Statistical Operating Characteristics of Current Early
Phase Dose Finding Designs with Toxicity and Efficacy
in Oncology

Hao Sun¹, Hsin-Yu Lin¹, Jieqi Tu², Revathi Ananthakrishnan¹, and Eunhee Kim *¹

¹Global Biometrics & Data Sciences, Bristol Myers Squibb, New Jersey, USA

²Division of Epidemiology and Biostatistics, School of Public Health, University of Illinois at Chicago, Illinois, USA

Abstract: Traditional phase I dose finding cancer clinical trial designs aim to determine the maximum tolerated dose (MTD) of the investigational cytotoxic agent based on a single toxicity outcome, assuming a monotone dose-response relationship. However, this assumption might not always hold for newly emerging therapies such as immuno-oncology therapies and molecularly targeted therapies, making conventional dose finding trial designs based on toxicity no longer appropriate. To tackle this issue, numerous early phase dose finding clinical trial designs have been developed to identify the optimal biological dose (OBD), which takes both toxicity and efficacy outcomes into account. In this article, we review the current model-assisted dose finding designs, BOIN-ET, BOIN12, UBI, TEPI-2,

^{*}Corresponding author: *Eunhee Kim, Global Biometrics & Data Sciences, Bristol Myers Squibb, New Jersey, USA. Email: Eunhee.Kim@bms.com

PRINTE, STEIN, and uTPI to identify the OBD and compare their operating characteristics. Extensive simulation studies and a case study using a CAR T-cell therapy phase I trial have been conducted to compare the performance of the aforementioned designs under different possible dose-response relationship scenarios. The simulation results demonstrate that the performance of different designs varies depending on the particular dose-response relationship and the specific metric considered. Based on our simulation results and practical considerations, STEIN, PRINTE, and BOIN12 outperform the other designs from different perspectives.

Keywords: Dose finding, early phase trial design, model-assisted design, optimal biological dose, random toxicity and efficacy probabilities

1 Introduction

Finding the right dose is crucial in early phase cancer clinical trials due to the safety concerns of adverse events induced by any oncology drug under investigation. There are numerous real world examples where the dose on the label of an approved oncology drug is not the dose administered in practice ^{17;18}. This implies that early phase dose finding oncology trials need to be carefully designed to determine the optimal dose of the investigational agent that can be used in late phase trials and the target population. In this regard, many early phase oncology dose finding methods have been proposed in the past few years.

The current existing dose finding methods can be categorized into algorithmic or rule-based, model-assisted, and model-based designs³⁶. The 3+3 design is an example of a rule-based design, which is commonly used because of its simplicity and ease of implementation. However, due to the limitations of the 3+3 design¹, many model-assisted and model-based designs were proposed to address them. For example, some popular model-assisted dose finding designs include the modified toxicity probability interval (mTPI) design⁸, the mTPI-2 design^{9;34;20}, and the Bayesian optimal interval (BOIN) design^{16;35}. Commonly used model-based dose finding designs include the continual reassessment method (CRM) design⁷, the escalation with overdose control (EWOC) design², and the Bayesian logistic regression model (BLRM) design¹⁹. All these designs only consider the toxicity of the drug to estimate the maximum tolerated dose (MTD). This works well for chemotherapies where both the efficacy and toxicity of these agents increase with dose.

However, traditional dose finding designs based on a single toxicity outcome may not be appropriate for immuno-oncology drugs and molecularly targeted therapies. For these novel anticancer agents, although the toxicity of the drug increases with dose, the efficacy of the drug may not always increase with dose. Additionally, in many immunotherapies, toxicities are usually low or moderate grade, preventing the observation of a dose-limiting toxicity (DLT). Also, severe toxicities are rare and often delayed in subsequent treatment cycles, which may prevent the MTD from being reached ^{21;32}. Hence, it is important to identify

the dose of such an oncology drug that is efficacious as well as tolerable, which is referred to as the optimal biological dose (OBD). Recognizing the importance of selecting the right dose, the FDA has recently launched the initiative Project Optimus⁴ to reform the dose optimization and dose selection paradigm in oncology drug development. There has been an ongoing effort to develop early phase clinical trial designs incorporating both toxicity and efficacy outcomes to determine the OBD. Model-based dose finding designs aimed at OBD detection include the Eff-Tox design³¹, the local logistic model (L-logistic) design³⁷, the change-point logistic (CP-logistic) model design²³, and the Bayesian dynamic model (Bdynamic) design ¹⁵. Eff-Tox ³¹ is an adaptive model-based design based on trade-offs between the treatment efficacy and toxicity probabilities. It is particularly sensitive to the efficacytoxicity trade-off contour. The L-logistic design³⁷ utilizes Bayesian logistic regression to model the dose-response curve. The CP-logistic method²³ introduces a change-point logistic model to account for the correlation between efficacy and toxicity outcomes. The B-dynamic method¹⁵ leverages a flexible Bayesian dynamic model that considers both toxicity and efficacy and borrows information across different dose levels. However, a limitation of the B-dynamic method is that it considers only monotone dose-response relationships.

Furthermore, an increasing number of model-assisted designs extend BOIN and mTPI by accounting for efficacy outcomes in the dose finding procedures. Some popular extensions of the BOIN design are BOIN-ET²⁷, BOIN12¹³, UBI¹⁰, and STEIN¹², while extensions of the mTPI design include TEPI⁹, TEPI-2¹⁰, PRINTE¹⁴, and uTPI²⁴. BOIN-ET derives optimal probability intervals for efficacy and toxicity by minimizing the joint probabilities of incorrect dosing decisions to determine the OBD. UBI, BOIN12, and uTPI use utility functions to assess the toxicity-efficacy trade-off and select the dose with the highest utility as the OBD. Unlike UBI, which bases dosing decisions on data from the current dose, BOIN12 and uTPI adaptively compare multiple doses to choose the optimal dose for the next group of patients. TEPI, TEPI-2, and UBI designs rely on pre-specified decision tables that map two-dimensional toxicity and efficacy probability intervals to a set of dosing decisions. On

the other hand, uTPI proposes a chessboard design that utilizes a two-dimensional square grid of toxicity and desirability to guide dosing decisions.

Multiple model-assisted designs mentioned above, such as BOIN-ET, BOIN12, and PRINTE, demonstrated that they are more robust than the Eff-Tox design as they do not make any model assumptions regarding the dose toxicity and efficacy curves. BOIN-ET was shown a higher OBD selection probability than CP-logistic ²⁷. Lin & Yin ¹² observed that STEIN outperforms both L-logistic and B-dynamic in OBD selection percentage. TEPI-2, UBI, PRINTE, BOIN-ET, and STEIN demonstrated superior performance compared to TEPI in terms of OBD selection probability and patient allocation at the optimal doses ^{10;14;33}. Yamaguchi et al. ³³ presented evidence that BOIN-ET and STEIN have similar OBD selection probabilities. Li et al. 11 found that BOIN-ET performs better than BOIN12 in more than half of the scenarios. Shi et al. 25 compared BOIN-ET, BOIN12, TEPI, PRINTE, Joint3+3, STEIN, uTPI, and Eff-Tox by applying the same utility score approach for OBD selection across all designs. They revealed that BOIN12 and uTPI outperformed other designs in OBD selection accuracy and minimizing poor dose allocation. Additionally, STEIN demonstrated the most effective overdose control. However, as discussed by Shi et al., the designs, except for BOIN12 and uTPI, can underperform when the OBD is defined based on the utility score method²⁵. Furthermore, they did not consider scenarios where the true OBD does not exist.

In this article, we compare seven model-assisted early phase dose finding oncology designs: BOIN-ET, BOIN12, TEPI-2, UBI, PRINTE, STEIN, and uTPI. Several model-based designs, including Eff-Tox, L-logistic, CP-logistic, and B-dynamic, as well as the model-assisted design TEPI, are not discussed in this article because their performance has been revealed to be less competitive compared to one or more of the aforementioned model-assisted designs. To the best of our knowledge, the performance of the selected seven designs, using their original dose exploration algorithms and OBD estimation approaches, has not been fully examined in the literature. The goal of our article is to provide comprehensive guid-

ance on selecting an appropriate early phase design under different scenarios in practice and to offer user-friendly, interactive software that can be used to implement these designs.

This article is organized as follows. Section 2 introduces some basic design parameters, OBD definitions, and safety and futility rules. Key concepts of different early phase dose finding designs are described in Section 3. Extensive simulation studies and sensitivity analyses are presented in Section 4, and a case study is featured in Section 5. Section 6 provides the design guidance and software information. Further discussion is provided in Section 7.

2 Notations

Consider an early phase trial with binary toxicity and efficacy outcomes and a total of D dose levels. The true toxicity probability and efficacy probability for the d-th dose level are defined as p_d and q_d , respectively, $d = 1, \ldots, D$, where p_d and q_d are assumed to be independent. The toxicity probabilities are assumed to be strictly increasing, i.e. $p_1 < \ldots < p_D$. However, the efficacy probabilities are assumed to have an unknown dose-response relationship. From Lin et al. 13, the combination of a binary toxicity outcome and a binary efficacy outcome consists of four different outcomes: 1 = (no toxicity, efficacy); 2 = (no toxicity, no efficacy); 3 =(toxicity, efficacy); 4 = (toxicity, no efficacy). The number of patients of the *i*-th outcome under dose level d is denoted by $y_{d,i}$, i = 1, 2, 3, 4. Let $y_{d,T} = y_{d,3} + y_{d,4}$ and $y_{d,E} = y_{d,1} + y_{d,3}$ be the number of patients experiencing a DLT and efficacy response under dose level d, respectively, and $n_d = \sum_{i=1}^{4} y_{d,i}$ be the total number of patients assigned to dose level d. We further define $\pi_{d,i}$ and $\hat{\pi}_{d,i} = y_{d,i}/n_d$ as the true probability and the observed probability of the *i*-th outcome, \hat{p}_d as the observed toxicity rate, and \hat{q}_d as the observed efficacy rate at dose level d. Then, $p_d = \pi_{d,3} + \pi_{d,4}$, $q_d = \pi_{d,1} + \pi_{d,3}$, $\hat{p}_d = y_{d,T}/n_d$, and $\hat{q}_d = y_{d,E}/n_d$. Assume that the maximum number of cohorts in the trial is C_M , where each cohort represents a group of patients joining the trial over a certain period. Let the sample size of each cohort be n_c , so that the maximum total sample size of the trial can be expressed as $N = C_M n_c$.

2.1 Optimal Biological Dose

The maximum acceptable toxicity probability and the minimum acceptable efficacy probability are defined as p_T and q_E , respectively. The MTD is the dose level d_{MTD} which has the largest toxicity probability not greater than p_T , i.e. $d_{MTD} = \arg \max_d p_d I(p_d \leq p_T)$, which can be any dose level among the D available dose levels. The MTD may not exist if all dose levels have a toxicity rate greater than p_T . Compared to the MTD, the OBD can be defined using one of three different methods:

(1) **Utility Score**: The utility score approach, employed by BOIN12 and uTPI, defines the OBD by assigning different utility values, $\{u_i\}_{i=1}^4$, to the combinations of toxicity and efficacy outcomes, as shown in Table 1. This method offers flexibility and can be extended to ordinal toxicity and efficacy outcomes. The OBD is defined as the dose yielding the highest expected utility calculated as $EU_d = \sum_{i=1}^4 \pi_{d,i} u_i$, with acceptable toxicity and efficacy probabilities, i.e.

$$d_{OBD}^{US} = \arg\max_{d} EU_{d}I(p_{d} \le p_{T}, q_{d} \ge q_{E}).$$

(2) Utility Function: The utility function approach involves a predefined utility function $U(p_d, q_d)$ based on toxicity and efficacy probabilities. This method is widely adopted in designs such as STEIN, TEPI-2, UBI, and PRINTE, although the specific utility functions may vary. The OBD is the dose level that maximizes this function with acceptable toxicity and efficacy probabilities, i.e.

$$d_{OBD}^{UF} = \arg\max_{d} U(p_d, q_d) I(p_d \le p_T, q_d \ge q_E).$$

(3) Maximum Efficacy: The maximum efficacy approach, adopted in BOIN-ET, selects the dose with the highest efficacy probability and acceptable toxicity probability given by

$$d_{OBD}^{ME} = \arg\max_{d} q_d I(p_d \le p_T, q_d \ge q_E).$$

When multiple doses share the highest utility value or efficacy probability, the OBD is the one with the lowest toxicity rate. If no dose level satisfies the criterion of having both acceptable toxicity and efficacy probabilities, specifically $p_d \leq p_T$ and $q_d \geq q_E$, then the true OBD does not exist. The maximum efficacy approach represents a specific case of both the utility score and utility function approaches. In the utility score approach, by setting $u_2 = u_4 = 0$ and $u_1 = u_3 > 0$, the expected utility simplifies to $EU_d = u_1q_d$, thereby designating the OBD as the dose level with the highest efficacy probability. Besides, in the utility function method, if the utility function, $U(q_d)$, depends only on the efficacy probability, then the OBD is the dose with the highest efficacy probability.

With the true toxicity and efficacy probabilities, each design has its own suggested OBD, which is the true OBD defined by each specific design, given its pre-specified design parameters. We adopted the same OBD definition approach as applied in the original paper for each design to ensure that the suggested OBD in our paper was consistent with the one defined in its original paper. In our simulation, we also ensured that the randomly chosen true OBD aligned with the suggested OBD of each design at the start of each replication. If a design had a different suggested OBD, we repeated the generation of true toxicity and efficacy probabilities until the alignment was achieved across all designs. A detailed discussion is provided in the supplementary document.

2.2 Safety and Futility Rules

A dose level d is considered admissible if the observed data indicate that dose level d is reasonably safe and efficacious, i.e. $p_d \leq p_T$ and $q_d \geq q_E$. After assigning each cohort of patients to a specific dose level, we check the following safety and futility rules:

(Safety) if $Pr(p_d > \phi_T \mid y_{d,T}, n_d) > \eta$, eliminate the current and all above doses from the dose list;

(Futility) if $Pr(q_d < \phi_E \mid y_{d,E}, n_d) > \xi$, eliminate the current dose from the dose list.

Only admissible doses can be used to treat patients. If there are no admissible doses to assign the next cohort of patients to, then the trial will be terminated. Different dose finding designs use different values for the parameters $\{\phi_T, \phi_E, \eta, \xi\}$. Specifically, $\{\phi_T, \phi_E, \eta, \xi\}$ equals to $\{0.35, 0.25, 0.95, 0.9\}$ for BOIN12, $\{0.4, 0.2, 0.9, 0.9\}$ for BOIN-ET, $\{0.33, 0.3, 0.95, 0.98\}$ for

STEIN, $\{0.3, 0.25, 0.95, 0.9\}$ for uTPI, and $\{0.4, 0.2, 0.9, 0.7\}$ for TEPI-2, UBI, and PRINTE. In this paper, we set $\{\phi_T, \phi_E, \eta, \xi\} = \{0.35, 0.25, 0.95, 0.8\}$ across all the designs to ensure a fair comparison among all designs.

3 Methods

3.1 BOIN-ET

BOIN-ET directly extends the idea of BOIN by utilizing both binary toxicity and efficacy outcomes²⁷. In BOIN-ET, there are two cut points on the toxicity interval, λ_1 and λ_2 , and one cut point, η_1 , on the efficacy interval, which slices the combination of toxicity and efficacy intervals into 6 regions as illustrated in Table 2. BOIN-ET determines the optimal values for $(\lambda_1, \lambda_2, \eta_1)$ under the restriction $\phi_1 < \lambda_1 < \phi_p < \lambda_2 < \phi_2$ and $\delta_1 < \eta_1 < \delta_E$, where ϕ_p and δ_E denote the target toxicity probability and efficacy probability, respectively, where $\phi_1=0.1p_T,\;\phi_2=1.4p_T,\;{\rm and}\;\delta_1=0.6\delta_E.$ Takeda et al.²⁷ developed 6 composite hypotheses and the following decision table as shown in Table 2. The values of these cut points are selected through a grid search to minimize the joint probability of incorrect dosing decisions. When $0 \le \hat{p}_d \le \lambda_1$ and $\eta_1 < \hat{q}_d \le 1$ are observed, the next cohort of the patients will be dosed at the current dose level using BOIN-ET, while BOIN chooses a higher dose level to dose the next cohort. In BOIN-ET, if $\lambda_1 < \hat{p}_d < \lambda_2$ and $0 \le \hat{q}_d \le \eta_1$, escalation, staying at the same dose, and de-escalation are all possible choices due to the unknown doseresponse relationship. On the contrary, BOIN only considers staying at the current dose if $\lambda_1 < \hat{p}_d < \lambda_2$. Let c = 1 denote the initial cohort and d = 1 the pre-specified initial dose level. The dose finding algorithm of BOIN-ET follows:

- 1. Treat the cohort c at the dose level d.
- 2. Calculate $\hat{p}_d = y_{d,T}/n_d$ and $\hat{q}_d = y_{d,E}/n_d$ at the current dose level d and follow the decision table in Table 2 to make the decision of escalation/stay/de-escalation. Specifically, if $\lambda_1 < \hat{p}_d < \lambda_2$ and $0 \le \hat{q}_d \le \eta_1$, define the admissible dose set $A_d = \{d-1, d, d+1\}$

and consider the following cases:

- (a) if dose level d+1 has never been used before, escalate to d+1;
- (b) if (a) is not applicable, choose the admissible dose that has the maximum probability of efficacy according to \hat{q}_{d-1} , \hat{q}_d , and \hat{q}_{d+1} ;
- (c) if (a) and (b) are not applicable because at least 2 doses have the same maximum probability of efficacy, randomly choose 1 dose among the doses that exhibit the maximum probability of efficacy.
- 3. Set c = c + 1 and update d based on the decision in Step 2.
- 4. Repeat Steps 1 3 until the maximum total sample size is reached.

BOIN-ET first applies isotonic regression on $\{\hat{p}_d\}_{d=1}^D$ to determine the MTD. For efficacy, BOIN-ET applies fractional polynomial regression with 2 degrees of freedom to estimate the efficacy probabilities which allows a non-monotonic dose-response relationship. The model deviance is used to select the best-fitting model containing 2 powers (k_1, k_2) from the set $\{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$. Among the tolerable doses, the OBD is determined by the best-fitting polynomial model.

3.2 BOIN12

BOIN12 is a flexible model-assisted design to find the OBD that optimizes the risk-benefit trade-off¹³. BOIN12 defines a utility u_i for each outcome i as shown in Table 1. The best outcome (no toxicity, efficacy) and the worst outcome (toxicity, no efficacy) have the utility values $u_1 = 100$ and $u_4 = 0$, respectively. The other two outcomes have utility values u_2 and u_3 within the range [0, 100]. As specified in Section 2, BOIN12 utilizes the expected utility to select the OBD, which is the dose level d with the largest mean utility $u_d = \sum_{i=1}^4 u_i \pi_{d,i}$. When $u_2 + u_3 = 100$, $u_d = u_2(1 - p_d) + u_3q_d$. Moreover, if $u_2 = 0$ and $u_3 = 100$, then $u_d = 100q_d$. In this special case, the OBD is the most efficacious dose level. We assume $u_2 = 40$ and $u_3 = 60$ as recommended by the authors to balance the risk-benefit trade-off.

Lin et al.¹³ introduced a rank-based desirability score (RDS) for the decisions during the trial. The standardized desirability of d is defined as $u_d^* = 100^{-1} \sum_{i=1}^4 u_i \pi_{d,i} \in [0,1]$, which

can be considered as a weighted average of $\{\pi_{d,i}\}_{i=1}^4$. Then suppose $x_d = 100^{-1} \sum_{i=1}^4 u_i y_{d,i}$ which can be interpreted as the number of events observed from n_d patients with event probability u_d^* . Under the Bayesian framework, assigning u_d^* a Beta (α, β) prior leads to the posterior distribution of u_d^* given the data to be $u_d^* \mid n_d, x_d \sim \text{Beta}(\alpha + x_d, \beta + n_d - x_d)$. The RDS is the rank of the posterior probabilities of $\Pr(u_d^* > u_b \mid n_d, x_d)$ for all possible situations, where the benchmark u_b for comparison is a constant within the range [0, 1]. In other words, the next dose level is selected based on the posterior probabilities of the dose levels in the admissible set. Table 3 of Lin et al. 13 is an example of an RDS Table. Let c=1 denote the initial cohort and d=1 the pre-specified initial dose level. The dose finding algorithm of BOIN12 follows from Figure 1 of Lin et al. 13 :

- 1. Treat the cohort c at the dose level d.
- 2. Calculate the observed DLT rate $\hat{p}_d = y_{d,T}/n_d$ at the current dose level d and compare it with two constants $0 < \lambda_e < \lambda_d < 1$, where λ_e and λ_d are from BOIN¹⁶.
 - (a) if $\hat{p}_d \geq \lambda_d$, de-escalate to the next lower dose level d-1;
 - (b) if λ_e < p̂_d < λ_d: (i) when n_d ≥ 6, select the dose from {d − 1, d} with larger RDS;
 (ii) when n_d < 6, select the dose from {d − 1, d, d + 1} with the largest RDS;
 - (c) if $\hat{p}_d \leq \lambda_e$, select the dose from $\{d-1, d, d+1\}$ with the largest RDS.
- 3. Set c = c + 1 and update d based on the decision in Step 2.
- 4. Repeat Steps 1 3 until the maximum total sample size is reached.

To prevent the dose finding process from getting stuck at a locally optimal dose, BOIN12 has an additional dose exploration rule: if $\hat{p}_d < \lambda_d$ and $n_d \geq 9$ for the current dose, and the next higher dose level has not been used, escalate to the next higher dose level d+1. To select the OBD, there are two following steps: (1) determine the MTD by applying isotonic regression to the observed toxicity rates $\{\hat{p}_d\}_{d=1}^D$ and choosing the dose level with the closest isotonically estimated toxicity rate to p_T ; (2) select the dose level with the highest estimated utility as the OBD among the doses with acceptable estimated toxicity probabilities.

3.3 UBI

Li et al. 10 developed UBI by combining BOIN with a utility function $U(\hat{p}_d, \hat{q}_d) = f_E(\hat{q}_d) - \theta f_T(\hat{p}_d)$ to incorporate efficacy with toxicity, where

$$f_{E}\left(\hat{q}_{d}\right) = \begin{cases} 0, & \hat{q}_{d} > \theta_{Eff} \\ \hat{q}_{d}, & \hat{q}_{d} \leq \theta_{Eff} \end{cases}, \qquad f_{T}\left(\hat{p}_{d}\right) = \begin{cases} 0, & \hat{p}_{d} \leq \theta_{Tox} \\ 1, & \hat{p}_{d} \geq \lambda_{d} \\ 0, & \hat{p}_{d} \in (\theta_{Tox}, \lambda_{e}], \text{ and } \hat{q}_{d} \leq \theta_{Eff} \\ \hat{p}_{d}, & \hat{p}_{d} \in (\lambda_{e}, \lambda_{d}) \text{ and } \hat{q}_{d} \leq \theta_{Eff} \\ \hat{p}_{d}/3, & \hat{p}_{d} \in (\theta_{Tox}, \lambda_{d}) \text{ and } \hat{q}_{d} > \theta_{Eff} \end{cases}$$

 λ_e and λ_d are the same parameters defined in BOIN and $\{\theta, \theta_{Eff}, \theta_{Tox}\} = \{2, 0.66, 0.15\}$ are the three design parameters defined in UBI. Specifically, when $\theta = 2$, the toxicity utility is considered to be twice as important as the efficacy utility. Let c = 1 denote the initial cohort and d = 1 be the pre-specified initial dose level. The dose finding algorithm of UBI is as follows:

- 1. Treat the cohort c at the dose level d.
- 2. Based on the observed data, calculate $\hat{p}_d = y_{d,T}/n_d$ and $\hat{q}_d = y_{d,E}/n_d$, and the utility function $U(\hat{p}_d, \hat{q}_d)$ at the current dose level d.
 - (a) if $U \ge 0$, escalate to the next higher dose;
 - (b) if U < -1/3, de-escalate to the next lower dose;
 - (c) otherwise, if $-1/3 \le U < 0$, stay at the current dose.
- 3. Set c = c + 1 and update d based on the decision in Step 2.
- 4. Repeat Steps 1 3 until the maximum total sample size is reached.

At the end of the trial, the OBD is selected as the dose with the highest utility score with a utility function defined as $U(\tilde{p}_d, \hat{q}_d) = g_E(\hat{q}_d) - \theta g_T(\tilde{p}_d)$, where \tilde{p}_d is the isotonically transformed estimate of the toxicity probability from the observed toxicity probability \hat{p}_d , $g_E(\hat{q}_d)$ and $g_T(\tilde{p}_d)$ are truncated functions given by

$$g_T(\hat{p}_i) = \begin{cases} p_{g1}^*, \hat{p}_i < p_{g1}^* \\ p_{g2}^*, \hat{p}_i > p_{g2}^*, & g_E(\hat{q}_i) = \begin{cases} q_{g1}^*, \hat{q}_i < q_{g1}^* \\ q_{g2}^*, \hat{q}_i > q_{g2}^* \end{cases}, \\ \hat{p}_i, \text{ else} & \hat{q}_i, \text{ else} \end{cases}$$

and $\{p_{g1}^*,\,p_{g2}^*,\,q_{g1}^*,\,q_{g2}^*\}=\{0.15,0.4,0.2,0.6\}$ are pre-specified cutoff values.

3.4 TEPI-2

Guo et al.⁶ proposed an extension of the mTPI design⁸, called the mTPI-2 design, by dividing the toxicity interval into subintervals of equal length. This modification aims to mitigate undesirable decisions that may arise from the mTPI design. Li et al.⁹ developed TEPI that extends the idea of the mTPI design by incorporating efficacy into the dose finding model. Similar to mTPI-2, TEPI-2 improves TEPI by dividing both the toxicity and efficacy intervals into subintervals with equal lengths, respectively ¹⁰. TEPI-2 partitions the toxicity unit into 4 subintervals, {Low, Moderate, High, Unacceptable}, and the efficacy unit into 4 subintervals, {Low, Moderate, High, Superb}. The TEPI-2 decision table needs to be pre-determined in consultation with physicians. In the example TEPI-2 decision table shown in Table 3, the lengths of the toxicity subintervals and efficacy subintervals are 0.08 and 0.2, respectively. Each interval rectangle constructed by a pair of toxicity and efficacy subintervals is assigned a decision. Unlike BOIN-ET, TEPI-2 escalates to a higher dose level with low toxicity but superb efficacy. With the observed data, TEPI-2 computes the joint utility probability mass (JUPM) for each interval rectangle, $(a_1, b_1) \times (a_2, b_2)$ as

$$JUPM_{(a_1,b_1)}^{(a_2,b_2)} = \frac{\Pr(p_d \in (a_1,b_1), q_d \in (a_2,b_2) \mid n_d, y_{d,T}, y_{d,E})}{(b_1 - a_1) \times (b_2 - a_2)}, \quad 0 < a_1 < b_1 < 1, 0 < a_2 < b_2 < 1.$$

The dosing decision for the next cohort is based on which interval rectangle has the largest JUPM. Let c=1 denote the initial cohort and d=1 the pre-specified initial dose level. The dose finding algorithm of TEPI-2 is as follows:

1. Treat the cohort c at the dose level d.

- 2. Based on the observed data, calculate the JUPMs for all interval rectangles and select the rectangle $(a_1, b_1) \times (a_2, b_2)$ with the largest JUPM. The dosing decision of the next cohort is based on what is given in the PRINTE decision table corresponding to the rectangle with the largest JUMP.
- 3. Set c = c + 1 and update d based on the decision in Step 2.
- 4. Repeat Steps 1 3 until the maximum total sample size is reached.

At the end of the trial, the OBD is selected as the dose with the highest estimated posterior expected utility. The utility score function is defined as $U(p,q) = f_1(p)f_2(q)$, where both $f_1(p)$ and $f_2(q)$ are truncated functions, given by

$$f_1(p) = \begin{cases} 1, & p \in (0, p_1^*] \\ 1 - \frac{p - p_1^*}{p_2^* - p_1^*}, & p \in (p_1^*, p_2^*), \end{cases} \qquad f_2(q) = \begin{cases} 0, & q \in (0, q_1^*] \\ \frac{q - q_1^*}{q_2^* - q_1^*}, & q \in (q_1^*, q_2^*). \\ 0, & p \in [p_2^*, 1) \end{cases}$$

The estimated posterior expected utility is given by $\hat{E}[U(p_d, q_d) \mid n_d, y_{d,T}, y_{d,E}] = T^{-1} \sum_{t=1}^T U^t(\hat{p}_d^t, q_d^t)$, where T is the Monte Carlo size, $\{p_d^t, q_d^t\}_{d,t}$ are generated from the posterior distributions of toxicity and efficacy rates, and \hat{p}_d^t is derived after isotonic transformation.

3.5 PRINTE

Similar to TEPI-2, PRINTE is an extension of TEPI and divides the efficacy and toxicity intervals into subintervals with equal lengths, respectively. Unlike TEPI-2, the decision table of PRINTE is pre-specified based on the values of the target toxicity rate p_T , target efficacy rate p_E , and a small fraction value ϵ . Per the PRINTE decision table, the dose needs to be de-escalated if p_d is greater than $p_T + \epsilon$. If p_d is not greater than $p_T + \epsilon$, the next cohort of patients is dosed at the current dose d if $q_d \geq p_d$ or at the next higher dose if $q_d < p_d$. An example PRINTE decision table is presented in Table 4. Let c = 1 denote the initial cohort and d = 1 the pre-specified initial dose level. The dose finding algorithm of PRINTE is as

follows:

- 1. Treat the cohort c at the dose level d.
- 2. Based on the observed data, calculate the JUPMs for all interval rectangles and select the rectangle $(a_1, b_1) \times (a_2, b_2)$ with the largest JUPM. The decision for the next cohort is the decision of the rectangle with the largest JUPM from the PRINTE decision table.
- 3. Set c = c + 1 and update d based on the decision in Step 2.
- 4. Repeat Steps 1 3 until the maximum total sample size is reached.

Once the maximum total sample size is attained, PRINTE determines the optimal dose by selecting the one with the highest estimated posterior expected utility. Let $A(p,q) = \{(p,q) \mid p \in (0,p_T], q \in [q_E + \delta, 1)\}$ be a graduate region and $p_{in} = \Pr(U(\hat{p}_d^t, \hat{q}_d^t) \in B)$, where B is the corresponding graduate utility region for $(p,q) \in A(p,q)$. The optimal dose will be selected as the OBD if $p_{in} \geq p_{graduate}$ where $p_{graduate}$ denotes a threshold value. Otherwise, the optimal dose will be rejected and the OBD will be considered as non-existent.

3.6 STEIN

STEIN¹² is another extension of BOIN, which utilizes a frequentist model averaging approach to estimate the efficacy probabilities. Similar to BOIN-ET, STEIN also has two cut points, ψ_L and ψ_U , on the toxicity intervals, and one cut point, ϕ , on the efficacy interval. Unlike BOIN-ET, which utilizes a grid search to estimate these cut points, STEIN can directly calculate these cut points using the same approach as BOIN³⁵. When $\hat{p}_d < \psi_U$ and $\hat{q}_d < \psi$, STEIN also constructs an admissible set for the dosing decision of the next cohort. Let c=1 denote the initial cohort and d=1 the pre-specified initial dose level. The dose finding algorithm of STEIN is as follows:

- 1. Treat the cohort c at the dose level d.
- 2. Calculate $\hat{p}_d = y_{d,T}/n_d$ and $\hat{q}_d = y_{d,E}/n_d$ at the current dose level d.
 - (a) if $\hat{p}_d \geq \psi_U$, escalate to d+1;

- (b) if $\hat{p}_d < \psi_L$ and $\hat{q}_d \ge \phi$, stay at the current dose d;
- (c) if $\hat{p}_d \leq \psi_L$ and $\hat{q}_d < \phi$, define the admissible set $A_d = \{d-1, d, d+1\}$ and select the dose d^{Next} for the next cohort with $d^{Next} = \arg\max_{d' \in A_d} \Pr(q_{d'} > \psi \mid n_{d'}, y_{d', E});$
- (d) if $\psi_L < \hat{p}_d < \psi_U$ and $\hat{q}_d < \phi$, define the admissible set $A_d = \{d-1, d\}$ and select the dose d^{Next} for the next cohort with $d^{Next} = \arg\max_{d' \in A_d} \Pr(q_{d'} > \psi \mid n_{d'}, y_{d', E})$.
- 3. Set c = c + 1 and update d based on the decision in Step 2.
- 4. Repeat Steps 1 3 until the maximum total sample size is reached.

At the end of the trial, STEIN uses isotonic regression on $\{\hat{p}_d\}_{d=1}^D$ to obtain the isotonically transformed values $\{\tilde{p}_d\}_{d=1}^D$. For the efficacy outcomes, STEIN performs D unimodal isotonic regressions on $\{\hat{q}_d\}_{d=1}^D$ by enumerating all possible models in the dose-efficacy curve. In the d'th model, $d'=1,\ldots,D$, the dose level d' attains the highest efficacy probability. With unimodal isotonic transformations, the transformed efficacy probabilities, $\{\tilde{q}_{d'd}\}_{d=1}^D$, have $\tilde{q}_{d'1} \leq \cdots \leq \tilde{q}_{d'd'} \geq \cdots \geq \tilde{q}_{d'D}$. The pseudo-likelihood based on the d'th unimodal isotonic regression is given by

$$\tilde{L}_{d'} = \prod_{d=1}^{D} \binom{n_d}{y_{d,E}} \tilde{q}_{d'd}^{y_{d,E}} (1 - \tilde{q}_{d'd}^{y_{d,E}})^{n_j - y_{d,E}}.$$

The final model averaging estimate of q_d is given by $\tilde{q}_d = \sum_{d'=1}^D \pi_{d'} \tilde{q}_{d'd}$, where $\pi_{d'} = \tilde{L}_{d'} / \sum_{d=1}^D \tilde{L}_d$. The OBD is the dose level with the highest utility, as defined by the utility function $U(\tilde{p}_d, \tilde{q}_d) = \tilde{q}_d - w_1 \tilde{p}_d - w_2 \tilde{p}_d I(\tilde{p}_d > p_T)^{15}$, where $w_1 = 0.33$ and $w_2 = 1.09$, as recommended in the original paper 12.

3.7 uTPI

uTPI combines the ideas of modeling dose desirability and the chessboard design method for dose finding²⁴. Similar to BOIN12, uTPI defines the expected utility or desirability as $EU_d = \sum_{i=1}^4 u_i \pi_{d,i}/100$ and the observed utility value as $OU_d = \sum_{i=1}^4 u_i y_{d,i}/100$, where u_i is the utility score of outcome i in Table 1. uTPI assumes that the numerical utility outcome has a pseudo binomial distribution so that the quasi-likelihood for EU_d is $L(OU_d \mid EU_d) = 1$

 $EU_d^{OU_d}(1-EU_d)^{n_d-OU_d}$. By assigning a noninformative beta prior $EU_d \sim \text{Beta}(1,1)$, the posterior distribution of EU_d is $EU_d \mid \text{Data} \sim \text{Beta}(1+OU_d,1+n_d-OU_d)$. Unlike the previous designs, uTPI constructs a two-dimensional chessboard by dividing the joint square of toxicity probability and desirability into equally sized squares. The toxicity interval and desirability interval are divided into subinverals, $\{\mathcal{I}_{k,T}\}_{k=1}^{10}$ and $\{\mathcal{I}_{k,U}\}_{k=1}^{10}$, where $\mathcal{I}_{k,T} = \mathcal{I}_{k,U} = [0.1(k-1), 0.1k)$ for $k = 1, \ldots, 9$ and $\mathcal{I}_{10,T} = \mathcal{I}_{10,U} = [0.9, 1]$. The strongest toxicity interval and the strongest desirability interval are defined as $k_d^T = \arg\max_k \Pr(p_d \in \mathcal{I}_{k,T} \mid \text{Data})$ and $k_d^U = \arg\max_k \Pr(EU_d \in \mathcal{I}_{k,U} \mid \text{Data})$, respectively. Let k^* denote the index of the toxicity subinterval such that $p_T \in \mathcal{I}_{k^*,T}$. Let c = 1 denote the initial cohort and d = 1 the pre-specified initial dose level. The dose finding algorithm of uTPI is as follows:

- 1. Treat the cohort c at the dose level d.
- 2. Identify the strongest toxicity interval index k_d^T and strongest desirability interval index k_d^U for all the dose levels.
 - (a) if $k_d^T > k^*$, de-escalate to dose level d-1;
 - (b) if $k_d^T < k^*$, choose dose level d' from the admissible set $\{d-1, d, d+1\}$ that has the largest k_d^U ;
 - (c) if $k_d^T = k^*$, choose dose level d' from the admissible set $\{d-1, d, d+1\}$ if $n_d < N*$ or from the admissible set $\{d-1, d\}$ if $n_d \ge N*$ that has the largest k_d^U .
- 3. Set c = c + 1 and update d based on the decision in Step 2.
- 4. Repeat Steps 1 3 until the maximum total sample size is reached.

In Step 2, if there is a tie, uTPI selects the dose level d' with the maximum $\Pr(OU_{d'} > \bar{\mathcal{I}}_{k_{d'}^U,U} \mid \text{Data})$, where $\bar{\mathcal{I}}_{k_{d'}^U,U}$ denotes the upper boundary of the $k_{d'}^U$ th desirability subinterval. At the end of the trial, uTPI also applies an isotonic regression on $\{\hat{p}_d\}_{d=1}^D$ to identify d_{MTD} , which is the dose level with the estimated toxicity rate \tilde{p}_d closest to p_T . The OBD is selected as the dose level that has the maximum posterior mean of EU_d , i.e. $d_{OBD} = \arg\max_{d \leq d_{MTD}} \widehat{EU}_d$.

4 Simulation Studies

4.1 Simulation Settings

We conducted simulation studies to compare the operating characteristics of the seven designs introduced in Section 3. The toxicity probabilities are assumed to increase monotonically with dose levels. For efficacy probabilities, we introduce four different dose-response relationships as follows:

Increasing (I): efficacy probabilities increase monotonically with dose levels, i.e. $q_1 < \cdots < q_D$;

Constant (C): efficacy probabilities are the same across all dose levels, i.e. $q_1 = \cdots = q_D$;

Unimodal (U): efficacy probabilities increase monotonically until a certain dose level $d^U < D$,

after which they decrease monotonically, i.e. $q_1 < \cdots < q_{d^U}$ and $q_{d^U} > \cdots > q_D$;

Increasing-Plateau (IP): efficacy probabilities increase monotonically until a certain dose level $1 < d^{IP} < D$, after which they remain constant, i.e. $q_1 < \cdots < q_{d^{IP}} = \cdots = q_D$.

In the unimodal case, we assumed that $d^U < D$ to exclude the increasing case. Note that the efficacy curve decreases monotonically when $d^U = 1$. In the increasing-plateau case, we set $d^{IP} \notin \{1, D\}$ to differentiate this relationship from the constant and increasing cases. Both d^U and d^{IP} were determined based only on the efficacy probabilities. Example toxicity and efficiency probability curves for all dose-response relationships are given in Figure 1.

Four dose levels were considered for all simulation studies, i.e., D=4. The maximum number of cohorts was set to $C_M=10$ with a cohort size of $n_c=3$, yielding a maximum total sample size of N=30. The maximum acceptable toxicity probability was set to $p_T=0.35$ and the minimum acceptable efficacy probability was set to $q_E=0.25$. Common design parameters were consistently applied across all designs to enable fair comparisons. Specifically, for the BOIN-based designs, BOIN12, BOIN-ET, UBI, and STEIN, it was assumed that $\lambda_e=0.276$ and $\lambda_d=0.419$, given $p_T=0.35$. For BOIN12 and uTPI, $u_2=40$, and $u_3=60$. Each trial, regardless of the design, continued until it reached the maximum total sample size or was terminated early due to the elimination of all dose levels during the trial.

We adopted the values provided in the original papers for design parameters specific to each design, such as those in each utility function. However, because our p_T and q_E differed from those values used in BOIN-ET, TEPI-2, and PRINTE, some of the design parameters were adjusted as follows: (1) we set $\{p_1^*, p_2^*, q_1^*, q_2^*\} = \{0.2, 0.35, 0.25, 0.6\}$ for TEPI-2 and PRINTE, which are design parameters in the truncated functions $f_1(p)$ and $f_2(q)$ for OBD selection as presented in Sections 3.4 and 3.5; (2) we modified TEPI-2's decision table to fit our simulation studies as presented in Table S1 because the length of the toxicity and efficacy subintervals in the decision table depends on p_T and q_E ; (3) ϕ_p was adjusted from 0.3 to 0.35 for BOIN-ET, which yields new optimal values of $(\lambda_1, \lambda_2, \eta_1) = (0.17, 0.44, 0.48)$ through a grid search. Note that "Table Si" denotes the *i*-th table in the supplementary document.

4.2 Generating Random Toxicity and Efficacy Probabilities

A total of 80,000 independent replications were carried out for each dose-response relationship, where 40,000 replications included the true OBD, while the remaining 40,000 did not. As outlined in Section 2.1, the OBD exists only if a dose d meets the criteria of $p_d \leq 0.35$ and $q_d \geq 0.25$. To ensure that we generated realistic toxicity and efficacy probabilities in our simulation, we assumed that $p_d \leq p_{max} = 0.7$ and $q_d \leq q_{max} = 0.9$ for all dose levels. In every replication where the true OBD existed, we generated the toxicity and efficacy probabilities as follows:

- 1. We first randomly selected one dose level, d^* , as the OBD (among those dose levels that qualify to be the OBD) with equal probability, and then generated $p_{d^*} \sim \text{Unif}(0, p_T)$, and $q_{d^*} \sim \text{Unif}(q_E, q_{max})$.
- 2. Let d^H be the dose level with the highest efficacy probability, where $q_{d^*} \leq q_{d^H}$. If more than one dose level has the highest efficacy probability, such as in the constant scenario, d^H is the lowest one among these dose levels. Given d^* , we selected d^H through a random process with equal probability among those qualified dose levels. If $d^H = d^*$, we set $p_{d^H} = p_{d^*}$ and $q_{d^H} = q_{d^*}$; if $d^H \neq d^*$, we generated $q_{d^H} \sim \text{Unif}(q_{d^*}, q_{max})$.

3. After d^* and d^H were selected, we generated the remaining probabilities of toxicity and efficacy by following the procedures outlined in Table 5.

Note that d^* and d^H can be different dose levels. For instance, in the increasing scenario of Figure 1, $d^* = 3$ and $d^H = 4$. However, d^* cannot exceed d^H , because if $d^* > d^H$, we would have $p_{d^*} > p_{d^H}$ under the assumption of monotonically increasing toxicity probabilities. This contradicts the definition of d^* . Besides, not all dose levels qualify as d^* or d^H . For example, only dose level 1 can be d^* and d^H in the constant scenario. Table S3 presents all the eligible combinations of (d^*, d^H) for each dose-response relationship. Additionally, Figure S1 shows example toxicity and efficacy probability curves for each combination. We also presented the algorithm for generating random probabilities of toxicity and efficacy in the replications where no OBD exists in Table S4.

For a fair comparison, it is crucial that the suggested OBD determined based on the given toxicity and efficacy probabilities is consistent across all designs. Therefore, in every replication, the process of generating these probabilities was repeated until they satisfied this requirement. Table S5 provides a breakdown of the frequency with which each dose level was chosen as the true OBD under each dose-response relationship. The table demonstrates that each dose level that qualifies to be the true OBD has an equal probability of being selected as the true OBD.

4.3 Results

We evaluated seven dose finding designs using seven distinct performance metrics:

- p_{OBD} : the proportion of trials that successfully select the OBD given the OBD exists;
- n_{OBD} : the average number of patients assigned to the OBD given the OBD exists;
- n_{over} : the average number of patients assigned to dose levels with toxicity probabilities greater than $p_T + 0.1$ given the OBD exists.
- p_{poor} : the proportion of trials that assign less than 20% of patients to the OBD given the OBD exists;

- p_{no} : the proportion of trials that successfully do not select any dose level when the OBD does not exist;
- $n_{no,over}$: the average number of patients assigned to dose levels with toxicity probabilities greater than $p_T + 0.1$ when the OBD does not exist;
- n_{no} : the average number of treated patients when the OBD does not exist.

The first four metrics are applicable when the OBD exists, while the last three are used when the OBD does not exist. For each metric under each dose-response relationship, the value corresponding to the best-performing design is highlighted in bold in Table 6. The results indicate that no single design consistently outperforms the others across all metrics and scenarios.

When the OBD exists, in scenario I (Increasing), STEIN exhibits the highest value of p_{OBD} , whereas uTPI has the second highest p_{OBD} . In scenario C (Constant), UBI leads with the highest probability of accurately identifying the OBD, with PRINTE closely following UBI. When the dose-response relationship follows a unimodal (U) or increasing-plateau (IP) pattern, STEIN has the highest p_{OBD} . Overall, STEIN has the best performance in terms of p_{OBD} . We compared the selected OBD from each design to the true OBD, averaging the results across all replications for each scenario. When the selected OBD deviates from the true OBD, Table S6 shows that, on average, the percentage of cases where the selected OBD was lower than the true OBD is around 37% for BOIN-ET, 53% for BOIN12, 60% for both uTPI and STEIN, and is as high as 80% for UBI, TEPI-2, and PRINTE. This result indicates that UBI, TEPI-2, and PRINTE are more conservative compared to the other designs, and tend to select lower dose levels as the OBD.

Regarding the metrics, $\{n_{OBD}, n_{over}, p_{poor}\}$, we found that: (1) PRINTE has the highest n_{OBD} in scenarios I, U, and, IP, while uTPI has the highest n_{OBD} in scenario C. (2) STEIN has the lowest n_{over} in scenarios C, U, and IP, while uTPI has the lowest n_{over} in scenario I. (3) In terms of p_{poor} , BOIN-ET performs the best in scenario I, while STEIN is the best design in scenarios C and IP, and BOIN12 is the best design in scenario U. (4) In terms of

these metrics, TEPI-2 and UBI underperform compared to the other designs. For example, in scenario C, while UBI and TEPI-2 have high p_{OBD} values, they also have notably low n_{OBD} values and high p_{poor} values. This is because both TEPI and UBI escalate to the next dose level if the current dose has a low observed toxicity probability and a high observed efficacy probability, as demonstrated in Table 3 and Table 3 of Li et al. ¹⁰.

When the true OBD does not exist, PRINTE consistently exhibits the highest accuracy in not choosing any dose level as the OBD, represented by p_{no} . TEPI-2 and UBI have smaller p_{no} values than PRINTE, but they perform better than the other four designs. While BOIN-ET and STEIN have high p_{OBD} values, they have much lower p_{no} values than the other remaining designs. For $n_{no,over}$, STEIN is the best-performing design across all scenarios, followed by TEPI-2 and UBI. TEPI-2 and UBI have a smaller n_{no} compared to other designs across all scenarios.

4.4 Sensitivity Analysis

To evaluate the robustness of our simulation findings, we performed five independent sensitivity analyses by changing the following:

- (SA1) the number of dose levels to D = 5;
- (SA2) the maximum number of cohorts to $C_M = 15$;
- (SA3) the correlation between the toxicity and efficacy outcomes to $\rho_1 = 0.2$;
- (SA4) the correlation between the toxicity and efficacy outcomes at $\rho_2 = 0.4$;
- (SA5) the initial dose level at the beginning of the trial to dose level 2.

In summary, the results of the first four sensitivity analyses, which used the first dose level as the initial dose, confirm the reliability and robustness of our findings in Section 4.3. However, in SA5 which sets a different initial dose level of dose level 2, all designs had a decrease in p_{OBD} and an increase in p_{poor} , which is particularly noticeable in the case of BOIN-ET. Detailed results of these analyses are provided in Section 7 of the supplementary document.

5 Case Study

Another simulation study was conducted using the CAR T-cell therapy phase I clinical trial example as illustrated in Li et al. ¹⁰ and Raje et al. ²². The patients with relapsed or refractory multiple myeloma (RRMM) were administered a single infusion at four distinct doses during the dose escalation phase: 50×10^6 , 150×10^6 , 450×10^6 , and 800×10^6 CAR-positive (CAR+) T cells. The four doses were defined as dose level 1 through dose level 4. In the dose escalation phase, 21 patients were assigned using the 3+3 design, with the number of patients at each dose level $n_1 = 3$, $n_2 = 6$, $n_3 = 9$, and $n_4 = 3$. Based on the observed DLTs at each dose level, the calculated DLT rates, $\{\hat{p}_d\}_{d=1}^4$, were 0%, 17%, 33%, and 67%. The efficacy outcome was assessed using the International Myeloma Working Group (IMWG) uniform response criteria for multiple myeloma. The responders included complete response, very good partial response, and partial response. The reported response rates, $\{\hat{q}_d\}_{d=1}^4$, were 33%, 75%, 95%, and $100\%^{10}$. The 150×10^6 and 450×10^6 CAR+ T cells were selected for the dose expansion phase. We assumed $p_T = 0.35$ and $q_E = 0.25$. Due to the high values of \hat{p}_4 and \hat{q}_4 , we set both p_{max} and q_{max} to be 1. The values of the other design parameters were the same as those specified in Section 4.1.

A random simulation was conducted by generating toxicity and efficacy probabilities based on their observed rates. It was assumed that all the toxicity and efficacy probabilities have the same uninformative prior Beta(1,1). The posterior distributions are given by Beta(1 + $n_d\hat{p}_d$, 1 + n_d (1 - \hat{p}_d)) for toxicity and Beta(1 + $n_d\hat{q}_d$, 1 + n_d (1 - \hat{q}_d)) for efficacy. We assumed that the true toxicity rates follow a monotonically increasing curve, while the true efficacy rates exhibit either an increasing (I) or an increasing-plateau (IP) pattern, with $q_3 = q_4$. We assumed that the OBD exists, given the toxicity and efficacy data of dose levels 2 and 3. Based on the observed toxicity and efficacy rates, dose level 3 is the suggested OBD for all designs. However, dose level 2 had a much lower observed toxicity probability than dose level 3. It had a high observed efficacy probability, resulting in a high observed utility score and a high observed utility function value. Due to the small sample size, dose

level 2 can potentially be the true OBD, as the actual toxicity and efficacy probabilities are unknown. Therefore, the true OBD was assumed to be either dose level 2 or 3. For each combination of the dose-response relationship (I or IP) and the dose level as the true OBD (dose levels 2 or 3), we conducted a simulation study with 10,000 independent replications for each of the four combinations. The approach for generating random toxicity and efficacy probabilities in this case study is detailed in Section 6 of the supplementary document.

Table 7 presents the percentage of times that each dose level is identified as the OBD in scenarios I and IP. When dose level 2 is the true OBD, all designs show a higher selection rate for dose level 2 as the OBD, compared to the selection rate of dose level 3 when it is the true OBD. When dose level 2 is the true OBD, STEIN has the highest values of p_{OBD} in both scenarios, followed by BOIN-ET. However, when the true OBD is dose level 3, BOIN-ET and BOIN12 identify the OBD more accurately than the other designs. In all scenarios, UBI, TEPI-2, and PRINTE select dose level 1 as the OBD around 20% of the time, which aligns with the conservative nature of these designs as discussed in Section 4. However, dose level 1 is evidently not an optimal choice for the OBD, given its lower efficacy rate of 33% compared to 75% at dose level 2. Our results regarding TEPI-2 and UBI align with those from Li et al. 10, who conducted a fixed simulation study based on the observed toxicity and efficacy rates. Table 8 shows that BOIN-ET, BOIN12, STEIN, and uTPI outperform the other three designs in terms of p_{OBD} . All designs exhibit similar n_{OBD} , with BOIN-ET having the highest n_{OBD} . BOIN-ET, PRINTE, STEIN, and uTPI have lower n_{poor} but higher p_{poor} values than BOIN-12, UBI, and TEPI-2. Overall, STEIN has the best performance among all the designs for this case study.

6 Practical Implementation

Based on our findings in Sections 4 and 5, no single design consistently outperforms the others across all performance metrics in every scenario. However, we identified three designs, STEIN, PRINTE, and BOIN12, that demonstrate strong performance under specific metrics

across all scenarios, serving as our guidance for practical users:

- (1) **STEIN**: It has the highest probability of identifying the true OBD and a low likelihood of poor allocation to the OBD when the true OBD exists. STEIN provides the best overdose control, regardless of the existence of the true OBD. Based on previous trials or other prior knowledge, STEIN can be used when there is at least one admissible dose level among the doses being considered. However, STEIN is not competitive when the true OBD does not exist.
- (2) **PRINTE**: It allocates a large number of patients to the OBD and few patients to toxic doses when the true OBD exists. In addition, PRINTE can terminate a trial without selecting any dose as the OBD with a high probability, if no dose level is admissible. This minimizes the risk of mistakenly advancing an unsuitable dose to the next trial phase. However, as a trade-off, PRINTE may expose more patients to toxic doses compared to other designs when the true OBD does not exist.
- (3) **BOIN12**: It is widely used as an extension of BOIN, which has been acknowledged as a "fit-for-purpose" design by the FDA. The software for implementing BOIN12 is publicly available at www.trialdesign.org, which features an interactive and user-friendly interface. The software also offers pre-defined language for the protocol and the statistical analysis plan, specifically tailored to the BOIN12 design. BOIN12 adopts intuitive utility scores and an RDS table for OBD selection, which can be easily implemented by practical users.

The other designs we have examined exhibit various limitations. For instance, BOIN-ET has a low probability of terminating the trial when the true OBD does not exist. Additionally, its performance can be affected by the choice of the initial dose level. TEPI-2 and UBI underperform compared to other designs across all four performance metrics, $\{p_{OBD}, n_{OBD}, n_{over}, p_{poor}\}$, when the OBD exists under scenarios I, U, and IP. uTPI applies the utility score method, similar to BOIN12, but uTPI has more design parameters and is more complicated to implement than BOIN12.

The BOIN-ET program can be accessed at https://github.com/yamagubed/boinet. An R package called "boinet", which includes BOIN-ET, is also available. The R code of uTPI can be obtained from https://github.com/haoluns/uTPI. We have shared our R programs for UBI, TEPI-2, PRINTE, and STEIN, which do not have either public software or R packages, at https://github.com/EugeneHao/phase-I-II-designs.

7 Discussion

In this article, we compared seven current innovative early phase dose finding designs, namely BOIN-ET, BOIN12, TEPI-2, UBI, PRINTE, STEIN, and uTPI. BOIN-ET serves as the most direct extension of the original BOIN design. Similar to BOIN-ET, STEIN divides the toxicity-efficacy square into six distinct regions for dosing decisions, but it utilizes unimodal isotonic regression models for OBD selection. BOIN12 is another BOIN-based design that applies different utility scores for each combination of toxicity and efficacy outcomes. uTPI merges the concepts of modeling dose desirability with the chessboard design method. These four designs formulate an admissible set for dosing decisions when the current dose has an acceptable observed toxicity rate and a less-than-superb observed efficacy rate. UBI, TEPI-2, and PRINTE only incorporate data from the current dose level to make dosing decisions, with the assistance of the corresponding decision tables. UBI uses a utility function that depends on a complex set of design parameters. TEPI-2 is characterized by a simple dose finding algorithm, although the procedure for specifying its decision table is subjective. PRINTE is superior in identifying situations where the OBD is absent, given its integration of an OBD double-validation criterion.

Our simulation results indicate that no single design consistently outperforms the other designs for all performance metrics across various dose-response relationships. To account for complex dose-response relationships, we examined four distinct dose-efficacy curves. When the OBD existed, STEIN was the best design in terms of accurately selecting the OBD in scenarios I, U, and IP while UBI had the best performance in scenario C. BOIN-ET, BOIN12,

and STEIN had low probabilities of poor allocation to the OBD. STEIN consistently assigned fewer patients to toxic dose levels, regardless of the existence of the true OBD. Furthermore, PRINTE had the highest probability of not selecting any dose level as the OBD when the true OBD did not exist, across all scenarios. These findings are consistent with those presented by Yamaguchi et al. ³³ and Li et al. ¹¹.

Although the list of designs we evaluated overlaps with that evaluated in Shi et al. 25, there are some key differences. Shi et al. only considered the unimodal and increasing-plateau relationships, whereas we also considered increasing and constant relationships. Additionally, they conducted simulations only under the assumption that the true OBD exists. In contrast, we conducted additional simulation studies to compare the designs where the true OBD does not exist. Furthermore, Shi et al. focused on comparing the dose exploration algorithms of the designs and applied the same utility score approach for OBD estimation across all designs. On the other hand, we used the OBD selection approach specific to each design, as outlined in the original papers. Note that only BOIN12 and uTPI directly adopt the utility score approach for both dose exploration and OBD estimation. Other designs can underperform when using this approach for OBD estimation. For example, UBI applies a utility function for both dose exploration and OBD selection. It is not consistent to change UBI's original OBD selection method to the utility score method without adjusting its dose finding algorithm. Shi et al. showed that with the utility score approach, BOIN12 excels in OBD selection accuracy and minimizing poor dose allocation, closely followed by uTPI²⁵. However, we found that STEIN had a higher probability of identifying the true OBD than BOIN12 with its original OBD estimation approach. Both Shi et al. and our findings conclude that STEIN has better overdose control compared to other designs.

Since the true toxicity and efficacy rates are unknown in real clinical studies, we implemented an objective random simulation approach that generates values from the assumed distributions of the toxicity and efficacy probabilities, rather than subjectively choosing potentially biased values as the true fixed rates. These distributions of toxicity and efficacy

probabilities take into account the inherent uncertainty in these probabilities. In our case study with the CAR-T example, we extended our random toxicity and efficacy probability generation procedure by incorporating prior information. In the CAR-T example, we examined scenarios with either dose levels 2 or 3 as the true OBD, considering both Increasing and Increasing-Plateau dose-efficacy relationships, of which only one scenario was analyzed through the fixed simulation studies conducted by Li et al.¹⁰.

The early phase designs investigated in this article have several limitations. Firstly, BOIN-ET, TEPI-2, UBI, PRINTE, STEIN, and uTPI assume that the safety and efficacy outcomes are independent, while the true relationship between safety and efficacy may be complex. However, it has been demonstrated that this independence assumption has negligible efficacy loss^{3,5}. We analyzed this issue via sensitivity analyses, where we considered the correlation between efficacy and toxicity probabilities, and reached the same conclusion. Secondly, designs other than BOIN12 do not consider ordinal outcomes. To the best of our understanding, gBOIN-ET 29 and TITE-gBOIN-ET 30 are the two model-assisted designs for handling both ordinal toxicity and efficacy outcomes. Additionally, all designs assume that both toxicity and efficacy outcomes will be observed quickly enough to inform the dosing decisions for the subsequent patient cohort. Implementing these designs may be challenging in certain immunotherapy trials, particularly those with late-onset or pending toxicity and efficacy responses or those experiencing rapid patient accrual. To address these challenges, time-to-event (TITE) designs, such as TITE-BOIN12³⁸ and TITE-BOIN-ET²⁸, have been developed. Finally, none of the seven designs incorporate pharmacokinetic (PK) or pharmacodynamic (PD) parameters, nor do they include biomarker data in their dose finding algorithms. For instance, PKBOIN-12 is an innovative model-assisted dose finding design that integrates PK, toxicity, and efficacy to optimize dose selection ²⁶. In future work, we will examine the performance of existing designs that incorporate these factors or consider developing a new design that enhances current methodologies. Clearly, the unique challenges of oncology drug development call for more integrated dose finding approaches. These should encompass innovative study designs, advanced statistical methods, and cross-functional collaborations.

Declaration of Interest Statement

The authors declare that they have no known competing financial interests or personal relationships that can have appeared to influence the work reported in this paper.

Data Availability

This paper does not use any real data. The simulation code will be available on request.

References

- [1] Daniel V Araujo, Marc Oliva, Kecheng Li, Rouhi Fazelzad, Zhihui Amy Liu, and Lillian L Siu. Contemporary dose-escalation methods for early phase studies in the immunotherapeutics era. *European Journal of Cancer*, 158:85–98, 2021.
- [2] James Babb, André Rogatko, and Shelemyahu Zacks. Cancer phase I clinical trials: efficient dose escalation with overdose control. Statistics in medicine, 17(10):1103–1120, 1998.
- [3] Chunyan Cai, Ying Yuan, and Yuan Ji. A Bayesian dose-finding design for oncology clinical trials of combinational biological agents. *Journal of the Royal Statistical Society*. Series C, Applied statistics, 63(1):159, 2014.
- [4] US Food, Drug Administration, et al. Project optimus: Reforming the dose optimization and dose selection paradigm in oncology, 2022.
- [5] Wentian Guo, Yang Ni, and Yuan Ji. TEAMS: Toxicity-and efficacy-based dose-insertion

- design with adaptive model selection for phase I/II dose-escalation trials in oncology. Statistics in biosciences, 7:432–459, 2015.
- [6] Wentian Guo, Sue-Jane Wang, Shengjie Yang, Henry Lynn, and Yuan Ji. A Bayesian interval dose-finding design addressingockham's razor: mTPI-2. Contemporary clinical trials, 58:23–33, 2017.
- [7] Alexia Iasonos and John O'Quigley. Continual reassessment and related designs in dose-finding studies. *Statistics in medicine*, 30(17):2057, 2011.
- [8] Yuan Ji, Ping Liu, Yisheng Li, and B Nebiyou Bekele. A modified toxicity probability interval method for dose-finding trials. *Clinical trials*, 7(6):653–663, 2010.
- [9] Daniel H Li, James B Whitmore, Wentian Guo, and Yuan Ji. Toxicity and efficacy probability interval design for phase I adoptive cell therapy dose-finding clinical trials. Clinical Cancer Research, 23(1):13–20, 2017.
- [10] Pin Li, Rachael Liu, Jianchang Lin, and Yuan Ji. TEPI-2 and UBI: designs for optimal immuno-oncology and cell therapy dose finding with toxicity and efficacy. *Journal of biopharmaceutical statistics*, 30(6):979–992, 2020.
- [11] Ran Li, Kentaro Takeda, and Alan Rong. Comparison between simultaneous and sequential utilization of safety and efficacy for optimal dose determination in Bayesian model-assisted designs. Therapeutic Innovation & Regulatory Science, pages 1–9, 2023.
- [12] Ruitao Lin and Guosheng Yin. STEIN: A simple toxicity and efficacy interval design for seamless phase I/II clinical trials. *Statistics in medicine*, 36(26):4106–4120, 2017.
- [13] Ruitao Lin, Yanhong Zhou, Fangrong Yan, Daniel Li, and Ying Yuan. BOIN12: Bayesian optimal interval phase i/ii trial design for utility-based dose finding in immunotherapy and targeted therapies. JCO precision oncology, 4:1393–1402, 2020.

- [14] Xiaolei Lin and Yuan Ji. Probability intervals of toxicity and efficacy design for dose-finding clinical trials in oncology. Statistical Methods in Medical Research, 30(3):843–856, 2021.
- [15] Suyu Liu and Valen E Johnson. A robust Bayesian dose-finding design for phase I/II clinical trials. *Biostatistics*, 17(2):249–263, 2016.
- [16] Suyu Liu and Ying Yuan. Bayesian optimal interval designs for phase I clinical trials. Journal of the Royal Statistical Society: Series C: Applied Statistics, pages 507–523, 2015.
- [17] Maurie Markman. Serious ethical dilemma of single-agent pegylated liposomal doxorubicin employed as a control arm in ovarian cancer chemotherapy trials. Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology, 28 (19):e319–20, 2010.
- [18] L Minasian, O Rosen, D Auclair, A Rahman, R Pazdur, and RL Schilsky. Optimizing dosing of oncology drugs. *Clinical Pharmacology & Therapeutics*, 96(5):572–579, 2014.
- [19] 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.
- [20] Haitao Pan, Ruitao Lin, Yanhong Zhou, and Ying Yuan. Keyboard design for phase I drug-combination trials. Contemporary Clinical Trials, 92:105972, 2020.
- [21] Sophie Postel-Vinay, Carlos Gomez-Roca, L Rhoda Molife, Bhavesh Anghan, Antonin Levy, Ian Judson, Johann De Bono, Jean-Charles Soria, Stan Kaye, and Xavier Paoletti. Phase I trials of molecularly targeted agents: should we pay more attention to late toxicities. J Clin Oncol, 29(13):1728–1735, 2011.

- [22] Noopur Raje, Jesus Berdeja, YI Lin, David Siegel, Sundar Jagannath, Deepu Madduri, Michaela Liedtke, Jacalyn Rosenblatt, Marcela V Maus, Ashley Turka, et al. Anti-bcma car t-cell therapy bb2121 in relapsed or refractory multiple myeloma. New England Journal of Medicine, 380(18):1726–1737, 2019.
- [23] Hiroyuki Sato, Akihiro Hirakawa, and Chikuma Hamada. An adaptive dose-finding method using a change-point model for molecularly targeted agents in phase I trials. Statistics in medicine, 35(23):4093–4109, 2016.
- [24] Haolun Shi, Jiguo Cao, Ying Yuan, and Ruitao Lin. uTPI: A utility-based toxicity probability interval design for phase I/II dose-finding trials. Statistics in Medicine, 40 (11):2626–2649, 2021.
- [25] Haolun Shi, Ruitao Lin, and Xiaolei Lin. Comparative review of novel model-assisted designs for phase I/II clinical trials. *Biometrical Journal*, 66(4):2300398, 2024.
- [26] Hao Sun and Jieqi Tu. Pkboin-12: A bayesian optimal interval phase i/ii design incorporating pharmacokinetics outcomes to find the optimal biological dose. *Pharmaceutical Statistics*.
- [27] Kentaro Takeda, Masataka Taguri, and Satoshi Morita. BOIN-ET: Bayesian optimal interval design for dose finding based on both efficacy and toxicity outcomes. *Pharmaceutical statistics*, 17(4):383–395, 2018.
- [28] Kentaro Takeda, Satoshi Morita, and Masataka Taguri. TITE-BOIN-ET: time-to-event Bayesian optimal interval design to accelerate dose-finding based on both efficacy and toxicity outcomes. *Pharmaceutical Statistics*, 19(3):335–349, 2020.
- [29] Kentaro Takeda, Satoshi Morita, and Masataka Taguri. gBOIN-ET: The generalized Bayesian optimal interval design for optimal dose-finding accounting for ordinal graded efficacy and toxicity in early clinical trials. *Biometrical Journal*, 64(7):1178–1191, 2022.

- [30] Kentaro Takeda, Yusuke Yamaguchi, Masataka Taguri, and Satoshi Morita. TITE-gBOIN-ET: Time-to-event generalized Bayesian optimal interval design to accelerate dose-finding accounting for ordinal graded efficacy and toxicity outcomes. *Biometrical Journal*, page 2200265, 2023.
- [31] Peter F Thall and John D Cook. Dose-finding based on efficacy-toxicity trade-offs. Biometrics, 60(3):684–693, 2004.
- [32] Jeffrey S Weber, James C Yang, Michael B Atkins, and Mary L Disis. Toxicities of immunotherapy for the practitioner. *Journal of Clinical Oncology*, 33(18):2092, 2015.
- [33] Yusuke Yamaguchi, Kentaro Takeda, Satoshi Yoshida, and Kazushi Maruo. Optimal biological dose selection in dose-finding trials with model-assisted designs based on efficacy and toxicity: a simulation study. *Journal of Biopharmaceutical Statistics*, pages 1–15, 2023.
- [34] Fangrong Yan, Sumithra J Mandrekar, and Ying Yuan. Keyboard: A novel Bayesian toxicity probability interval design for phase I clinical trials. *Clinical Cancer Research*, 23(15):3994–4003, 2017.
- [35] Ying Yuan, Kenneth R Hess, Susan G Hilsenbeck, and Mark R Gilbert. Bayesian optimal interval design: A simple and well-performing design for phase I oncology trials. Clinical Cancer Research, 22(17):4291–4301, 2016.
- [36] Ying Yuan, J Jack Lee, and Susan G Hilsenbeck. Model-assisted designs for early-phase clinical trials: simplicity meets superiority. *JCO Precision Oncology*, 3:1–12, 2019.
- [37] Yong Zang, J Jack Lee, and Ying Yuan. Adaptive designs for identifying optimal biological dose for molecularly targeted agents. *Clinical Trials*, 11(3):319–327, 2014.
- [38] Yanhong Zhou, Ruitao Lin, J Jack Lee, Daniel Li, Li Wang, Ruobing Li, and Ying Yuan.

TITE-BOIN12: A Bayesian phase I/II trial design to find the optimal biological dose with late-onset toxicity and efficacy. *Statistics in medicine*, 41(11):1918–1931, 2022.

Table 1: Utility score table for binary toxicity and efficacy outcomes

	Efficacy						
Toxicity	Yes	No					
No	$u_1 = 100$	u_2					
Yes	u_3	$u_4 = 0$					

Table 2: Decision Table for BOIN-ET

	$0 \le \hat{p}_d \le \lambda_1$	$\lambda_1 < \hat{p}_d < \lambda_2$	$\lambda_2 < \hat{p}_d \le 1$
$\eta_1 < \hat{q}_d \le 1$	Stay	Stay	De-escalate
$0 \le \hat{q}_d \le \eta_1$	Escalate	Escalate/Stay/De-escalate	De-escalate

Table 3: An example of a TEPI-2 decision table based on $p_T = 0.4$ and $q_E = 0.2$

Table 5. All example of a 1E1 1-2 decision table based on $p_T = 0.4$ and $q_E = 0.2$									
			Efficacy Rate						
			Low	Low Moderate High S		Supe	ıperb		
	(0, 0.2)	(0.2, 0.4)	(0.4, 0.6)	(0.6, 0.8)	(0.8, 1)				
	Low	(0, 0.08)	Е	E	E	Е	E		
	LOW	(0.08, 0.16)	E	E	Ε	E	E		
	Moderate	(0.16, 0.24)	Е	E	E	S	S		
		(0.24, 0.32)	E	E	E	S	S		
Toxicity Rate	High	(0.32, 0.4)	D	S	S	S	S		
Toxicity Itale	Unacceptable	(0.4, 0.48)	D	D	D	D	D		
		(0.48, 0.56)	D	D	D	D	D		
		• • •	D	D	D	D	D		
		(0.88, 0.96)	D	D	D	D	D		
		(0.96, 1)	D	D	D	D	D		

E = escalation, S = stay, D = de-escalation

Table 4: An example of a PRINTE decision table based on $p_T = 0.4$, $p_E = 0.4$, and $\epsilon = 0.05$

		Efficacy Rate							
	(0, 0.2)	(0.2, 0.4)	(0.4, 0.6)	(0.6, 0.8)	(0.8, 1)				
	(0, 0.05)	E	E	S	S	S			
	(0.05, 0.15)	E	E	S	S	S			
	(0.15, 0.25)	E	E	S	S	S			
	(0.25, 0.35)	E	E	S	S	S			
Toxicity Rate	(0.35, 0.45)	E	E	S	S	S			
	(0.45, 0.55)	D	D	D	D	D			
	• • •	D	D	D	D	D			
	(0.85, 0.95)	D	D	D	D	D			
	(0.95, 1)	D	D	D	D	D			

E = escalation, S = stay, D = de-escalation

Table 5: Generating random toxicity and efficacy probabilities under each dose-response

relationship in the presence of the true OBD

Scenario	Generating Random Probabilities
Increasing	Toxicity: generate an ascending vector of toxicity probabilities from $\operatorname{Unif}(0, p_{d^*})$ for dose levels $\{1, \ldots, d^* - 1\}$ and another ascending vector of toxicity probabilities from $\operatorname{Unif}(p_T, p_{max})$ for dose levels $\{d^* + 1, \ldots, D\}$.
	Efficacy: generate an ascending vector of efficacy probabilities from Unif $(0, q_{d^*})$ for dose levels $\{1, \ldots, d^* - 1\}$ and another ascending vector of efficacy probabilities from Unif (q_{d^*}, q_{max}) for dose levels $\{d^* + 1, \ldots, D\}$.
Constant	Toxicity: generate an ascending vector of toxicity probabilities from $\mathrm{Unif}(p_{d^*}, p_{max})$ for dose levels $\{2, \ldots, D\}$ where $d^* = 1$.
	Efficacy: assign q_1 as the efficacy probabilities for dose levels $\{2, \ldots, D\}$.
Unimodal	 Toxicity: (a) generate an ascending vector of toxicity probabilities from Unif(0, p_{d*}) for dose levels {1,, d* - 1}; (b1) if d* = d^H, generate another ascending vector of toxicity probabilities from Unif(p_{d*}, p_{max}) for dose levels {d* + 1,, D}; (b2) if d* ≠ d^H, generate another ascending vector of toxicity probabilities from Unif(p_T, p_{max}) for dose levels {d* + 1,, D}.
	 Efficacy: (a) generate an ascending vector of efficacy probabilities from Unif(0, q_{d*}) for dose levels {1,, d* - 1}; (b1) if d* = d^H, generate a decreasing vector of efficacy probabilities from Unif(0, q_{d*}) for dose levels {d* + 1,, D}; (b2) if d* ≠ d^H, generate an ascending vector of efficacy probabilities from Unif(q_{d*}, q_{dH}) for dose levels {d* + 1, d^H - 1} and a decreasing vector of efficacy probabilities from Unif(0, q_{dH}) for dose levels {d^H + 1, D}.
Increasing- Plateau	Toxicity: the procedure for generating toxicity probabilities follows the same method used in the unimodal scenario.
	 Efficacy: (a) generate an ascending vector of efficacy probabilities from Unif(0, q_{d*}) for dose levels {1,, d* - 1}; (b1) if d* = d^H, assign q_{d*} as the efficacy probabilities for dose levels {d* + 1,, D}; (b2) if d* ≠ d^H, generate an ascending vector of efficacy probabilities from Unif(q_{d*}, q_d) for dose levels {d* + 1,, d^H - 1} and assign q_d as the efficacy probabilities for dose levels {d^H + 1,, D}.

- (1) d^* : true OBD; (2) d^H : the lowest dose level with the highest efficacy probability;
- (3) (p_{d^*}, q_{d^*}) : toxicity and efficacy probabilities of d^* ;
- (4) (p_{d^H}, q_{d^H}) : toxicity and efficacy probabilities of d^H ;
- (5) (p_{max}, q_{max}) : upper bounds of the generated toxicity and efficacy probabilities.

Table 6: Comparison of seven different performance metrics across the seven early phase designs under four scenarios for the simulation study

		OBD Exists				No OBD		
Scenario	Design	$p_{OBD}(\%)$	n_{OBD}	n_{over}	$p_{poor}(\%)$	$p_{no}(\%)$	$n_{no,over}$	n_{no}
	BOIN-ET	64.11	16.29	2.69	12.46	20.89	7.80	27.90
	BOIN12	63.04	14.55	3.10	14.53	37.61	7.03	25.43
	UBI	59.49	12.99	4.36	16.66	45.96	5.03	22.79
Increasing (I)	TEPI-2	61.68	13.24	4.23	14.64	45.89	5.04	22.80
	PRINTE	64.31	17.12	2.35	13.44	52.14	7.95	25.77
	STEIN	69.72	15.06	2.39	14.40	29.46	4.90	26.18
	uTPI	67.52	15.93	2.29	17.22	41.01	7.06	25.56
	BOIN-ET	62.02	19.68	2.15	16.90	22.64	5.97	28.35
	BOIN12	70.02	18.49	2.10	10.62	41.97	4.78	26.38
	UBI	78.81	11.64	3.38	36.41	48.17	3.83	24.00
Constant (C)	TEPI-2	75.41	11.93	3.31	34.90	48.03	3.87	24.01
	PRINTE	77.22	20.13	2.08	21.30	60.78	6.64	26.42
	STEIN	70.25	20.46	1.29	10.30	22.08	3.50	27.77
	uTPI	72.89	20.89	1.71	14.62	47.73	4.92	26.06
	BOIN-ET	72.04	18.33	2.32	9.93	37.24	6.82	27.38
	BOIN12	72.12	16.74	2.60	7.79	36.12	6.42	25.01
	UBI	67.72	13.37	4.01	18.07	41.03	5.19	23.42
Unimodal (U)	TEPI-2	70.44	13.57	3.93	16.69	40.96	5.17	23.42
	PRINTE	71.06	18.69	2.19	13.18	50.40	6.75	25.29
	STEIN	73.63	18.11	1.76	8.27	30.86	4.47	25.48
	uTPI	72.04	18.61	1.81	10.66	38.68	5.83	25.03
	BOIN-ET	67.30	18.04	2.67	10.51	22.76	7.24	27.84
	BOIN12	67.67	16.12	3.17	9.14	38.79	6.58	25.35
Increasing-	UBI	66.78	13.15	4.49	18.15	46.57	4.92	22.85
Plateau	TEPI-2	69.61	13.45	4.37	16.01	46.50	4.92	22.85
(IP)	PRINTE	70.25	18.42	2.49	13.37	54.54	7.52	25.70
	STEIN	73.64	17.52	2.22	8.86	29.81	4.61	26.09
	uTPI	70.26	18.09	2.27	11.90	42.12	6.48	25.46

⁽¹⁾ p_{OBD} : OBD selection rate; (2) n_{OBD} : average number of patients assigned to the OBD; (3) n_{over} : average number of overdose patients; (4) p_{poor} : poor allocation proportion; and two metrics when the OBD does not exist: (5) p_{no} : proportion of correctly not selecting any dose; (6) $n_{no,over}$: average number of overdose patients; (7) n_{no} : average number of treated patients.

Table 7: The selection probability of each dose level (DL) as the OBD (%) for all the designs under two scenarios: Increasing (I) and Increasing-Plateau (IP), and two designated true OBDs, for the case study

		Selection Probability (%)							
		Т	rue OBI	$O = DL_2$	2	True $OBD = DL3$			
Scenario	Design	DL1	DL2	DL3	DL4	DL1	DL2	DL3	DL4
	BOIN-ET	10.58	77.71	11.25	0.10	5.54	24.01	68.36	1.89
	BOIN12	12.19	71.75	14.90	0.27	12.30	17.69	66.18	3.12
	UBI	26.30	68.42	3.51	0.01	20.93	33.83	43.33	0.38
Increasing (I)	TEPI-2	20.97	73.95	3.33	0.00	17.77	30.99	49.45	0.26
	PRINTE	21.57	74.39	2.54	0.01	16.37	33.74	48.72	0.26
	STEIN	12.51	80.34	6.40	0.02	9.55	25.14	63.96	0.97
	uTPI	16.90	74.33	7.79	0.07	11.68	27.78	58.75	1.15
	BOIN-ET	10.86	81.97	6.77	0.04	5.57	24.63	67.90	1.60
	BOIN12	9.60	76.61	12.79	0.13	11.60	16.43	68.97	2.26
Increasing-	UBI	26.06	70.42	1.85	0.01	20.96	32.86	43.92	0.31
Plateau	TEPI-2	18.57	77.38	2.38	0.02	17.34	30.71	49.69	0.33
(IP)	PRINTE	20.96	76.39	1.32	0.02	16.67	33.70	47.98	0.26
	STEIN	12.03	83.51	3.81	0.01	9.48	25.82	63.41	0.84
	uTPI	16.56	78.03	4.69	0.05	11.57	26.88	59.91	0.96

Table 8: Comparison of four performance metrics across the seven early phase designs under two scenarios: Increasing (I) and Increasing-Plateau (IP), for the case study

Scenario	Design	$p_{OBD}(\%)$	n_{OBD}	n_{over}	$p_{poor}(\%)$
	BOIN-ET	73.03	10.74	0.69	23.13
	BOIN12	68.97	9.43	1.92	18.75
	UBI	55.88	9.19	2.92	19.14
Inceasing (I)	TEPI-2	61.70	9.19	2.91	18.39
	PRINTE	61.56	10.29	0.58	29.05
	STEIN	72.15	10.36	0.67	21.28
	uTPI	66.54	10.16	0.71	27.70
	BOIN-ET	74.94	10.99	0.45	22.94
	BOIN12	72.79	9.76	1.76	17.04
Increasing-	UBI	57.17	9.25	2.85	19.93
Plateau	TEPI-2	63.53	9.24	2.84	19.23
(IP)	PRINTE	62.19	10.46	0.39	28.78
	STEIN	73.46	10.68	0.46	20.96
	uTPI	68.97	10.47	0.50	27.02

(1) p_{OBD} : OBD selection rate; (2) n_{OBD} : average number of patients assigned to the OBD; (3) n_{over} : average number of overdose patients; (4) p_{poor} : poor allocation proportion.

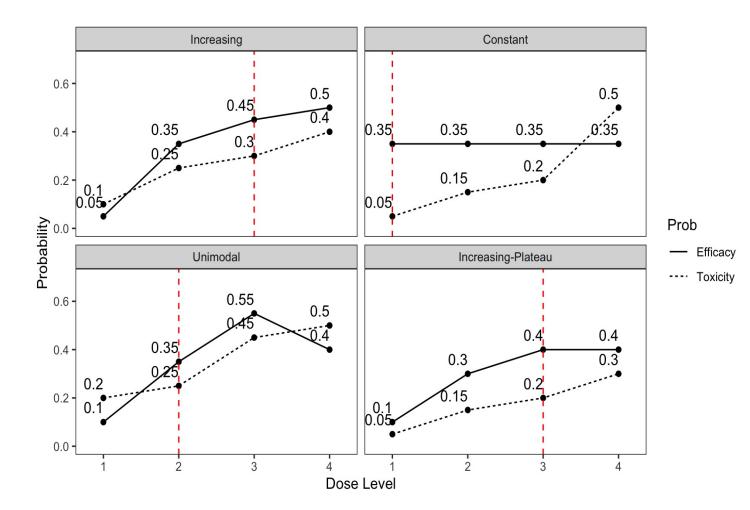


Figure 1: Example toxicity (dashed lines) and efficacy (solid lines) curves with true OBD (red vertical dashed lines) under four different dose-response relationships

(1) $d^U = 3$ for Scenario U; (2) $d^{IP} = 3$ for Scenario IP; (3) $(d^*, d^H) = \{(3, 4), (1, 1), (2, 3), (3, 3)\}$ for Scenarios I, C, U, and IP, respectively.