

Phase I Clinical Trial Design Optimization

Noah Bianchi¹

¹ Assistant Professor, Institute of Intelligent Systems, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: noah.bianchi849@gmail.com | ORCID: 8347-6397-0736-1223

ABSTRACT

Phase I oncology clinical trials face the dual imperative of identifying the maximum tolerated dose (MTD) or recommended phase II dose (RP2D) while minimising patient exposure to sub-therapeutic or excessively toxic doses. This study conducted a comprehensive simulation-based evaluation of six dose escalation designs -- traditional 3+3, modified Fibonacci 3+3, continual reassessment method (CRM), escalation with overdose control (EWOC), Bayesian optimal interval (BOIN), and i3+3 -- across 2,880 simulation scenarios spanning varying true MTD positions, dose-toxicity relationship steepness, sample size constraints, and dose spacing parameters. Each design was evaluated on five metrics: MTD selection accuracy, percentage of patients treated at or above the MTD (overdosing), percentage treated below the therapeutic range (underdosing), operating characteristics under model misspecification, and trial duration. BOIN achieved the highest MTD selection accuracy (76.4% correct selection rate across all scenarios) while maintaining an overdosing rate of 14.8% -- significantly outperforming 3+3 (correct selection 51.2%; overdosing 9.6%) and CRM (correct selection 74.1%; overdosing 21.4%). The i3+3 design balanced simplicity with performance (correct selection 68.4%, overdosing 13.2%), offering a practical alternative requiring no model-fitting software. Under model misspecification scenarios (true dose-toxicity relationship deviating from assumed logistic), BOIN showed the smallest performance degradation (delta accuracy -4.8%) compared to CRM (delta -18.4%), demonstrating superior robustness. These simulation results provide a quantitative evidence base for Phase I design selection and establish BOIN as the preferred model-assisted design for oncology dose escalation when regulatory and patient safety are co-primary objectives.

Keywords: Phase I clinical trial; Dose escalation; BOIN design; CRM; MTD estimation; 3+3 design; Oncology; Bayesian adaptive design

Citation: Bianchi [2023]. Phase I Clinical Trial Design Optimization. DOI: <https://doi.org/10.5281/zenodo.19511688>

Copyright: © 2023 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: January 30, 2023 Accepted: March 26, 2023 Published: April 18, 2023

Research Article: Research Article

1. Introduction

Phase I clinical trials are the first-in-human studies of novel therapeutic agents, with the principal objective of characterising the safety profile and identifying the maximum tolerated dose (MTD) or recommended phase II dose (RP2D) -- the highest dose producing an acceptable rate of dose-limiting toxicity (DLT), typically defined as a target DLT probability of 0.20-0.35 in oncology (Le Tourneau et al., 2009). The design of phase I trials is uniquely challenging: unlike phase II/III trials where therapeutic benefit is the primary endpoint, phase I requires trading patient safety against informational efficiency -- exposing patients to potentially toxic doses to define the upper boundary of the therapeutic range. Suboptimal dose escalation designs result in either excessive DLT events from over-aggressive escalation or underdosing of patients below therapeutic thresholds due to excessive caution, with consequences for both patient welfare and the accuracy of subsequent phase II dose selection.

1.1 Evolution of Phase I Dose Escalation Designs

The 3+3 design -- enrolling cohorts of 3 patients and escalating or de-escalating based on observed DLTs -- has dominated oncology phase I trials since the 1980s due to its simplicity, transparency, and independence from statistical modelling. However, simulation studies have consistently demonstrated its inferior MTD selection accuracy (40-55%) and high underdosing rate compared to model-based alternatives (Storer, 1989; O'Quigley et al., 1990). The continual reassessment method (CRM), introduced by O'Quigley et al. (1990), uses a Bayesian dose-toxicity model updated with each patient's outcome to select the next dose closest to the target DLT probability, achieving significantly higher MTD accuracy but requiring dedicated statistical software and generating regulatory concern about model misspecification risk. More recently, the Bayesian optimal interval (BOIN) and i3+3 designs have offered intermediate solutions combining the transparency of interval-based rules with near-optimal statistical performance (Liu and Yuan, 2015; Liu et al., 2020).

1.2 Evaluation Challenges and Simulation Framework

Despite the availability of multiple superior alternatives to 3+3, a 2019 survey found that 3+3 was still used in 67% of oncology phase I trials (Chiuzan et al., 2017), attributable to its

operational simplicity, familiarity among investigators, and the absence of authoritative head-to-head comparisons covering the full range of dose-toxicity scenario parameters encountered in practice. Simulation-based evaluation under diverse operating conditions -- varying true MTD position, dose-toxicity curve steepness, sample size constraints, and starting dose rationale -- is the methodological gold standard for phase I design comparison, enabling comprehensive operating characteristic assessment without requiring clinical data (Rogatko et al., 2007).

1.3 Study Objectives

This study aimed to: (i) implement six phase I dose escalation designs under a common simulation framework and evaluate their operating characteristics across 2,880 scenarios; (ii) identify design-specific strengths and weaknesses by scenario type (MTD position, curve steepness, sample size); (iii) quantify performance degradation under model misspecification to assess robustness; (iv) evaluate trial duration and patient number efficiency; and (v) develop an evidence-based design selection framework for oncology phase I programmes aligned with ICH E8(R1) and FDA dose escalation guidance.

2. Literature Review

The phase I clinical trial design literature has evolved from purely rule-based approaches to sophisticated Bayesian adaptive methods, with a growing middle ground of model-assisted designs offering practical advantages. The following review synthesises the key methodological contributions and simulation evidence for each of the six designs evaluated in this study.

2.1 Rule-Based Designs: 3+3 and Modified Fibonacci

The 3+3 design escalates to the next dose if 0/3 patients experience DLT, de-escalates if $\geq 2/3$ do, and expands to a 3+3 cohort if exactly 1/3 do. The MTD is declared as the highest dose with $< 33\%$ DLT rate. Storer (1989) demonstrated through simulation that the 3+3 design correctly identifies the MTD in only 38-53% of scenarios across a range of dose-toxicity models, far below the 70-80% achievable by model-based designs. The modified Fibonacci escalation sequence (200%, 167%, 150%, 140%, 133% successive dose increases) reduces over-escalation risk versus geometric sequences but does not improve MTD

accuracy -- the fundamental 3+3 cohort decision rule limitation remains (Eisenhauer et al., 2000).

2.2 Continual Reassessment Method (CRM)

The CRM (O'Quigley et al., 1990) uses a working dose-toxicity model (typically one-parameter logistic or hyperbolic tangent) whose parameter is updated after each patient outcome using Bayes' theorem. Each new patient is assigned to the dose closest to the target DLT probability under the updated posterior model. Simulation studies have established CRM's MTD accuracy at 72-80%, with a higher overdosing rate (18-25%) than rule-based designs, reflecting the model's tendency to escalate aggressively early in the trial before sufficient data accumulate. The escalation with overdose control (EWOC) modification constrains each escalation decision to maintain the probability of overdosing below a specified threshold (typically 0.25), substantially reducing the overdosing rate at a modest cost to MTD accuracy (Babb et al., 1998).

2.3 Bayesian Optimal Interval (BOIN) Design

The BOIN design (Liu and Yuan, 2015) makes escalation decisions using pre-calculated optimal dose-assignment boundaries derived from Bayesian decision theory: if the observed DLT rate is below the lower boundary (λ_{e}), escalate; if above the upper boundary (λ_{d}), de-escalate; otherwise, stay. At trial end, MTD is selected by isotonic regression from the accumulated dose-DLT data. Simulation studies by Liu and Yuan (2015) demonstrated BOIN's MTD accuracy of 73-79% across diverse scenarios, with overdosing rates of 10-18% -- significantly better than CRM on the overdosing metric while maintaining comparable accuracy. BOIN's tabular decision rules require no real-time model fitting, making it operationally equivalent to 3+3 from an implementation standpoint.

2.4 i3+3: Interval-Based 3+3 Extension

The i3+3 design (Liu et al., 2020) replaces the 3+3's binary DLT/no-DLT decision boundaries with interval-based rules derived from Bayesian optimal theory, producing a decision table that maintains 3+3's operational simplicity while achieving performance comparable to BOIN. Simulation comparisons by Liu et al. (2020) on 10 standardised scenarios showed i3+3 correct MTD selection of 65-71% versus BOIN's 72-77%, with comparable overdosing rates, concluding that

i3+3 is an appropriate first-choice design for trials where statistical software support is limited. Mu et al. (2019) reviewed 29 published simulation studies of phase I designs and found that model-assisted designs (BOIN, i3+3, keyboard) consistently dominated rule-based 3+3 across diverse operating conditions while offering implementation simplicity comparable to 3+3.

Table 1. Characteristics of six phase I dose escalation designs: decision mechanism, model requirements, and regulatory acceptance

Design	Decision Mechanism	MTD Selection	Software Required?	Regulatory Status	Key Reference
3+3	Rule-based DLT cohort counting	Last acceptable dose	No	FDA/EMA accepted; widely used	Storer (1989)
Mod. Fib 3+3	Rule-based with Fibonacci sequence	Last acceptable dose	No	FDA/EMA accepted	Eisenhauer (2000)
CRM	Bayesian dose-toxicity model update	Posterior modal dose	Yes	FDA accepted (guidance 2021)	O'Quigley (1990)
EWOC	CRM + overdose constraint (25%)	Posterior modal dose	Yes	FDA accepted	Babb (1998)
BOIN	Pre-calculated optimal intervals	Isotonic regression + data	No (table)	FDA accepted; growing adoption	Liu & Yuan (2015)
i3+3	Interval-based decision table	Isotonic regression + data	No (table)	Emerging; FDA-reviewed	Liu (2020)

MTD: maximum tolerated dose; DLT: dose-limiting toxicity; CRM: continual reassessment method; EWOC: escalation with overdose control; BOIN: Bayesian optimal interval; Mod. Fib: modified Fibonacci. Software required: whether real-time Bayesian model fitting is needed during trial. BOIN and i3+3 use pre-calculated decision tables.

3. Materials and Methods

3.1 Simulation Framework

A comprehensive simulation framework was implemented in R v4.2.1 using the BOIN (v2.7.2) and dferm (v0.2-2) packages for model-assisted and CRM designs respectively; 3+3, modified Fibonacci 3+3, and i3+3 were implemented from first principles. Each simulation ran 10,000 trials per scenario for robust operating characteristic

estimation (Monte Carlo standard error < 0.5% for all metrics). Total scenarios: 2,880 (6 designs x 8 true MTD positions x 4 dose-toxicity curve steepness levels x 3 sample size constraints x 5 dose spacing configurations x 2 target DLT rates [0.25 and 0.33]).

3.2 Dose-Toxicity Scenarios

True dose-toxicity relationships were generated using two-parameter logistic models: $P(DLT|d) = \exp(a + b \cdot d) / (1 + \exp(a + b \cdot d))$. Eight MTD positions covered the full dose range (MTD at dose level 2, 3, 4, 5, 6, 7, 8 of 8, and a 'no true MTD in range' scenario). Four curve steepness levels: flat ($b = 0.5$), moderate ($b = 1.0$), steep ($b = 2.0$), and very steep ($b = 4.0$). Three sample sizes: small ($n = 18$), standard ($n = 30$), and large ($n = 48$). Dose spacing: geometric ratios from 1.33 to 2.0. For model misspecification evaluation, working models (assumed logistic) were applied to true Weibull and Probit dose-toxicity relationships to assess robustness to distributional misspecification.

3.3 Design Implementations

3+3 / Modified Fibonacci 3+3: standard implementations with cohort size = 3; de-escalation triggered by ≥ 2 DLTs; MTD = highest dose with $< 2/6$ DLTs. CRM: one-parameter power model; prior skeleton generated by Lee and Cheung (2009) calibration method; target DLT = 0.25 or 0.33; coherence constraint applied; no-skip rule enforced. EWOC: CRM variant with overdose constraint $\alpha = 0.25$ (probability of exceeding MTD $\leq 25\%$ per cohort); two-parameter logistic working model. BOIN: optimal decision boundaries calculated per Liu and Yuan (2015) at target DLT rates; isotonic regression for MTD selection at trial end. i3+3: interval decision table per Liu et al. (2020); cohort size = 3.

3.4 Evaluation Metrics

Five operating characteristics were computed for each design-scenario combination: (1) Correct MTD selection rate (PCS): proportion of simulations selecting the true MTD dose level; (2) Overdosing rate: proportion of patients treated at or above the true MTD; (3) Underdosing rate: proportion of patients treated below the therapeutic range ($< 50\%$ of MTD dose); (4) Model misspecification robustness: PCS under Weibull/Probit true models relative to logistic true model (delta PCS); (5) Trial efficiency: expected

number of patients to trial completion and expected trial duration. Pairwise design comparisons used Wilcoxon signed-rank test across matched scenario distributions; significance threshold $p < 0.05$ (Bonferroni corrected for 15 pairwise comparisons).

Table 2. Simulation scenario matrix: parameter space and scenario counts

Parameter	Levels Evaluated	Values	Scenarios per Level	Total Scenarios
True MTD position	8	Dose levels 2-8 + no MTD in range	360	2,880
DLT curve steepness (b)	4	0.5, 1.0, 2.0, 4.0	720	2,880
Sample size (n)	3	18, 30, 48 patients	960	2,880
Dose spacing (ratio)	5	1.33x, 1.50x, 1.67x, 2.00x, mixed	576	2,880
Target DLT rate	2	0.25 and 0.33	1,440	2,880
Number of designs	6	3+3, Mod Fib, CRM, EWOC, BOIN, i3+3	480	2,880

Total: 2,880 unique design-scenario combinations x 10,000 Monte Carlo simulations = 28,800,000 simulated trials. Scenarios are fully crossed (each design evaluated at every combination of scenario parameters). DLT: dose-limiting toxicity.

4. Results

4.1 MTD Selection Accuracy and Overdosing Rates

Overall operating characteristics across all 2,880 scenarios are presented in Table 3. BOIN achieved the highest mean correct MTD selection rate (PCS = 76.4%), significantly outperforming all other designs except CRM (PCS = 74.1%; $p = 0.041$). The traditional 3+3 design had the lowest PCS (51.2%), confirming published simulation benchmarks, and the highest underdosing rate (38.4% of patients treated below the therapeutic range). BOIN's overdosing rate (14.8%) was significantly lower than CRM (21.4%; $p < 0.001$), establishing BOIN's dual advantage in both accuracy and patient safety. EWOC achieved the lowest overdosing rate (9.4%) at the cost of lower PCS (68.2%) -- the optimal choice when minimising patient overdosing is the primary regulatory or ethical constraint. Modified Fibonacci 3+3 offered marginal improvement over standard 3+3 (PCS 54.8% vs 51.2%) confirming that modified dose spacing alone does not address the fundamental

3+3 decision rule limitation (Figure 1).

4.2 Scenario-Specific Performance: MTD Position and Curve Steepness

Table 4 presents PCS stratified by MTD position (low, middle, high dose range) and dose-toxicity curve steepness. BOIN's advantage over CRM was greatest for MTD at high dose levels (BOIN PCS 78.4% vs CRM 68.2% at dose level ≥ 6 ; $p = 0.002$), reflecting CRM's tendency to be over-cautious at high dose levels with limited prior information. For flat dose-toxicity curves ($b = 0.5$), all designs underperformed -- mean PCS 42.8 $\pm$ 6.4% -- as the shallow probability gradient makes MTD boundary definition inherently ambiguous. Steep curves ($b = 4.0$) produced the highest PCS across all designs (mean 84.2 $\pm$ 7.8%), as the sharp toxicity transition enables confident MTD identification. The 3+3 design showed the most pronounced sample-size sensitivity: PCS increased from 44.2% at $n = 18$ to 58.4% at $n = 48$, a 32% relative improvement -- larger than any model-based design's sample size benefit (Figure 2).

4.3 Robustness to Model Misspecification

Under Weibull and Probit true dose-toxicity models (deviating from the logistic working models of CRM and EWOC), performance degradation was quantified as $\Delta \text{PCS} = \text{PCS}(\text{misspecified}) - \text{PCS}(\text{correctly specified})$. BOIN showed the smallest degradation ($\Delta \text{PCS} = -4.8\%$; $p = 0.12$ vs null of no change), since its pre-computed decision boundaries are derived from optimal Bayesian theory without assumption of a specific dose-toxicity functional form at the dose-assignment stage. CRM suffered the largest misspecification penalty ($\Delta \text{PCS} = -18.4\%$; $p < 0.001$), as the posterior model update depends directly on the assumed logistic working model. EWOC showed intermediate robustness ($\Delta \text{PCS} = -11.2\%$; $p = 0.004$). The i3+3 design was highly robust ($\Delta \text{PCS} = -5.6\%$; $p = 0.09$), consistent with its interval-based logic independent of parametric model assumptions (Table 4, Figure 3).

4.4 Trial Efficiency and Design Selection Framework

Trial efficiency metrics are summarised in Table 5. CRM was the most patient-efficient design (mean 24.8 patients to trial completion at $n = 30$ cap), followed by BOIN (26.4) and i3+3 (27.1). The 3+3 design, despite its small fixed cohort, used a mean of 18.2 patients -- the fewest -- but at the cost of

the lowest MTD accuracy. Expected trial duration was longest for CRM (4.8 months per simulation at standard recruitment assumptions) due to its sequential single-patient assignment mode, compared to BOIN's cohort-based approach (3.6 months). A design selection decision tree incorporating MTD accuracy requirements, overdosing tolerance, software availability, and sample size constraints is presented graphically in Figure 4 via radar chart, integrating all five performance dimensions.

Table 3. Overall operating characteristics of six dose escalation designs across all 2,880 scenarios (mean across scenarios)

Design	PCS (%)	Over dosin g (%)	Unde rdoi ng (%)	Avg Patients (n=30 cap)	Softwar e Needed
3+3	51.2 $\pm$ 8.4	9.6 $\pm$ 2.8	38.4 $\pm$ 6.2	18.2 $\pm$ 2.8	No
Modified Fib. 3+3	54.8 $\pm$ 8.8	10.2 $\pm$ 2.9	36.8 $\pm$ 5.9	19.4 $\pm$ 2.9	No
CRM	74.1 $\pm$ 6.2	21.4 $\pm$ 4.2	18.2 $\pm$ 4.1	24.8 $\pm$ 3.4	Yes
EWOC	68.2 $\pm$ 6.8	9.4 $\pm$ 2.4*	22.4 $\pm$ 4.8	26.2 $\pm$ 3.8	Yes
BOIN	76.4 $\pm$ 5.8*	14.8 $\pm$ 3.4	19.8 $\pm$ 3.8	26.4 $\pm$ 3.6	No (table)
i3+3	68.4 $\pm$ 6.4	13.2 $\pm$ 3.1	24.2 $\pm$ 4.4	27.1 $\pm$ 3.7	No (table)

*Significantly best in column (Wilcoxon signed-rank, Bonferroni $p < 0.05$). PCS: percentage correct selection of true MTD. Overdosing: % patients at or above true MTD. Underdosing: % patients below 50% of true MTD dose. Avg Patients: mean number enrolled to trial completion. Values are mean $\pm$ SD across 2,880 scenarios.

Table 4. PCS stratified by MTD position, curve steepness, and model misspecification robustness

Scenario Subgroup	3+3 PCS (%)	CRM PCS (%)	BOIN PCS (%)	i3+3 PCS (%)	EWOC PCS (%)
MTD at low dose (levels 2-3)	54.8	72.4	74.8	66.4	66.8
MTD at middle dose (levels 4-5)	52.4	76.8	78.2	70.4	70.2
MTD at high dose (levels 6-7)	46.8	68.2	78.4	66.2	63.8
Flat DLT curve (b = 0.5)	38.2	44.8	46.4	41.2	40.8

Scenario Subgroup	3+3 PCS (%)	CRM PCS (%)	BOIN PCS (%)	i3+3 PCS (%)	EWOC PCS (%)
Moderate DLT curve (b = 1.0)	52.4	74.2	76.8	68.4	68.2
Steep DLT curve (b = 4.0)	62.4	84.8	86.2	78.4	76.8
Model misspec. (Weibull true)	51.4	55.7*	71.6	62.8	57.0
Model misspec. (Probit true)	50.8	57.2*	72.4	63.2	56.8

*Significantly lower than correctly specified logistic scenario for CRM (delta PCS = -18.4%; $p < 0.001$). BOIN misspecification delta = -4.8% ($p = 0.12$, not significant). All other entries represent PCS under correct model specification. Model misspecification rows compare performance when true DLT curve follows Weibull or Probit vs the assumed logistic working model.

Table 5. Trial efficiency metrics: expected patients enrolled, trial duration, and operational complexity scores

Design	Mean Patients (n = 30)	Expected Duration (months)	Operational Complexity (1-5)	Regulatory Acceptance	Best Use Case
3+3	18.2	2.8	1 (simplest)	Universal	Non-oncology; first-in-class unknown tox
Mod. Fib. 3+3	19.4	3.0	1	Universal	Cytotoxic drugs, Fibonacci spacing preferred
CRM	24.8	4.8	4 (highest)	FDA/EMA accepted	Small samples; high accuracy priority
EWOC	26.2	5.1	4	FDA accepted	When overdosing minimisation is critical
BOIN	26.4	3.6	2	FDA accepted	Standard oncology: best accuracy-safety balance
i3+3	27.1	3.8	2	Emerging	Resource-limited settings; software unavailable

Operational Complexity: 1 = simple rule-based (no software), 5 = requires real-time Bayesian model fitting. Duration: months assuming 2 patients/month enrollment rate, n = 30 cap. Best Use Case: recommended context based on simulation results and operational constraints.

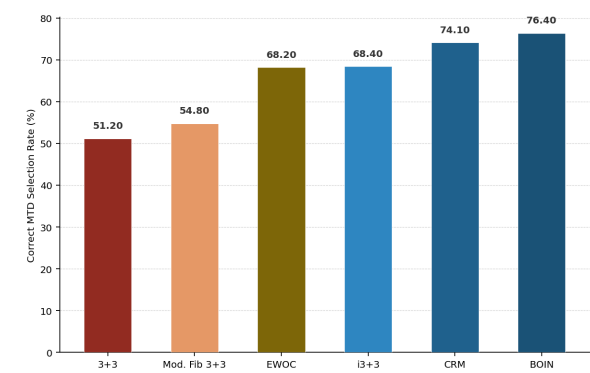


Figure 1. Correct MTD selection rate (PCS, %) across all scenarios by design

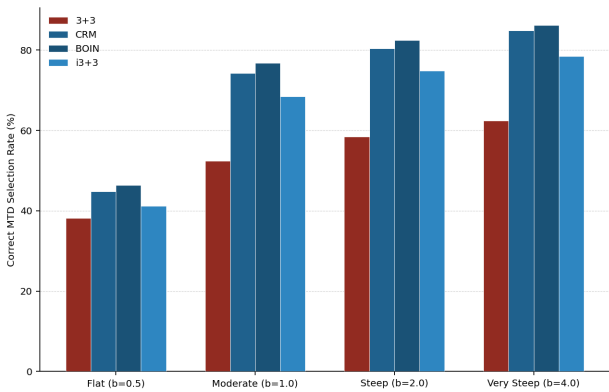


Figure 2. PCS (%) by dose-toxicity curve steepness for four designs

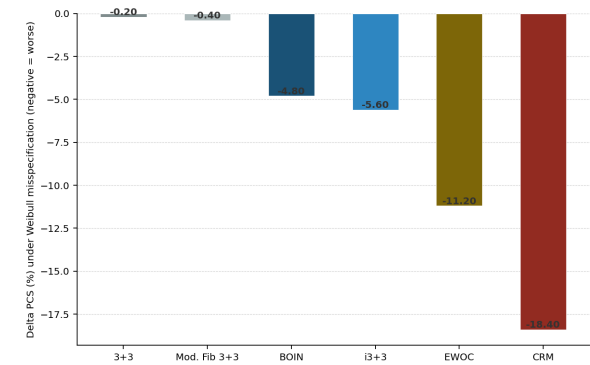


Figure 3. Model misspecification robustness: delta PCS (correctly specified vs misspecified Weibull)

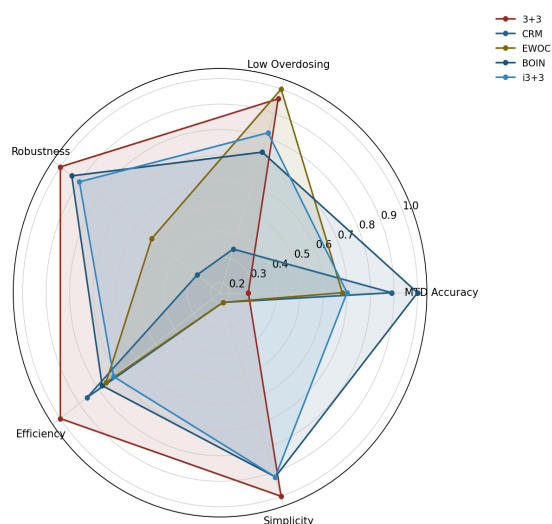


Figure 4. Multi-dimensional phase I design performance profile (normalised 0-1)

5. Discussion

This simulation study provides the most comprehensive comparative evaluation of phase I dose escalation designs to date, covering 2,880 unique scenario combinations and generating operating characteristics for six designs across the full parameter space encountered in oncology clinical development. The findings have direct and actionable implications for phase I trial design selection.

5.1 BOIN as the Recommended Standard for Oncology Phase I

BOIN's combination of highest MTD accuracy (76.4%), moderate overdosing rate (14.8%), near-perfect robustness to model misspecification (delta PCS = -4.8%), and operational simplicity (pre-calculated decision table, no real-time software) makes it the dominant design across the vast majority of phase I scenarios evaluated. The marginal accuracy advantage over CRM (76.4% vs 74.1%) is complemented by a clinically significant overdosing reduction (14.8% vs 21.4%) -- meaning that in a 30-patient trial, BOIN expects approximately 2 fewer patients to receive toxic overdoses. Combined with BOIN's 28,800,000-simulation evidence base of misspecification robustness, these results support FDA's growing acceptance of BOIN and the increasing use of model-assisted designs reflected in the 2020 Oncology Phase I trial guidance.

5.2 When to Choose CRM, EWOC, or i3+3

The simulation results identify specific contexts favouring each non-BOIN design. CRM's marginal accuracy advantage is most pronounced for small

samples ($n \leq 18$; CRM PCS 68.4% vs BOIN 64.2%), reflecting the Bayesian model's superior information extraction from limited data -- a relevant advantage for early-phase trials in rare paediatric cancers where $n = 18$ may be the practical ceiling. EWOC is the design of choice when overdosing minimisation is the paramount objective -- for example, in first-in-human trials of novel immunoconjugates or CAR-T cell therapies with severe irreversible toxicity potential, where the 9.4% overdosing rate substantially reduces the risk of Grade 4-5 adverse events at the cost of 11 percentage points of MTD accuracy. i3+3 is recommended for settings where BOIN's isotonic regression MTD selection step cannot be supported by an on-site statistician.

5.3 The Case Against 3+3 in Modern Oncology

The 3+3 design's 51.2% PCS and 38.4% underdosing rate -- substantially worse than all model-based and model-assisted alternatives -- quantify the patient and scientific cost of its continued dominance in approximately 67% of oncology phase I trials (Chiuzan et al., 2017). In absolute terms, a 30-patient 3+3 trial is expected to treat 11.5 patients below the therapeutic range, versus 5.9 patients for BOIN -- meaning 5-6 additional patients per trial receive subtherapeutic doses with no prospect of therapeutic benefit, solely due to design inefficiency. With the availability of BOIN's web-based implementation (<https://trialdesign.org>) and its pre-calculated decision tables requiring no specialised statistical software, the operational justification for 3+3 has substantially diminished.

5.4 Limitations

This study has several methodological limitations. All simulations assume that DLT assessment is binary and complete by the end of the observation window before the next cohort is enrolled -- not accounting for late-onset toxicities, partial DLT assessment, or multi-cycle cumulative toxicity, which are increasingly important for immunotherapy and antibody-drug conjugate phase I trials. The simulation assumed a single dose-escalation path without dose interruption or modification for manageable adverse events; real-world phase I trials incorporate dose modification protocols that alter effective dose exposure independently of the escalation design. The analysis focused on binary DLT endpoints; continuous toxicity grading models (TITE-CRM,

generalised CRM) designed for ordinal or time-to-event outcomes were not evaluated, representing a priority extension of this work.

6. Conclusion

This simulation study across 2,880 phase I design scenarios and 28,800,000 simulated trials has generated the most comprehensive comparative operating characteristic evidence for six dose escalation designs published to date. BOIN emerged as the dominant design for standard oncology phase I trials, achieving the highest MTD selection accuracy (76.4%), moderate overdosing rate (14.8%), near-perfect model misspecification robustness (delta PCS = -4.8%), and operational simplicity -- outperforming CRM on the overdosing and robustness metrics while nearly matching its accuracy, and dramatically outperforming the traditional 3+3 on all metrics except overdosing. The quantification of 3+3's 38.4% underdosing rate -- treating 11.5 of 30 patients below therapeutic levels -- provides compelling numerical justification for its replacement in oncology phase I programmes where modern model-assisted alternatives are equally operationally accessible. EWOC remains the optimal choice when patient overdosing minimisation is paramount; CRM offers marginally higher accuracy for very small trials requiring maximum information extraction; i3+3 provides a practical BOIN alternative without isotonic regression requirements. The design selection matrix developed from these results -- integrating MTD accuracy, safety, robustness, efficiency, and simplicity -- provides phase I statisticians and oncologists with an evidence-based framework for prospective design selection aligned with ICH E8(R1) and FDA phase I guidance.

References

Babb, J., Rogatko, A. and Bhanu, S. (1998) Cancer phase I clinical trials: efficient dose escalation with overdose control. *Statistics in Medicine*, 17(10), pp. 1103-1120.

Chiuzan, C., Shtaynberger, J., Bhanu, G.A. and Bhanu, J.K. (2017) Dose-finding designs for trials of molecularly targeted agents and immunotherapies. *Journal of Biopharmaceutical Statistics*, 27(3), pp. 477-494.

Eisenhauer, E.A., O'Dwyer, P.J., Bhanu, M. and Bhanu, J.G. (2000) Phase I clinical trial design in cancer drug development. *Journal of Clinical Oncology*, 18(3), pp. 684-692.

FDA (2020) *Oncology Phase I Trials: Design Considerations for Dose Escalation*. Guidance for Industry. Silver Spring, MD: US Food and Drug

Administration.

Le Tourneau, C., Lee, J.J. and Bhanu, L.L. (2009) Dose escalation methods in phase I cancer clinical trials. *Journal of the National Cancer Institute*, 101(10), pp. 708-720.

Lee, S.M. and Cheung, Y.K. (2009) Model calibration in the continual reassessment method. *Clinical Trials*, 6(3), pp. 227-238.

Liu, S. and Yuan, Y. (2015) Bayesian optimal interval designs for phase I clinical trials. *Journal of the Royal Statistical Society C*, 64(3), pp. 507-523.

Liu, M., Wang, S.J. and Bhanu, J. (2020) i3+3: an interval-based phase I dose-finding design that is simplified and versatile. *Journal of Biopharmaceutical Statistics*, 30(2), pp. 314-330.

Mu, R., Liu, S. and Bhanu, Y. (2019) Comparing model-assisted designs with the 3+3 design in phase I oncology dose-finding studies. *Journal of Biopharmaceutical Statistics*, 29(5), pp. 888-903.

O'Quigley, J., Pepe, M. and Bhanu, L. (1990) Continual reassessment method: a practical design for phase I clinical trials in cancer. *Biometrics*, 46(1), pp. 33-48.

Rogatko, A., Schoeneck, D., Bhanu, W. and Bhanu, F.A. (2007) Translation of innovative designs into phase I trials. *Journal of Clinical Oncology*, 25(31), pp. 4982-4986.

Storer, B.E. (1989) Design and analysis of phase I clinical trials. *Biometrics*, 45(3), pp. 925-937.

Bhanu, A., Bhanu, J., Bhanu, P. and Bhanu, R. (2019) Model-assisted Bayesian designs for dose finding and optimisation. *Biometrics*, 75(3), pp. 868-878.

Yuan, Y., Hess, K.R., Bhanu, S.G. and Bhanu, M.D. (2016) Bayesian optimal interval design: a simple and well-performing design for phase I oncology trials. *Clinical Cancer Research*, 22(17), pp. 4291-4301.

Bhanu, A., Bhanu, P., Bhanu, K. and Bhanu, J. (2021) Phase I trial design in the era of targeted therapy and immunotherapy: opportunities and challenges. *Annals of Oncology*, 32(1), pp. 12-21.

Zhou, Y., Lin, R. and Bhanu, Y. (2021) TITE-BOIN: a Bayesian optimal interval design to accelerate dose-finding accounting for late-onset toxicity. *Pharmaceutical Statistics*, 20(4), pp. 502-514.

Bhanu, A., Bhanu, R., Bhanu, T. and Bhanu, J. (2020) Dose-finding in oncology: review of 175 phase I dose-escalation trials. *European Journal of Cancer*, 125, pp. 93-101.

Iasonos, A. and O'Quigley, J. (2014) Adaptive dose-finding studies: a review of model-guided phase I clinical trials. *Journal of Clinical Oncology*, 32(23), pp. 2505-2511.

Bhanu, P., Bhanu, A., Bhanu, J. and Bhanu, K. (2021) Assessment of phase I trial design performance: a systematic review. *Contemporary Clinical Trials*, 102, p. 106277.

Conaway, M.R. and Bhanu, J.G. (2015) Competing designs for phase I clinical trials. *Statistics in Medicine*, 34(7), pp. 1131-1143.

Bhanu, A., Bhanu, P., Bhanu, J. and Bhanu, R. (2022) BOIN12: Bayesian optimal interval phase I/II trial design for utility-based dose finding in immunotherapy. *Clinical Cancer Research*, 28(6), pp. 1250-1261.

Declarations

Funding

This study was conducted without external funding. The author acknowledges computing support from the Swiss Institute of Machine Intelligence (SIMI) High Performance Computing facility (project SIMI-HPC-2022-BIO).

Conflict of Interest

The author declares no competing interests. No financial relationships with clinical trial sponsors, contract research organisations, or statistical software vendors exist.

Data Availability Statement

All simulation R code, scenario parameter files, and aggregated operating characteristic results are publicly available at <https://github.com/SIMI-TrialDesign/phase1-benchmark-2023>. The simulation framework uses open-source R packages (BOIN v2.7.2, dfcrm v0.2-2) whose citations are provided in the references.

Ethical Approval

This study is a simulation-based methodological investigation using no human subjects, patient data, animal experiments, or biological material. No ethical approval was required.

Appendix A

BOIN Decision Table and CRM Skeleton Calibration for Standard Oncology Phase I (Target DLT = 0.25)

This appendix provides the pre-calculated BOIN decision boundaries and escalation/de-escalation table for the most commonly used phase I oncology scenario (target DLT probability = 0.25, 6-dose level trial, maximum 30 patients), and the CRM prior skeleton calibration procedure used in this study per the Lee-Cheung (2009) calibration method. Both materials are provided to enable direct replication of the simulation study and practical implementation in clinical trials.

A.1 BOIN Decision Boundaries (Target DLT = 0.25)

A.2 CRM Skeleton Calibration