VIA MONOTONE BAYESIAN TREES

Anonymous authors
Paper under double-blind review

ABSTRACT

Phase I cancer trials seek the maximum tolerated dose (MTD) while protecting patients from excessive toxicity. Dose assignments must therefore balance patient safety with learning the dose–toxicity relationship as data accrue. We model continuously measured toxicity outcomes using two forms of Bayesian additive regression trees (BART): isotonic BART projects posterior response curves onto nondecreasing functions, whereas monotone BART constrains the model. Joint curve and variance draws induce an MTD posterior that guides dose selection through escalation with overdose control (EWOC). We compare these methods with three parametric procedures across seven dose–toxicity curves in simulation. We assess dose-limiting toxicity (DLT) counts, above-MTD assignments, signed last-dose error, and relative absolute error. The tree procedures attained the lowest mean RAE on four nonlinear curves and jointly minimized mean DLT counts and above-MTD assignments on four curves. A Bayesian reinforcement learning perspective formulates these sequential decisions as a finite-horizon planning problem. Dose restrictions yield a lower bound on last-dose error; under exact posterior-predictive evaluation, an EWOC-based rollout policy has no greater expected weighted loss than its baseline. **Dose Trial Lab**, a desktop simulator for all five procedures, accompanies the supplementary material.

1 INTRODUCTION

Phase I cancer trials seek to identify the maximum tolerated dose (MTD), the highest dose associated with an acceptable toxicity risk (Tighiouart & Rogatko, 2010). Traditional phase I cancer dose-escalation trials generally enroll roughly 30–50 patients (Dahlberg et al., 2014). Responses from earlier patients or cohorts guide later dose assignments. Because participants often have advanced cancer and limited treatment options, dose escalation must pursue precise MTD estimation while carefully limiting exposure to excessive toxicity (Tighiouart & Rogatko, 2010; Bartroff & Lai, 2011).

In this paper, we model continuously measured toxicity outcomes, such as a composite toxicity score derived from adverse events graded using the Common Terminology Criteria for Adverse Events (CTCAE)[1] or a surrogate biomarker (Lee et al., 2022; Lee, 2023). In our framework, a response meeting or exceeding a prespecified threshold defines a dose-limiting toxicity (DLT) event. The probability of this event determines the MTD.

We use escalation with overdose control (EWOC) as our dose-allocation rule. EWOC selects a conservative posterior quantile of the MTD, with a feasibility bound α limiting the posterior probability that the assigned dose exceeds the MTD (Babb et al., 1998; Zacks et al., 1998). For cytotoxic agents, toxicity risk is presumed nondecreasing with dose, although the shape of the dose–toxicity relationship is unknown (Tighiouart & Rogatko, 2010; Pantoja et al., 2022; Lee, 2023). Given the small sample sizes of phase I cancer trials, the dose-finding literature has typically relied on parametric models. Our comparators are the one-parameter linear dose finder (1PLD), which estimates a slope; the two-parameter linear dose finder (2PLD), which also estimates residual variance; and the three-parameter nonlinear dose finder (3PND), which estimates scale, variance,

[1] https://dctd.cancer.gov/research/ctep-trials/for-sites/adverse-events

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

and a curvature exponent (Eichhorn & Zacks, 1973; Lee et al., 2022; Lee, 2023). We also study two constructions of Bayesian additive regression trees (BART): isotonic BART (**iBART**) projects posterior draws onto nondecreasing functions, whereas monotone BART (**mBART**) constrains the model itself (Chipman et al., 2010; 2022). Figure 6 illustrates the mean dose–toxicity curves for all five procedures.

We also study a Bayesian reinforcement learning (RL) formulation that connects these response models to sequential dose decisions under the trial protocol. The state comprises the joint posterior, last administered dose, enrollment, and cumulative DLT count. Under the upward cap and Only Escalation restriction, an excessive dose increase cannot be reversed, while holding the dose can leave the target unreachable within the remaining patient budget. We derive a lower bound on the error of the terminal last-dose estimate and develop a finite-candidate Bellman formulation with an EWOC-based rollout policy. Under exact posterior-predictive evaluation and the stated conditions, rollout has no greater expected weighted loss than the EWOC baseline. Appendix H gives the proofs and planning algorithm; Appendix I reports the numerical 1PLD rollout comparison. This formulation accounts for dose controls, stopping after complete cohorts, and the use of the last administered dose as the terminal estimate.

In simulation experiments, we compare all five procedures under the same EWOC allocation, dose restrictions, and cumulative-DLT stopping, with 45 planned patients and seven dose–toxicity curves. The last dose estimates the MTD; the four criteria are DLT counts, above-MTD assignments, signed last-dose error, and relative absolute error. The accompanying desktop software supports EWOC and rollout with all five models, a user-selected rollout budget, paired policy comparisons, trial playback, posterior plots, and exports. Its bundled runtime permits offline use; Appendix L provides an illustrated guide.

Novelty. Table 2 and Appendix A establish our methodological distinction. To our knowledge, ours is the first phase I oncology framework combining continuously measured toxicity, monotone Bayesian trees, EWOC, and EWOC rollout within a Bayesian reinforcement learning formulation, with MTD-posterior consistency, a protocol-specific reachability bound, and exact expected-loss improvement of rollout over EWOC under the stated conditions. Alongside the methodological contribution, we developed a desktop simulator for users.

2 Bayesian dose finding

2.1 Problem setup

Let N denote the planned enrollment, $n \leq N$ the number of patients observed so far, and $\mathcal{F}_n$ the information contained in the observed dose–response pairs $\{(X_i, Y_i)\}_{i=1}^n$. The MTD posterior guides the next dose after each patient or completed cohort. Clinicians prespecify a fixed dose interval $\mathcal{X} = (x_{\min}, x_{\max}) \subset (0, \infty)$ in original dose units, with width $L_x = x_{\max} - x_{\min}$. Preclinical studies typically inform these limits. Larger continuous responses $Y(x)$ indicate more severe toxicity; a DLT event occurs when $Y(x) \geq \eta$ for prespecified $\eta > 0$, with cumulative count $D_n = \sum_{i=1}^n \mathbf{1}\{Y_i \geq \eta\}$. Sponsors and investigators seek the MTD while limiting excessive DLT risk (Bartroff & Lai, 2011).

Definition 1 (Maximum tolerated dose; Lee, 2023)**.** For prespecified $\eta > 0$ and $\gamma \in (1/2, 1)$, a dose $x \in \mathcal{X}$ is acceptable if

$$\Pr\{Y < \eta \mid X = x\} \geq \gamma.$$

Equivalently, its DLT probability is at most $\theta = 1 - \gamma$. The MTD ξ is the supremum of acceptable doses, with $\xi = x_{\min}$ if none are acceptable and $\xi = x_{\max}$ if all are acceptable.

We model the continuously measured outcome with a joint prior on the mean dose–toxicity curve f and residual standard deviation (SD) σ:

$$Y_i = f(X_i) + \epsilon_i, \quad \epsilon_i \mid \mathcal{F}_{i-1}, X_i \sim \mathcal{N}(0, \sigma^2),$$
$$d\Pi_n(f, \sigma) \propto \left[\prod_{i=1}^n \frac{\phi\{(Y_i - f(X_i))/\sigma\}}{\sigma}\right] d\Pi(f, \sigma). \tag{1}$$

Here $X_i \in \mathcal{X}$ is the dose administered to patient i. Subsection B.2 derives the posterior under adaptive allocation.

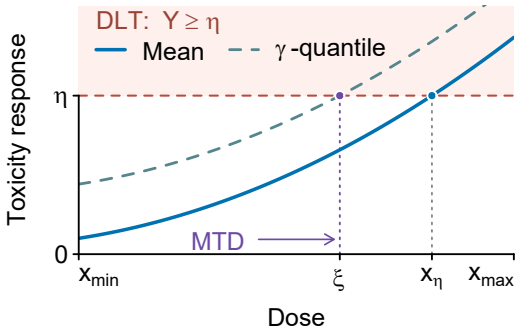


Figure 1: A continuously measured toxicity outcome and the MTD. For an increasing Gaussian mean, the dashed response quantile $f + \sigma\Phi^{-1}(\gamma)$ reaches η at ξ; the mean reaches it at x_η. DLT probabilities are $1-\gamma$ and $1/2$, respectively. The dashed curve describes outcome variability, not posterior uncertainty. Increasing σ lowers the MTD when both crossings are interior. Joint curve–variance draws propagate uncertainty into dose selection.

2.2 Monotone Bayesian trees: iBART and mBART

iBART. Fit the unconstrained BART mean $g = b_0 + \sum_{j=1}^{m} h_j$, replacing f by g in (1). On the dose cells $B_1, \ldots, B_J$, project each draw onto the nondecreasing cone:

$$(\mathcal{P}g)_j = \left[\underset{z_1 \le \cdots \le z_J}{\arg\min} \sum_{k=1}^{J} |B_k| (g_k - z_k)^2 \right]_j .$$

The pool-adjacent-violators algorithm (PAVA) solves this weighted least-squares problem (de Leeuw et al., 2009). Cell-length weights give the Lebesgue L_2 projection on $[x_{\min}, x_{\max}]$. Pair each projected draw with its sampled σ for MTD inference.

mBART. Constrain each tree's leaf values to be nondecreasing from left to right (Chipman et al., 2022). The sum f is then nondecreasing, so no projection is needed. The constrained-product prior imposes this monotonicity in the fitted model. Appendix C distinguishes that continuous prior from its numerical leaf-grid approximation.

Both methods target a nondecreasing true mean f_0, but the dose-finding estimand is the MTD rather than the curve itself. Thus f (together with σ) is an intermediate, or nuisance, component for MTD inference, unlike in conventional curve fitting. Monotonicity concerns the mean response and toxicity risk, not individual outcomes. For MTD estimation, iBART uses $f = \mathcal{P}g$, while mBART uses f directly. Define the toxicity-risk functional $p_{f,\sigma}$ and MTD functional $\xi(f, \sigma)$ by

$$p_{f,\sigma}(x) = \Phi\{(f(x) - \eta)/\sigma\}, \qquad \xi(f, \sigma) = \sup\big(\{x_{\min}\} \cup \{x \in \mathcal{X} : p_{f,\sigma}(x) \le \theta\}\big).$$

For mBART, $p_{f,\sigma}$ is conditional toxicity risk; for iBART it is a projected tail functional, with the unprojected curve retained in the likelihood. Definition 1 permits jumps and plateaus without requiring $p_{f,\sigma}(\xi) = \theta$; assigned doses remain in $\mathcal{X}$. Joint draws of f and σ induce the MTD posterior.

Priors and computation. We use the BART and mBART tree-depth defaults, 200 trees, and shrinkage $k = 2$ (Chipman et al., 2010; 2022). A common prior on residual variance completes each model. Before enrollment, we fix all priors and 100 equally spaced candidate tree-split points on $(x_{\min}, x_{\max})$, the dose thresholds available for splitting tree nodes. Doses and split points retain their physical units. Appendix C gives the response calibration, variance prior, cut locations, and numerical approximation. Appendix E compares the graphical models of all five procedures: 1PLD, 2PLD, 3PND, iBART, and mBART.

2.3 EWOC allocation and stopping

After observing $\mathcal{F}_n$, let $\Pi_n^{\xi}(x) = \Pr(\xi \le x \mid \mathcal{F}_n)$ denote the posterior cumulative distribution function (CDF) of the MTD. EWOC (Babb et al., 1998; Zacks et al., 1998; Tighiouart & Rogatko, 2010; Lee et al., 2022; Lee, 2023) proposes the next dose at its lower α-quantile:

$$D_\alpha(\mathcal{F}_n) = (\Pi_n^{\xi})^{-1}(\alpha).$$

For a continuous MTD posterior, $\Pr\{\xi \le D_\alpha(\mathcal{F}_n) \mid \mathcal{F}_n\} = \alpha$. A smaller α therefore selects a more conservative dose. The clinical target $\theta = 1 - \gamma$ specifies acceptable toxicity risk, whereas α controls allocation under posterior uncertainty.

Tree-based MTD posteriors can place mass at cut points or endpoints. We inward-adjust the EWOC quantile to obtain a proposal $\widetilde{x}_{n+1}$, using the numerical interior floor $x_{\min} + \delta_{\text{num}}$, and assess the nominal condition $\Pr\{p_{f,\sigma}(\widetilde{x}_{n+1}) > 1 - \gamma \mid \mathcal{F}_n\} \leq \alpha$. For iBART, this probability concerns the projected tail functional.

The assigned dose also respects the protocol's dose restrictions. With a prespecified upward cap $\Delta_{\max} > 0$ and Only Escalation enabled, the next cohort receives

$$X_{n+1} = \max\left\{X_n,\, \min\left(\widetilde{x}_{n+1}, X_n + \Delta_{\max}\right)\right\}.$$

The cap limits the size of an increase, while Only Escalation prevents a decrease. Thus, the implemented allocation combines an EWOC proposal with dose restrictions that prevent abrupt escalation: even when the posterior α-quantile suggests a substantially higher dose, the increase for the next patient or cohort cannot exceed $\Delta_{\max}$.

Let K denote the prespecified DLT-count limit and D_n the cumulative number of observed DLTs. Enrollment ends at the first completed cohort for which $D_n > K$, or when the planned enrollment N is reached. Clinical thresholds, dose range, α, initial dose, patient budget, cohort size, dose restrictions, stopping limit, and priors are specified before enrollment. Appendix B gives the full rule.

This comparison uses $N = 45$, cohorts of three, $\Delta_{\max} = 3.5$, and $K = \lfloor 0.1N \rfloor = 4$, with Only Escalation enabled. The generating curve, variance, and MTD remain unknown to the allocation rule. **Dose Trial Lab** allows users to disable either dose restriction and explore alternative settings.

A common adaptive design for five models. Each model supplies an MTD posterior; the common rule $\mathcal{D}$ applies the resulting EWOC proposal, dose restrictions, and count-based stopping. Algorithm 1 gives the common sequence. Appendix B specifies MTD-posterior inversion and the common restrictions; Appendix G gives the model-specific posterior updates.

Algorithm 1 EWOC trial loop for any of the five response models

Require: Prespecified initial dose x_1, budget N, cohort size c, threshold $K_N = \lfloor 0.1N \rfloor$, and the chosen model's rule $\mathcal{D}$, including priors, clinical inputs, α, inward margin, floor, upward cap, and Only Escalation setting.

Ensure: Reported MTD estimate $\widehat{\xi}_{n_{\text{end}}} = X_{n_{\text{end}}}$ and information $\mathcal{F}_{n_{\text{end}}}$, where $n_{\text{end}} \leq N$.

Set $n = 0$, $D_0 = 0$, and $x = x_1$.
while $n < N$ **do**
  Set $c' = \min(c, N - n)$.
  Assign $X_{n+1} = \cdots = X_{n+c'} = x$.
  Observe all c' responses, append them to the history, and set $n \leftarrow n + c'$.
  Update the posterior using all of $\mathcal{F}_n$.
  Set $D_n = \sum_{i=1}^{n} \mathbf{1}\{Y_i \geq \eta\}$.
  if $n = N$ or $D_n > K_N$ **then**
    break
  end if
  Set $x = \mathcal{D}(\mathcal{F}_n)$ using the inward EWOC proposal, upward cap, and Only Escalation hold.
end while
Set $n_{\text{end}} = n$ and retain the final posterior.

Each cohort shares one dose. All of its responses enter the posterior before stopping is checked; the next dose is computed only if the trial continues.

MTD posterior consistency. Appendix F proves consistency of the MTD posterior under a fixed-cell Gaussian model, prior support at the truth, persistent cell sampling, and a strict risk-threshold margin. The proof uses a Schwartz-type likelihood numerator–denominator argument (Schwartz, 1965). Under these conditions, posterior mass concentrates at the true MTD as observations accumulate. This asymptotic result does not establish finite-sample operating characteristics.

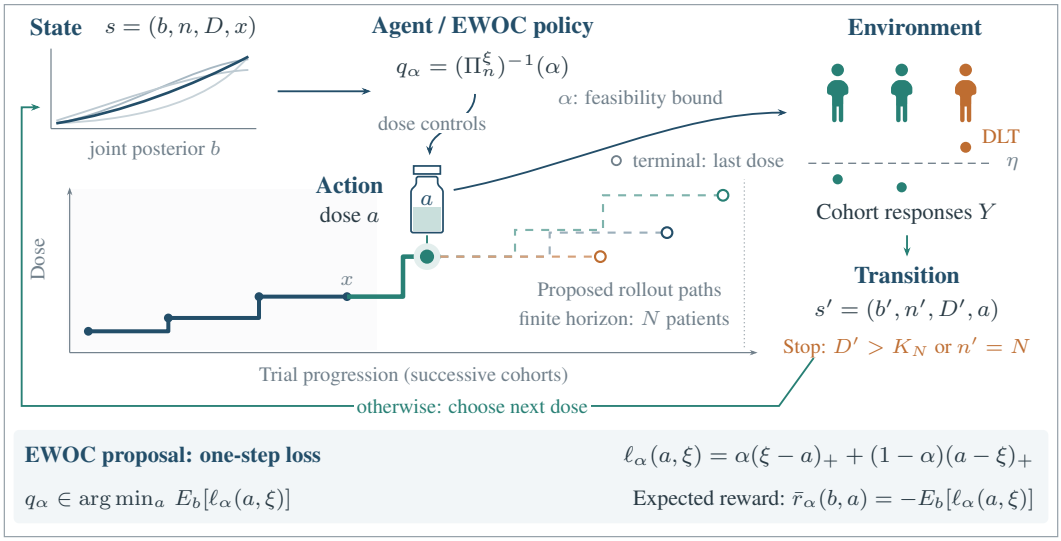


Figure 2: **Dose finding as a Bayesian RL decision process.** The state combines the joint posterior and trial status. The EWOC policy selects a dose; complete cohort responses update the state before the stopping check. Idealized EWOC minimizes posterior expected asymmetric dose loss (Bartroff & Lai, 2011), whose realized value depends on the unknown MTD. The implemented policy additionally applies the inward adjustment and dose controls, including the protocol's floor and hold fallback. Dashed paths illustrate rollout continuations; the five-model comparison uses the EWOC rule.

3 Bayesian reinforcement learning under the trial protocol

Dose selection changes both patient exposure and the information available for subsequent decisions. A Bayesian RL perspective represents this coupling through a joint posterior state and a finite enrollment horizon (Ross et al., 2011). Bayesian dynamic programming and EWOC-based rollout have clinical precedents (Bartroff & Lai, 2010); here we incorporate continuously measured toxicity outcomes, monotone tree inference, and the last-dose estimate. The upward cap, Only Escalation, and completed-cohort stopping apply to both actual and hypothetical trajectories. Figure 2 summarizes the resulting loop. We formulate EWOC and Bayesian planning within this belief-state decision process. Section 4 compares five EWOC procedures; Appendix H provides the planning proofs, and Appendix I reports the numerical rollout comparison.

State and admissible doses. After a completed cohort, use the state $s = (b, n, D, x)$: the joint parameter posterior, patient count, cumulative DLT count, and last assigned dose. The joint posterior is needed because an MTD distribution alone does not determine the next response distribution. For example, different positive slopes can have the same Gaussian tail crossing but different responses below that crossing. The same MTD belief can consequently produce different information from the next cohort. The last dose x determines the no-decrease constraint; n and D determine the remaining budget and proximity to a compulsory stop. A state is terminal exactly when $D > K_N$ or $n = N$. With Only Escalation and the upward cap active after the prespecified initial cohort, the protocol interval is

$$I(s) = [x, \min\{x + \Delta_{\max}, x_{\max} - \delta_{\text{num}}\}].$$

Set $r_b(a) = b\{p_{f,\sigma}(a) > 1 - \gamma\}$ and $F(s) = \{a \in I(s) : r_b(a) \le \alpha\}$. The planning action set is

$$\mathcal{A}(s) = \begin{cases} F(s), & F(s) \neq \varnothing, \\ \{x\}, & F(s) = \varnothing. \end{cases} \tag{2}$$

The second case enforces a hold and records a feasibility exception; it does not stop enrollment. Monotonicity implies that if the current dose is infeasible, every allowed higher dose is also infeasible. With Only Escalation disabled, replace the lower endpoint by $x_{\min} + \delta_{\text{num}}$ and use that floor as the fallback. With the cap disabled, replace the upper endpoint by $x_{\max} - \delta_{\text{num}}$. Thus the same formulation accommodates all four option combinations. For the exact-posterior EWOC policy, the

inward proposal is below its lower α-quantile unless the lower floor binds. Below that quantile, $r_b(a) \le b\{\xi \le a\} < \alpha$; an upward cap can only lower the proposal. Any infeasible baseline assignment therefore occurs at the permitted lower endpoint and agrees with the fallback in (2).

The one-step loss underlying the EWOC proposal. The agent applies a dose-selection policy, its action is the next cohort's dose, and the patient response model defines the environment. EWOC is a specified policy in this decision process. Its idealized, unadjusted posterior quantile $q_\alpha = (\Pi_n^\xi)^{-1}(\alpha)$ minimizes the posterior expected asymmetric dose loss (Bartroff & Lai, 2011):

$$\ell_\alpha(a,\xi) = \alpha(\xi - a)_+ + (1-\alpha)(a-\xi)_+, \qquad q_\alpha \in \arg\min_a \mathbb{E}_b[\ell_\alpha(a,\xi)].$$

The two terms penalize underdosing and overdosing, respectively; a small α gives greater weight to overdosing. Thus the quantile maximizes the one-step expected reward $-\mathbb{E}_b[\ell_\alpha(a,\xi)]$. The implemented policy also applies the inward adjustment and dose controls. The planning extension below evaluates the full trial under a separate cumulative loss.

Learning transition. Write ϑ for the model parameter and μ_ϑ for its likelihood mean: the raw g for iBART, and f for mBART. For $m = \min(c, N-n)$ patients at dose a, where c is cohort size,

$$P_b(d\mathbf{y} \mid a) = \int \prod_{j=1}^{m} \mathcal{N}(dy_j; \mu_\vartheta(a), \sigma^2)\, b(d\vartheta), \quad s' = \big(\mathcal{U}(b,a,\mathbf{y}),\, n+m,\, D + d(\mathbf{y}),\, a\big),$$

where $d(\mathbf{y}) = \sum_{j=1}^m \mathbf{1}\{y_j \ge \eta\}$ and $\mathcal{U}$ is the joint Bayesian update. Integrating a product of conditional densities accounts for the shared unknown parameter; multiplying marginal predictive densities would not. The final cohort is fully observed and its posterior update is completed before checking the stopping rule. Under the numerical-study protocol, $N = 45$, $c = 3$, and $K_N = 4$: five or more cumulative DLTs trigger the count rule after the cohort, and enrollment also ends at 45 patients. Neither the upward cap nor Only Escalation is a stopping rule. These variables and the joint posterior suffice for the conditional law of future responses, costs, and states (Proposition 4).

Planning with the last dose as the estimate. The planning extension uses a different loss from ℓ_α: it combines DLT counts, above-MTD assignments, and terminal relative absolute error with prespecified nonnegative weights w_D, w_O, w_R. EWOC supplies a baseline policy; it is not assumed to minimize this loss at one step or over the trial. Signed error retains its role as a reporting criterion. Let $o_b(a) = b\{\xi < a\}$. This probability defines above-MTD exposure and may differ from $r_b(a)$ at a tree jump. Write

$$R(b,x) = w_R \int \frac{|x - \xi(\vartheta)|}{\xi(\vartheta)}\, b(d\vartheta).$$

The continuous-action value $V(s)$ is the infimum of posterior expected continuation loss over admissible policies: future DLT and above-MTD counts plus terminal last-dose error. Past count contributions are fixed at the current state. The following finite-horizon Bellman recursion holds when measurable ε-optimal selectors exist for the admissible action sets:

$$\begin{aligned} V(s) &= R(b,x), \quad D > K_N \text{ or } n = N, \\ V(s) &= \inf_{a \in \mathcal{A}(s)} \left\{ w_O m\, o_b(a) + \int \big[w_D d(\mathbf{y}) + V(s')\big] P_b(d\mathbf{y} \mid a) \right\}, \quad \text{otherwise.} \end{aligned} \tag{3}$$

Its terminal estimate is the dose actually administered to the last cohort, including after a DLT stop. There is no additional terminal dose choice. The terminal expectation quantifies loss under current uncertainty; it does not replace the last dose by a posterior median. The weights express the relative importance assigned to the existing clinical criteria; they are not additional reported performance metrics. In RL terms, maximizing expected undiscounted return is equivalent to minimizing this loss: each cohort contributes reward $-w_D d(\mathbf{y}) - w_O m \mathbf{1}\{\xi < a\}$, followed by terminal reward $-w_R|x-\xi|/\xi$. MTD-dependent rewards are latent and are evaluated through posterior expectations. The DLT term in (3) uses the predictive distribution from the likelihood. In particular, iBART uses its raw mean g to generate responses, while its projected mean defines the MTD and posterior feasibility. Substituting the projected tail for the raw predictive DLT probability would change this planning problem.

The horizon is finite because every action enrolls at least one patient. With $x_{\min} > 0$, bounded doses, and fixed nonnegative weights, all costs are bounded. For a measurable finite candidate set $C(s) \subseteq \mathcal{A}(s)$, the infimum becomes a minimum; backward induction with a fixed tie rule yields a Markov optimal policy on those candidates (Proposition 5). No assertion that a minimizer exists on the full continuous action set is needed for the finite-candidate planner.

Reachability and terminal error. The allocation restrictions determine the range of terminal doses reachable from a given state. Let $H(s) = \lceil (N - n)/c \rceil$ at a nonterminal state and $H(s) = 0$ at a terminal state, and put

$$u_H = \min\{x + H(s)\Delta_{\max}, x_{\max} - \delta_{\text{num}}\}.$$

Proposition 1 (Reachability bound)**.** *Under Only Escalation and the upward cap, every admissible continuation from s has terminal dose $X_{n_{\text{end}}} \in [x, u_H]$. If $\xi(\vartheta) \in [x_{\min}, x_{\max}]$ and $x_{\min} > 0$, then for every such policy π,*

$$\mathbb{E}_b^{\pi}\left[\frac{|X_{n_{\text{end}}} - \xi(\vartheta)|}{\xi(\vartheta)}\right] \geq \int \frac{(x - \xi(\vartheta))_+ + (\xi(\vartheta) - u_H)_+}{\xi(\vartheta)}\, b(d\vartheta),$$

where $(z)_+ = \max(z, 0)$ and n_{end} is terminal enrollment.

Each cohort can increase the dose by at most $\Delta_{\max}$ and cannot decrease it, so the bound follows by measuring the distance from the latent MTD to the reachable interval. The first term is the error from overshooting a target to which the trial cannot return. The second is the error from a target too high to reach in the remaining cohorts. DLT stopping can shorten the path and cannot invalidate the bound. After the initial cohort in our protocol, $n = 3$, $x = 6$, and $H = 14$; the largest reachable last dose is $6 + 14(3.5) = 55$. Feasibility and DLT stopping can further narrow the realized path. For example, at $n = 33$ and $x = 36$, four cohorts remain and the cap permits a last dose of at most 50. A hold leaves three cohorts, lowering this upper bound to 46.5; if the true MTD is 50, every continuation then has relative absolute error at least 7%. This bound isolates lost reachability, while the Bellman objective jointly accounts for toxicity and learning. With four DLTs already observed, any additional DLT in the next completed cohort ends the trial, potentially making that cohort's dose the terminal estimate.

Policy improvement over the EWOC baseline. Let π_0 denote the exact-posterior EWOC policy, including inward adjustment, dose controls, fallback, and mandatory stopping. Choose a finite candidate set $C(s) \subseteq \mathcal{A}(s)$ containing $\pi_0(s)$. Define

$$Q^{\pi_0}(s, a) = w_O m\, o_b(a) + \int \left[w_D d(\mathbf{y}) + V^{\pi_0}(s')\right] P_b(d\mathbf{y} \mid a),$$
$$\pi_+(s) \in \arg\min_{a \in C(s)} Q^{\pi_0}(s, a).$$

Thus a candidate is evaluated using its immediate cohort and the complete future EWOC continuation, including a possible early DLT stop. The improved policy recomputes this comparison at every reached state and therefore governs the full realized trial.

Proposition 2 (Exact rollout improvement)**.** *For fixed nonnegative weights, exact posterior predictive expectations, and measurable finite candidate sets containing the baseline action, the policy π_+ with a fixed tie rule satisfies $V^{\pi_+}(s) \leq V^{\pi_0}(s)$ at every state. Both policies use the same response model, dose restrictions, and terminal rules.*

The baseline candidate gives $Q^{\pi_0}(s, \pi_+(s)) \leq V^{\pi_0}(s)$; induction on the remaining cohorts extends that comparison to repeated rollout decisions. For the same candidate sets, $V_C(s) \leq V^{\pi_+}(s) \leq V^{\pi_0}(s)$: rollout has no greater expected loss than the baseline, while V_C records the finite-candidate optimum. This ordering concerns exact Bayesian evaluation of the specified weighted loss.

Rollout evaluates candidate doses through hypothetical trials, updating the posterior after every simulated cohort until stopping. Bartroff & Lai (2010) implemented EWOC rollout in phase I trials and noted its computational cost. Algorithm 3 and Figure 8 in Appendix H.7 give the full procedure and schematic. Appendix I and Table 4 give a 1PLD rollout illustration with analytic posterior updates. Across upward caps 3.5, 7, and 10.5, rollout lowered observed mean weighted loss under all three weight profiles at both $B = 1,000$ and $2,000$. Safety emphasis reduced NPD and NPO, precision emphasis reduced RAE, and results changed little when B doubled. **Dose Trial Lab** also implements rollout for 2PLD, 3PND, iBART, and mBART, whose posterior updates are numerical.

The final-cohort decision. When only one cohort remains, every candidate dose becomes the terminal estimate. Write $\rho_b(a)$ for the posterior predictive DLT probability, computed from the likelihood mean μ_ϑ. The candidate's expected loss then reduces to

$$Q(s,a) = mw_D\rho_b(a) + mw_O o_b(a) + R(b,a).$$

This follows by averaging the terminal posterior loss over the cohort's possible responses: the law of iterated expectation gives $\int R\{\mathcal{U}(b,a,\mathbf{y}),a\}P_b(d\mathbf{y} \mid a) = R(b,a)$. The final observations can update the posterior but cannot change the already assigned dose. Earlier in the trial, information can change subsequent assignments, and its value depends on which doses remain reachable. The Bellman recursion accounts for this difference without adding a separate reward for reducing posterior uncertainty.

4 Experiments

4.1 Design and performance criteria

The study compares 1PLD, 2PLD, 3PND, iBART, and mBART across seven dose–toxicity curves: linear, piecewise linear, quadratic, square root, sigmoid, exponential, and logarithmic. Figure 9 in Appendix J.2 shows their shapes and shared MTD calibration. The dose interval is $(5, 80)$, initial dose 6, toxicity threshold $\eta = 3$, true residual SD 0.1, and true MTD $\xi_0 = 50$. Each trial plans $N = 45$ patients in cohorts of three, with $\gamma = 0.9$ and $\alpha = 0.05$. Both controls are enabled: the upward cap 3.5 and Only Escalation (Section 2.3). They are required components of the evaluated protocol and the software defaults. The posterior uses every response in each completed cohort. The stopping rule is checked after every cohort: enrollment ends when cumulative DLTs exceed $K_N = 4$, or when 45 patients have been observed.

The study uses 100 paired replications per curve; each replication evaluates all five methods with shared presampled patient innovations, for 3,500 trials. Posterior sampling uses separate streams. Priors are specified in Appendices C and D; 1PLD knows the residual SD, whereas the other models estimate it. Induced MTD priors differ, so this is a comparison of complete procedures.

The reported MTD estimate is the last administered dose $\widehat{\xi}_{n_{\text{end}}} = X_{n_{\text{end}}}$. The four performance criteria are

$$\text{NPD} = \sum_{i=1}^{n_{\text{end}}} \mathbf{1}\{Y_i \geq \eta\},\ \ \text{NPO} = \sum_{i=1}^{n_{\text{end}}} \mathbf{1}\{X_i > \xi_0\},\ \ \text{BTM} = X_{n_{\text{end}}} - \xi_0,\ \ \text{RAE} = |\text{BTM}|/\xi_0.$$

NPD and NPO assess safety performance. The precision criteria BTM and RAE describe MTD estimation error: BTM retains its direction, whereas RAE measures its relative magnitude without cancellation. All trials remain in the summaries with their actual enrollment and early stopping status.

4.2 Seven-curve comparison

Table 1 reports replicate means, medians, and Monte Carlo standard errors (MCSEs) for the four criteria. No trial ended before $N = 45$, so the early-stopping percentage was zero. Appendices J and G give the simulation design, enrollment and stopping summaries, and posterior computation.

Table 1: Entries are *mean* | *median* | *MCSE*; RAE is a percentage. Bold, underlined methods minimize mean RAE (p), both mean NPD and NPO (s), or both categories (s, p) within each curve, using unrounded means and exact ties. A zero MCSE may reflect rounding.

		Safety performance		**Precision performance**	
Curve	Method	NPD	NPO	BTM	RAE (%)
Linear	**1PLD**$^{\text{p}}$	0.34 | 0.00 | 0.06	0.33 | 0.00 | 0.13	-0.57 | -0.58 | 0.04	1.20 | 1.16 | 0.07
	2PLD	0.26 | 0.00 | 0.05	0.24 | 0.00 | 0.11	-0.75 | -0.75 | 0.05	1.53 | 1.49 | 0.09
	3PND	0.20 | 0.00 | 0.04	0.27 | 0.00 | 0.11	-0.89 | -0.87 | 0.06	1.83 | 1.74 | 0.11
	iBART$^{\text{s}}$	0.03 | 0.00 | 0.02	0.00 | 0.00 | 0.00	-1.99 | -1.93 | 0.06	3.98 | 3.86 | 0.12
	mBART	0.06 | 0.00 | 0.02	0.06 | 0.00 | 0.06	-1.61 | -1.93 | 0.06	3.24 | 3.86 | 0.11

Table 1 (continued)

		Safety performance						Precision performance					
Curve	Method	NPD			NPO			BTM			RAE (%)		
Piecewise linear	1PLD	3.86	4.00	0.09	6.00	6.00	0.00	5.00	5.00	0.00	10.00	10.00	0.00
	2PLD	3.86	4.00	0.09	6.00	6.00	0.00	5.00	5.00	0.00	10.00	10.00	0.00
	3PND	3.86	4.00	0.09	6.00	6.00	0.00	4.98	5.00	0.01	9.97	10.00	0.01
	iBART^{s}	2.02	2.00	0.12	6.00	6.00	0.00	1.51	1.50	0.00	3.02	3.00	0.01
	mBART^{s,p}	2.02	2.00	0.12	6.00	6.00	0.00	1.50	1.50	0.00	3.00	3.00	0.00
Quadratic	1PLD	5.21	5.00	0.07	6.00	6.00	0.00	5.00	5.00	0.00	10.00	10.00	0.00
	2PLD	4.36	4.00	0.10	6.00	6.00	0.00	1.51	1.50	0.01	3.02	3.00	0.01
	3PND^{p}	0.93	1.00	0.10	3.60	6.00	0.29	0.12	0.09	0.04	0.55	0.42	0.05
	iBART^{s}	0.39	0.00	0.09	0.99	0.00	0.22	-0.46	-0.45	0.05	1.19	0.89	0.07
	mBART	1.05	0.00	0.15	2.73	0.00	0.30	0.04	-0.45	0.07	1.20	0.89	0.07
Square root	**1PLD**^{s}	0.00	0.00	0.00	0.00	0.00	0.00	-10.02	-10.03	0.04	20.05	20.07	0.08
	2PLD^{s}	0.00	0.00	0.00	0.00	0.00	0.00	-17.78	-17.81	0.06	35.56	35.62	0.13
	3PND^{p}	0.11	0.00	0.03	0.00	0.00	0.00	-3.45	-3.40	0.12	6.90	6.80	0.24
	iBART	0.03	0.00	0.02	0.00	0.00	0.00	-4.75	-4.90	0.09	9.49	9.80	0.18
	mBART	0.06	0.00	0.02	0.00	0.00	0.00	-4.37	-4.16	0.09	8.73	8.32	0.19
Sigmoid	1PLD	4.73	5.00	0.09	6.00	6.00	0.00	5.00	5.00	0.00	10.00	10.00	0.00
	2PLD	3.52	4.00	0.12	6.00	6.00	0.00	1.50	1.50	0.00	2.99	3.00	0.01
	3PND^{s}	0.12	0.00	0.04	0.00	0.00	0.00	-0.78	-0.76	0.03	1.55	1.52	0.07
	iBART	0.14	0.00	0.04	0.27	0.00	0.12	-0.84	-1.19	0.05	1.74	2.38	0.09
	mBART^{p}	0.53	0.00	0.09	1.71	0.00	0.27	-0.40	-0.45	0.07	1.43	0.89	0.09
Exponential	1PLD	5.67	6.00	0.06	6.00	6.00	0.00	5.00	5.00	0.00	10.00	10.00	0.00
	2PLD	5.28	5.00	0.07	6.00	6.00	0.00	1.59	1.50	0.02	3.18	3.00	0.04
	3PND	3.97	4.00	0.17	5.94	6.00	0.06	1.11	1.21	0.04	2.23	2.43	0.08
	iBART^{s,p}	0.72	0.00	0.13	2.25	0.00	0.29	-0.18	-0.45	0.06	1.08	0.89	0.07
	mBART	2.24	2.00	0.21	4.47	6.00	0.26	0.47	0.30	0.08	1.49	0.89	0.10
Logarithmic	**1PLD**^{s}	0.00	0.00	0.00	0.00	0.00	0.00	-8.24	-8.22	0.04	16.49	16.45	0.08
	2PLD^{s}	0.00	0.00	0.00	0.00	0.00	0.00	-12.63	-12.65	0.05	25.26	25.30	0.10
	3PND	0.03	0.00	0.02	0.00	0.00	0.00	-4.72	-4.83	0.09	9.44	9.66	0.18
	iBART	0.02	0.00	0.01	0.00	0.00	0.00	-4.46	-4.16	0.08	8.93	8.32	0.16
	mBART^{p}	0.04	0.00	0.02	0.00	0.00	0.00	-3.95	-4.16	0.09	7.90	8.32	0.19

For safety, Table 1 shows that iBART jointly minimized the observed mean NPD and NPO on four of seven curves, including a tie with mBART for the piecewise linear curve. On the square-root and logarithmic curves, 1PLD and 2PLD recorded zero NPD and NPO but substantially underestimated the MTD; iBART recorded zero mean NPO and mean NPD of only 0.03 and 0.02, respectively, with lower RAE than those zero-count procedures. The sigmoid curve instead favored 3PND on safety.

For precision, the lowest observed mean RAE was attained by 1PLD on the linear curve, 3PND on the quadratic and square-root curves, iBART on the exponential curve, and mBART on the piecewise linear, sigmoid, and logarithmic curves. mBART on the piecewise linear curve and iBART on the exponential curve attained both safety and precision minima. No method attained both the safety and precision minima across all seven curves.

5 Discussion

Flexible nonparametric and machine learning methods often need more data than correctly specified parametric models, a particular obstacle in phase I cancer trials with limited enrollment. Our design uses the available structure: dose is the sole predictor, both Bayesian tree models enforce monotonicity, and controlled EWOC limits escalation. Across seven dose–toxicity curves with 45 patients, these models balance safety and MTD precision without prespecifying curve shape, showing that flexible MTD estimation remains practical in this small-sample setting. We also applied all five methods to the O^6-benzylguanine example to estimate an AGT-based dose target. Their final doses of 80–100 mg/m^2 are broadly consistent with the 100 mg/m^2 biochemical modulatory dose reported by Friedman et al. (1998); see Supplement Section K.

Future work will compare deep reinforcement learning (Matsuura et al., 2022; 2023), Super Learner ensembles (van der Laan et al., 2007), and Gaussian processes (Gotovos et al., 2013; Losalka & Scarlett, 2023; Chien et al., 2024) under state-dependent dose controls, and quantify the effects of finite rollout budgets and approximate posterior updates.

ETHICS STATEMENT

The study uses simulated trials and previously reported pharmacodynamic measurements. It enrolls no participants.

REPRODUCIBILITY STATEMENT

The supplement specifies the models, priors, allocation and stopping rules, simulation scenarios, posterior computation, and evaluation criteria. It provides the RL formulation, proofs, rollout algorithm, and policy comparison. The accompanying anonymous software package provides a Windows installer for **Dose Trial Lab**, which supports EWOC, rollout, and paired policy comparisons for all five models. Appendix L gives an illustrated usage guide.

AI USE STATEMENT

Large language models (LLMs) assisted with grammatical checking and language polishing of the manuscript, and with software development for **Dose Trial Lab**. The authors are responsible for the mathematical claims, empirical results, citations, and final text.

REFERENCES

J. Babb, A. Rogatko, and S. Zacks. Cancer phase I clinical trials: Efficient dose escalation with overdose control. *Statistics in Medicine*, 17(10):1103–1120, 1998. doi: 10.1002/(SICI)1097-0258(19980530)17:10⟨1103::AID-SIM793⟩3.0.CO;2-9. URL https://pubmed.ncbi.nlm.nih.gov/9618772/.

Jay Bartroff and Tze Leung Lai. Approximate dynamic programming and its applications to the design of phase I cancer trials. *Statistical Science*, 25(2):245–257, 2010. doi: 10.1214/10-STS317. URL https://arxiv.org/abs/1011.6509.

Jay Bartroff and Tze Leung Lai. Incorporating individual and collective ethics into phase I cancer trial designs. *Biometrics*, 67(2):596–603, 2011. doi: 10.1111/j.1541-0420.2010.01471.x. URL https://arxiv.org/abs/1108.1223.

Zhengjia Chen, Mourad Tighiouart, and Jeanne Kowalski. Dose escalation with overdose control using a quasi-continuous toxicity score in cancer phase I clinical trials. *Contemporary Clinical Trials*, 33(5):949–958, 2012. doi: 10.1016/j.cct.2012.04.007. URL https://pubmed.ncbi.nlm.nih.gov/22561391/.

Isabel Chien, Wessel P. Bruinsma, Javier Gonzalez, and Richard E. Turner. Safe exploration in dose finding clinical trials with heterogeneous participants. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 8436–8467. PMLR, 2024. URL https://proceedings.mlr.press/v235/chien24a.html.

Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010. doi: 10.1214/09-AOAS285. URL https://doi.org/10.1214/09-AOAS285.

Hugh A. Chipman, Edward I. George, Robert E. McCulloch, and Thomas S. Shively. mBART: Multidimensional monotone BART. *Bayesian Analysis*, 17(2):515–544, 2022. doi: 10.1214/21-BA1259. URL https://doi.org/10.1214/21-BA1259.

Yu-Che Chung, Yunqi Zhao, Meizi Liu, Jianchang Lin, and Rachael Liu. DOD-PRO-BART: Dose optimization design incorporating patient-reported outcomes via machine learning with Bayesian additive regression trees. *Statistics in Biopharmaceutical Research*, 17(3):347–356, 2025. doi: 10.1080/19466315.2025.2491596. URL https://doi.org/10.1080/19466315.2025.2491596.

Suzanne E. Dahlberg, Geoffrey I. Shapiro, Jeffrey W. Clark, and Bruce E. Johnson. Evaluation of statistical designs in phase I expansion cohorts: The Dana-Farber/Harvard cancer center experience. *Journal of the National Cancer Institute*, 106(7):dju163, 2014. doi: 10.1093/jnci/dju163. URL https://academic.oup.com/jnci/article/106/7/dju163/1010404.

Jan de Leeuw, Kurt Hornik, and Patrick Mair. Isotone optimization in R: Pool-adjacent-violators algorithm (PAVA) and active set methods. *Journal of Statistical Software*, 32(5):1–24, 2009. doi: 10.18637/jss.v032.i05. URL https://www.jstatsoft.org/article/view/v032i05.

B. H. Eichhorn and S. Zacks. Sequential search of an optimal dosage, I. *Journal of the American Statistical Association*, 68(343):594–598, 1973. doi: 10.1080/01621459.1973.10481389.

H. S. Friedman, D. M. Kokkinakis, J. Pluda, A. H. Friedman, I. Cokgor, M. M. Haglund, D. M. Ashley, J. Rich, M. E. Dolan, A. E. Pegg, R. C. Moschel, R. E. McLendon, T. Kerby, J. E. Herndon, D. D. Bigner, and S. C. Schold, Jr. Phase I trial of O6-benzylguanine for patients undergoing surgery for malignant glioma. *Journal of Clinical Oncology*, 16(11):3570–3575, 1998. doi: 10.1200/JCO.1998.16.11.3570. URL https://ascopubs.org/doi/10.1200/JCO.1998.16.11.3570.

Alkis Gotovos, Nathalie Casati, Gregory Hitz, and Andreas Krause. Active learning for level set estimation. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, pp. 1344–1350, 2013. URL https://www.ijcai.org/Proceedings/13/Papers/202.pdf.

Arthur Guez, David Silver, and Peter Dayan. Efficient Bayes-adaptive reinforcement learning using sample-based search. In *Advances in Neural Information Processing Systems*, volume 25, pp. 1025–1033, 2012. URL https://proceedings.neurips.cc/paper/2012/hash/35051070e572e47d2c26c241ab88307f-Abstract.html.

Yu Inatsu, Shogo Iwazaki, and Ichiro Takeuchi. Active learning for distributionally robust level-set estimation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4574–4584. PMLR, 2021. URL https://proceedings.mlr.press/v139/inatsu21a.html.

Anastasia Ivanova and Se Hee Kim. Dose finding for continuous and ordinal outcomes with a monotone objective function: A unified approach. *Biometrics*, 65(1):307–315, 2009. doi: 10.1111/j.1541-0420.2008.01045.x. URL https://pubmed.ncbi.nlm.nih.gov/18479486/.

Se Yoon Lee. A flexible dose-response modeling framework based on continuous toxicity outcomes in phase I cancer clinical trials. *Trials*, 24:745, 2023. doi: 10.1186/s13063-023-07793-0. URL https://link.springer.com/article/10.1186/s13063-023-07793-0.

Se Yoon Lee, Alain Munafo, Pascal Girard, and Kosalaram Goteti. Optimization of dose selection using multiple surrogates of toxicity as a continuous variable in phase I cancer trial. *Contemporary Clinical Trials*, 113:106657, 2022. doi: 10.1016/j.cct.2021.106657. URL https://pubmed.ncbi.nlm.nih.gov/34954097/.

Suyu Liu and Ying Yuan. Bayesian optimal interval designs for phase I clinical trials. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 64(3):507–523, 2015. doi: 10.1111/rssc.12089. URL https://doi.org/10.1111/rssc.12089.

Arpan Losalka and Jonathan Scarlett. Benefits of monotonicity in safe exploration with Gaussian processes. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pp. 1304–1314. PMLR, 2023. URL https://proceedings.mlr.press/v216/losalka23a.html.

Arpan Losalka and Jonathan Scarlett. No-regret algorithms for safe Bayesian optimization with monotonicity constraints. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 3232–3240. PMLR, 2024. URL https://proceedings.mlr.press/v238/losalka24a.html.

Kentaro Matsuura, Junya Honda, Imad El Hanafi, Takashi Sozu, and Kentaro Sakamaki. Optimal adaptive allocation using deep reinforcement learning in a dose-response study. *Statistics in Medicine*, 41(7):1157–1171, 2022. doi: 10.1002/sim.9247. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC9298337/.

Kentaro Matsuura, Kentaro Sakamaki, Junya Honda, and Takashi Sozu. Optimal dose escalation methods using deep reinforcement learning in phase I oncology trials. *Journal of Biopharmaceutical Statistics*, 33(5):639–652, 2023. doi: 10.1080/10543406.2023.2170402. URL https://www.tandfonline.com/doi/abs/10.1080/10543406.2023.2170402.

Radford M. Neal. Slice sampling. *The Annals of Statistics*, 31(3):705–767, 2003. doi: 10.1214/aos/1056562461. URL https://doi.org/10.1214/aos/1056562461.

Beat Neuenschwander, Michael Branson, and Thomas Gsponer. Critical aspects of the Bayesian approach to phase I cancer trials. *Statistics in Medicine*, 27(13):2420–2439, 2008. doi: 10.1002/sim.3230. URL https://pubmed.ncbi.nlm.nih.gov/18344187/.

John O'Quigley and Mark Conaway. Continual reassessment and related dose-finding designs. *Statistical Science*, 25(2):202–216, 2010. doi: 10.1214/10-STS332. URL https://pubmed.ncbi.nlm.nih.gov/21731191/.

John O'Quigley, Margaret Pepe, and Lloyd Fisher. Continual reassessment method: A practical design for phase I clinical trials in cancer. *Biometrics*, 46(1):33–48, 1990. doi: 10.2307/2531628. URL https://pubmed.ncbi.nlm.nih.gov/2350571/.

Kristyn Pantoja, Shankar Lanke, Alain Munafo, Anja Victor, Christina Habermehl, Armin Schueler, Karthik Venkatakrishnan, Pascal Girard, and Kosalaram Goteti. Designing phase I oncology dose escalation using dose–exposure–toxicity models as a complementary approach to model-based dose–toxicity models. *CPT: Pharmacometrics & Systems Pharmacology*, 11(10):1371–1381, 2022. doi: 10.1002/psp4.12851. URL https://doi.org/10.1002/psp4.12851.

Stéphane Ross, Joelle Pineau, Brahim Chaib-draa, and Pierre Kreitmann. A Bayesian approach for learning and planning in partially observable Markov decision processes. *Journal of Machine Learning Research*, 12(48):1729–1770, 2011. URL https://jmlr.org/papers/v12/ross11a.html.

Lorraine Schwartz. On Bayes procedures. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 4(1):10–26, 1965. doi: 10.1007/BF00535479. URL https://doi.org/10.1007/BF00535479.

Yanan Sui, Alkis Gotovos, Joel Burdick, and Andreas Krause. Safe exploration for optimization with Gaussian processes. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 997–1005. PMLR, 2015. URL https://proceedings.mlr.press/v37/sui15.html.

Mourad Tighiouart and André Rogatko. Dose finding with escalation with overdose control (EWOC) in cancer clinical trials. *Statistical Science*, 25(2):217–226, 2010. doi: 10.1214/10-STS333. URL https://arxiv.org/abs/1011.6479.

Mark J. van der Laan, Eric C. Polley, and Alan E. Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1):Article 25, 2007. doi: 10.2202/1544-6115.1309. URL https://doi.org/10.2202/1544-6115.1309.

S. Zacks, A. Rogatko, and J. Babb. Optimal Bayesian-feasible dose escalation for cancer phase I trials. *Statistics & Probability Letters*, 38(3):215–220, 1998. doi: 10.1016/S0167-7152(98)00016-9. URL https://www.sciencedirect.com/science/article/pii/S0167715298000169.

Yunqi Zhao, Rachael Liu, Jianchang Lin, Andy Chi, and Simon Davies. DOD-BART: machine learning-based dose optimization design incorporating patient-level prognostic factors via Bayesian additive regression trees. *Journal of Biopharmaceutical Statistics*, pp. 1–16, 2024. doi: 10.1080/10543406.2024.2429463. URL https://doi.org/10.1080/10543406.2024.2429463. Advance online publication.

GUIDE TO THE SUPPLEMENTARY MATERIAL

Section	Contents	Page
A. Related methods		
A	Clinical outcomes and decision targets	13
B–E. Models and allocation		
B	MTD definition, posterior inversion, and EWOC allocation	15
C	Tree priors, calibration, and isotonic projection	17
D	1PLD, 2PLD, and 3PND likelihoods and priors	19
E	Mean dose–toxicity curves and graphical models	21
F–G. Posterior theory and computation		
F	MTD posterior consistency	22
G	Posterior computation	25
H–I. Reinforcement learning and rollout		
H	Belief-state planning, reachability, and rollout proofs	28
I	Linear 1PLD rollout illustration and budget sensitivity	36
J. Numerical study		
J	Dose–toxicity scenarios, enrollment, and performance metrics	38
K. Pharmacodynamic case study		
K	Sequential pharmacodynamic example	40
L. Software manual		
L	Installation, settings, trial playback, visualization, and exports	43

A RELATED METHODS

The literature on dose finding in phase I cancer trials is extensive and diverse. We review the work most closely related to our methods.

A.1 BINARY TOXICITY RESPONSES

Parametric models. Several widely used designs learn from the same binary DLT indicator but translate it into a dose decision differently. The continual reassessment method (CRM) updates a one-parameter monotone working model; common implementations encode prior DLT probabilities across dose levels in a skeleton and choose the dose whose posterior toxicity estimate is closest to the target (O'Quigley et al., 1990; O'Quigley & Conaway, 2010). Classical EWOC instead chooses a lower posterior quantile of the MTD, with a feasibility bound that directly limits the posterior probability of assigning a dose above the MTD (Babb et al., 1998; Zacks et al., 1998; Tighiouart & Rogatko, 2010). The Bayesian logistic regression model (BLRM) uses a two-parameter logistic curve in log dose, a posterior target interval, and an overdose-probability screen (Neuenschwander et al., 2008). The Bayesian optimal interval (BOIN) design compares the observed DLT rate at the current dose with prespecified escalation and de-escalation boundaries (Liu & Yuan, 2015). Thus, CRM, EWOC, BLRM, and BOIN all use binary DLT feedback, but their dose-selection rules are distinct.

Nonparametric models. Gaussian-process (GP) models allow more flexible dose–toxicity relationships. SAFE-T uses multiple-output GPs with a probit link to model binary toxicity and efficacy in heterogeneous participants (Chien et al., 2024). The latent function is continuous-valued, but the observed toxicity response is binary. Thus, flexibility in the probability curve does not change the type of toxicity information entering the likelihood.

A.2 CONTINUOUSLY MEASURED TOXICITY OUTCOMES

Parametric models. Earlier designs retained toxicity severity in forms other than a binary DLT indicator. Ivanova & Kim (2009) give a unified local allocation rule for binary, ordinal, or continuous monotone objectives. EWOC-NETS places a normalized equivalent toxicity score in a quasi-Bernoulli likelihood and applies EWOC to the resulting quasi-continuous outcome (Chen et al., 2012). These

approaches establish that graded toxicity information can guide phase I allocation, while using a local monotone rule or a parametric score model rather than a flexible tree-based mean.

Methods using the measured toxicity outcome retain more information than its binary DLT indicator. The one-parameter linear dose finder (1PLD) estimates a slope with residual variance treated as known, while the two-parameter linear dose finder (2PLD) estimates both slope and residual variance (Eichhorn & Zacks, 1973; Lee et al., 2022). The three-parameter nonlinear dose finder (3PND) additionally allows curvature through a power exponent (Lee, 2023). Although 3PND accommodates nonlinear mean curves, it remains parametric because the curve belongs to a specified finite-dimensional family. In these models, a response threshold defines DLT events, and the response distribution determines their dose-dependent probability.

Nonparametric models. Flexible modeling is central when toxicity is continuously measured because the MTD depends on a tail probability of the response distribution; misspecifying the mean curve can therefore shift the clinical target. BART represents the mean as a sum of regression trees and can recover nonlinear shapes without committing to a linear or power-law curve (Chipman et al., 2010). We develop two monotone Bayesian tree procedures for this setting: iBART projects each posterior draw onto nondecreasing functions, whereas mBART imposes monotonicity within the tree ensemble (Chipman et al., 2022). Both retain flexible nonlinear learning while enforcing the clinical requirement that toxicity not decrease with dose.

The core contribution is to turn these monotone tree models into a complete phase I decision framework. Each joint posterior draw of the mean curve and residual standard deviation induces an MTD draw under the specified DLT threshold and acceptable toxicity probability. The resulting MTD posterior directly drives EWOC and EWOC rollout, so uncertainty in curve shape and outcome variability propagates into each dose decision. This framework uses the full continuously measured toxicity outcome, accommodates nonlinear relationships beyond the 1PLD, 2PLD, and 3PND forms, and supports MTD-posterior consistency, a reachability bound, and exact expected-loss rollout improvement under the stated conditions.

Table 2 shows how representative phase I dose-finding methods differ in toxicity outcome type, dose–toxicity model, dose-selection approach, and theoretical scope. The final row highlights the combination developed here: continuously measured toxicity, monotone Bayesian trees, EWOC and EWOC rollout, and formal guarantees for MTD learning and policy improvement under the stated conditions.

Table 2: **Representative phase I dose-finding methods and their relation to the present study.**

Selected method	Toxicity outcome type	Dose–toxicity model	Dose-selection approach	Main scope or result
CRM (O'Quigley et al., 1990)	Binary	One-parameter monotone working model; calibrated skeleton	Skeleton	Foundational continual reassessment
Classical EWOC (Babb et al., 1998)	Binary	Parametric DLT-risk and MTD model	EWOC	Direct posterior overdose control
BLRM (Neuenschwander et al., 2008)	Binary	Two-parameter logistic model in log standardized dose	Target interval	Prior-calibrated, interval-based decisions
BOIN (Liu & Yuan, 2015)	Binary	Model-assisted interval construction	Interval rule	Simple decision rule with finite- and large-sample properties
Graded-outcome designs (Ivanova & Kim, 2009; Chen et al., 2012)	Ordinal / continuous	Monotone objective or quasi-Bernoulli NETS model	Up/down or EWOC	Retains severity beyond one DLT indicator
Bartroff–Lai rollout (Bartroff & Lai, 2010)	Binary	Two-parameter logistic model	EWOC rollout	Combines cumulative patient loss and terminal estimation loss
1PLD, 2PLD, and 3PND (Eichhorn & Zacks, 1973; Lee et al., 2022; Lee, 2023)	Continuous	Linear (1PLD/2PLD) or nonlinear power (3PND)	EWOC	Parametric mean; residual variation learned in 2PLD and 3PND
This paper	**Continuous**	**Monotone Bayesian Trees (iBART/mBART) with 1PLD, 2PLD, and 3PND as comparators**	**EWOC and EWOC rollout**	**MTD-posterior consistency, a reachability bound, and exact expected-loss rollout improvement under stated conditions**

A.3 Clinical objectives and sequential planning

Beyond the response type and curve model, dose-finding designs differ in their clinical objectives. Related BART-based approaches address broader phase I/II dose optimization. DOD-BART combines prognostic factors and accruing outcomes in seamless phase I/II allocation (Zhao et al., 2024), while DOD-PRO-BART incorporates patient-reported outcomes alongside clinician-reported toxicity and efficacy (Chung et al., 2025). Our study focuses on phase I dose finding with a single continuously measured toxicity outcome, where the primary objective is to estimate the MTD and guide dose allocation under a tightly limited patient budget.

Bayesian planning represents accumulated information through a posterior belief state, allowing decisions to account for immediate outcomes and subsequent learning. Ross et al. (2011) develop this approach for partially observed systems with unknown models, while Guez et al. (2012) study sample-based Bayes-adaptive search. In phase I design, Bartroff & Lai (2010) formulate Bayesian dynamic programming with cumulative patient losses and terminal estimation loss, and study rollout with EWOC as the base policy. Deep reinforcement learning has also been used for adaptive allocation in dose–response studies (Matsuura et al., 2022) and dose escalation in phase I oncology trials (Matsuura et al., 2023).

Safe optimization and level-set learning provide related approaches to sequential exploration using GP confidence bounds (Sui et al., 2015; Gotovos et al., 2013). Monotone variants address safety, regret, and safe-boundary learning (Losalka & Scarlett, 2023; 2024), while robust level-set methods consider environmental uncertainty (Inatsu et al., 2021). Their function-value feedback and function-level safety objectives differ from the patient-level toxicity probability induced by a continuous response distribution.

Our planning state contains the joint posterior, cumulative DLT count, enrollment, and last administered dose; these variables determine the next transition and permitted actions. The upward cap and no-decrease restriction yield the terminal-error bound in Proposition 1, and the final loss evaluates the last administered dose. The rollout result specializes exact expected-loss policy improvement to these dose controls, cohort stopping rules, and terminal estimation convention.

A.4 Clinical risk and posterior feasibility

The clinical target is $p_0(x) = \Pr_0(Y \geq \eta \mid X = x) \leq 1 - \gamma$. For a continuously measured outcome, this probability depends on both the mean response and residual variability. The EWOC feasibility criterion concerns posterior uncertainty about the modeled risk, with nominal bound α. The clinical risk limit $1 - \gamma$ and the posterior feasibility bound α therefore have different roles. For iBART, the modeled risk is a projected tail functional. A no-decrease restriction can require a dose exceeding the nominal feasibility bound, which is recorded as an allocation exception. The observed DLT-count stopping threshold is separate from both probability bounds. Figure 3 illustrates the measured-outcome target and the effect of residual variability on its dose location.

B The MTD posterior and allocation rule

B.1 Clinical specification

Let x denote dose in its original units and $Y(x)$ a continuously measured toxicity outcome. Before enrollment, clinicians specify the interval $\mathcal{X} = (x_{\min}, x_{\max}) \subset (0, \infty)$, the maximum toxicity level $\eta > 0$, and the homogeneity constant $\gamma \in (1/2, 1)$ (Lee et al., 2022; Lee, 2023). Set $\theta = 1 - \gamma$. The criterion $\Pr\{Y(x) < \eta\} \geq \gamma$ defines an acceptable dose. These quantities determine the target; the feasibility bound α determines how doses are selected under posterior uncertainty.

B.2 Likelihood

Write $\mathcal{F}_n$ for the history of the first n dose–response pairs $\{(X_i, Y_i)\}_{i=1}^n$, where i indexes patients and N is planned enrollment. For a cohort, doses are selected using the history at the preceding completed-cohort boundary and remain fixed until all cohort responses are observed. For iBART, let

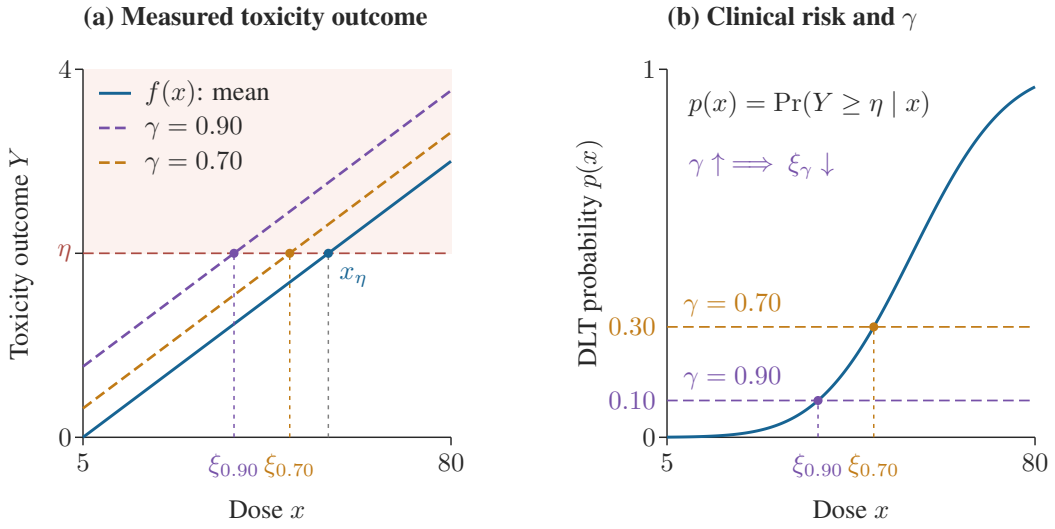


Figure 3: **A continuously measured toxicity outcome, the clinical target, and the MTD.** Gaussian schematic with $f(x) = 0.04(x - 5)$, $\sigma = 0.6$, and $\eta = 2$. The subscript in ξ_γ identifies the required non-DLT probability. (a) The response quantile $f(x) + \sigma\Phi^{-1}(\gamma)$ reaches η at the MTD; the mean reaches it at x_η, where DLT probability is $1/2$. (b) Increasing γ lowers the acceptable risk $1 - \gamma$ and the MTD; the underlying risk curve is unchanged. Dashed response curves describe outcome variability, not posterior uncertainty. The schematic follows the interpretation in Lee (2023, Figure 3).

g denote the unconstrained mean. The sequential working model is

$$Y_i \mid \mathcal{F}_{i-1}, X_i, g, \sigma \sim \mathcal{N}\{g(X_i), \sigma^2\}, \qquad \sigma > 0. \tag{4}$$

The allocation kernel is the same function of past observations under every candidate parameter. Prespecified dose restrictions are also functions of that history and do not change this property. The allocation factors therefore cancel from the posterior:

$$d\Pi(g, \sigma \mid \mathcal{F}_n) \propto \prod_{i=1}^{n} \sigma^{-1}\phi\{(Y_i - g(X_i))/\sigma\}\, d\Pi(g, \sigma). \tag{5}$$

For mBART, replace g by its monotone mean f in (4) and (5). In iBART, the unconstrained draw g yields $f = \mathcal{P}g$ by projection onto the nondecreasing cone in $L_2([x_{\min}, x_{\max}])$. The pool-adjacent-violators algorithm (PAVA) uses the lengths of the constant dose intervals as weights. In mBART, each tree is nondecreasing and no projection is needed. In both cases the mean draw retains its corresponding residual standard deviation (SD).

B.3 MTD POSTERIOR

For the projected curve $f = \mathcal{P}g$ in iBART or the directly sampled monotone curve f in mBART, let

$$\begin{aligned} p_{f,\sigma}(x) &= \Phi\{(f(x) - \eta)/\sigma\}, \\ \xi(f, \sigma) &= \sup\big(\{x_{\min}\} \cup \{x \in (x_{\min}, x_{\max}) : p_{f,\sigma}(x) \leq 1 - \gamma\}\big). \end{aligned} \tag{6}$$

The lower endpoint $\xi = x_{\min}$ represents the absence of an acceptable admissible dose; $\xi = x_{\max}$ represents an entirely acceptable interval. The posterior of ξ is supported on the ordinary dose interval $[x_{\min}, x_{\max}]$. Neither endpoint is administered.

Applying (6) to every paired mean and variance draw propagates their joint uncertainty to the MTD. Inverting a posterior mean curve generally produces a different quantity.

B.4 THE FEASIBILITY BOUND α

Write $\Pi_n^\xi(x) = \Pr(\xi \leq x \mid \mathcal{F}_n)$. The common EWOC proposal is the posterior α-quantile:

$$D_\alpha(\mathcal{F}_n) = (\Pi_n^\xi)^{-1}(\alpha) = \inf\{x \in [x_{\min}, x_{\max}] : \Pi_n^\xi(x) \geq \alpha\}.$$

When the posterior is continuous, $\Pi_n^{\xi}\{D_\alpha(\mathcal{F}_n)\} = \alpha$. With atoms, the generalized quantile satisfies $\Pi_n^{\xi}\{D_\alpha(\mathcal{F}_n)-\} \le \alpha \le \Pi_n^{\xi}\{D_\alpha(\mathcal{F}_n)\}$. A smaller α selects a lower posterior quantile. It changes the allocation rule while leaving the MTD definition and the response likelihood unchanged.

The feasibility condition for the next dose is

$$\Pr\{p_{f,\sigma}(x_{n+1}) > 1-\gamma \mid \mathcal{F}_n\} \le \alpha. \tag{7}$$

Right-continuous tree curves can jump at the MTD. We move inward from the quantile $q_n = D_\alpha(\mathcal{F}_n)$, using a prespecified numerical margin $\delta_{\text{num}} > 0$:

$$\widetilde{x}_{n+1} = \max\{x_{\min} + \delta_{\text{num}}, \min(x_{\max}, q_n) - \delta_{\text{num}}\}.$$

The numerical study uses $\delta_{\text{num}} = 7.5 \times 10^{-7}$ dose units. Apply an optional prespecified upward cap first, then the optional *Only Escalation* restriction:

$$\begin{aligned} x_{n+1}^{\text{cap}} &= \min\{\widetilde{x}_{n+1}, X_n + \Delta_{\max}\}, \qquad 0 \le \Delta_{\max} \le \infty, \\ X_{n+1} &= \begin{cases} \max\{X_n, x_{n+1}^{\text{cap}}\}, & \text{Only Escalation enabled,} \\ x_{n+1}^{\text{cap}}, & \text{otherwise.} \end{cases} \end{aligned} \tag{8}$$

The cap limits increases only; Only Escalation permits a hold but forbids a decrease. These options can be used separately or together. The primary five-model comparison uses $\Delta_{\max} = 3.5$ with Only Escalation enabled.

We evaluate (7) at the actual assigned dose. An inward proposal strictly below q_n satisfies the nominal bound; the upward cap alone can only lower this proposal and preserves that property. A binding lower floor can violate the bound. Only Escalation can also violate it by holding the current dose above a newly lowered proposal. We record the assigned-dose posterior exceedance probability and flag these exceptions to nominal α; they do not trigger stopping. If the lower floor binds, it is assigned unless Only Escalation requires holding the current dose. In particular, posterior mass above α at $\xi = x_{\min}$ implies that no interior dose satisfies the bound, but does not terminate this protocol. For iBART the feasibility calculation concerns the projected tail functional, not the unprojected model's predictive risk.

The first dose x_1 is prespecified in $(x_{\min}, x_{\max})$, and $0 < \delta_{\text{num}} < \min\{x_1 - x_{\min}, x_{\max} - x_1\}$. The numerical study uses $x_{\min} = 5$ and $x_1 = 6$. Initialization is included in the patient budget and remains an exception to the posterior allocation rule. Optional dose restrictions can accompany the quantile proposal (Lee et al., 2022, Section 4.2).

B.5 Cumulative DLT stopping

Set $D_n = \sum_{i=1}^{n} \mathbf{1}\{Y_i \ge \eta\}$ and $K_N = \lfloor 0.1N \rfloor$ before enrollment. After observing every response in a completed cohort and updating the posterior, stop if $D_n > K_N$ or $n = N$. Thus $N = 45$ gives $K_N = 4$, and the fifth DLT triggers the count rule. A cohort can cross the threshold by more than one DLT. Neither the posterior endpoint mass nor the numerical distance of the EWOC proposal from $x_{\min}$ is a stopping criterion. The terminal posterior includes all responses from the final cohort. Figure 4 shows the history updates used for subsequent decisions.

Algorithm 2 gives the complete cohort sequence for the two tree procedures, including the final posterior update.

The primary five-model comparison uses $N = 45$, cohorts of three, $x_1 = 6$, $\delta_{\text{num}} = 7.5 \times 10^{-7}$, and $\Delta_{\max} = 3.5$, with Only Escalation enabled. All assignments remain in the open interval. The first cohort is prespecified; subsequent exceedance probabilities are estimated from posterior draws. The terminal posterior includes the final cohort even when its DLT count triggers stopping.

C Tree priors and isotonic projection

C.1 Prior specification

Both procedures use $m = 200$ trees and shrinkage $k = 2$. At depth d, a node splits with probability $0.95(1+d)^{-2}$ in BART and $0.25(1+d)^{-0.8}$ in mBART (Chipman et al., 2010; 2022). The response

Algorithm 2 Tree-model implementation of EWOC: iBART and mBART

Require: Dose interval $(x_{\min}, x_{\max})$, initial dose x_1, prior, cuts, thresholds η, θ, allocation parameter α, cohort size c, budget N, optional upward cap $\Delta_{\max}$ and Only Escalation setting, numerical margin δ_{num}, and posterior sample size S.

Set $\mathcal{H} = \varnothing$, $n = 0$, $D_0 = 0$, $K_N = \lfloor 0.1N \rfloor$, and $x = x_1$.
while $n < N$ **do**
  Assign $c' = \min(c, N - n)$ patients to x and observe all their responses.
  Append the observations to $\mathcal{H}$; set $n \leftarrow n + c'$ and $D_n = \sum_{i=1}^{n} \mathbf{1}\{Y_i \geq \eta\}$.
  if using iBART **then**
    Draw $(g^{(s)}, \sigma^{(s)})_{s=1}^{S}$; set $f^{(s)} = \mathcal{P}g^{(s)}$ for each s.
  else
    Draw $(f^{(s)}, \sigma^{(s)})_{s=1}^{S}$ directly from the mBART posterior.
  end if
  for $s = 1, \ldots, S$ **do**
    Compute $\xi^{(s)} \in [x_{\min}, x_{\max}]$ by (6).
  end for
  Record posterior summaries and last-assigned-dose metrics.
  if $n = N$ or $D_n > K_N$ **then**
    break
  end if
  Let q_α be the lower empirical α-quantile of the MTD draws.
  Set $\widetilde{x} = \max\{x_{\min} + \delta_{\text{num}}, \min(x_{\max}, q_\alpha) - \delta_{\text{num}}\}$.
  Set $x' = \min\{\widetilde{x}, x + \Delta_{\max}\}$.
  If Only Escalation is enabled, set $x' \leftarrow \max\{x, x'\}$.
  Report $\widehat{u} = S^{-1} \sum_s \mathbf{1}\{p_{f^{(s)},\sigma^{(s)}}(x') > \theta\}$ and flag any floor or held-dose exception $\widehat{u} > \alpha$.
  Set $x \leftarrow x'$; the nominal α check does not determine stopping.
end while
Set $n_{\text{end}} = n$ and retain the final posterior.

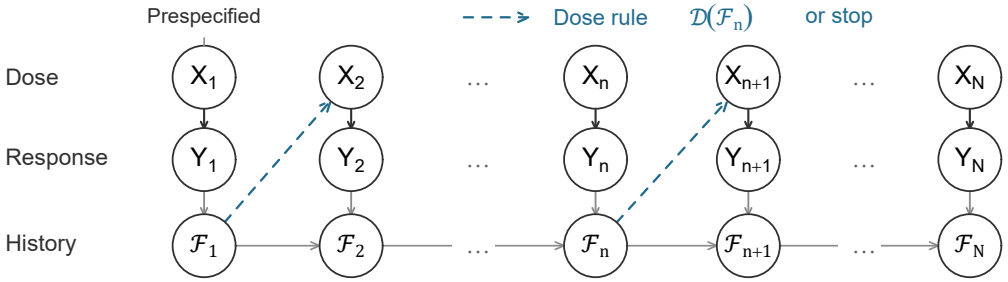


Figure 4: **Sequential learning during a dose-finding trial.** The initial dose is prespecified. After n patients, the cumulative history $\mathcal{F}_n$ contains the first n dose–response pairs and determines the updated posterior belief. The selected rule $\mathcal{D}(\mathcal{F}_n)$ chooses the next dose X_{n+1} or stops enrollment; it may use 1PLD, 2PLD, 3PND, iBART, or mBART with feasibility bound α. Black arrows denote observation, gray arrows information updates, and dashed arrows dose decisions. Ellipses omit intermediate updates; the displayed path ends at the budget N and may stop earlier. With cohorts, doses are assigned before that cohort's responses are observed, and the rule updates after the cohort is complete.

reference range $[0, 4]$ gives mean offset $b_0 = 2$ and Gaussian base leaf standard deviations $\tau_{\text{iBART}} = 1/\sqrt{200}$ and $\tau_{\text{mBART}} = \sqrt{1.467}/\sqrt{200}$. This range calibrates the prior; Gaussian responses are not restricted to it. Our implementation applies the inflated mBART base scale to all leaves, including single-leaf trees. Section 3.3 of Chipman et al. (2022) specifies inflation only for means subject to monotonicity constraints. Independently of the mean, the residual variance has prior $\sigma^2 \sim \nu_\sigma \lambda / \chi^2_{\nu_\sigma}$, with $\nu_\sigma = 3$ and $\lambda = 0.20^2 F^{-1}_{\chi^2_3}(0.10)/3$, so $\Pr(\sigma < 0.20) = 0.90$.

The offset, leaf scales, variance prior, and cut set are fixed before enrollment. Each tree node may split only at one of the 100 candidate dose thresholds $x_{\min} + j(x_{\max} - x_{\min})/101$, $j = 1, \ldots, 100$; in the simulation these are $5 + 75j/101$. Both samplers use doses and cuts in their original units. A split is eligible whenever its prespecified cut lies inside the node, including when a resulting child has no observations.

C.2 Isotonic projection of BART draws

iBART fits the unconstrained Gaussian BART likelihood. For each posterior draw, let g_j be its value on dose cell B_j. Its Lebesgue $L_2([x_{\min}, x_{\max}])$ projection is obtained by weighted isotonic regression:

$$(z_1, \ldots, z_J) = \arg \min_{z_1 \le \cdots \le z_J} \sum_{j=1}^{J} |B_j|(g_j - z_j)^2, \qquad (\mathcal{P}g)|_{B_j} = z_j. \tag{9}$$

PAVA solves this problem with physical cell lengths as weights (de Leeuw et al., 2009). The projected mean retains its paired residual-SD draw when computing risk and the MTD. The projection is a posterior functional; it does not replace g in the fitted likelihood.

C.3 Monotone tree priors

For reference, the two continuous normalizations are

$$\underbrace{\rho(\mathcal{T})K! \prod_{j=1}^{K} \phi_\tau(\mu_j)\mathbf{1}_{\text{ordered}}}_{\text{conditional ordered-leaf prior}} \quad \text{and} \quad \underbrace{\widetilde{\pi}(\mathcal{T}, \mu) \propto \rho(\mathcal{T}) \prod_{j=1}^{K} \phi_\tau(\mu_j)\mathbf{1}_{\text{ordered}}}_{\text{constrained-product prior}}. \tag{10}$$

mBART orders the leaves of each univariate tree, making their sum nondecreasing. For a tree with probability $\rho(\mathcal{T})$ and K independent and identically distributed (iid) continuous Gaussian base heights, conditioning on their order multiplies the product density by $K!$ and retains tree marginal $\rho(\mathcal{T})$. The constrained-product formulation is instead

$$\widetilde{\pi}(\mathcal{T}, \mu) \propto \rho(\mathcal{T}) \prod_{j=1}^{K} \phi_\tau(\mu_j)\mathbf{1}\{\mu_1 \le \cdots \le \mu_K\}. \tag{11}$$

Global normalization gives tree marginal proportional to $\rho(\mathcal{T})/K!$. The computation uses this formulation, approximating leaf integration with 50 points in $[-3\tau, 3\tau]$. Discrete ties change the ordering probability. The implemented grid is a numerical approximation; its posterior sampling is described in Appendix G.

D Parametric dose finders

All three parametric procedures observe the measured outcome and use the same dose interval and feasibility bound as the tree procedures. Write $q_\gamma = \Phi^{-1}(\gamma)$ and $L_x = x_{\max} - x_{\min}$.

1PLD. The Bayesian search using a continuously measured toxicity outcome from Eichhorn & Zacks (1973), reproduced as the one-parameter linear dose finder in Lee et al. (2022, Eq. (2)), uses

$$Y_i \mid X_i, \beta \sim \mathcal{N}\{\beta(X_i - x_{\min}), s^2\}, \qquad \beta \sim \mathcal{N}(a, b^2),$$

with known s. The Gaussian slope prior is untruncated. For $(x_{\min}, x_{\max}) = (5, 80)$, we set $a = 0.04$, $b = 0.02$, and $s = 0.1$ before simulation. These are study calibrations.

Let $A = \eta - s\Phi^{-1}(\gamma) > 0$ and $L_x = x_{\max} - x_{\min}$. The MTD is

$$\xi(\beta) = \begin{cases} x_{\max}, & \beta \le A/L_x, \\ x_{\min} + A/\beta, & \beta > A/L_x. \end{cases} \tag{12}$$

Negative slopes contribute to the mass at $x_{\max}$. The risk near $x_{\min}$ is below the target for every slope, so this calibration has no posterior mass at the lower endpoint.

Posterior calculation. Write $z_i = X_i - x_{\min}$, $P_n = b^{-2} + s^{-2}\sum_i z_i^2$, and $S_n = a/b^2 + s^{-2}\sum_i z_i Y_i$. Completing the square gives

$$\beta \mid \mathcal{F}_n \sim \mathcal{N}(\mu_n, V_n), \qquad \mu_n = S_n/P_n, \quad V_n = P_n^{-1}. \tag{13}$$

For $x \in (x_{\min}, x_{\max})$,

$$\Pr\{p_\beta(x) > \theta \mid \mathcal{F}_n\} = 1 - \Phi\left\{\frac{A/(x - x_{\min}) - \mu_n}{\sqrt{V_n}}\right\},$$
$$\Pr(\xi = x_{\max} \mid \mathcal{F}_n) = \Phi\left\{\frac{A/L_x - \mu_n}{\sqrt{V_n}}\right\}.$$

For $0 < u < 1$, put $B_u = \mu_n + \sqrt{V_n}\Phi^{-1}(1-u)$. The lower posterior u-quantile is $x_{\max}$ if $B_u \le A/L_x$ and $x_{\min} + A/B_u$ otherwise. The allocation uses $u = \alpha$ and the common dose restrictions.

The known noise SD equals the simulation value 0.1. This supplies 1PLD with information that the remaining methods estimate.

D.1 2PLD AND 3PND

The likelihoods of Lee et al. (2022), Eq. (3), and Lee (2023), Eq. (4), are

$$Y_i \mid X_i, \beta, \nu, \sigma \sim \mathcal{N}\{\beta(X_i - x_{\min})^\nu, \sigma^2\}.$$

2PLD fixes $\nu = 1$; 3PND estimates $\nu > 0$. Both fix the mean at zero at $x_{\min}$. Their toxicity risk and MTD are

$$p_{\beta,\nu,\sigma}(x) = \Phi\left\{\frac{\beta(x - x_{\min})^\nu - \eta}{\sigma}\right\}, \qquad \xi = x_{\min} + \left\{\frac{\eta - \sigma q_\gamma}{\beta}\right\}^{1/\nu}. \tag{14}$$

The priors in Lee et al. (2022), Eqs. (4)–(5), and Lee (2023), Eqs. (7)–(9), are

$$\pi_\sigma(s) = \frac{\mathbf{1}\{0 < s < \eta/q_\gamma\}}{\arctan(\eta/q_\gamma)(1+s^2)}, \quad \beta \mid \sigma, \nu \sim \mathrm{Unif}\{l(\sigma,\nu), u(\sigma,\nu)\},$$
$$l(\sigma,\nu) = \frac{\eta - \sigma q_\gamma}{L_x^\nu}, \qquad u(\sigma,\nu) = \frac{\eta}{L_x^\nu} + \sigma q_\gamma. \tag{15}$$

For 3PND, $\log\nu \sim \mathcal{N}(0, \delta^2)$ independently of σ, with $\delta = 0.1$ in the simulation. The scale-one half-Cauchy prior is truncated on the standard deviation. The uniform density includes the normalizer

$$\{u - l\}^{-1} = \{\sigma q_\gamma(1 + L_x^{-\nu})\}^{-1}.$$

In particular, the σq_γ term in the upper endpoint is not divided by L_x^ν.

Independent $U, V \sim \mathrm{Unif}(0,1)$ and $Z \sim \mathcal{N}(0,1)$ give the exact prior coordinates

$$\sigma = \tan\{V \arctan(\eta/q_\gamma)\}, \qquad \nu = \exp(\delta Z), \qquad \beta = l(\sigma,\nu) + U\{u(\sigma,\nu) - l(\sigma,\nu)\}. \tag{16}$$

For 2PLD, omit Z and set $\nu = 1$. These constraints imply $x_{\min} < \xi < x_{\max}$ almost surely. The priors exclude both endpoint states, whereas tree priors permit them.

D.2 ALLOCATION AND EVALUATION

The posterior dose rule is $D_\alpha(\mathcal{F}_n) = F^{-1}_{\xi\mid\mathcal{F}_n}(\alpha)$. For the increasing parametric curves, the event that dose x is excessive equals $\{\xi < x\}$. Under the 1PLD convention in Eq. (12), the same equivalence holds for its allocated positive doses; negative slopes contribute to the all-acceptable endpoint. The tree procedures assess the toxicity-risk event directly to account for step boundaries. Posterior feasibility is a statement under each method's own posterior and is not a common fixed-truth frequentist bound.

In Eq. (20), BTM and RAE use the last assigned dose; NPD and NPO use individual patient records. Posterior summaries incorporate all observed responses, including the final cohort, and are distinct from the last-dose estimator.

E DOSE–TOXICITY CURVES AND GRAPHICAL MODELS

Figure 5 shows what each method learns from the measured outcome and how those unknowns determine the MTD. Shaded circles denote observed responses and assigned doses; unshaded circles denote unknown quantities. Rectangles contain fixed inputs, and diamonds are deterministic functions of their parents. Solid arrows specify stochastic dependencies; dashed arrows lead to deterministic functions. The patient plate repeats the response factor for $i = 1, \ldots, n$.

The clinical inputs $C = (x_{\min}, x_{\max}, \eta, \gamma)$ are fixed throughout the trial. To keep the diagrams readable, arrows from C are suppressed, including its contribution to the priors and the MTD. The dose nodes show the values used in the sequential likelihood $p(Y_i \mid X_i, \mathcal{F}_{i-1}, \vartheta)$, where ϑ denotes a method's unknown parameters. The graphs omit the allocation history and do not assert conditional independence after conditioning on the entire random sequence of adaptive doses. The allocation kernels cancel from the posterior for the reason given in Section B. These are probability-model diagrams, not causal diagrams.

Parametric procedures. For 1PLD, the only unknown is the slope β; the residual standard deviation s is known. Its prior is the untruncated $\mathcal{N}(a, b^2)$ distribution. For 2PLD, σ is unknown and the support of the slope prior depends on it. For 3PND, the exponent ν also changes the slope-prior support. Conditional on the fixed inputs, the respective prior factorizations are

$$\begin{aligned}
\pi_{\mathrm{1PLD}}(\beta) &= \phi_b(\beta - a), \\
\pi_{\mathrm{2PLD}}(\beta, \sigma) &= \pi_\sigma(\sigma)\, \pi_\beta(\beta \mid \sigma, \nu = 1), \\
\pi_{\mathrm{3PND}}(\beta, \sigma, \nu) &= \pi_\sigma(\sigma)\, \pi_\nu(\nu \mid \delta)\, \pi_\beta(\beta \mid \sigma, \nu).
\end{aligned}$$

Here π_σ and π_β are exactly the truncated half-Cauchy and conditional uniform densities in Eq. (15), and $\log \nu \sim \mathcal{N}(0, \delta^2)$. Thus σ and ν are independent before observing data in 3PND, but β is independent of neither. The functions represented by the f nodes are $\beta(x - x_{\min})$ for 1PLD and 2PLD, and $\beta(x - x_{\min})^\nu$ for 3PND. All three use a Gaussian response likelihood. The MTD node uses Eq. (12) for 1PLD, including its upper-endpoint convention for small or negative slopes, and Eq. (14) for 2PLD and 3PND. Figure 6(a)–(c) illustrates their mean-curve families.

Tree procedures. In the lower row of Figure 5, T denotes the collection of tree structures and μ their leaf heights. The fixed inputs H_T specify the tree count, depth probabilities, cut set, and leaf scale; H_σ specifies the residual-variance prior. The fixed mean offset is also conditioned on. These settings are given in Section C. The forest and residual variance are independent under the prior; observing the responses generally makes them dependent in the posterior.

For iBART, the forest produces the unconstrained mean g. This g enters the Gaussian likelihood. The separate deterministic branch applies the weighted isotonic projection $f = \mathcal{P}g$ and then computes $\xi(f, \sigma)$. There is no arrow from f to Y_i: projection changes the posterior quantity used for dose selection, not the likelihood that produced the draw. The corresponding σ draw is retained through both calculations.

For mBART, the forest node represents the *joint* constrained-product prior for structures and leaf heights in Eq. (11). Each tree is nondecreasing, so the sum f enters both the likelihood and the MTD calculation. Grouping structures and heights in one node is deliberate: using the unconstrained topology marginal followed by a normalized ordered-leaf conditional would instead give the different prior in Eq. (10). The numerical implementation approximates the constrained-product formulation on the stated leaf grid; the diagram does not identify that finite approximation with the continuous prior. Figure 6(d,e) illustrates projection and monotone summation.

The common EWOC decision. The graphs end at the MTD because α is a decision parameter, not a response-model parameter. After conditioning on the available patient data, each method computes an MTD posterior and uses its lower α-quantile to propose the next dose. The tree procedures also check the posterior risk event at the proposed dose, as in Eq. (7), to handle step boundaries. The initial dose, cohort size, dose restrictions, and stopping rules belong to this allocation stage. Placing that stage after the posterior update keeps the diagrams acyclic while preserving the shared EWOC structure.

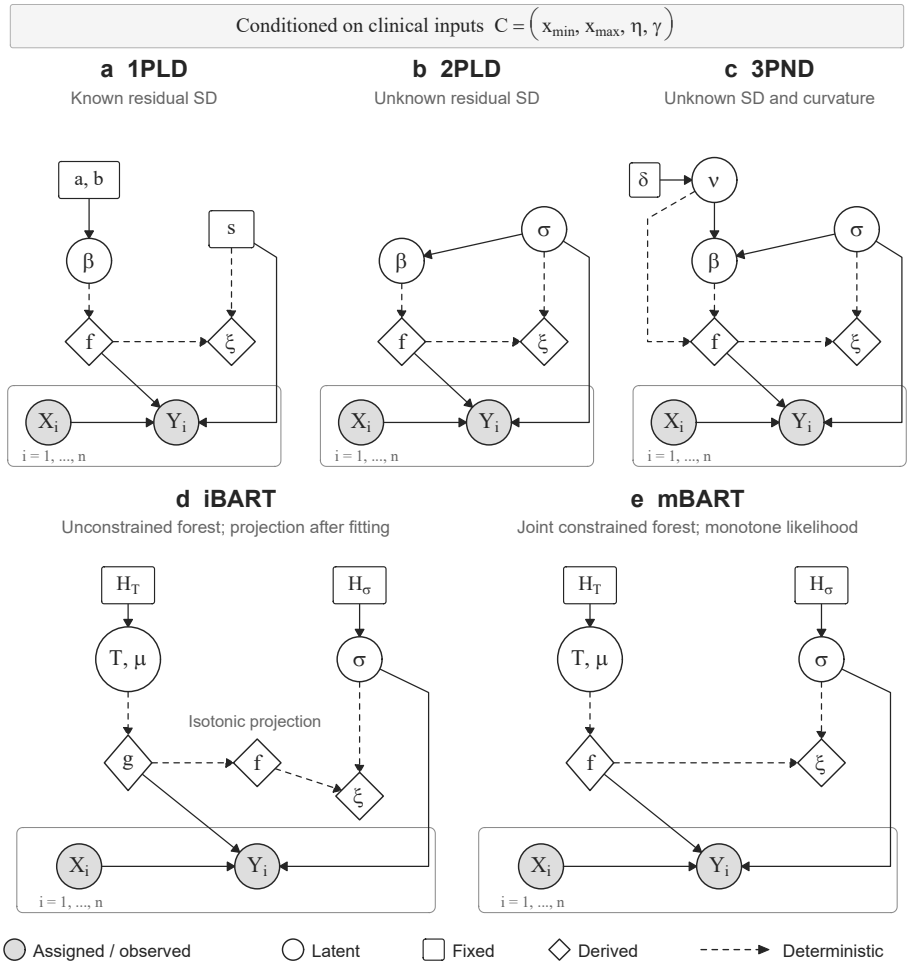


Figure 5: **Graphical models of the five procedures.** Top: 1PLD, 2PLD, and 3PND; bottom: iBART and mBART. X_i and Y_i are patient i's assigned dose and response. β is the mean-curve coefficient; a, b are its 1PLD prior mean and SD; s is the known residual SD and σ an unknown residual SD. The 3PND exponent is ν, with prior SD δ for $\log \nu$. T collects tree structures and μ their leaf heights. H_T fixes the tree count, depth probabilities, cut set, and leaf scale; H_σ fixes the residual-variance prior's degrees of freedom and scale. g is the unconstrained forest mean, f the curve defining the MTD ξ. The fixed clinical inputs $C = (x_{\min}, x_{\max}, \eta, \gamma)$ are the dose bounds, DLT threshold, and required non-DLT probability. The arrows $\sigma \to \beta$ and $\nu \to \beta$ encode conditional priors. In iBART, g enters the likelihood while its isotonic projection $f = \mathcal{P}g$ determines the MTD; in mBART, the joint constrained forest gives the monotone mean f used by both. Every MTD retains its paired residual SD. Allocation history and arrows from C are omitted.

F MTD POSTERIOR CONSISTENCY

For the fixed-cell model below, we establish posterior concentration at the true MTD under prior support, persistent cell sampling, and a strict risk-threshold margin. The proof uses a Schwartz-type numerator–denominator argument (Schwartz, 1965). The margin makes the MTD locally unchanged, allowing likelihood comparisons to exclude parameters giving a different MTD.

Fix positive-length dose cells $B_1, \ldots, B_J$, with boundaries $c_0 = x_{\min} < c_1 < \cdots < c_J = x_{\max}$ and a fixed convention at each boundary. Write $\vartheta = (u, v)$ for the cell means $u = (u_1, \ldots, u_J)$ and residual variance $v > 0$. The parameter space Θ is $\mathbb{R}^J \times (0, \infty)$ for the unconstrained iBART likelihood, or its subset with nondecreasing means for mBART. The prior may have smaller support. Throughout, $\|\cdot\|$ is the Euclidean norm on these identifiable parameters.

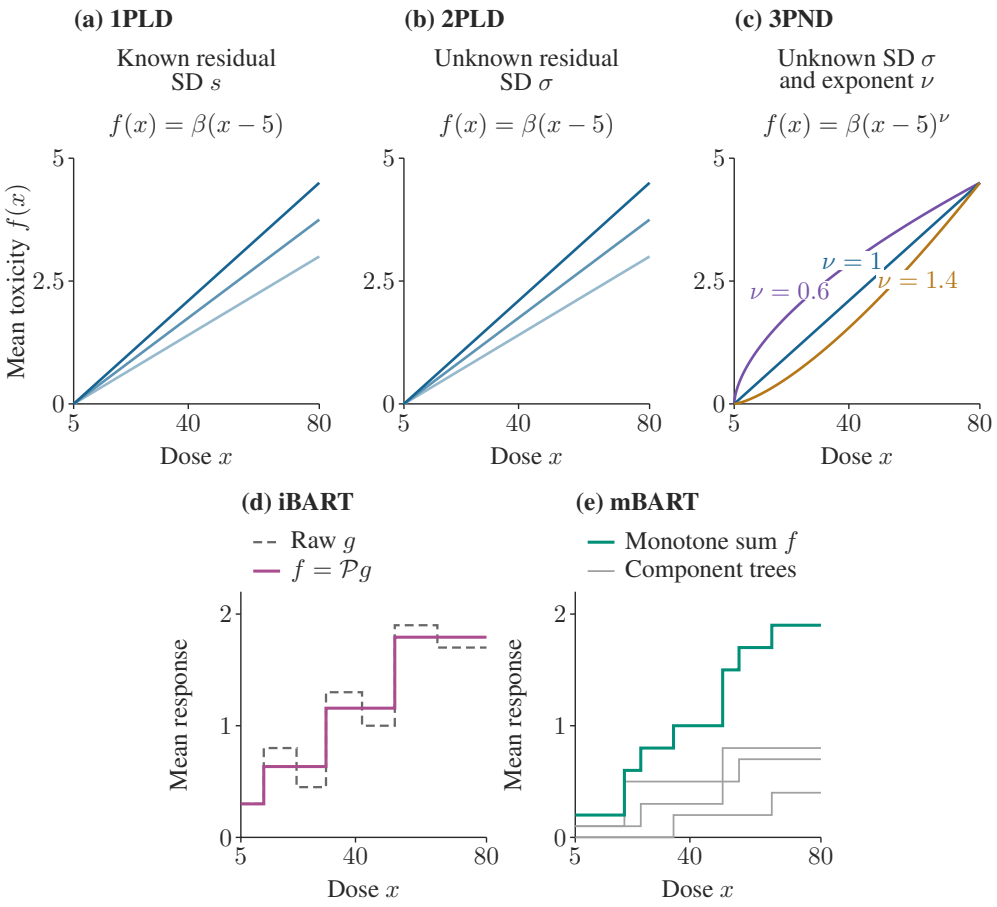


Figure 6: **Illustrative mean dose–toxicity curves for the five procedures.** (a,b) 1PLD and 2PLD share a linear mean family, differing in whether residual SD is known or estimated; the examples use $\beta \in \{0.04, 0.05, 0.06\}$. (c) 3PND also estimates the exponent ν; the examples use $\nu \in \{0.6, 1, 1.4\}$ and $\beta = 4.5/75^{\nu}$, giving a common mean at the upper dose. Only positive slopes are illustrated; the 1PLD prior also permits negative slopes. (d) iBART applies cell-length-weighted isotonic projection to the unconstrained step curve g (dashed gray), giving $f = \mathcal{P}g$ (purple). The likelihood retains g, while f and its paired residual SD determine the MTD. (e) Three nondecreasing component trees (thin gray) sum to a monotone curve f (green), used in both the likelihood and MTD calculation. These schematic curves are not fitted estimates.

For MTD inference, let $f(u) = u$ for mBART and $f(u) = \mathcal{P}u$ for iBART, where $\mathcal{P}$ is the fixed cell-length-weighted isotonic projection. The cell risks and MTD functional are

$$r_j(\vartheta) = \Phi\{(f_j(u) - \eta)/\sqrt{v}\},$$
$$\xi(\vartheta) = \sup\big(\{x_{\min}\} \cup \{x \in \mathcal{X} : r_{j(x)}(\vartheta) \le \theta\}\big),$$

where $j(x)$ identifies the dose cell and $\theta = 1 - \gamma$. The MTD takes values among the boundaries $\{c_0, \ldots, c_J\}$, with the endpoint conventions of Definition 1. Write Π_n^{ξ} for its induced posterior distribution.

Fix a true parameter $\vartheta_0 = (u_0, v_0) \in \Theta$ and its MTD $\xi_0 = \xi(\vartheta_0)$. For iBART, assume additionally that u_0 is nondecreasing, so $\mathcal{P}u_0 = u_0$ and ξ_0 is the MTD of the data-generating response law. Suppose an indefinitely continued, predictable allocation satisfies

$$Y_i \mid \mathcal{F}_{i-1}, X_i \sim \mathcal{N}(u_{0j}, v_0) \quad \text{when } X_i \in B_j,$$
$$\liminf_{n\to\infty} \frac{N_{j,n}}{n} \ge \kappa > 0 \qquad (j = 1, \ldots, J)$$

almost surely under P_{ϑ_0}, where $N_{j,n} = \sum_{i=1}^{n} \mathbf{1}\{X_i \in B_j\}$ and κ is fixed. The allocation rule is the same function of observed history under all parameters. Let Π be a fixed proper prior with $\Pi\{\|\vartheta - \vartheta_0\| < \delta\} > 0$ for every $\delta > 0$. Assume also that the true projected or monotone cell risks have a strict margin from the clinical target:

$$\min_{1 \le j \le J} |r_j(\vartheta_0) - \theta| > 0.$$

Theorem 3 (Posterior consistency of the MTD)**.** *Under the model and assumptions stated above, the posterior distribution of the MTD ξ is consistent at its true value ξ_0: for every $\epsilon > 0$,*

$$\Pi_n^{\xi}\{z : |z - \xi_0| \ge \epsilon\} \longrightarrow 0 \qquad P_{\vartheta_0}\text{-almost surely}.$$

Proof. The weighted isotonic projection and Gaussian tail are continuous. There are finitely many cells, and none of their true risks equals θ. Hence there is a radius $r > 0$ such that $\|\vartheta - \vartheta_0\| < r$ preserves every acceptable/unacceptable cell classification and therefore gives $\xi(\vartheta) = \xi_0$. Define the set of parameters giving the wrong MTD by $A = \{\vartheta \in \Theta : \xi(\vartheta) \ne \xi_0\}$. Every parameter in A is outside this radius-r neighborhood. It suffices to show that its posterior probability tends to zero. We compare the two integrals in

$$\Pi_n(A) = \frac{\int_A L_n(\vartheta)\,\Pi(d\vartheta)}{\int_\Theta L_n(\vartheta)\,\Pi(d\vartheta)}. \tag{17}$$

The dose assigned to patient i is determined by the preceding history. Its allocation factor therefore has no unknown-parameter term and cancels from the posterior ratio. Multiplying the conditional Gaussian response densities gives

$$L_n(u, v) = (2\pi v)^{-n/2} \exp\left\{ -\frac{1}{2v} \sum_{j=1}^{J} \sum_{i:X_i \in B_j} (Y_i - u_j)^2 \right\}.$$

This factorization uses the sequential conditional model; it does not require the adaptive observations to be iid.

Every cell is visited infinitely often by the sampling assumption. For all sufficiently large n, define the cell proportions, cell means, and pooled residual variance by

$$p_{j,n} = \frac{N_{j,n}}{n}, \qquad \bar{Y}_{j,n} = \frac{1}{N_{j,n}} \sum_{i:X_i \in B_j} Y_i, \qquad \widehat{v}_n = \frac{1}{n} \sum_{j} \sum_{i:X_i \in B_j} (Y_i - \bar{Y}_{j,n})^2.$$

Let $\varepsilon_i = Y_i - u_{0j}$ when $X_i \in B_j$. Predictability makes $\mathbf{1}\{X_i \in B_j\}\varepsilon_i$ and $\varepsilon_i^2 - v_0$ martingale differences, with conditional variances at most v_0 and $2v_0^2$, respectively. Dividing these increments by i gives summable variances. The martingale convergence theorem and Kronecker's lemma therefore give

$$\frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{X_i \in B_j\}\varepsilon_i \to 0, \qquad \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i^2 \to v_0 \quad \text{almost surely}.$$

Since $p_{j,n} \ge \kappa/2$ eventually, the first limit implies $\bar{Y}_{j,n} \to u_{0j}$. Completing the square within each cell also shows

$$\widehat{v}_n = \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i^2 - \sum_{j} p_{j,n}(\bar{Y}_{j,n} - u_{0j})^2 \longrightarrow v_0 > 0.$$

For the rest of the proof, work on the probability-one event where these limits and the allocation assumption hold. No convergence of the proportions themselves is needed.

Let $\widehat{L}_n = L_n(\bar{Y}_n, \widehat{v}_n)$, using the unrestricted cell estimates even when $\bar{Y}_n$ is not nondecreasing. Completing the square in the likelihood gives

$$\frac{L_n(u, v)}{\widehat{L}_n} = \exp\{-nH_n(u, v)/2\},$$
$$H_n(u, v) = h\left(\frac{v}{\widehat{v}_n}\right) + \frac{\sum_j p_{j,n}(u_j - \bar{Y}_{j,n})^2}{v}, \qquad h(t) = \log t + t^{-1} - 1.$$

The function h is nonnegative, vanishes only at $t = 1$, and tends to infinity as t tends to zero or infinity. Thus $H_n \geq 0$ and, eventually, $0 < L_n(u, v) \leq \widehat{L}_n < \infty$ for every parameter. A proper prior then gives a finite, positive posterior denominator. The common factor $\widehat{L}_n$ can now be canceled from (17).

A small value of H_n forces both the variance and the cell means to be close to their true values. To see this, suppose $H_n(u, v) \leq c$. Then $h(v/\widehat{v}_n) \leq c$. The properties of h imply that $v/\widehat{v}_n$ is as close to one as desired when c is small. Because $\widehat{v}_n \to v_0$, we can choose $c > 0$ small enough that, eventually, this entails $v \leq 2v_0$ and $|v - v_0| < r/2$, uniformly over all such parameters.

The remaining term of H_n, together with $p_{j,n} \geq \kappa/2$, gives the explicit bound

$$\|u - \bar{Y}_n\|^2 \leq \frac{2v}{\kappa} H_n(u, v) \leq \frac{4v_0 c}{\kappa}.$$

Choose c smaller if necessary so the last expression is at most $r^2/16$. Eventually $\|\bar{Y}_n - u_0\| < r/4$, so the triangle inequality yields $\|u - u_0\| < r/2$. Together with the variance bound, this implies $\|\vartheta - \vartheta_0\| < r$. Taking the contrapositive, there is a fixed $c_r > 0$ such that $H_n(\vartheta) \geq c_r$ for every $\vartheta \in A$ and all sufficiently large n. Hence the normalized numerator satisfies

$$\int_A e^{-nH_n(\vartheta)/2}\, \Pi(d\vartheta) \leq e^{-nc_r/2}.$$

The bound applies to every parameter giving the wrong MTD, including arbitrarily large means and variances.

Choose a small fixed ball $B = \{\vartheta \in \Theta : \|\vartheta - \vartheta_0\| < \delta\}$. For small enough δ, and eventually in n, all its variances are bounded below by $v_0/2$ and its cell means are close to $\bar{Y}_n$. More explicitly, since the proportions sum to one,

$$\frac{\sum_j p_{j,n}(u_j - \bar{Y}_{j,n})^2}{v} \leq \frac{2}{v_0}\|u - \bar{Y}_n\|^2 \quad (\vartheta \in B).$$

This bound can be made uniformly smaller than $c_r/4$ by choosing δ small and using $\bar{Y}_n \to u_0$. Continuity of h at one and $\widehat{v}_n \to v_0$ similarly make the variance term smaller than $c_r/4$. Thus $\sup_{\vartheta \in B} H_n(\vartheta) \leq c_r/2$ eventually. The support assumption gives $\Pi(B) > 0$, so

$$\int_\Theta e^{-nH_n(\vartheta)/2}\, \Pi(d\vartheta) \geq \Pi(B) e^{-nc_r/4}.$$

Combining the numerator and denominator bounds in the posterior ratio gives

$$\Pi_n(A) \leq \Pi(B)^{-1} e^{-nc_r/4} \longrightarrow 0.$$

For every $\epsilon > 0$, the event $|\xi - \xi_0| \geq \epsilon$ is contained in the event $\xi \neq \xi_0$. Thus

$$\Pi_n^{\xi}\{z : |z - \xi_0| \geq \epsilon\} \leq \Pi_n(A) \longrightarrow 0.$$

This holds on the probability-one event established above and proves the stated MTD consistency.□

G Posterior computation

The five procedures use different posterior calculations. The 1PLD posterior is analytic. For 2PLD and 3PND, we use coordinate slice sampling in transformed prior coordinates. For the tree models, we use Bayesian backfitting: Metropolis–Hastings (MH) updates of tree structure, followed by conditional draws of leaf values and residual variance. Thus, *Gibbs sampling* alone would not specify the tree algorithm. This appendix describes posterior computation for the $N = 45$ study.

G.1 Parametric procedures

1PLD: analytic updating. Let $d_i = X_i - x_{\min}$. With the untruncated prior $\beta \sim \mathcal{N}(a, b^2)$ and known residual SD s, conjugacy gives $\beta \mid \mathcal{F}_n \sim \mathcal{N}(\mu_n, V_n)$, where

$$V_n = \left(b^{-2} + s^{-2}\sum_{i=1}^{n} d_i^2\right)^{-1}, \qquad \mu_n = V_n\left(ab^{-2} + s^{-2}\sum_{i=1}^{n} d_i Y_i\right).$$

We use $(a, b, s) = (0.04, 0.02, 0.1)$. For $A = \eta - sq_\gamma > 0$, where $q_\gamma = \Phi^{-1}(\gamma)$, the MTD is $x_{\max}$ if $\beta \le A/L_x$ and $x_{\min} + A/\beta$ otherwise, with $L_x = x_{\max} - x_{\min}$. MTD quantiles, upper-endpoint mass and assigned-dose overdose probabilities follow analytically, including the negative-slope probability; see Eqs. (12)–(13). No MCMC, burn-in or thinning is needed.

2PLD and 3PND: target in prior coordinates. The likelihood and normalized hierarchy are given in Appendix D. The coefficient bounds in Eq. (15) have width $u(\sigma, \nu) - l(\sigma, \nu) = \sigma q_\gamma(1 + L_x^{-\nu})$, which depends on both σ and ν. To retain this normalizer and the parameter-dependent support, put $a_\sigma = \arctan(\eta/q_\gamma)$ and use Eq. (16):

$$U, V \sim \mathrm{Unif}(0,1), \quad Z \sim \mathcal{N}(0,1) \quad \text{independently},$$
$$\sigma = \tan(a_\sigma V), \qquad \nu = \exp(\delta Z), \qquad \beta = l(\sigma,\nu) + U\{u(\sigma,\nu) - l(\sigma,\nu)\}.$$

For 3PND, $\delta = 0.1$; for 2PLD, fix $\nu = 1$ and omit Z. The sampler updates $z_1 = \mathrm{logit}(U)$, $z_2 = \mathrm{logit}(V)$ and, for 3PND, $z_3 = Z$. Including the transformation Jacobian gives the log target, up to an additive constant,

$$\ell_n(z) = -n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \{Y_i - \beta d_i^\nu\}^2 + \log\{U(1-U)V(1-V)\} - \frac{Z^2}{2}.$$

The last term is omitted for 2PLD. The physical prior normalizers cancel with the corresponding change-of-variables factors; no additional physical-coordinate Jacobian is applied.

Slice transitions and physical parameters. Each sweep visits z_1, z_2, z_3 in that order, omitting the third for 2PLD. For one coordinate, draw $\log h = \ell_n(z) + \log U_0$, $U_0 \sim \mathrm{Unif}(0,1)$. Initialize a randomly positioned interval of width 1.5 around its current value, then step out while its endpoints lie above this height. A random integer J_0 uniform on $\{0, \ldots, 39\}$ assigns J_0 additional outward increments to the left and $39 - J_0$ to the right. Uniform proposals from the resulting interval are accepted when their log target exceeds $\log h$; otherwise shrink the interval toward the current point and propose again (Neal, 2003). The width is fixed, and failure to accept within 1,000 shrinkage proposals is reported as an error.

These are slice-within-Gibbs moves in the *transformed* coordinates. Updating z_2 changes both σ and β; updating z_3 changes both ν and β. They do not reproduce the physical-parameter transitions in Algorithms 3–5 of Lee (2023), whose Gibbs cycle itself includes slice moves. Our specification concerns the normalized hierarchy stated here. For example, a physical update of ν holding (β, σ) fixed would retain

$$\pi(\nu \mid \beta, \sigma, \mathcal{F}_n) \;\propto\; \exp\left[-\frac{\sum_i \{Y_i - \beta d_i^\nu\}^2}{2\sigma^2}\right] \frac{\mathrm{logN}(\nu; 0, \delta^2)}{1 + L_x^{-\nu}} \, \mathbf{1}\{l(\sigma,\nu) < \beta < u(\sigma,\nu)\}.$$

The displayed normalizer and support are not constants in ν; the prior-coordinate target retains both automatically. The change-of-variables identity does not assert identical finite chains or adaptive dose paths for different samplers. Each retained draw gives an MTD via Eq. (14).

G.2 Tree backfitting and conditional updates

One ensemble sweep. The likelihood mean is $b_0 + \sum_{j=1}^{m} h_j(x)$, with $m = 200$ and fixed $b_0 = 2$. For tree j, remove its old contribution from the current ensemble and form the partial residuals

$$R_{ij} = Y_i - b_0 - \sum_{k \neq j} h_k(X_i), \qquad i = 1, \ldots, n.$$

Update its structure against these residuals, refresh its leaf values, and restore its contribution. After visiting all trees, update σ^2. The priors, response calibration and physical-dose cuts in Appendix C remain fixed as data accrue.

Both tree samplers use birth and death proposals, without separate change-rule or swap moves. Birth selects an eligible leaf and an available split threshold uniformly; death selects an internal node with two leaf children uniformly. With one predictor, the split variable is dose. Birth probability is one for a splittable stump, zero if no leaf can split, and one half otherwise. For a proposed birth $T \to T'$, let

p_d be the parent's prior split probability, p_L, p_R the children's split probabilities, $B(T)$ the eligible leaves and $D(T')$ the removable sibling pairs. The topology-prior and proposal factor is

$$A_T = \frac{p_d(1-p_L)(1-p_R)}{1-p_d}\,\frac{P_D(T')/|D(T')|}{P_B(T)/|B(T)|}.$$

The uniform threshold prior and proposal factors cancel. If I_{parent} and I_{children} are the corresponding leaf-integrated likelihood-times-prior masses, accept with probability $\min\{1, A_T I_{\text{children}}/I_{\text{parent}}\}$. Death uses the reverse factor. For mBART, these masses also retain the order constraints imposed by unchanged leaves.

iBART: Gaussian leaves and posterior projection. For iBART, the Gaussian leaf update uses unit observation weights, giving the usual homoskedastic sufficient statistics. For leaf ℓ, let n_ℓ be its count and $S_\ell = \sum_{i\in\ell} R_{ij}$. With Gaussian leaf prior variance τ^2, the conditional update is

$$V_\ell = (\tau^{-2} + n_\ell/\sigma^2)^{-1}, \qquad \mu_\ell \mid \cdots \sim \mathcal{N}(V_\ell S_\ell/\sigma^2, V_\ell).$$

Put $b_\ell = n_\ell/\sigma^2$ and $M_\ell = S_\ell/\sigma^2$. Dropping likelihood factors common to the compared partitions, the integrated log factor for a topology move is

$$H_\ell = -\tfrac{1}{2}\log(1+b_\ell\tau^2) + \frac{M_\ell^2\tau^2}{2(1+b_\ell\tau^2)}.$$

Thus the birth likelihood ratio is $\exp(H_L + H_R - H_{\text{parent}})$. Accepted new leaves receive conditional Gaussian draws, followed by a refresh of all leaves. Empty children are allowed under the fixed cut prior: for $n_\ell = S_\ell = 0$, the integrated factor is one and the draw is exactly $\mathcal{N}(0,\tau^2)$.

After sampling the unconstrained likelihood mean g, evaluate each draw on every cell of the common dose-cut grid and apply weighted PAVA, Eq. (9), to obtain $f = \mathcal{P}g$ using physical cell lengths as weights (de Leeuw et al., 2009). Retain the original paired σ draw. Projection neither replaces g in the likelihood nor refits variance from projected residuals.

Residual variance in both tree models. With residual sum of squares $\text{SSE} = \sum_i\{Y_i - b_0 - \sum_j h_j(X_i)\}^2$, draw

$$\sigma^2 \mid \cdots \sim \text{IG}\left(\frac{n+\nu_\sigma}{2}, \frac{\nu_\sigma\lambda + \text{SSE}}{2}\right), \qquad \sigma^2 = \frac{\nu_\sigma\lambda + \text{SSE}}{U_\sigma}, \quad U_\sigma \sim \chi^2_{n+\nu_\sigma}.$$

Here $\text{IG}(a,b)$ is the inverse-gamma distribution with density proportional to $v^{-a-1}\exp(-b/v)$, $\nu_\sigma = 3$, and λ is fixed as in Appendix C. The SSE uses g for iBART and the constrained likelihood mean f for mBART.

G.3 mBART: ORDERED FINITE-GRID LEAF UPDATES

The numerical target uses $Q = 50$ fixed leaf values

$$u_q = -3\tau + \frac{6\tau q}{Q+1}, \qquad p_q = \frac{\phi(u_q/\tau)}{\sum_{r=1}^{Q}\phi(u_r/\tau)}, \qquad q = 1,\ldots,Q,$$

with $\tau = \sqrt{1.467}/\sqrt{200}$. These masses are proportional to Gaussian density values, not Gaussian-bin probabilities; the endpoints $\pm 3\tau$ are excluded. Other leaves determine the allowable interval $[L_\ell, U_\ell]$: the current leaf lies above all lower-dose leaf values and below all higher-dose leaf values, with ties allowed. Its finite-grid conditional masses are

$$w_{\ell q} = p_q \exp\left\{\frac{u_q S_\ell - \frac{1}{2}n_\ell u_q^2}{\sigma^2}\right\}\mathbf{1}\{L_\ell \le u_q \le U_\ell\}.$$

Normalize these masses and draw categorically, visiting leaves sequentially and recomputing the constraints after each draw. This is a Gibbs update for the finite-grid model, not a continuous truncated-normal draw. An empty leaf has constant likelihood and uses the restricted prior masses.

For a birth, the single-leaf mass and ordered pair mass are

$$I_1 = \sum_q w_{\text{parent},q}, \qquad I_2 = \sum_{q \le r} w_{Lq} w_{Rr},$$

where each child's weights include constraints from unchanged leaves. The MH ratio uses I_2/I_1. An accepted pair is drawn jointly from the normalized pair weights; death uses I_1/I_2 and draws the merged leaf from its normalized weights. The restricted prior masses remain in I_1, I_2: there is no division by a separate order-cone probability for each tree. Normalizing conditional draw probabilities is a different operation. This implements Eq. (11)'s constrained-product formulation on the finite grid.

The pair calculation is linear in Q. For numerically scaled $a_q = w_{Lq}$ and $b_r = w_{Rr}$, the suffix recurrence is

$$B_{Q+1} = 0, \qquad B_q = b_q + B_{q+1}, \qquad I_2 = \sum_{q=1}^{Q} a_q B_q.$$

Choose q with probability proportional to $a_q B_q$, then $r \ge q$ with probability proportional to b_r. The implementation uses the residual of the same uniform variate to preserve the lexicographic pair distribution. Scaled likelihood weights and a log-sum-exp fallback avoid underflow, with scale factors restored in the MH ratio. For equally spaced leaf points, successive quadratic-likelihood ratios change by $\exp(-n_\ell \Delta_u^2/\sigma^2)$, where Δ_u is the spacing. These recurrences preserve the fixed numerical target; they do not refine its support.

G.4 Refitting, retention and inversion

At every completed cohort, including the final one, each method uses all accrued observations. The sampled methods start two fresh, separately seeded chains; no preceding chain is used as a warm start and no retained sweeps are thinned (Table 3). For 2PLD and 3PND, each chain starts from the stated prior. iBART starts from zero-valued stump contributions with offset 2 and residual SD 0.2. mBART starts from stump values summing to the centered sample mean and the sample residual SD, with positive fallback $\sqrt{\lambda}$ for a degenerate initial sample. These starting values do not alter the fixed priors.

Table 3: Full posterior computation at each cohort in the $N = 45$ study. Burn-in and retained counts are complete sweeps per chain. Retained draws from both chains are pooled, without thinning.

Method	Chains	Burn-in	Retained	Main transition
1PLD	0	0	0	Analytic Gaussian posterior
2PLD	2	4,000	64,000	Two-coordinate slice sweep
3PND	2	4,000	64,000	Three-coordinate slice sweep
iBART	2	1,000	16,000	Tree MH, Gaussian leaves, variance
mBART	2	1,000	4,000	Tree MH, ordered grid leaves, variance

For a retained tree draw, calculate $p_k = \Phi\{(f_k - \eta)/\sigma\}$ on the 101 dose cells and count the acceptable prefix with $p_k \le \theta = 1 - \gamma$. No acceptable cell gives $\xi = x_{\min}$; all acceptable cells give $\xi = x_{\max}$; otherwise ξ is the cut at the prefix's right boundary. This implements the supremum in Eq. (6), including jumps and endpoint atoms, without imposing risk equality at ξ. MTD quantiles use lower empirical quantiles (type 1). Assigned-dose risk is recomputed after the interior adjustment and dose restrictions, with a dose equal to a cut entering the cell on its right. The last administered dose remains the reported MTD estimate; the final posterior is an additional inferential summary.

H Bayesian planning under dose restrictions and count-based stopping

This appendix develops the planning formulation in Section 3 for the protocol used in this paper. The unknown response model is static; doses affect both patient outcomes and what can be learned before subsequent assignments. Joint posterior states are a standard device for Bayesian planning

(Ross et al., 2011; Guez et al., 2012), and dynamic programming for phase I trials has a precedent in Bartroff & Lai (2010). Here the two dose controls make the last administered dose consequential: Only Escalation prevents reversing an upward assignment, and the cap limits the doses reachable within the remaining cohorts. The results below specialize the decision problem to those controls, the cumulative DLT stopping rule, and a terminal estimate equal to the actual last dose. We first derive exact-policy guarantees and then evaluate finite-B rollout under the conjugate 1PLD model using a fixed linear response curve (Section I). The main simulation study compares the five EWOC procedures.

H.1 From the agent–environment loop to dose finding

The *agent–environment loop* describes sequential decisions: an agent uses its state and policy to choose an action, receives feedback from the environment, and updates its knowledge for the next decision. Reward defines the objective; return accumulates rewards over an episode (Ross et al., 2011). Figure 7 maps this loop to dose finding.

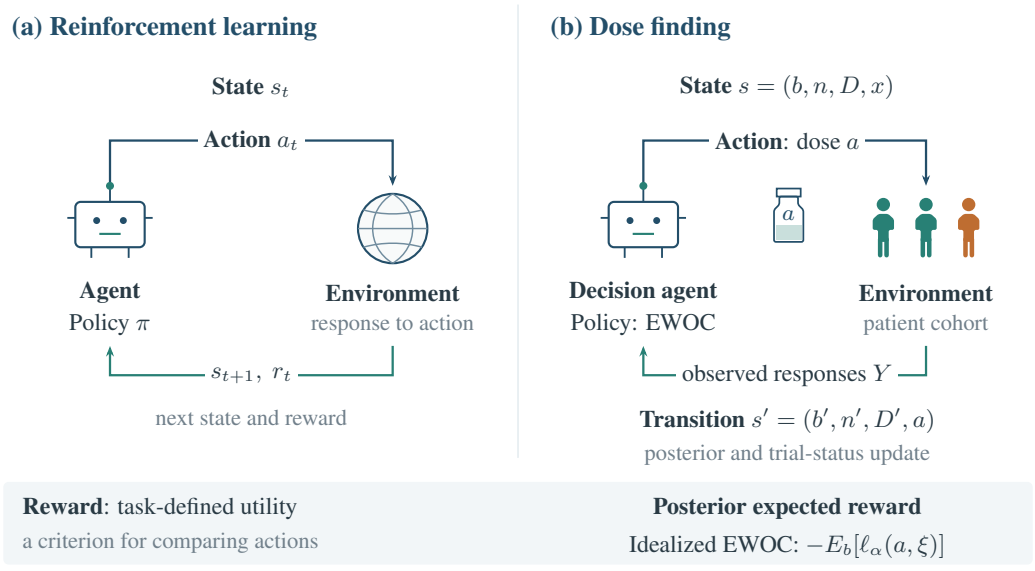


Figure 7: **The RL loop and its dose-finding counterpart.** An action changes patient exposure and supplies observations for the next decision. Cohort responses update the posterior and trial state; posterior expected reward evaluates dose choices. The reward's MTD-dependent component is latent.

Agent, environment, state, and action. The *agent* is the dose-selection procedure; the *environment* is the patient response process with a fixed unknown response model. After a completed cohort, the *state* $s = (b, n, D, x)$ records the joint parameter posterior, enrollment, cumulative DLT count, and last dose. The *action* a is the next cohort's dose. The implemented *policy* maps this state to a dose using EWOC and the enabled controls. The joint posterior supports MTD inference and predicts responses to candidate doses.

Feedback, transition, and episode. At dose a, the full cohort supplies observed responses $\mathbf{Y}$. These update the posterior, enrollment, and DLT count; the new last dose is a. This *transition* links patient exposure to subsequent learning. The trial *episode* ends when the completed-cohort stopping check gives $D > K_N$ or $n = N$.

Reward and the EWOC connection. The idealized EWOC proposal uses asymmetric dose loss and posterior expected reward

$$\ell_\alpha(a, \xi) = \alpha(\xi - a)_+ + (1 - \alpha)(a - \xi)_+, \qquad \bar{r}_b(a) = -\mathbb{E}_b[\ell_\alpha(a, \xi)].$$

The unadjusted posterior α-quantile maximizes this one-step expected reward; smaller α weights overdosing more heavily (Bartroff & Lai, 2011). The implemented policy subsequently applies its numerical adjustment and dose controls. The observed response $\mathbf{Y}$ supplies learning feedback, not this reward: the true MTD ξ is unknown, so dose quality is assessed through its posterior.

Learning and planning roles. Value-based RL uses expected future return to compare actions, while policy-based RL optimizes an action-selection rule; these approaches can be combined. Here, Bayesian updating learns the response model within a model-based belief-state formulation (Ross et al., 2011). EWOC is a prescribed posterior-dependent policy, with no policy-gradient training. The proposed rollout uses a different objective from the EWOC one-step loss: it compares expected remaining DLT and above-MTD counts plus terminal last-dose error. This connects the formulation to action-value planning (Bartroff & Lai, 2010). The main five-model experiments evaluate EWOC.

H.2 Joint posterior state and the cohort transition

Fix the model, prior, dose interval, and protocol. Let ϑ denote the static model parameter, μ_ϑ its likelihood mean, and f_ϑ the curve used to define the MTD. For iBART these are $\vartheta = (g, \sigma)$, $\mu_\vartheta = g$, and $f_\vartheta = \mathcal{P}g$; for mBART they are $\vartheta = (f, \sigma)$ and $\mu_\vartheta = f_\vartheta = f$. The parametric models use their stated mean, risk, and endpoint conventions. Write $\xi(\vartheta) = \xi(f_\vartheta, \sigma)$ and $p_\vartheta(a) = \Phi\{(f_\vartheta(a) - \eta)/\sigma\}$. The physical bounds satisfy $0 < x_{\min} < x_{\max} < \infty$, so $\xi(\vartheta) \in [x_{\min}, x_{\max}]$ is positive even at an endpoint.

After a completed cohort the state is $s = (b, n, D, x)$, where $b = \Pi(d\vartheta \mid \mathcal{F}_n)$, n is accrued enrollment, D is the cumulative DLT count, and x is the last administered dose. At dose a, the next cohort contains $m = \min(c, N - n)$ patients. For $\mathbf{y} = (y_1, \ldots, y_m)$ and $\phi_\sigma(v) = \sigma^{-1}\phi(v/\sigma)$, its likelihood, predictive density, and posterior are

$$L_a(\vartheta; \mathbf{y}) = \prod_{j=1}^{m} \phi_\sigma\{y_j - \mu_\vartheta(a)\}, \qquad q_b(\mathbf{y} \mid a) = \int L_a(\vartheta; \mathbf{y}) b(d\vartheta),$$

$$b^{a,\mathbf{y}}(d\vartheta) = \frac{L_a(\vartheta; \mathbf{y}) b(d\vartheta)}{q_b(\mathbf{y} \mid a)}, \qquad T(s, a, \mathbf{y}) = (b^{a,\mathbf{y}}, n + m, D + d(\mathbf{y}), a), \tag{18}$$

where $d(\mathbf{y}) = \sum_j \mathbf{1}\{y_j \geq \eta\}$. The same parameter generates the entire cohort. The integral therefore encloses the product; a product of marginal predictive densities would instead redraw the latent parameter for each patient. The raw g appears in the iBART likelihood and update, even though its projected curve defines the target and feasibility event.

Proposition 4 (Sufficiency of the joint belief state)**.** *Suppose the parameter space is standard Borel, the likelihood and parameter functionals are measurable, and within each cohort responses are conditionally independent given the preceding history, assigned dose, and static parameter, with the Gaussian densities above. Suppose policy randomization is independent of that parameter conditional on the observed history. Then $s = (b, n, D, x)$ is a controlled Markov state for continuation decisions under the specified protocol, with transition* (18)*. A state is absorbing when $D > K_N$ or $n = N$.*

Proof. Conditional on the observed history and chosen action, the parameter law is still b: prescribed or independently randomized actions add no parameter information. Conditional independence gives L_a, and integration gives q_b. Bayes' formula gives $b^{a,\mathbf{y}}$ wherever the denominator is positive and finite, which holds predictive-almost surely; choose an arbitrary measurable version on the remaining null set. Probability measures on a standard Borel space admit the required regular conditional distributions, and this dominated likelihood gives a measurable update. The cohort size, count increment, and new last dose depend only on $(s, a, \mathbf{y})$. The full cohort, including its posterior update, precedes the terminal check. Hence no further history is needed for the future transition or permitted actions. □

Why the MTD marginal is insufficient. For an interior candidate cut z, fixed $\sigma > 0$, and distinct $v_1, v_2 > 0$, consider the nondecreasing step means

$$f_k(a) = \eta - \sigma\Phi^{-1}(\gamma) + v_k\{2\mathbf{1}(a \geq z) - 1\}, \qquad k = 1, 2.$$

Both have MTD z: doses below the cut are acceptable and those at or above it are not. Degenerate beliefs at these two parameters have the same MTD marginal, but different predictive response distributions. Their DLT probabilities also differ. This counterexample to a general state reduction applies to the tree construction as well as to monotone Gaussian mean models. Keeping only the MTD posterior discards information needed to simulate future observations and update the model.

H.3 Two independent dose controls and the EWOC baseline

Let $\ell_0 = x_{\min} + \delta_{\text{num}}$ and $u_0 = x_{\max} - \delta_{\text{num}}$. At a post-initial, nonterminal state define

$$\ell(s) = \begin{cases} x, & \text{Only Escalation on,} \\ \ell_0, & \text{Only Escalation off,} \end{cases} \qquad u(s) = \begin{cases} \min(x + \Delta_{\max}, u_0), & \text{cap on,} \\ u_0, & \text{cap off,} \end{cases}$$

Set $I(s) = [\ell(s), u(s)]$. Every reached last dose is in $[\ell_0, u_0]$, so $I(s)$ is nonempty. Define three different posterior quantities:

$$\rho_b(a) = \int \Phi\{(\mu_\vartheta(a) - \eta)/\sigma\} b(d\vartheta),$$
$$r_b(a) = b\{p_\vartheta(a) > 1 - \gamma\}, \qquad o_b(a) = b\{\xi(\vartheta) < a\}.$$

These are, respectively, predictive DLT probability, posterior probability of exceeding the toxicity-risk target, and posterior probability of assigning above the MTD. In particular r_b is not ρ_b; at tree jumps it need not equal o_b either. For iBART, r_b uses the projected curve whereas ρ_b uses the raw likelihood mean.

The nominal set is $F(s) = \{a \in I(s) : r_b(a) \le \alpha\}$. Use $F(s)$ if nonempty and the singleton $\{\ell(s)\}$ otherwise. Thus the default fallback holds x; disabling Only Escalation instead permits the lower floor. The fallback's actual risk is recorded as an exception, not converted into a stopping condition. Initialization is the separate prescribed action $x_1 = 6$. Termination occurs only after an observed cohort when $D > K_N$ or $n = N$, regardless of the dose options.

For the exact posterior lower α-quantile $q_\alpha(b)$, write the complete EWOC baseline as

$$\widetilde{a}(b) = \max\{\ell_0, \min(x_{\max}, q_\alpha(b)) - \delta_{\text{num}}\}, \qquad \pi_0(s) = \max\{\ell(s), \min(\widetilde{a}(b), u(s))\}.$$

This composition applies the inward margin, floor, cap, and hold in that order. The admissibility result below is stated for the exact posterior quantile.

For nondecreasing curves, excessive risk at a implies $\xi(\vartheta) \le a$, including an unacceptable boundary point of a step. Consequently, for $a < q_\alpha(b)$,

$$r_b(a) \le b\{\xi \le a\} < \alpha.$$

The same implication holds under the stated 1PLD endpoint convention: its negative slopes are all acceptable under the fixed calibration. The inward proposal is below q_α unless the floor binds, and a cap only lowers it. If $\pi_0(s)$ fails the nominal bound, it must therefore equal $\ell(s)$. Monotonicity of the exceedance event then makes all of $I(s)$ infeasible, so this is exactly the required fallback. Otherwise it belongs to $F(s)$. Thus $\pi_0(s) \in \mathcal{A}(s)$ for all four option combinations, with the lower-endpoint fallback when $F(s) = \varnothing$.

H.4 Last-dose loss and protocol-imposed reachability

Fix nonnegative planning weights w_D, w_O, w_R before comparing policies. For a fixed parameter ϑ and current state s, define the continuation risk of policy π by

$$\mathcal{R}^\pi_\vartheta(s) = \mathbb{E}^\pi_\vartheta\left[w_D \sum_{i=n+1}^{n_{\text{end}}} \mathbf{1}\{Y_i \ge \eta\} + w_O \sum_{i=n+1}^{n_{\text{end}}} \mathbf{1}\{X_i > \xi(\vartheta)\} + w_R \frac{|X_{n_{\text{end}}} - \xi(\vartheta)|}{\xi(\vartheta)} \right]. \quad (19)$$

The policy continues to update its belief and obey the count rule while this expectation holds the generating parameter fixed. The Bayes continuation value is $V^\pi(s) = \int \mathcal{R}^\pi_\vartheta(s) b(d\vartheta)$. At initialization the sums are the full NPD and NPO. At later states, past loss contributions are fixed additive constants and may be omitted from the continuation value without changing future decisions.

Signed BTM remains a reporting criterion for the direction of error; the scalar planning loss uses its relative absolute magnitude.

Let

$$c_b(s,a) = m\{w_D\rho_b(a) + w_O o_b(a)\}, \qquad R(b,x) = w_R \int \frac{|x-\xi(\vartheta)|}{\xi(\vartheta)} b(d\vartheta).$$

Let $s_t = (b_t, n_t, D_t, x_t)$, with $s_0 = s$, and let τ count the remaining cohorts until termination. Averaging (19) over the current posterior and applying iterated expectation gives

$$V^\pi(s) = \mathbb{E}_b^\pi \left[\sum_{t=0}^{\tau-1} c_{b_t}(s_t, a_t) + R(b_\tau, x_\tau) \right].$$

Thus the count terms accrue to the realized stopping time, and the terminal term scores last-dose accuracy; the objective assigns no separate value to enrollment itself. The terminal posterior includes the final responses, but x_τ remains the dose selected beforehand, with no replacement estimate or terminal dose optimization.

Proof of Proposition 1. Write $H = H(s)$, equal to zero at terminal states and to $\lceil (N-n)/c \rceil$ otherwise, and $u_H = \min\{x + H\Delta_{\max}, u_0\}$. There are at most H remaining cohort assignments. Each is at least the preceding dose and at most that dose plus $\Delta_{\max}$, with upper limit u_0. Induction gives the asserted interval. Early DLT stopping can only reduce the number of changes. For every fixed ϑ, distance from $\xi(\vartheta)$ to this interval is $(x - \xi(\vartheta))_+ + (\xi(\vartheta) - u_H)_+$. Divide the pathwise distance bound by the positive $\xi(\vartheta)$ and integrate over the fixed initial belief and future outcomes. Later adaptive posterior updates do not change that joint-expectation argument. □

The bound isolates irreversible overshoot and a target too far above the current dose to reach in time. Feasibility and count-based stopping can further restrict a path. For the default $N = 45$, $c = 3$, and $\Delta_{\max} = 3.5$, the first completed cohort leaves $n = 3$, $x = 6$, and $H = 14$, giving $u_H = 55$. Cohort j cannot exceed $6 + (j-1)3.5$; the first 39 patients cannot exceed 48. For the evaluated truths with MTD 50, NPO is therefore at most six irrespective of the model. This bound follows directly from the dose schedule. Disabling Only Escalation removes the irreversible lower bound x; disabling the cap removes its incremental upper bound.

H.5 A FINITE-CANDIDATE BELLMAN PROBLEM

For existence and implementation, take a fixed finite list of measurable dose functions $a_1(s), \ldots, a_M(s)$ in $I(s)$, including both $\ell(s)$ and the complete baseline action $\pi_0(s)$. At non-terminal post-initial states retain candidates with $r_b(a) \le \alpha$; if none pass, retain only $\ell(s)$. Denote the resulting nonempty set by $C(s)$. Because the lower endpoint is included and r_b is nondecreasing, this fallback occurs exactly when $F(s)$ is empty. The baseline proof above gives $\pi_0(s) \in C(s)$ everywhere. For the initial cohort the permitted set is simply $\{x_1\}$, without nominal feasibility certification, and $\pi_0(b_0, 0, 0, x_1) = x_1$. Its last-dose coordinate only becomes an administered dose after that forced cohort.

Proposition 5 (Bellman recursion and an optimal finite-candidate policy)**.** *Under Proposition 4's measurability assumptions and the finite candidate construction, a measurable deterministic Markov policy attains the optimal Bayes continuation value V_C. It obeys*

$$V_C(s) = R(b,x), \qquad D > K_N \text{ or } n = N,$$
$$V_C(s) = \min_{a \in C(s)} Q_C(s,a), \qquad \text{otherwise},$$

$$Q_C(s,a) = c_b(s,a) + \int V_C\{T(s,a,\mathbf{y})\} q_b(\mathbf{y} \mid a) d\mathbf{y}.$$

The same statement uses the singleton prescribed action at initialization.

Proof. All continuation costs are nonnegative and bounded by $(w_D + w_O)(N-n) + w_R(x_{\max} - x_{\min})/x_{\min}$. The forced terminal value is measurable. Induct on $N - n$; each action observes at least one patient, so successor values are already defined. Integration against the measurable transition kernel preserves measurability. A finite minimum with measurable eligibility indicators is

measurable; choose the first minimizer in the fixed enumeration. Conditional expectation bounds the value of any randomized or history-dependent continuation below by this minimum. The selected action followed by the induction policies attains it. This proves both the recursion and optimality. Count-based terminal states encountered before N use the same forced terminal value. □

The finite formulation admits the measurable optimal policy in Proposition 5; continuous-action variants follow under suitable measurable-selection conditions.

H.6 Exact rollout improvement

Let V^{π_0} be the continuation cost of the complete EWOC baseline under the same candidate and stopping conventions. Define

$$Q^{\pi_0}(s,a) = c_b(s,a) + \int V^{\pi_0}\{T(s,a,\mathbf{y})\} q_b(\mathbf{y} \mid a) d\mathbf{y}.$$

This evaluates one candidate cohort followed by baseline continuation to termination, rather than only the next response.

Proof of Proposition 2. The values agree at terminal states. Assume the inequality holds for states with fewer remaining patients. At $a = \pi_+(s)$ it gives

$$\begin{aligned} V^{\pi_+}(s) &\le c_b(s,a) + \int V^{\pi_0}\{T(s,a,\mathbf{y})\} q_b(\mathbf{y} \mid a) d\mathbf{y} \\ &= \min_{v \in C(s)} Q^{\pi_0}(s,v) \le Q^{\pi_0}\{s, \pi_0(s)\} = V^{\pi_0}(s). \end{aligned}$$

The second inequality uses inclusion of the complete baseline action, including its forced fallback. Induction finishes the proof. □

The next section implements finite-B rollout under the same transition model, candidate sets, stopping rule, and weighted loss.

H.7 Monte Carlo rollout specification and approximation error

Algorithm 3 gives one repeated-rollout policy for all five response models. Model-specific routines supply joint posterior draws, likelihood predictions, posterior updates, and MTD inversion; $T_{\mathcal{M}}$ is the whole-cohort update in (18) for the chosen model. The algorithm starts after the prescribed initial cohort. It samples one latent model per simulated continuation and retains it throughout that path. The latent draw generates responses and scores the path; subsequent EWOC decisions use only the updated posterior and observed trial state. Each hypothetical update conditions on all actual and earlier simulated observations. For iBART, the raw g supplies the likelihood and update, while $\mathcal{P}g$ with the same σ determines the MTD and feasibility. For mBART, the constrained f supplies both. The candidate list below uses the two enabled dose controls of the numerical study. Posterior updates and EWOC quantiles are analytic in the 1PLD evaluation in Section I; the other models require numerical posterior procedures. The accompanying **Dose Trial Lab** application supports EWOC and rollout with all five response models and a user-selected simulation budget.

Algorithm 3 Repeated rollout with any of the five response models

Require: Model $\mathcal{M} \in \{\text{1PLD}, \text{2PLD}, \text{3PND}, \text{iBART}, \text{mBART}\}$; joint posterior state $s = (b, n, D, x)$ after the prescribed initial cohort.
Require: Shared protocol and EWOC baseline π_0, fixed weights (w_D, w_O, w_R), and B simulated continuations per candidate.
while $n < N$ and $D \leq K_N$ **do**
  Set $a_0 = \pi_0(s)$, including inward adjustment, dose bounds, upward cap, and Only Escalation.
  Form $(a_0, x, x + \frac{1}{4}(a_0 - x), x + \frac{1}{2}(a_0 - x), x + \frac{3}{4}(a_0 - x))$ and remove duplicates.
  Retain candidates with $r_b(a) \leq \alpha$; if none pass, retain the permitted hold x. Call this ordered list $C(s)$.
  if $|C(s)| = 1$ **then**
    Set a to its sole entry.
  else
    Draw independent joint parameters $\vartheta_r \sim b$, retaining their paired σ_r, and innovations $e_{r,n+1:N} \sim \mathcal{N}(0, I)$, for $r = 1, \ldots, B$; share them across candidates.
    for each $a \in C(s)$ **do**
      for $r = 1, \ldots, B$ **do**
        Copy $s_r = s$; set $a_r = a$ and $Z_r = 0$.
        while $n_r < N$ and $D_r \leq K_N$ **do**
          Set $m_r = \min(c, N - n_r)$; generate $y_{r,j} = \mu_{\vartheta_r}(a_r) + \sigma_r e_{r,n_r+j}$ for $j = 1, \ldots, m_r$.
          Add $w_D d(\mathbf{y}_r) + w_O m_r \mathbf{1}\{a_r > \xi(\vartheta_r)\}$ to Z_r.
          Update $s_r = T_{\mathcal{M}}(s_r, a_r, \mathbf{y}_r)$ using the complete simulated cohort.
          If s_r is nonterminal, set $a_r = \pi_0(s_r)$ using only its updated posterior and trial state.
        end while
        Add $w_R|x_r - \xi(\vartheta_r)|/\xi(\vartheta_r)$ to Z_r.
      end for
      Set $\widehat{Q}(s, a) = B^{-1} \sum_{r=1}^{B} Z_r$.
    end for
    Set a to the first minimizer of $\widehat{Q}(s, a)$ in $C(s)$.
  end if
  Administer $\min(c, N - n)$ patients at a, observe the complete actual cohort, and update $s = T_{\mathcal{M}}(s, a, \mathbf{y})$.
end while
Ensure: Terminal estimate x, the final administered dose.

With exact posterior primitives and independent continuations, $\widehat{Q}(s, a)$ is unbiased for $Q^{\pi_0}(s, a)$ at a fixed state and candidate. Selecting the smallest Monte Carlo estimate introduces selection error, motivating a conditional bound under a uniform approximation assumption.

Proposition 6 (Rollout under uniform approximation error)**.** *Let $\mathcal{S}_{\text{nt}}$ denote the nonterminal states in the continuation problem. Suppose a fixed measurable approximation, using the same legal candidate sets, satisfies*

$$\sup_{s \in \mathcal{S}_{\text{nt}}} \max_{a \in C(s)} |\widehat{Q}(s, a) - Q^{\pi_0}(s, a)| \leq \varepsilon$$

Its first-minimizing policy $\widehat{\pi}$ satisfies

$$V^{\widehat{\pi}}(s) \leq V^{\pi_0}(s) + 2H(s)\varepsilon.$$

Here $H(s)$ is zero at terminal states and $\lceil (N - n)/c \rceil$ otherwise, the maximum remaining cohort count.

Proof. The baseline candidate and the two approximation errors give $Q^{\pi_0}(s, \widehat{\pi}(s)) \leq V^{\pi_0}(s) + 2\varepsilon$. Induct on $H(s)$, with zero error at compulsory terminal states. The successor continuation contributes at most $2\{H(s) - 1\}\varepsilon$, and the current candidate comparison contributes at most 2ε. Early stopping only reduces these remaining stages, giving the stated bound. □

The bound requires uniform approximation error over the continuation domain with the same legal candidate sets. The numerical comparison below reports finite-Monte-Carlo rollout performance.

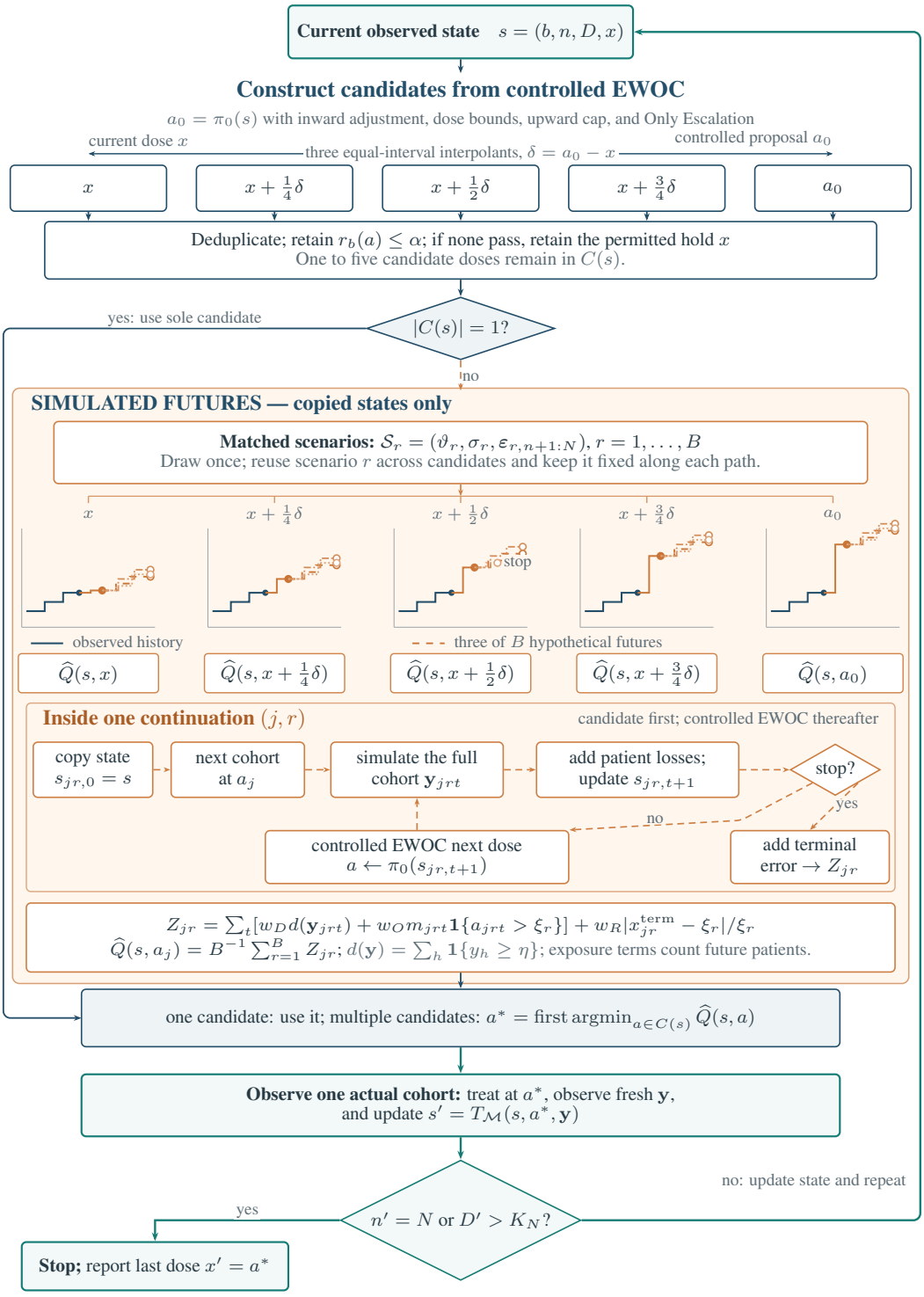


Figure 8: **Repeated EWOC-based rollout at one observed trial state.** Controlled EWOC generates up to five feasible candidates. A sole candidate is used directly; otherwise, B matched posterior-predictive continuations estimate candidate losses, using each candidate first and controlled EWOC thereafter. The minimum-loss candidate treats the next cohort, the state is updated, and the comparison repeats until stopping.

I ROLLOUT UNDER A LINEAR 1PLD MODEL

We use the fixed linear scenario in Table 5, for which the 1PLD working model is correctly specified and analytic posterior updates isolate rollout planning from numerical posterior approximation. The mean response has true MTD 50 and Gaussian noise with standard deviation 0.1. All policies fit the conjugate 1PLD working model $Y_i = \beta(X_i - 5) + \epsilon_i$, with $\epsilon_i \sim \mathcal{N}(0, 0.1^2)$ and the untruncated prior $\beta \sim \mathcal{N}(0.04, 0.02^2)$.

The study uses 1,000 paired trial replications. For each loss profile, rollout is evaluated with $B = 1{,}000$ and $B = 2{,}000$ simulated continuations per candidate, with the latter assessing sensitivity to the planning budget. All policies use initial dose 6, dose bounds 5 and 80, inward adjustment 7.5×10^{-7}, feasibility level $\alpha = 0.05$, target probability $\gamma = 0.9$, and toxicity threshold $\eta = 3$. We compare upward caps $\Delta_{\max} \in \{3.5, 7, 10.5\}$, with Only Escalation enabled. With $N = 45$ and cohorts of three patients, the whole cohort is observed and the posterior updated before stopping at cumulative DLT count above four or enrollment 45. The terminal estimate is the final administered dose. Write $d = \mathrm{NPD}/N$, $o = \mathrm{NPO}/N$, and $e = \mathrm{RAE}$, where planned N remains the denominator if a trial stops early. We use the equal-component objective $L_B = d + o + e$, corresponding to $(w_D, w_O, w_R) = (1/N, 1/N, 1)$, and two sensitivity profiles:

$$L_S = 2d + 2o + e, \qquad L_P = d + o + 2e.$$

Thus L_S emphasizes safety, whereas L_P increases the relative weight on MTD precision. Each weighted loss is compared only within its own profile.

Rollout follows Algorithm 3 and is recomputed after each observed cohort. The same patient-indexed innovations pair all policies and all cap settings within each replication, while planning draws are generated independently of the observed outcomes. The policy calculation uses only the current posterior and protocol state.

Table 4 reports the mean and Monte Carlo standard error (sample standard deviation divided by $\sqrt{1{,}000}$) for each policy. Weighted loss is the objective in Proposition 2; the component metrics show how its weights affect safety and precision.

Results and sensitivity analyses. Across all three cap settings, rollout attained lower observed mean weighted loss than EWOC under every loss profile and planning budget. Reductions relative to EWOC ranged from 4.1% to 25.9% with $\Delta_{\max} = 3.5$, from 55.8% to 80.9% with $\Delta_{\max} = 7$, and from 62.6% to 84.3% with $\Delta_{\max} = 10.5$. Safety emphasis reduced NPD and NPO, whereas precision emphasis reduced RAE relative to balanced rollout. Results changed little when B increased from 1,000 to 2,000.

Relaxing the cap exposed the safety–precision trade-off. Under EWOC, mean NPD increased from 0.318 to 1.364 and 1.759, and mean NPO from 0.384 to 2.460 and 3.309, as the cap increased from 3.5 to 7 and 10.5; mean RAE decreased from 1.250% to 0.855% and 0.795%. At the 10.5 cap, rollout kept mean NPD between 0.096 and 0.118 and mean NPO between 0.189 and 0.198 across the three profiles and two planning budgets. All rollout policies completed enrollment. EWOC stopped early in 1.1% of trials under the 7 cap and 3.8% under the 10.5 cap.

Computational scope and software. Let K bound the decision stages, $A_k \le A$ the candidates at stage k, $H_k \le K - k + 1$ the remaining cohorts, and c_m the cost of one simulated cohort for model m, including posterior updating, baseline-dose selection, response generation, and loss accounting. Full rollout planning per trial has work

$$T_{\mathrm{plan}} = O\left(B \sum_{k=1}^{K} A_k(1 + H_k)c_m\right) = O(BAK^2 c_m).$$

Early stopping shortens the realized paths. For fixed cohort size, analytic 1PLD updates and EWOC quantiles have constant arithmetic cost. A numerical update with I sampling or move iterations instead costs $O(Ic_{\mathrm{iter},m})$. Table 4 therefore uses 1PLD; **Dose Trial Lab** implements rollout for all five models with a user-selected planning budget.

Table 4: EWOC and rollout under the fitted 1PLD model for the fixed linear response curve, with 1,000 paired replications under each upward cap. Entries are mean (MCSE); RAE and early stopping are displayed as percentages, and RAE enters the losses as a proportion. Loss reduction is $100(1 - \overline{L}_{\text{policy}}/\overline{L}_{\text{EWOC}})$ within each cap and profile. Only Escalation is enabled in all three panels.

				Safety performance		Precision performance			
Loss profile	Policy	Weighted loss	Loss reduction vs. EWOC (%)	NPD	NPO	BTM	RAE (%)	Mean enrollment	Early stop (%)
(a) Upward cap $\Delta_{\max} = 3.5$									
$L_B = d + o + e$		0.0281		0.3180	0.3840	-0.593	1.250	45.00	0.0
(balanced)	EWOC	(0.0011)	0.0	(0.0175)	(0.0449)	(0.013)	(0.022)	(0.00)	(0.0)
		0.0249		0.1320	0.1500	-0.907	1.860	45.00	0.0
	Rollout, $B = 1{,}000$	(0.0006)	11.5	(0.0113)	(0.0207)	(0.021)	(0.040)	(0.00)	(0.0)
		0.0247		0.1320	0.1440	-0.906	1.858	45.00	0.0
	Rollout, $B = 2{,}000$	(0.0006)	12.0	(0.0113)	(0.0203)	(0.021)	(0.040)	(0.00)	(0.0)
$L_S = 2d + 2o + e$		0.0437		0.3180	0.3840	-0.593	1.250	45.00	0.0
(safety-emphasized)	EWOC	(0.0023)	0.0	(0.0175)	(0.0449)	(0.013)	(0.022)	(0.00)	(0.0)
		0.0328		0.1000	0.0690	-1.254	2.534	45.00	0.0
	Rollout, $B = 1{,}000$	(0.0009)	24.8	(0.0101)	(0.0142)	(0.029)	(0.057)	(0.00)	(0.0)
		0.0324		0.1010	0.0600	-1.248	2.522	45.00	0.0
	Rollout, $B = 2{,}000$	(0.0009)	25.9	(0.0102)	(0.0133)	(0.029)	(0.057)	(0.00)	(0.0)
$L_P = d + o + 2e$		0.0406		0.3180	0.3840	-0.593	1.250	45.00	0.0
(precision-emphasized)	EWOC	(0.0011)	0.0	(0.0175)	(0.0449)	(0.013)	(0.022)	(0.00)	(0.0)
		0.0388		0.1440	0.1590	-0.778	1.603	45.00	0.0
	Rollout, $B = 1{,}000$	(0.0007)	4.4	(0.0118)	(0.0213)	(0.016)	(0.030)	(0.00)	(0.0)
		0.0389		0.1420	0.1590	-0.782	1.611	45.00	0.0
	Rollout, $B = 2{,}000$	(0.0007)	4.1	(0.0117)	(0.0213)	(0.016)	(0.030)	(0.00)	(0.0)
(b) Upward cap $\Delta_{\max} = 7$									
$L_B = d + o + e$		0.0935		1.3640	2.4600	-0.367	0.855	44.92	1.1
(balanced)	EWOC	(0.0050)	0.0	(0.0388)	(0.2104)	(0.011)	(0.017)	(0.03)	(0.3)
		0.0266		0.1160	0.1590	-0.994	2.053	45.00	0.0
	Rollout, $B = 1{,}000$	(0.0006)	71.5	(0.0108)	(0.0213)	(0.020)	(0.037)	(0.00)	(0.0)
		0.0268		0.1140	0.1650	-0.999	2.062	45.00	0.0
	Rollout, $B = 2{,}000$	(0.0006)	71.3	(0.0107)	(0.0216)	(0.020)	(0.037)	(0.00)	(0.0)
$L_S = 2d + 2o + e$		0.1785		1.3640	2.4600	-0.367	0.855	44.92	1.1
(safety-emphasized)	EWOC	(0.0100)	0.0	(0.0388)	(0.2104)	(0.011)	(0.017)	(0.03)	(0.3)
		0.0341		0.1060	0.1560	-1.090	2.242	45.00	0.0
	Rollout, $B = 1{,}000$	(0.0011)	80.9	(0.0104)	(0.0211)	(0.024)	(0.045)	(0.00)	(0.0)
		0.0342		0.1050	0.1560	-1.097	2.256	45.00	0.0
	Rollout, $B = 2{,}000$	(0.0011)	80.9	(0.0104)	(0.0211)	(0.025)	(0.046)	(0.00)	(0.0)
$L_P = d + o + 2e$		0.1021		1.3640	2.4600	-0.367	0.855	44.92	1.1
(precision-emphasized)	EWOC	(0.0049)	0.0	(0.0388)	(0.2104)	(0.011)	(0.017)	(0.03)	(0.3)
		0.0451		0.1210	0.1710	-0.934	1.931	45.00	0.0
	Rollout, $B = 1{,}000$	(0.0007)	55.8	(0.0110)	(0.0220)	(0.019)	(0.034)	(0.00)	(0.0)
		0.0450		0.1210	0.1680	-0.933	1.930	45.00	0.0
	Rollout, $B = 2{,}000$	(0.0007)	55.9	(0.0110)	(0.0218)	(0.019)	(0.034)	(0.00)	(0.0)
(c) Upward cap $\Delta_{\max} = 10.5$									
$L_B = d + o + e$		0.1206		1.7590	3.3090	-0.323	0.795	44.66	3.8
(balanced)	EWOC	(0.0063)	0.0	(0.0444)	(0.2674)	(0.011)	(0.016)	(0.06)	(0.6)
		0.0293		0.0980	0.1890	-1.106	2.292	45.00	0.0
	Rollout, $B = 1{,}000$	(0.0006)	75.7	(0.0101)	(0.0231)	(0.022)	(0.041)	(0.00)	(0.0)
		0.0292		0.0960	0.1920	-1.099	2.275	45.00	0.0
	Rollout, $B = 2{,}000$	(0.0006)	75.8	(0.0100)	(0.0232)	(0.022)	(0.040)	(0.00)	(0.0)
$L_S = 2d + 2o + e$		0.2332		1.7590	3.3090	-0.323	0.795	44.66	3.8
(safety-emphasized)	EWOC	(0.0127)	0.0	(0.0444)	(0.2674)	(0.011)	(0.016)	(0.06)	(0.6)
		0.0367		0.0970	0.1890	-1.160	2.401	45.00	0.0
	Rollout, $B = 1{,}000$	(0.0012)	84.3	(0.0101)	(0.0231)	(0.023)	(0.042)	(0.00)	(0.0)
		0.0369		0.0960	0.1980	-1.149	2.381	45.00	0.0
	Rollout, $B = 2{,}000$	(0.0012)	84.2	(0.0100)	(0.0236)	(0.023)	(0.042)	(0.00)	(0.0)
$L_P = d + o + 2e$		0.1285		1.7590	3.3090	-0.323	0.795	44.66	3.8
(precision-emphasized)	EWOC	(0.0063)	0.0	(0.0444)	(0.2674)	(0.011)	(0.016)	(0.06)	(0.6)
		0.0481		0.1170	0.1890	-0.996	2.063	45.00	0.0
	Rollout, $B = 1{,}000$	(0.0008)	62.6	(0.0108)	(0.0231)	(0.020)	(0.036)	(0.00)	(0.0)
		0.0478		0.1180	0.1890	-0.988	2.049	45.00	0.0
	Rollout, $B = 2{,}000$	(0.0008)	62.8	(0.0109)	(0.0231)	(0.020)	(0.036)	(0.00)	(0.0)

J SIMULATION DESIGN

J.1 ENROLLMENT, ALLOCATION, AND STOPPING

The study compares all five procedures with $N = 45$, cohort size $c = 3$, initial dose 6, interval $(5, 80)$, $\eta = 3$, $\gamma = 0.9$, and $\alpha = 0.05$. Both the upward cap $\Delta_{\max} = 3.5$ and Only Escalation are enabled. A cohort receives a common dose, and all its responses are observed before posterior updating and the next allocation. The procedure stops at a completed cohort if $D_n > \lfloor 0.1N \rfloor = 4$, or if $n = 45$. The fifth observed DLT triggers the count rule; a complete cohort may overshoot it. Posterior feasibility exceptions do not stop the trial. The full dose calculation is (8).

Starting at 6, a complete trial has 14 adaptive cohort changes and therefore a maximum reachable dose of 55. Cohort j receives at most $6 + (j - 1)3.5$, giving a maximum of 48 at cohort 13. Only the final six patients can be assigned above the true MTD 50, so $\mathrm{NPO} \leq 6$ for every path, including early stops. This reachability constraint is part of the study design, not information supplied to a fitted model about the true MTD.

J.2 SEVEN TRUE DOSE–TOXICITY CURVES

Let $t = (x - 5)/45$ and $A = 3 - 0.1\Phi^{-1}(0.9)$. Responses satisfy

$$Y_i = f_0(X_i) + \epsilon_i, \qquad f_0(x) = Ah(x), \qquad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, 0.1^2).$$

The seven shapes are defined in Table 5 and displayed in Figure 9. Each is nondecreasing with $h(5) = 0$ and $h(50) = 1$, giving true risk 0.1 and MTD 50 at dose 50. The mean there is below the response threshold $\eta = 3$ because residual variation is positive. All models receive dose inputs in physical units; t only defines the data-generating curves.

Table 5: The seven data-generating shapes, exactly matching the software.

Curve	$h(x)$
Linear	t
Piecewise linear	Linear interpolation of the knots below
Quadratic (convex)	t^2
Square root (concave)	$\sqrt{t}$
Sigmoid	$\{L(x) - L(5)\}/\{L(50) - L(5)\}$, $L(x) = \{1 + \exp[-(x - 42.5)/8]\}^{-1}$
Exponential (convex)	$(e^{2t} - 1)/(e^2 - 1)$
Logarithmic (concave)	$\log(1 + 4t)/\log(5)$

The piecewise linear knots are

x	5	20	35	45	50	60	80
$h(x)$	0	0.04	0.08	0.5	1	1.2	1.25.

J.3 PRIORS, REPLICATION, AND COMPUTATION

There are 100 replications for each of the seven curves and five methods, totaling 3,500 trials. Patient innovations are paired across methods within curve and replication; posterior random streams are separate. No truth or target location is supplied to allocation. The 1PLD slope prior is $\mathcal{N}(0.04, 0.02^2)$ with known $\sigma = 0.1$. The 2PLD and 3PND priors follow Appendix D, with $\delta = 0.1$ for 3PND. Tree models use 200 trees, shrinkage $k = 2$, response reference range $[0, 4]$, 100 physical-dose cuts, and the inverse-χ^2_3 variance prior with $\Pr(\sigma < 0.2) = 0.9$. The reference range does not truncate observed responses. Full sampling settings appear in Appendix G.

Exact random-stream schedule. The study uses R 4.5.1 with the Mersenne–Twister generator, Inversion normal generation, and Rejection sampling. Number the curves $s = 1, \ldots, 7$ in Table 5 order, the replications $r = 1, \ldots, 100$, and methods $m = 1, \ldots, 5$ as 1PLD, 2PLD, 3PND, iBART, and mBART. The job index is

$$j = 5\{100(s - 1) + (r - 1)\} + m.$$

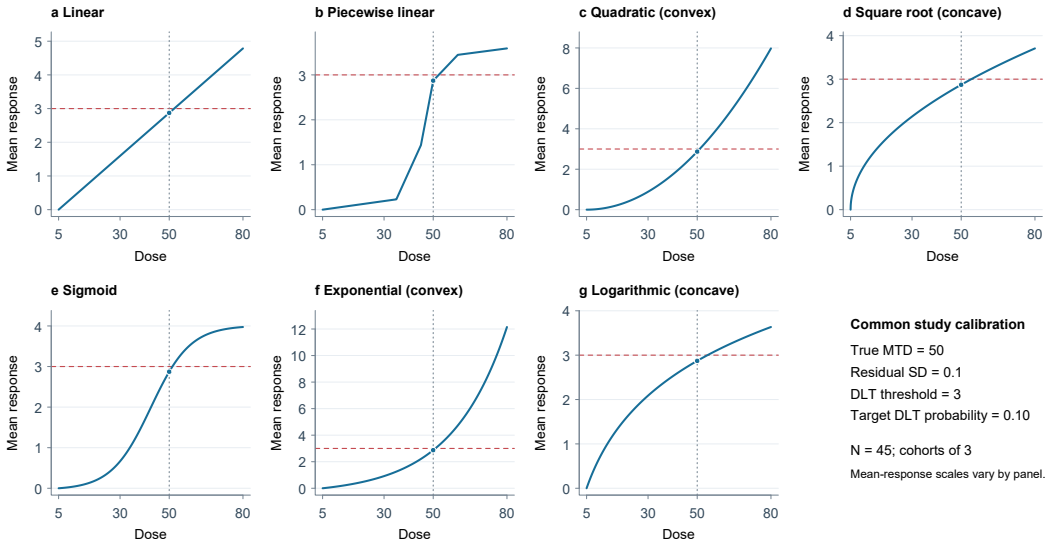


Figure 9: Seven true mean dose–toxicity curves used in the $N = 45$ study. All panels share the physical dose axis $(5, 80)$; mean-response axes are scaled separately to show each shape. The red horizontal line is the toxicity-response threshold $\eta = 3$, and the dotted vertical line is the true MTD 50. At that dose the mean is $3 - 0.1\Phi^{-1}(0.9) < 3$, so the Gaussian response tail has DLT probability 0.1. The curve is a mean, not a posterior interval or an individual outcome.

The patient-innovation seed is $801000000 + 10000s + r$, shared by all five methods within a curve and replication. The posterior seed for cohort k and chain $\ell \in \{1, 2\}$ is $810000000 + 100000j + 100k + \ell$. The analytic 1PLD posterior requires no MCMC stream.

J.4 Four performance criteria

At the realized endpoint $n = n_{\text{end}}$, define

$$\begin{aligned} \text{NPD}(n) &= \sum_{i=1}^{n} \mathbf{1}\{Y_i \geq \eta\}, \quad & \text{NPO}(n) &= \sum_{i=1}^{n} \mathbf{1}\{X_i > \xi_0\}, \\ \text{BTM}(n) &= X_n - \xi_0, & \text{RAE}(n) &= |X_n - \xi_0|/\xi_0. \end{aligned} \tag{20}$$

These follow the last-dose convention of Lee (2023). Reported RAE is multiplied by 100. Counts include every patient in the final cohort. Trials stopped early remain in all summaries with their actual endpoint and final dose. Each model–curve cell reports the replicate mean, median, and MCSE, where MCSE is the sample standard deviation divided by $\sqrt{100}$. Mean enrollment and early-stop frequency provide context for the four criteria.

All 3,500 trials enrolled 45 patients, with no early stops. Of these, 2,942 reached the enrollment limit without crossing the DLT-count limit; 558 also crossed that limit in the final cohort. Thus, lower NPD or NPO cannot be attributed to shorter trials.

J.5 Relation to the other analyses

Section I reports a separate 1PLD comparison of EWOC and rollout at two planning budgets under the linear scenario; its replications are not pooled with the five-model comparison in Table 1. The pharmacodynamic analysis in Appendix K is a separate 30-patient adaptive application and contributes no replications to either simulation comparison.

K PHARMACODYNAMIC EXAMPLE

K.1 PUBLISHED DATA AND COMMON TARGET

We compare five models using the O^6-benzylguanine example of Lee (2023, pp. 12–15), which specifies $(x_{\min}, x_{\max}) = (20, 140)$ mg/m^2 and reports 24 tumor O^6-alkylguanine-DNA alkyltransferase (AGT) measurements at 40, 60, 80, and 100 mg/m^2 (Table 6).

Table 6: Published AGT measurements used as dose-specific resampling pools.

Dose (mg/m^2)	Count	AGT (fmol/mg)
40	3	26.35, 42.00, 15.00
60	3	23.00, 13.50, 11.00
80	9	31.67, 8.00, 9.00, 14.50, 11.50, 7.00, 11.70, 9.03, 8.00
100	9	4.07, 5.00, 8.70, 2.50, 4.07, 6.13, 3.60, 5.00, 5.00

Following Lee (2023), let the continuous response be $R = 60 - \mathrm{AGT}$, with threshold 55. For 1PLD and the two tree models we use the equivalent score $Y = R/60$ and $\eta = 55/60$; 2PLD and 3PND retain the original response units. Doses remain in mg/m^2. With $\gamma = 0.6$ and $\theta = 1 - \gamma = 0.4$, the common pharmacodynamic boundary is

$$p_{\mathrm{dep}}(x) = \Phi\left\{\frac{f(x) - 55/60}{\sigma}\right\}, \qquad \xi = \sup\big(\{20\} \cup \{x \in (20, 140) : p_{\mathrm{dep}}(x) \le 0.4\}\big).$$

Here $p_{\mathrm{dep}}(x) = \Pr(\mathrm{AGT} < 5 \mid x)$; f and σ in the display use the score scale. For iBART, $f = \mathcal{P}g$ is the projected mean functional with each draw's residual SD, whereas its likelihood uses g. The other methods compute the tail from their fitted Gaussian mean and SD. The endpoints 20 and 140 represent an empty and an entirely acceptable range.

Although AGT depletion measures pharmacodynamic activity rather than clinical toxicity, the analysis retains the sequential dose-finding task of learning a monotone dose–response boundary while assigning doses. These published measurements make the application reproducible; comparable patient-level toxicity data are difficult to access and publish because of confidentiality restrictions. The inferred AGT boundary is a pharmacodynamic target, not a toxicity MTD or a validated clinical dose recommendation.

For each method we generate a separate 30-patient allocation path in ten cohorts of three, beginning at 40 mg/m^2. At the assigned dose, each cohort draws three AGT values without replacement from its published pool and adds independent centered uniform perturbations to $R = 60 - \mathrm{AGT}$. The half-widths are 8, 6, 4, and 2 fmol/mg at 40, 60, 80, and 100 mg/m^2, respectively, following the redesign of Lee (2023). A source value may reappear in a later cohort. A common precomputed cohort-by-dose response bank gives two methods identical outcomes if they assign the same dose in the same cohort. Each posterior update uses only that method's accrued cohort responses. The 24 source values provide the resampling pools and 1PLD's fixed SD but are not counted as an initial cohort.

K.2 PRIOR SPECIFICATION AND COMPUTATION

The five models share the target above but require model-specific calibration of the response scale and residual variation. For iBART and mBART we retain the 200-tree architecture, method-specific tree-depth settings, and $k = 2$ described in Appendices C and G. For the normalized AGT score, we use mean center 0.5, response span 1, and residual-SD reference 0.20 on that scale; the leaf scale is thereby recalibrated to the score units. The mean center and span correspond to the score units rather than transferring the simulation outcome's numerical 2 and 4 unchanged. The 100 equally spaced interior split points are placed on the case's physical interval, $20 + 120j/101$, $j = 1, \ldots, 100$. This is an outcome-scale calibration of the BART prior, not a change to the clinical threshold or dose domain.

For 2PLD and 3PND, we use the normalized hierarchical priors in Appendix D on the source's raw response $60 - \mathrm{AGT}$, with threshold 55, and use $\delta = 0.5$ for 3PND as in the source's pharmacodynamic

example (Lee, 2023). This retains the scale-one truncated half-Cauchy prior in its original response units. Dividing fitted means and SDs by 60 then gives the common display scale; applying a scale-one half-Cauchy directly to the score-scale SD would specify a different prior. All three parametric mean functions are anchored at zero at dose 20; the tree means are not constrained to that endpoint value.

For 1PLD, the residual SD is estimated once by pooling within-dose variation across the four groups and then treated as fixed:

$$s^2_{\text{pool}} = \frac{\sum_{d\in\{40,60,80,100\}} \sum_{i:X_i=d} (Y_i - \bar{Y}_d)^2}{24-4}, \qquad s_{\text{pool}} = 0.1142754.$$

This is 6.8565 fmol/mg before division by 60. The untruncated score-scale slope prior is $\beta \sim \mathcal{N}(0.005, 0.0025^2)$, using the fixed calibration $a = 0.60/(140-20)$ and $b = 0.30/(140-20)$. These are implementation prior constants, not estimates from the AGT observations. The analytic 1PLD posterior conditions on the plug-in SD and does not propagate uncertainty from estimating it.

At each cohort, the 2PLD and 3PND posterior calculations use 4,096 sequential Monte Carlo particles with eight rejuvenation sweeps and a 0.75 effective-sample-size resampling threshold. Each tree posterior is refitted using 1,000 burn-in iterations and 4,000 retained draws; the 1PLD update is analytic. Independent posterior random seeds are fixed by method and cohort. Posterior calculation follows Appendix G.

K.3 Sequential EWOC allocation

All five methods use the same decision rule and differ only in their dose–response posterior. Let $q_{m,t}$ be model m's posterior α-quantile of ξ after cohort t, with $\alpha = 0.05$. We adopt the published 3PND application's strict upward grid mapping, while adding a 20 mg/m^2 upward cap and Only Escalation. We first cap the *continuous* EWOC proposal, then choose the first eligible dose strictly above it. Specifically, let

$$\begin{aligned} u_{m,t+1} &= \min\{q_{m,t}, d_{m,t} + 20\}, \\ \mathcal{D}_{m,t} &= \{d \in \mathcal{D} : d \le d_{m,t} + 20\}, \\ \mathcal{D} &= \{40, 60, 80, 100\}\,\text{mg/m}^2. \end{aligned}$$

Choose $r_{m,t+1}$ as the smallest dose in $\mathcal{D}_{m,t}$ strictly above $u_{m,t+1}$; if none exists, use $\max \mathcal{D}_{m,t}$. Then $d_{m,t+1} = \max\{d_{m,t}, r_{m,t+1}\}$, so the assignment rises by at most one level and never decreases. Strict rounding advances an exact-grid proposal when the cap permits it. The published pool contains no AGT values at 120 mg/m^2, so 100 is the highest available action. The first cohort is assigned 40 mg/m^2 by protocol. Every method completes ten cohorts; an AGT-depletion event does not trigger toxicity stopping. This application uses no rollout.

The posterior α-quantile is a continuous-dose EWOC proposal. Upward grid rounding or holding a previous dose can give an assigned dose with $\Pr\{p_{\text{dep}}(d_{m,t+1}) > 0.4 \mid \mathcal{F}_{m,3t}\} > \alpha$. Consequently, this discrete rule is ***EWOC-guided*** but does not guarantee the nominal posterior feasibility bound for the administered doses. We record every such post-initial exception.

K.4 Five-method sequential results

Figure 10 shows the administered doses by cohort, and Table 7 reports the last-cohort dose and terminal posterior summary for each method.

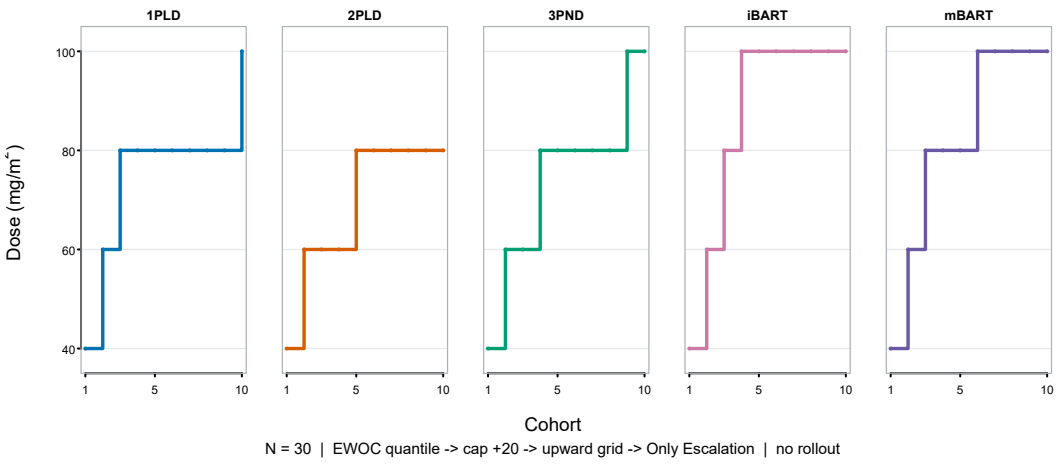


Figure 10: Administered-dose paths for five 30-patient runs under a common $\alpha = 0.05$ EWOC quantile, a 20 mg/m^2 cap, strict upward-grid assignment, and Only Escalation. Table 7 reports posterior-bound exceptions.

Table 7: Single-path AGT results. Last dose follows 27 outcomes; terminal $q_{0.05}(\xi)$ uses all 30. "Bound exceeded" counts post-initial assignments with $\Pr\{p_{\text{dep}}(d_{\text{next}}) > 0.4 \mid \mathcal{F}\} > 0.05$. AGT < 5 is pharmacodynamic, not a clinical DLT.

Method	Last dose (mg/m^2)	AGT < 5 (of 30)	Terminal $q_{0.05}(\xi)$ (mg/m^2)	Bound exceeded (of 9)
1PLD	100	3	82.93	8
2PLD	80	1	76.17	9
3PND	100	6	85.38	9
iBART	100	14	82.97	7
mBART	100	10	82.97	8

The methods escalate at different times. 3PND reaches 60, 80, and 100 mg/m^2 in cohorts 2, 4, and 9; iBART, mBART, and 1PLD reach 100 in cohorts 4, 6, and 10; 2PLD ends at 80. These last doses are allocation outputs, not terminal posterior boundaries. AGT-depletion counts reflect the distinct paths; a single path per method cannot establish comparative operating characteristics.

K.5 Interpretation

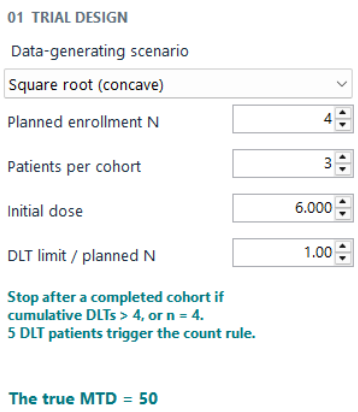
The original glioma trial selected 100 mg/m^2 as a biochemical modulatory dose using an AGT threshold of 10 fmol/mg (Friedman et al., 1998). In a 30-patient redesign with a threshold of 5 fmol/mg, Lee (2023, pp. 13–15) reported 80 mg/m^2 for standard 3PND escalation and 100 mg/m^2 for accelerated escalation. Our five paths use a common resampled response bank, five posteriors, and capped upward-grid allocation with Only Escalation; they are not numerical replications of either published trajectory.

The nominal $\alpha = 0.05$ posterior bound is exceeded at 7–9 of nine post-initial assignments per model (Table 7). For example, 3PND's first continuous proposal is 42.84 mg/m^2; upward rounding assigns 60, with posterior exceedance probability 0.501. Neither the cap nor Only Escalation restores feasibility at the *administered* dose. Because AGT depletion is pharmacodynamic and the true boundary is unknown, we cannot compute NPO, BTM, or RAE against a known target or infer a validated clinical optimal dose. Within these limits, the five-model comparison places terminal doses in the 80–100 mg/m^2 region reported in the earlier AGT analyses, while showing how the posterior model changes escalation timing and depletion outcomes.

L SOFTWARE MANUAL

Dose Trial Lab 0.1.4 provides EWOC, rollout, and paired comparisons for all five models and seven scenarios. The controls illustrate policy selection; the dashboard and summaries show saved EWOC examples from the linear scenario with 45 patients. These examples are not simulation-study results.

(a) Trial design



(b) Policy and planning budget

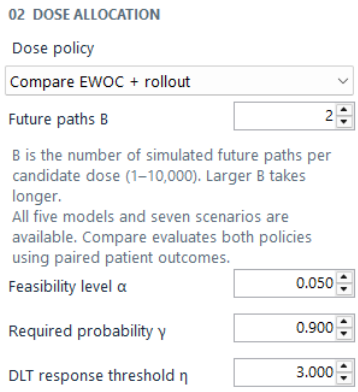


Figure 11: Trial and policy controls. Choose **EWOC**, **Rollout**, or **Compare EWOC + rollout**. **Future paths B** is the number of hypothetical continuations per candidate; larger values require more computation. Scroll down for the dose controls and model choices.

L.1 INSTALL, CHOOSE SETTINGS, AND RUN

1. Run `DoseTrialLab-0.1.4-Setup.exe` on 64-bit Windows with .NET Framework 4.8, then open **Dose Trial Lab**. Other required components are bundled.
2. Choose the scenario, planned enrollment, cohort size, and initial dose. Defaults are $N = 45$, cohorts of three, and initial dose 6. Under these defaults and with DLT-limit fraction 0.10, observe the whole cohort, then stop at five or more DLTs or enrollment 45.
3. Choose the **Dose policy**. For rollout or comparison, set **Future paths B** from 1 to 10,000 (default 30). Comparison pairs patient innovations between policies within each trial. Set the feasibility level and response thresholds as required.
4. Both dose controls are on by default, with upward cap 3.5. Turn off **Only Escalation** to permit dose decreases.
5. Select models and replicates; retain or change the random seed. Select **Run simulation**; **Cancel** interrupts a run.

Standalone EWOC uses manuscript computation settings. Rollout and comparison use faster settings, matched between comparison policies. These settings apply to interactive software runs and differ from those used in the reported experiments.

L.2 VIEW THE TRIAL AND REPLAY ITS UPDATES

Select the model, policy, and replicate from **Trial**. The **Trial dashboard** shows doses, DLTs, responses, the mean curve, and MTD posterior. Use the cohort slider or **Play/Pause** to inspect updates. **Show γ-response quantile** adds the optional curve overlay.

Primary Trial Summaries reports the last-dose estimate, NPD, NPO, BTM, RAE, weighted loss, enrollment, and stopping reason. **Additional Trial Summaries** contains posterior summaries. The header identifies the selected model, policy, and endpoint. Paired summaries can coincide when both policies choose the same doses.

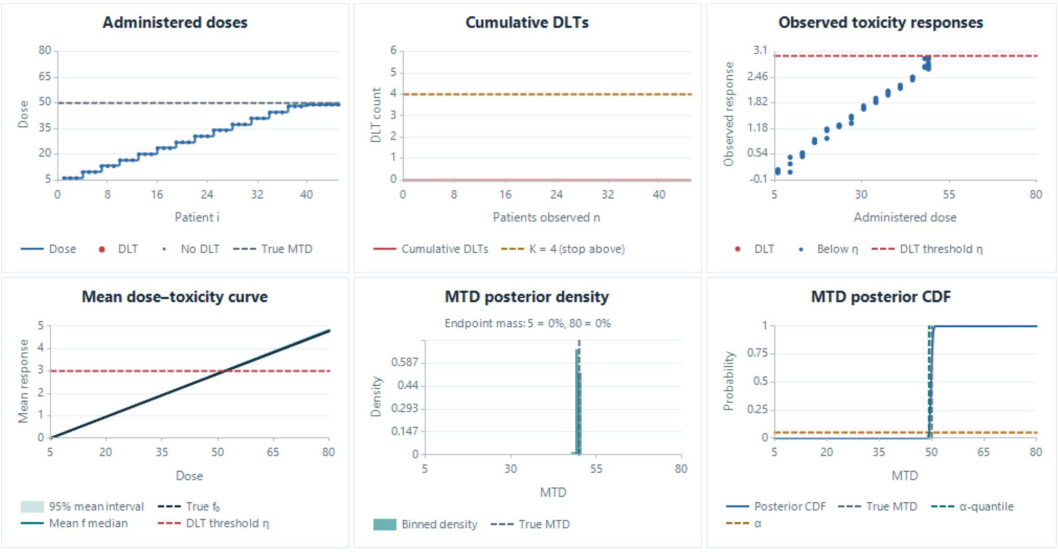
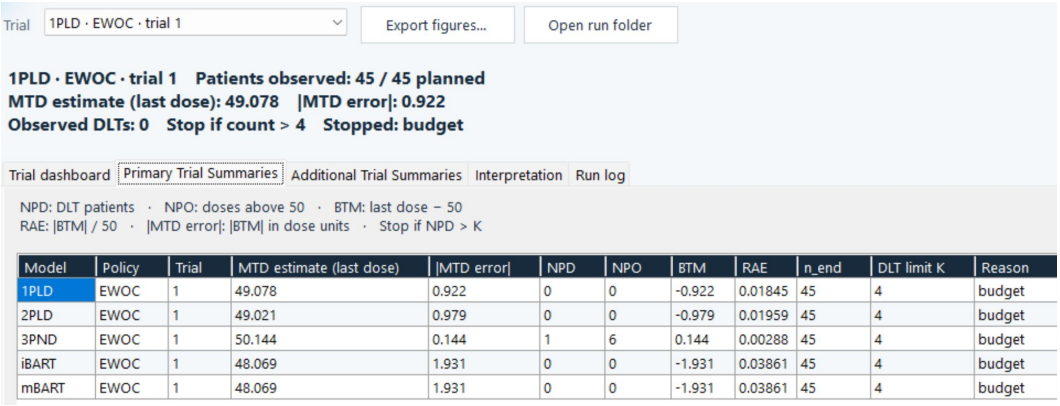


Figure 12: Six-panel dashboard for a saved 1PLD EWOC trial under the linear scenario after 45 patients: administered doses, cumulative DLTs, observed responses, mean dose–toxicity curve, and MTD posterior density and CDF. Change **Trial** or the cohort slider to inspect another saved update.

L.3 Export figures and reopen a run

Choose **Export figures** to save all six plots for the selected trial and cohort. **Open run folder** locates saved results and paired exports; **Load saved result** reopens them. **Save configuration** retains the settings.



Model	Policy	Trial	MTD estimate (last dose)	|MTD error|	NPD	NPO	BTM	RAE	n_end	DLT limit K	Reason
1PLD	EWOC	1	49.078	0.922	0	0	-0.922	0.01845	45	4	budget
2PLD	EWOC	1	49.021	0.979	0	0	-0.979	0.01959	45	4	budget
3PND	EWOC	1	50.144	0.144	1	6	0.144	0.00288	45	4	budget
iBART	EWOC	1	48.069	1.931	0	0	-1.931	0.03861	45	4	budget
mBART	EWOC	1	48.069	1.931	0	0	-1.931	0.03861	45	4	budget

Figure 13: EWOC-only **Primary Trial Summaries** for the five models in a saved linear-scenario run with 45 patients per trial. The selected 1PLD trial is the dashboard example above. These individual trials illustrate the interface, not a comparison of average performance.