Title: Bayesian ordered lattice design for Phase I clinical trials

Authors:

1. Gi-Ming Wang, Case Comprehensive Cancer Center, Case Western Reserve University, OH 44106; Email: gxw174@case.edu
2. Curtis Tatsuoka (Corresponding author), Department of Epidemiology and Public Health, Division of Biostatistics and Bioinformatics, University of Maryland-Baltimore; Email: ctatsuoka@som.umaryland.edu

Summary:

We develop a new framework specifically for early Phase I clinical trials called Bayesian Ordered Lattice Design (BOLD). This study is motivated by two key factors. First, Phase I clinical trials typically involve relatively small sample sizes, making prior information regarding dose-limiting toxicity (DLT) highly significant. To address this challenge, the proposed Bayesian methodology incorporates prior information and posterior updating based on experimental results to guide dose selection, toxicity monitoring, early stopping, and identification of the maximum tolerable dose (MTD). Second, a natural ordering among toxicity probabilities across different dose levels can be utilized, with the idea being that analysis of dose-level posterior probabilities can and should acquire insights from data obtained at other dose levels, by leveraging their order relationship. Our proposed approach employs straightforward and intuitive dose-level Bayesian specifications and relies on intuitive and clinically interpretable toxicity-related posterior probabilities for decision making. Importantly, we show that it is either comparable to or outperforms popular methods, in terms of accuracy in determining the MTD and efficiency in

number of patients needed in testing. This Bayesian approach is also computationally simple and avoids simulation.

Key words: early Phase I clinical trial, Bayesian design, lattices, order, dose selection, maximum tolerable dose.

1. Introduction

Phase I clinical trials are designed to determine the maximum tolerable dose (MTD), which is the dose that has a probability of dose-limiting toxicity (DLT) nearest to a predetermined target rate (Liu and Yuan, 2015). The design comprises two primary elements. The first element involves a set of criteria for making decisions regarding the allocation of dose level to each participant. The second element entails a method for determining the MTD at the end of the trial (Braun, 2014). The process of identifying the appropriate dose for targeted therapies presents several challenges, including the need to enhance the efficiency of dose-finding efforts to reduce human and financial costs and a shorten development timelines (Ji et al., 2018; Xu, Zhang and Mu, 2022). From both practical and ethical perspectives, it is crucial to minimize decision-making errors and reduce the risk of exposing patients to either sub-therapeutic or excessively toxic doses (Liu and Yuan, 2015).

A variety of Phase I trial designs have been established to determine MTD for single-agent studies. These designs can be categorized into three distinct types based on their statistical principles and methods of execution: algorithm-based, model-based, and model-assisted designs (Zhou et al., 2018). The traditional 3+3 design is the most widely utilized algorithm-based approach in clinical trials, employing a straightforward set of predetermined rules to guide dose escalation and de-escalation (Zhou, 2021). This design continues to be the predominant choice for Phase I trials, primarily due to its ease of use (Yuan et al., 2016). Nevertheless, the 3+3 design has faced significant criticism for its inadequate operating characteristics (O'Quigley, Pete and Fisher, 1990), and its limited effectiveness in accurately identifying MTD and estimating DLT rates. Its heightened likelihood of administering sub-therapeutic doses to patients has also been noted (Yuan, Wu and Gilbert, 2021).

Model-based designs utilize a statistical model, such as a logistic model, to characterize the dose-toxicity relationship and facilitate dose adjustments (O'Quigley, Pete and Fisher, 1990). A prominent example of this design is the Continuous Reassessment Method (CRM), which initiates with an established dose-toxicity curve (O'Quigley, Pete and Fisher, 1990). As new data is collected, the CRM method iteratively refines the model estimates after each cohort, aiding in the determination of appropriate dose selection (Yuan, Wu and Gilbert, 2021). Compared to the traditional 3+3 design, CRM demonstrates superior accuracy in identifying and assigning a greater number of patients to MTD (Iasonos et al., 2008). However, it is important to note that the implementation of CRM involves inherent statistical and computational complexities, such as in the estimation and validation of the model for dose-toxicity (Yuan, Wu and Gilbert, 2021).

The model-assisted designs were created to integrate the benefits of both algorithm-based and model-based designs. Like model-based designs, model-assisted designs utilize a statistical model (e.g., the binomial model to estimate DLT rate) to aid in formulating decision rules. In a manner akin to algorithm-based designs, model-assisted designs can outline the rules for dose escalation and de-escalation, thereby simplifying implementation (Zhou et al., 2021). Common examples of model-assisted designs include TPI (toxicity probability interval design), mTPI (modified toxicity probability interval design) (Ji et al., 2010), Keyboard design (specifically, mTPI-2) (Yan, Mandrekar and Yuan, 2017), and Bayesian Optimal Interval Design (BOIN) (Yuan et al., 2016). Even though BOIN is presented as a Bayesian design, it juxtaposes Bayesian toxicity monitoring with empirical proportions that serve as the basis for dose selection. Its boundaries for escalation/de-escalation are derived from a Bayesian 3-class problem posed at each dose level, and Bayesian posteriors for the specific DLT rate values play no explicit role in

dose selection. Instead, intervals of DLT rates are boundaries for the empirical proportion that determine dose selection decisions.

We propose a Bayesian framework aiming to enhance the precision in determining the MTD, an unresolved issue in dose escalation trials. The primary motivation behind the proposed design is that Phase I clinical trials typically involve relatively small sample sizes, making prior information regarding dose toxicity highly significant. In clinical trials, the majority of conventional (frequentist) statistical techniques typically incorporate data from prior studies solely during the design phase, treating this information as supplementary rather than integral to the formal analysis (Gupta, 2012). By utilizing prior information in a Bayesian framework, the accuracy of dose selection can be enhanced.

Furthermore, since a natural ordering of toxicity rates emerges across different dose levels, this can be recognized and utilized to enhance both the precision and potentially the efficiency of determining MTD. The underlying premise is that higher dose levels are anticipated to have higher DLT rates (Conaway, Dunbar, and Peddada, 2004). The associated order constraints may be explicitly recognized through the posterior probabilities related to DLT rates across different doses (Yin and Yuan, 2009). This linkage of posterior probabilities to reflect the expectation of a natural order among dose DLT values allows for “learning” from response data across dose levels.

The objective of this research is to create a Bayesian Ordered Lattice Design (BOLD) that integrates prior knowledge and ordering constraints with empirical data of toxicity for the purposes of adaptive dose selection, early stopping and classification of MTD in Phase I clinical trials. For the proposed BOLD, the incorporation of even weakly informative prior knowledge

regarding dose toxicity in an order-constrained Bayesian manner can improve accuracy in scenarios where sample sizes are small.

The term "lattice" originates from the idea that the dose levels in combination-drug trials are considered as partially ordered within a lattice structure, while the dose levels in single-drug trials can be perceived as a linearly ordered lattice. Web Figure 1 illustrates the lattice diagram featuring five dose levels, which also includes top and bottom states. The bottom state signifies that a MTD is lower than the minimum dose. The top state signifies that the MTD is higher than the maximum dose. Classification is to one state in the lattice; to a specific dose level indicates that the dose is MTD. The term “lattice” also signifies that BOLD is relevant to both single-drug and combination-drug trials, and that the notion of order is central. We only consider single drug trials here.

2. Methodology

Prior and conjugate distribution

The implementation of BOLD begins by assigning dose-level prior distributions for the prior respective DLT rate. With binary outcomes, we employ Beta distributions, which are represented as $\pi_j \sim Beta(\alpha_j, \beta_j)$ for dose level j. Importantly, Beta distributions have conjugate posteriors $\pi_j | x_j, n_j \sim Beta\ (\alpha_j + x_j,\ \ \beta_j + n_j - x_j)$, where n_j represents the total number of patients who have received treatment at dose level *j,* and x_j denotes the total number of observed DLTs at that dose *j*, with the condition $0 \leq x_j \leq n_j$. Furthermore, x_j is modeled as following a Binomial distribution, specifically $x_j \sim Binomial(n_j,\ \pi_j)$. Bayesian updating is conducted after each stage, which includes response data from the selected dose’s administration with the pre-selected administration cohort size.

In prior specification, we adopt the following guiding principles: 1) It is assumed that the prior distributions of respective DLT rates reflects the ordering of the dose levels, e.g. such as in terms of the prior means; 2) Priors should not be excessively informative, in that each dose must have the potential to be identified as MTD, even when only limited sample sizes are available. This principle supports equipoise. Specifically, this entails that the variances of the prior DLT rates should not be too small. 3) The prior means should also fall within a reasonable proximity to the target rate ϕ.

In the case of trials assuming non-informative priors, we define the parameters of the prior distribution (α and β) by setting the prior mean equal to the target rate, while the prior effective sample size (PESS), which is α+β, is suggested to match the standard cohort size at a given stage of administration. This would lead to a relatively weak prior in the sense that PESS is equal to one stage of dose administrations. It is interesting to note that if we were to assume the priors adhere to a uniform distribution, specifically $Beta(1,1)$, the prior mean would be 0.5, which significantly exceeds target toxicity values. PESS would be equal to 2, which is weak. For informative priors, we will for instance specify the PESS of the dose considered most likely to be the MTD to be larger than other doses, which is equivalent to assigning a smaller prior variance.

Order constraints on posterior probabilities

A key idea of BOLD is to impose order constraints to “share” updated information across dose levels after each stage. This contrasts with BOIN, the design with which we will make a comparison, which only imposes order constraints across the estimated DLT rates at the end of the trial, for the classification of the MTD. Since the collection of dose levels under consideration comprise a linear lattice, each dose has a cover and/or anti-cover dose. The cover

dose is the dose level directly above the current dose with no dose level in-between, and anti-cover is the dose level directly below. These terms will be utilized in the text below.

Another important feature is that inference for dose selection, toxicity monitoring, and early stopping rely on the same statistical basis, which we refer to as the Conjugate Probability of the DLT rate being Above the Target (CPAT), expressed as

$$CPAT = P\left(\pi_j > \phi \mid x_j, n_j\right) \qquad (1)$$

Note that this posterior probability value provides the most direct and up-to-date insight on the most pertinent clinical question: Is the dose's DLT rate safe or not, relative to the target ϕ? Larger values of equation (1) indicate that a dose's DLT rate is more likely to be unacceptably high.

Order constraints on CPAT values can be implemented using the Pool-Adjacent-Violators Algorithm (PAVA) (Barlow et al., 1972), an isotonic regression method that is appropriate for toxicity which is presumed to be non-decreasing with dose (Leung and Wang, 2001; Ivanova and Flournoy, 2009). PAVA substitutes any neighboring values that violate the non-decreasing ordering with a weighted average, with the number of treated patients as a weighting factor. This ensures that the resulting estimates exhibit monotonicity (Liu and Yuan, 2015). We apply PAVA in a "local" manner, in the sense that only the CPATs of the current dose and its cover and anti-cover doses are included. This is reasonable, since we only consider escalation/de-escalation in one dose level increments, as does BOIN. The corresponding modified CPAT value following PAVA is referred to as the Posterior Probability of DLT rates being Above the Target (PPAT).

Dose selection

In contrast to approaches like BOIN that conduct the decisions of dose selection (escalation/de-escalation) solely relying on a current dose's empirical DLT proportion value, BOLD jointly considers multiple dose levels in selection decisions based on the "local" PPAT values. Typically, at a given stage, multiple patients receive the current dose prior to the selection of the subsequent dose (e.g., $n = 3$). We restrict dose selection to either escalation, de-escalation, or maintaining the current dose. The proposed selection criterion involves identifying the dose for which the PPAT is nearest to 0.5 among the current dose and the cover/anti-cover doses. If we denote this set of dose levels as *J*, then specifically among *J*, dose *j* is selected if it minimizes equation (2):

$$|PPAT_j - 0.5| \qquad (2)$$

A PPAT value of 0.5 signifies the highest level of uncertainty regarding the safety of that dose level (akin to a coin flip). Following observation from this stage, no matter the responses, we should gain improved insights one way or the other into whether the dose's DLT rate is above or below the target.

When two values of $|PPAT - 0.5|$ are equal, i.e., $\left|PPAT_j - 0.5\right| = |PPAT_k - 0.5|$, the proposed decision rule is as follows:

Case 1: If both PPAT values are below 0.5, we will proceed with the higher dose.

Case 2: In the event that both of the PPAT values exceed 0.5, we will opt for the lower dose.

Case 3: If one PPAT is below 0.5 while the other exceeds this threshold, we opt for the lower dose.

Toxicity control

During experimentation, we will remove doses that are deemed excessively toxic. A dose is classified as excessively toxic if its CPAT exceeds a predetermined threshold, for instance, 0.95. The threshold for the lowest dose may be adjusted to 0.9, to help enable the identification that all doses are too toxic, which is a difficult classification problem due to data not directly being observed from that condition. Should a dose be identified as excessively toxic, it, along with any higher doses, will be excluded from further evaluation.

Stopping rule

The trial is halted upon the fulfillment of either of the following criteria:

Case 1: The minimum dose is excessively toxic, as indicated by its CPAT exceeding the toxicity threshold, suggesting that the MTD is lower than this minimum dose level.

Case 2: The cumulative count of patients who have received treatment attains a predetermined figure (e.g., $N = 30$).

Case 3: The overall number of patients undergoing treatment at any dose has reached a specified maximum threshold (e.g., 12), and the decision has been made to maintain the current dose level once this threshold has been met or exceeded. The threshold for the number of patients receiving the minimum dose is established at 15, to allow for distinguishing whether this lowest dose is too toxic or not.

MTD identification

The MTD will be determined upon the conclusion of the trial. The identification of the MTD relies on isotonically-regressed posterior means of DLT rates. Specifically, we apply PAVA “locally” to the conjugate posterior means of the final dose, and its cover and anti-cover doses, and consider these as the candidate doses. We have found that this can give improved overall accuracy versus considering all posterior means after isotonic regression. The candidates

for MTD must possess trial data (e.g., at least 3 patients) by the final stage. The dose *j* with its value of $\hat{\mu}_{pava(j)}$ being nearest to ϕ and less than the specified upper boundary (e.g., 1.4ϕ) will be designated as the MTD. The determination of MTD is not feasible if the lowest dose administered is excessively toxic, which is reported as all doses being deemed as too toxic. If a tie occurs (i.e., $|\hat{\mu}_{pava(k)} - \phi| = |\hat{\mu}_{pava(k)} - \phi|$), the proposed method for identification is as follows:

Case 1: If $\hat{\mu}_{pava(j)}$ is less than ϕ and $\hat{\mu}_{pava(k)}$ is greater than ϕ, then dose *j* is selected as the MTD

Case 2: If $\hat{\mu}_{pava(j)}$ equals $\hat{\mu}_{pava(k)}$ and both are less than ϕ, with level *k* being higher than level *j*, then dose *k* is selected.

Case 3: If $\hat{\mu}_{pava(j)}$ equals $\hat{\mu}_{pava(k)}$ and both are greater than ϕ, with level *k* being higher than level *j*, then dose *j* is selected.

Flowchart of BOLD

The flowchart that outlines and summarizes the procedures and guidelines for dose selection is presented in Figure 1. A concise description of the flowchart is as follows: 1) Specify the target DLT rate ϕ. 2) Assess prior information, determining whether the prior distribution will be informative or non-informative, and subsequently assign the Beta prior parameters that determine prior mean and variance, as well as toxicity threshold for each dose level. 3) Initiate the trial at the lowest dose. 4) Administer the dose level to a small cohort of patients (e.g., 3). 5) Update the CPAT values. 6) Implement overdose control and consider an early termination by assessing the CPAT values across doses. 7) Monitor the overall sample size ceiling. 8) Implement order constraints on the updated CPAT values to obtain PPAT values locally (anti-

cover, current and cover dose). 9) Select the next dose to either escalate, de-escalate or maintain the current dose based on minimizing the PPAT criterion in equation (2). 10) Monitor the sample size ceiling for the current dose for early stopping if the decision in step 9 is to maintain the current dose. 11) Repeat another stage of experimentation from steps 4 to 10 until the trial concludes. 12) Once stopping occurs, determine the MTD using the posterior mean DLT rates (with PAVA) except when the lowest dose exhibits excessive toxicity, so that it is determined that there is no MTD and that all doses are excessively toxic.

3. Comparators

To assess the performance of the BOLD design, a comparative analysis is conducted against the BOIN design. Furthermore, the 3+3 design (Le Tourneau, Lee and Siu, 2009), which is commonly adopted and shares a design structure with both BOIN and BOLD, will also be part of the comparison.

Justification of comparison to BOIN

We will focus on comparing BOLD to BOIN, in terms of methodological differences, due to the following: 1) In contrast to algorithm-based design, BOIN is more inclined than the 3+3 design to accurately identify the MTD and to assign a greater number of patients to this dose (Yuan, Wu and Gilbert, 2021). 2) In contrast to CRM, BOIN employs a nonparametric methodology, rendering it more robust (Yuan, Wu and Gilbert, 2021). 3) BOIN demonstrates superior performance over the mTPI design, significantly reducing the risk of patient overdose while enhancing the likelihood of accurately identifying the MTD (Yuan et al., 2016). BOIN also exhibits operating characteristics that are equal to or superior to those of the mTPI-2 design (i.e., keyboard design), which is an evolution of the original mTPI design (Ananthakrishnan et al., 2022). 4) Finally, BOIN is widely adopted.

Outline of BOIN

For the BOIN design, one of three dose-selection actions is executed based on the observed data: either escalating, de-escalating, or maintaining the current dose. BOIN evaluates the current dose's observed DLT rate against predetermined boundaries for escalation (λ_e) and de-escalation (λ_d) (Liu and Yuan, 2015). The boundaries are established within the Bayesian hypothesis framework by choosing an interval of indifference (e.g. $[0.6\phi, 1.4\phi]$), and considering a 3-state classification of intervals: the true DLT for the dose is in $(0, 0.6\phi)$, within $[0.6\phi, 1.4\phi]$, or in $(1.4\phi, 1)$. This framework operates under the prior assumption that the dose DLT rate has an equal probability of being in any of these intervals.

BOLD follows BOIN in how stopping is implemented through Bayesian toxicity monitoring and use of dose and study level sample size limits. BOIN by default relies on a uniform prior distribution for DLT rate, specifically $Beta(1,1)$, while BOLD relies on the specified dose-level priors. This is how results appear to differ in classification performance to the bottom state.

Methodological comparison of BOLD to BOIN

Briefly, BOLD differs from BOIN in the following ways: 1) BOLD specifies prior probabilities for specific DLT values, not on intervals. Bayesian analysis and prior specification is thus more precise with BOLD, and this is reflected by improved performance when informative priors are warranted. 2) No indifference interval needs to be specified or relied upon; 3) Decisions on dose selection and toxicity are conducted in a unified manner by relying on the same statistic, CPAT, which has the natural interpretation as the posterior probability that the DLT rate for a dose exceeds the target ϕ. In contrast, BOIN mixes together statistics and models in the monitoring of toxicity (Beta posterior) with the empirical proportion of DLT in dose

selection; 4) Dose selection is conducted by comparing order-constrained PPAT values from across different dose levels, while BOIN conducts dose selection based on the current dose's empirical DLT rate, and hence, one at a time, only focuses on a single dose's data. 5) Isotonic regression and implementation of order constraints is conducted by BOLD at each stage of dose selection to enhance shared learning across dose levels. In contrast, BOIN utilizes isotonic regression only at the MTD selection stage.

We will now illustrate the performance of BOLD designs through simulations in the following section and perform a comparative analysis with BOIN and 3+3 designs, to assess their respective accuracy and efficiency in a range of early phase 1 clinical trial settings.

4. Simulation

With non-informative prior specification

To evaluate the performance of BOLD with non-informative prior, we perform a comparative analysis versus BOIN and 3+3 designs through simulations conducted in R(4.0.4). In general, our default prior *Beta*(α, β) distribution specification is to set for all doses the prior mean to be ϕ, which reflects equipoise, and PESS (i.e., α+β) = 3, the standard cohort size. These are thus weak priors. For □=0.2, this is equivalent to *Beta*(0.6, 2.4); when ϕ= 0.25, it is *Beta*(0.75, 2.25). We assume that each cohort per dose administration consists of a cohort size of 3 participants. The maximum number of patients to be treated throughout a trial is capped at 30. Stopping occurs when the number of patients receiving any specific dose reaches 12 or greater and if concurrently the next-stage dose selection decision is to continue with the current dose. An exception is made for the lowest dose, which stops at 15. A trial will also be stopped early if the toxicity threshold for dose 1 is exceeded. This threshold is set at 0.95 for all doses, except for the lowest dose (dose 1). For dose 1, we set the threshold to be 0.9 to enhance classification to dose

0, which has clinical importance by denoting all doses as too toxic. Moreover, dose 0 is only classified indirectly, since it is not a true dose. This less stringent threshold at dose 1 facilitates classification to dose 0.

The matrices representing the true DLT rates for simulation across dose levels corresponding to trials with 5, 4, 3, and 2 dose levels, respectively, at □ values of 0.2 and 0.25 are displayed in Table 1. Two sets of true DLT rates matrices are employed: one is an example of matrix, and the other is the default matrix suggested by the BOIN app. Multiple scenarios are utilized for simulation purposes. For *b* doses ($b = 2, 3, 4$ or 5), *b*+2 scenarios are studied, including scenario 0 when all doses are too toxic, and scenario *b*+1 when all doses have DLT rates below target. The MTD of scenario *k* corresponds to dose level *k*, signifying that the DLT rate at dose level *k* is equal to ϕ.

A difference between the example true DLT rate matrix and the BOIN's default matrix is that the true DLT rates for lower doses selected by the BOIN app are markedly lower than those of the example we propose. In the case of a 5-dose lattice with □=0.25, the true DLT rates of the example for the lowest dose in scenarios 3 to 6 are set at 0.1. In contrast, the default rates given by the BOIN app are 0.02 for scenario 3 and 0.01 for scenarios 4 to 6. The rationale behind the example true DLT rates for lower doses being relatively higher stems from our assumption of parsimony, in that each dose is selected to be a viable candidate with some probability of DLT.

The comparative analysis focuses on both accuracy and efficiency, primarily on the dose level. Accuracy is quantified by the proportion of trials that successfully identify the MTD, while efficiency is assessed by the mean and standard deviation of the number of patients treated in each trial. Secondarily, we also consider overdosing rates (the proportion of the patients treated with doses higher than the actual MTD), the underestimation rates (the proportion of trials with

identified MTD being lower than the actual MTD for a given scenario), and overestimation rates (the proportion of trials with identified MTD being higher than the actual MTD for a given scenario). The simulation executes 1,000 iterations for BOLD, BOIN (via the BOIN app) and 3+3 design.

With informative prior specification

With informative priors, we conduct a comparative analysis with iBOIN, which incorporates historical data into BOIN. This historical information regarding DLT rates is termed the "skeleton" (Zhou et al., 2021), representing the anticipated prior DLT values for each of the doses with a corresponding PESS value. Importantly, this step does not involve specifying associated prior distributions for dose-level DLT rate values and hence does not directly employ standard Bayesian updating. The iBOIN design maintains the same decision-making framework as BOIN; however, it distinguishes itself by incorporating historical data into the dose selection boundaries, which can be tailored to specific doses. In the absence of any prior information, iBOIN becomes equivalent to BOIN in that the boundaries align when the skeleton values are all equal to ϕ and PESS=3.

Two types of informative prior specifications that we consider are: 1) A preferred dose is associated with a greater PESS value, such as 6 as opposed to 3, in contrast to the other doses, while the prior means are all equal to □. Note that the larger PESS value indicates relatively smaller posterior variance, given the same prior mean, and hence is more informative. 2) Assume that the prior means (skeletons for iBOIN) are mildly distinct, with the preferred dose having the prior mean equal to □ while other prior means (skeletons for iBOIN) being within a close range (e.g., ± 0.05 from □) as the values satisfies the linear order, and PESS = 3 for all doses. Both approaches preserve equipoise and are only mildly informative. For each type of informative

prior specifications, simulations for 5 dose levels are utilized to conduct comparisons. The toxicity threshold is set at 0.95 for all doses and 0.9 for dose 1, as utilized for non-informative prior. The simulation executes 1,000 iterations for BOLD and iBOIN (via the iBOIN app).

5. Results

Figure 2 illustrates the bar chart representation of accuracy and efficiency for BOLD, BOIN and 3+3 with non-informative prior and 5-dose lattice by scenario at $\phi = 0.2$ and 0.25. The matrices representing the example true DLT rates for the 5-dose lattice at =0.15 and 0.3, along with the associated bar charts illustrating accuracy and efficiency, are available in Web Table 1 and Web Figure 2 of the Supplementary Material, respectively. The results obtained from the simulations demonstrate that the overall performance of BOLD, when assuming non-informative priors, clearly exceeds that of the 3+3 design and is generally superior to BOIN in terms of accuracy, particularly for the example true DLT matrices. The superiority of BOLD over both the BOIN and 3+3 designs in accuracy also remains apparent when we calculate the average accuracy rates for scenarios 1 through 5 only, while omitting scenarios 0 and 6. BOIN performs poorly at the higher dose levels for the example DLT matrices that we proposed.

The findings from the simulation also indicate that, regarding efficiency, the 3+3 design requires a considerably smaller patient sample than the BOLD and BOIN designs. Nevertheless, it exhibits a markedly lower accuracy, particularly at the intermediate dose levels. The standard deviations (SD) of the number of patients are also indicated on the bar charts corresponding to dose levels. Note that the SD's for BOIN are actually determined by our own code, as the BOIN app does not offer this information (see also the overdosing, understimation, and overestimation rates below).

One can see from the bar charts that BOIN can require a bit fewer number of patients when the MTD is at lower doses, while BOLD requires less when the MTD is at higher dose levels. This could be due to BOLD being quicker to escalate, which does lead to some relatively higher overdosing. Web Tables 2 and 3 present the overdosing, underestimation, and overestimation rates associated with BOLD, BOIN, and 3+3 designs by scenario with non-informative priors and 5-dose lattice at ϕ values of 0.15, 0.2, 0.25, and 0.3, utilizing the example true DLT rate matrices and BOIN's default matrices, respectively. Note that for scenario 5 where dose 5 is the true MTD, neither overdosing nor overestimating occurs. In scenario 6, there is also no risk of overdosing or overestimating, as no dose is toxic. Similarly, there is no possibility of underestimating in scenario 0, where every dose is toxic. Conversely, the tables indicate that the overdosing and overestimation rates for BOLD are higher than those of BOIN design, while the underestimation rates of BOLD are comparatively lower. Note that differences in overestimation are small for the most part, while the corresponding discrepancies in underestimation rates are almost always much larger, in favor of BOLD.

Table 2 illustrates the average accuracy and efficiency of BOLD, BOIN and 3+3 design with non-informative priors across scenarios by lattice at ϕ=0.2 and 0.25, utilizing both the example and BOIN's default DLT rate matrices. Web Tables 4 to 7 illustrate the results at the dose level. The results obtained from the simulation demonstrate that the performance of BOLD exceeds that of the 3+3 design and is either superior to or at least on par with BOIN in terms of accuracy across lattice dimensions (i.e., for 2, 3, 4 and 5 doses). This is strikingly clear with the example DLT rate matrices that we proposed, which are quite realistic. In addition to accuracy rates, these tables also present the efficiency, illustrated by the average number of patients treated along with their respective average standard deviations for different lattices. The number

of patients depicts the "cost" associated with the increase in the number of dose levels. Note that there is a decline in the accuracy rate as the number of dose levels rises, accompanied by a reduction in efficiency, as evidenced by the larger average number of patients receiving treatment. This phenomenon is observed across all designs, including BOLD, BOIN, and 3+3 design.

Figure 3 illustrates the bar charts representation of accuracy and efficiency of BOLD and iBOIN with informative PESS and 5-dose lattice by scenario (1 to 5) at ϕ=0.2 and 0.25, utilizing both the example and BOIN's default DLT matrices. It is assumed that the preferred dose is the true MTD and corresponds to a PESS value of 6, whereas the PESS for the other doses is set at 3. The results of the simulation suggest that BOLD, when assuming informative priors, generally outperforms iBOIN in terms of accuracy, and that there is greater sensitivity to this information as reflected by an even greater difference in accuracy compared with non-informative settings. Another approach to informative prior specification is to adjust the prior means. Figure 4 illustrates the bar charts representation of accuracy and efficiency of BOLD and iBOIN with informative prior means (skeletons for iBOIN) and 5-dose lattice by scenario (1 to 5) at ϕ=0.2 and 0.25. It is also presumed that the preferred dose represents the true MTD and aligns with the prior mean (skeleton) designated as ϕ. In this context, the prior means (skeletons) are uniformly established at ϕ+0.05 for doses exceeding ϕ and at ϕ-0.05 for doses below ϕ. The results of the simulation also suggest that BOLD generally outperforms iBOIN in terms of accuracy, and strikingly so for the example DLT matrices.

We also perform sensitivity analyses related to prior specifications. In the trials with informative prior, we have assumed so far in our simulations that the preferred dose corresponds to the true MTD, so that our prior specifications lean "correctly". As these informative

specifications may incorrectly favor the wrong dose in real-world applications, we perform sensitivity analyses to evaluate the effectiveness of the proposed design under misspecification. Two possible situations concerning the preferred doses are examined: either the anti-cover dose or the cover dose of the preferred dose is indicative of the true MTD. The findings are detailed in Web Tables 8 to 11 of the supplementary material for $\phi = 0.2$ and 0.25, utilizing both the example and BOIN's default matrices, respectively. The simulation demonstrates that BOLD generally exhibits superior accuracy compared to iBOIN in most scenarios even when the preferred dose diverges from the true MTD.

6. Discussion

For the respective non-informative settings, the simulations indicate that BOLD can match or decisively outperform both BOIN and 3+3 design across a range of scenarios in terms of accuracy. A clear relative advantage emerges depending on the "shape" of the toxicity-dose response. Particularly, when there is a non-zero "floor" for the lowest DLT rate across the dose levels, say 0.10, as in the example set of DLT rates, then BOIN's compartmentalized strategy of testing the same dose level until an escalation/de-escalation decision can be costly, particularly when there are several dose levels to consider. Note that in the BOIN default set of DLT rates, when the highest dose is MTD, the lowest doses have small (near zero) DLT rates. Consequently, BOIN's approach is not penalized, as testing at the lower doses is swift. However, note in the example set of DLT rates that the lowest value is 0.10, reflecting a non-zero "floor". BOIN can then require more patients to correctly escalate. In this case, and in general, we see that the "shared learning" implemented in BOLD, as reflected by imposing order constraints at each stage of dose selection, and the direct comparison of posterior toxicity probabilities across doses for selection, lead to faster decisions to escalate/de-escalate. This has some drawback for

accuracy relating to the case when all doses have DLT rates above the target (Scenario 0 in the simulations), but this can be compensated for by allowing for a higher ceiling on patient numbers and setting slightly less stringent toxicity thresholds for the lowest dose level, as we did in the simulations. Note that even for BOIN's default DLT rates, BOLD generally performs just as good or even slightly better, and only when the target is high (ϕ= 0.30) are there certain scenarios when BOIN has a slight advantage.

In terms of efficiency (i.e., the number of required patients), we see that there is no clear advantage between BOLD and BOIN at the given settings. Any differences, while not large, depend on dose and target. Note that we compensate for the scenarios when all doses are too toxic by allowing for a higher ceiling of patients at lowest dose, which affects the overall average. However, when the highest dose is MTD, the average number of patients in the simulated trials is lower for BOLD. This is in line with BOLD's proclivity to change dose levels more quickly. We also monitor for overdosing, underestimation, and overestimation rates. A trade-off that results from the tendency for BOLD to switch doses faster is a higher overdosing rate while achieving a lower underestimation rate.

In Table 2 and Web Tables 4 to 7, we note the similar patterns in relative performance arise, across lattice sizes (number of dose levels). BOLD holds the clear advantage in accuracy for the example set of DLT rates across lattice sizes and also for BOIN's default DLT rates at ϕ values of 0.20 and 0.25. Accuracies are comparable for BOIN's default DLT rates in conjunction with higher ϕ values, and perhaps a slight edge for BOIN when a high level of ϕ = 0.30. Given concerns about sustainability of treatments from FDA Project Optimus (Thall et. al, 2024), a DLT rate of 0.30 may at some point be viewed as unacceptably high in many disease scenarios. Moreover, given the reduction in number of patients needed for MTD classification as the

number of doses decreases, it is important not to include superfluous dose levels for consideration. Trying to only include doses with DLT rates above a non-zero floor benefits overall efficiency. Hence, the accuracy advantages of BOLD importantly apply to practically useful sets of DLT rates.

As mentioned, BOLD involves specifying prior Beta probability distributions on specific DLT values, which is recognizable, intuitive and in line with formal Bayesian analysis. In contrast, iBOIN uses skeletons with PESS values to reflect prior information, and prior probability distributions are not explicitly specified. In terms of relative performance with informative priors, one can see that the benefits in accuracy become even more apparent with BOLD's approach. BOLD generally outperforms iBOIN when employing informative priors, even under sensitivity analysis.

In contrast to BOIN (and iBOIN), which utilizes posteriors solely for the purpose of toxicity monitoring independent of its dose selection criteria, BOLD incorporates the Beta posteriors into all decision-making processes. BOIN (and iBOIN) recognizes order constraints only at the final stage for MTD identification, yet BOLD employs order constraints throughout each stage. Wang and Ivanova (2014) employ constraints at each stage across sampled dose-level means as well in a Bayesian continuous response model, but their approach involves simulating order constrained posteriors using PAVA for each sampled set of values. We instead only apply PAVA on the dose-level posterior probability values for the event that the DLT rate exceeds threshold (Equation 1). Importantly, this provides a computational advantage. PAVA is only applied to one set of values, and extensive simulation is averted. Computations are thus quite fast. This toxicity-related posterior-based statistic encompasses information from the full posterior distribution in sense that all the posterior mass is reflected (the complement is known as well).

This appears to be more informative than posterior probabilities of an interval around target, which reflect only a portion of the posterior mass. Hence, our approach is novel.

6. Conclusion

In sum, we propose a new, computationally fast approach that demonstrates clear conceptual and performance advantages in Bayesian dose finding trials over popular and widely adopted methods such as BOIN and 3+3. The primary characteristic of the proposed Bayesian Ordered Lattice Design (BOLD) is its integration of posterior information with order constraints that are imposed at each decision stage. This enables the sharing of information across doses.

Another innovation of BOLD is the application of a unified posterior probability-based criterion (probability that dose DLT exceeds the toxicity threshold) for dose selection and early stopping. Simulation studies validate that the performance of BOLD can surpass BOIN with non-informative priors and surpasses iBOIN when informative priors are utilized. This advantage depends on the toxicity-dose relationship. It is quite manifest when each of the doses has a DLT probability clearly bounded away from zero, which is a natural and parsimonious setting, as inclusion of unnecessary lower doses in a trial adversely affects efficiency.

Furthermore, the BOLD design framework can be generalized and adapted for combination-drug trials and continuous toxicity responses, along with augmented backfill designs. These designs will be elaborated upon in forthcoming work.

7. References

1. Ananthakrishnan, R., Lin R., He C., et al. *An overview of the BOIN design and its current extensions for novel early-phase oncology trials.* Contemp Clin Trials Commun, 2022. **28**: p. 100943.

2. Barlow, R. E., Bartholomew, D. J., Bremner, J. M. and Brunk, H. D. *Statistical Inference under Order Restrictions*. 1972: London, New York: John Wiley & Sons.
3. Braun T. M. *The current design of oncology Phase I clinical trials: progressing from algorithms to statistical models.* Chin Clin Oncol. 2014 Mar; 3(1):2.
4. Conaway, M. R., Dunbar S., and Peddada, S. D. *Designs for single- or multiple-agent Phase I trials.* Biometrics, 2004. **60**(3): p. 661-9.
5. Gupta S. K. *Use of Bayesian statistics in drug development: Advantages and challenges*. Int J Appl Basic Med Res. 2012 Jan; 2(1):3-6.
6. Iasonos, A., Wilton A. S., Riedel, E. R., et al. *A comprehensive comparison of the continual reassessment method to the standard 3 + 3 dose escalation scheme in Phase I dose-finding studies.* Clin Trials, 2008. **5**(5): p. 465-77.
7. Ivanova, A. and Flournoy N. *Comparison of Isotonic Designs for Dose-Finding.* Stat Biopharm, Res. 2009 Feb 1; 1(1):101.
8. Ivanova, A. and Kim S. H. *Dose finding for continuous and ordinal outcomes with a monotone objective function: a unified approach*. Biometrics. 2009 Mar; 65(1):307-15.
9. Ji, Y., Jin J. Y., Hyman, D.M., et al. *Challenges and Opportunities in Dose Finding in Oncology and Immuno-oncology.* Clin Transl Sci, 2018. 11(4): p. 345-351.
10. Ji, Y., Liu P., Li Y., et al. *A modified toxicity probability interval method for dose-finding trials.* Clin Trials, 2010. 7(6): p. 653-63.
11. Le Tourneau C, Lee J. J., and Siu L. L. *Dose escalation methods in Phase I cancer clinical trials. J Natl Cancer Inst*. 2009 May 20;101(10):708-20
12. Lee S. Y. *A flexible dose-response modeling framework based on continuous toxicity outcomes in Phase I cancer clinical trials.* Trials. 2023 Nov 21;24(1):745.

13. Leung D. H. and Wang Y. *Isotonic designs for Phase I trials*. Control Clin Trials. 2001 Apr; 22(2):126-38.

14. Liu S, Yuan, Y. *Bayesian optimal interval designs for Phase I clinical trials.* Journal of the Royal Statistical Society, 2015. 64: p. 507-523.

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

16. Thall, P. F., Garrett-Mayer, E., Wages, N. A., et. al. *Current issues in dose-finding designs: A response to the US Food and Drug Administration's Oncology Center of Excellence Project Optimus.* Clin Trials. 2024 Jun; 21(3): 267-272.

17. Wang, Y., Ivanova, A. *Dose finding with continuous outcome in Phase I oncology trials*. Pharm Stat. 2014 Mar-Apr; 14(2):102-7

18. Xu, J., Zhang, D. and Mu, R. *A dose-finding design for Phase I clinical trials based on Bayesian stochastic approximation.* BMC Med Res Methodol, 2022. 22(1): p. 258.

19. Yan, F., Mandrekar, S. J. and Yuan, Y. *Keyboard: A Novel Bayesian Toxicity Probability Interval Design for Phase I Clinical Trials.* Clin Cancer Res, 2017. 23(15): p. 3994-4003.

20. Yin, G., Yuan, Y., *Bayesian dose finding in oncology for drug combinations by copula regression.* Royal Statistical Society Series C: Applied Statistics, 2009. 58: p. 211-224.

21. Yuan, Y., Hess, K.R., Hilsenbeck, S.G., et al. *Bayesian Optimal Interval Design: A Simple and Well-Performing Design for Phase I Oncology Trials.* Clin Cancer Res, 2016. 22(17): p. 4291-301.

22. Yuan, Y., Wu, J. and Gilbert, M. R. *BOIN: a novel Bayesian design platform to accelerate early phase brain tumor clinical trials.* Neurooncol Pract, 2021. 8(6): p. 627-638.

23. Zhou Y, Lee J. J., Wang, S., et al. *Incorporating historical information to improve Phase I clinical trials*. Pharm Stat. 2021 Nov; 20(6):1017-1034

24. Zhou, Y., Li, R., Yan, F., et al. *A comparative study of Bayesian optimal interval (BOIN) design with interval 3+3 (i3+3) design for Phase I oncology dose-finding trials.* Stat Biopharm Res, 2021. 13(2): p. 147-155.

25. Zhou, H., Murray, T. A., Pan, H., et al. *Comparative review of novel model-assisted designs for Phase I clinical trials.* Stat Med, 2018. 37(14): p. 2208-2222.

26. Zhou, Y., Lin, R., Kuo, Y. W., et al. *BOIN Suite: A Software Platform to Design and Implement Novel Early-Phase Clinical Trials.* JCO Clin Cancer Inform, 2021. 5: p. 91-101.

Figure 1. Flowchart of BOLD design

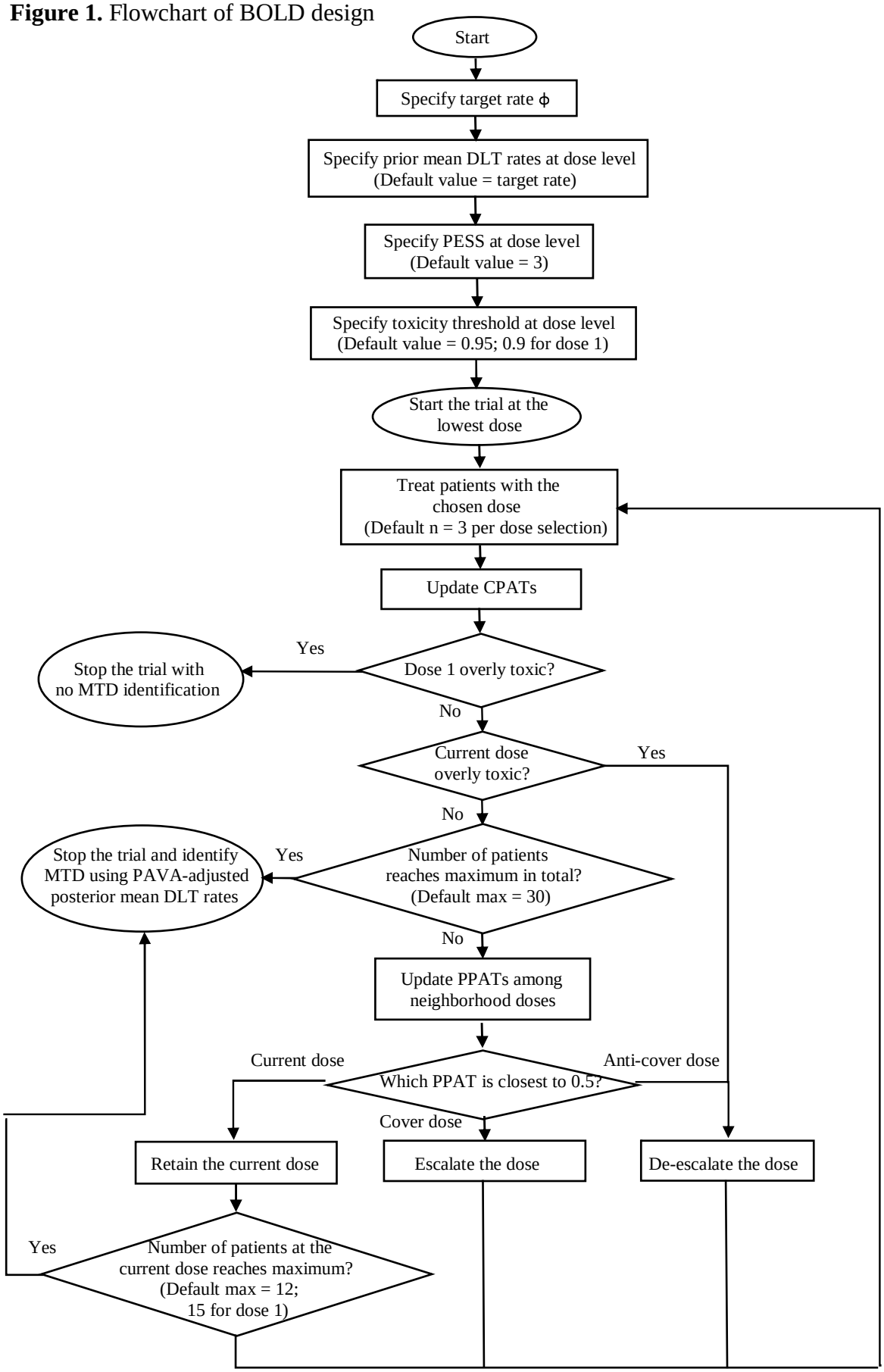

Table 1. Matrices of true DLT rates for simulations at dose (D) levels by lattice size at ϕ=0.2 and 0.25

Target rate & Lattice	Scenario	Example true DLT rate matrix					BOIN's default DLT rate matrix				
		D1	D2	D3	D4	D5	D1	D2	D3	D4	D5
□=0.2 & 5-dose lattice	0	0.45	0.50	0.55	0.60	0.65	0.36	0.40	0.44	0.48	0.52
	1	0.20	0.45	0.50	0.55	0.60	0.20	0.36	0.40	0.44	0.48
	2	0.10	0.20	0.45	0.50	0.55	0.10	0.20	0.37	0.44	0.50
	3	0.10	0.11	0.20	0.45	0.50	0.02	0.10	0.20	0.38	0.58
	4	0.10	0.11	0.12	0.20	0.45	0.01	0.04	0.08	0.20	0.35
	5	0.10	0.11	0.12	0.13	0.20	0.01	0.03	0.06	0.08	0.20
	6	0.10	0.11	0.12	0.13	0.14	0.01	0.02	0.03	0.06	0.08
□=0.25 & 5-dose	0	0.50	0.55	0.60	0.65	0.70	0.41	0.45	0.49	0.53	0.57
	1	0.25	0.50	0.55	0.60	0.65	0.25	0.41	0.45	0.49	0.53
	2	0.10	0.25	0.50	0.55	0.60	0.12	0.25	0.42	0.49	0.55
	3	0.10	0.11	0.25	0.50	0.55	0.04	0.12	0.25	0.43	0.63
	4	0.10	0.11	0.12	0.25	0.50	0.02	0.06	0.10	0.25	0.40
	5	0.10	0.11	0.12	0.13	0.25	0.02	0.05	0.08	0.11	0.25
	6	0.10	0.11	0.12	0.13	0.14	0.01	0.02	0.05	0.08	0.11
□=0.2 & 4-dose	0	0.45	0.50	0.55	0.60	--	0.37	0.45	0.54	0.61	--
	1	0.20	0.45	0.50	0.55	--	0.20	0.37	0.45	0.54	--
	2	0.11	0.20	0.45	0.50	--	0.07	0.20	0.35	0.57	--
	3	0.11	0.12	0.20	0.45	--	0.01	0.08	0.20	0.37	--
	4	0.11	0.12	0.13	0.20	--	0.02	0.06	0.09	0.20	--
	5	0.11	0.12	0.13	0.14	--	0.01	0.02	0.06	0.09	--
□=0.25 & 4-dose	0	0.50	0.55	0.60	0.65	--	0.42	0.50	0.59	0.62	--
	1	0.25	0.50	0.55	0.60	--	0.25	0.42	0.50	0.59	--
	2	0.11	0.25	0.50	0.55	--	0.10	0.25	0.40	0.62	--
	3	0.11	0.12	0.25	0.50	--	0.02	0.10	0.25	0.42	--
	4	0.11	0.12	0.13	0.25	--	0.05	0.08	0.12	0.25	--
	5	0.11	0.12	0.13	0.14	--	0.02	0.05	0.08	0.12	--
□=0.2 & 3-dose	0	0.45	0.50	0.55	--	--	0.38	0.57	0.61	--	--
	1	0.20	0.45	0.50	--	--	0.20	0.38	0.57	--	--
	2	0.12	0.20	0.45	--	--	0.08	0.20	0.34	--	--
	3	0.12	0.13	0.20	--	--	0.02	0.08	0.20	--	--
	4	0.12	0.13	0.14	--	--	0.01	0.02	0.08	--	--
□=0.25 & 3-dose	0	0.50	0.55	0.60	--	--	0.43	0.62	0.62	--	--
	1	0.25	0.50	0.55	--	--	0.25	0.43	0.62	--	--
	2	0.12	0.25	0.50	--	--	0.11	0.25	0.39	--	--
	3	0.12	0.13	0.25	--	--	0.05	0.11	0.25	--	--
	4	0.12	0.13	0.14	--	--	0.02	0.05	0.11	--	--
□=0.2 & 2-dose	0	0.45	0.50	--	--	--	0.38	0.56	--	--	--
	1	0.20	0.45	--	--	--	0.20	0.38	--	--	--
	2	0.13	0.20	--	--	--	0.10	0.20	--	--	--
	3	0.13	0.14	--	--	--	0.02	0.10	--	--	--
□=0.25 & 2-dose	0	0.50	0.55	--	--	--	0.43	0.62	--	--	--
	1	0.25	0.50	--	--	--	0.25	0.43	--	--	--
	2	0.13	0.25	--	--	--	0.12	0.25	--	--	--
	3	0.13	0.14	--	--	--	0.02	0.12	--	--	--

Figure 2. Accuracy and efficiency of BOLD, BOIN and 3+3 design at 5-dose model scenarios with non-informative priors, by scenario (0 to 6) at ϕ=0.2 and 0.25, assuming both an example and BOIN's default DLT rate matrix

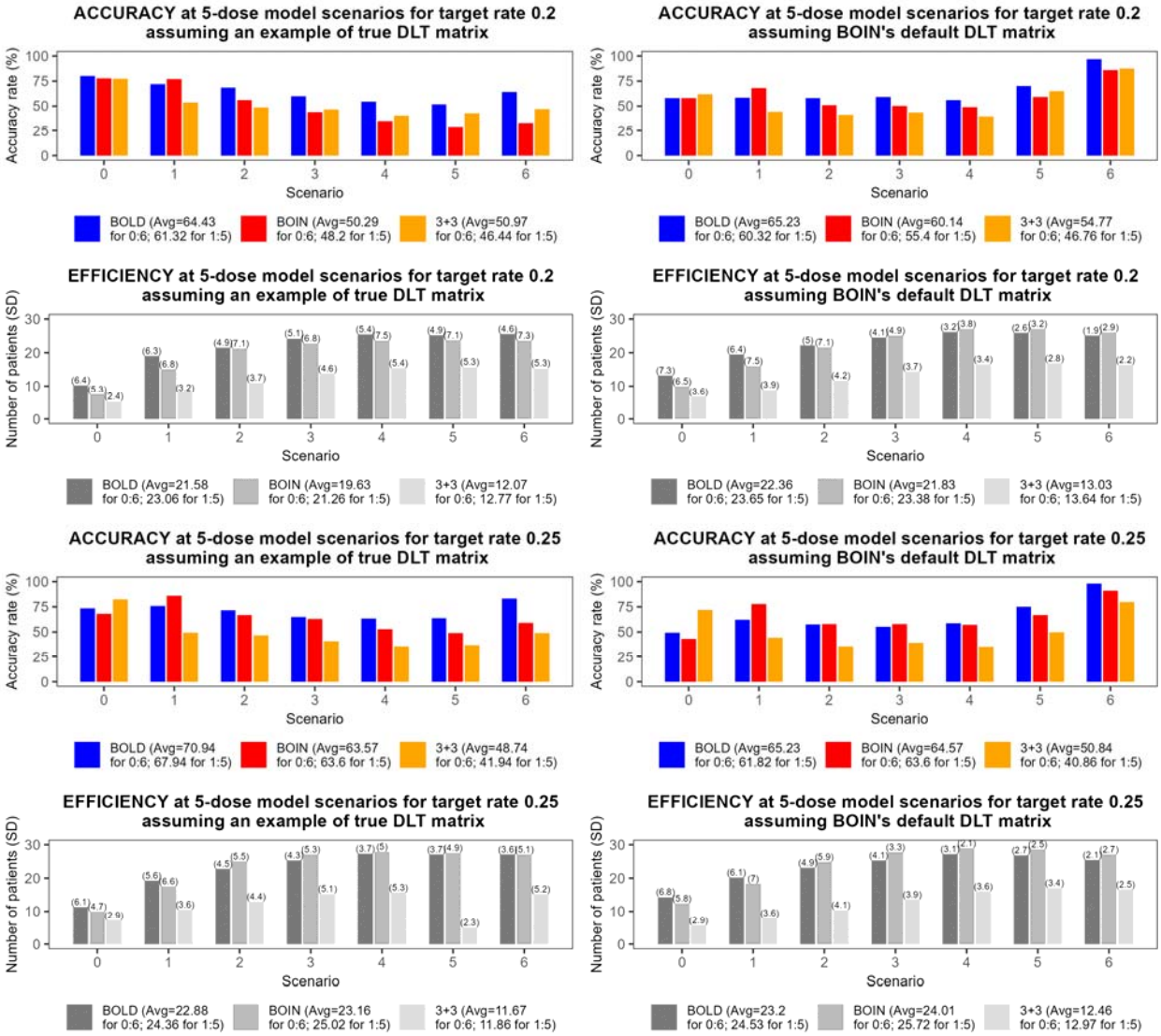

Table 2. Summary of average accuracy and efficiency of BOLD, BOIN and 3+3 design with non-informative priors across scenarios by lattice at ϕ=0.2 and 0.25, assuming both an example and BOIN's default DLT rate matrix

Target rate ϕ	True DLT rate matrix	Dose levels	Accuracy (%)			Number of treated patients					
						BOLD		BOIN		3+3	
			BOLD	BOIN	3+3	Avg.	SD	Avg.	SD	Avg.	SD
0.2	Example	2	77.45	69.75	65.58	15.40	4.86	13.98	4.42	6.82	2.16
		3	71.92	60.80	58.58	17.88	5.14	16.56	5.24	8.70	2.98
		4	66.27	54.83	55.42	19.95	5.43	18.38	5.80	10.47	3.63
		5	64.43	50.29	50.97	21.58	5.36	19.63	6.84	12.07	4.27
	BOIN's	2	77.15	72.75	66.35	15.52	4.30	14.78	3.88	6.85	2.05
		3	73.52	69.60	62.50	18.08	4.36	17.72	4.20	8.93	2.37
		4	67.68	64.67	57.37	20.47	4.50	20.35	4.50	11.11	2.91
		5	65.23	60.14	54.77	22.36	4.36	21.83	5.14	13.03	3.39
0.25	Example	2	80.60	78.25	63.10	16.02	4.45	16.25	3.57	6.65	2.16
		3	76.46	72.20	56.20	18.93	4.79	19.38	4.46	8.48	2.83
		4	73.33	67.50	52.28	21.15	4.72	21.63	4.89	10.10	3.53
		5	70.94	63.57	48.74	22.88	4.50	23.16	5.30	11.67	4.12
	BOIN's	2	76.03	75.50	63.98	16.18	4.07	16.58	3.42	6.69	2.04
		3	71.56	71.20	59.06	18.81	4.32	19.88	3.86	8.75	2.53
		4	67.57	67.83	52.48	21.35	4.52	22.52	4.07	10.71	3.04
		5	65.23	64.57	50.84	23.20	4.26	24.01	4.21	12.46	3.41

Figure 3. Accuracy and efficiency of BOLD and iBOIN at 5-dose model scenarios with informative PESS (preferred dose is true with PESS=6; other doses with PESS=3), by scenario (1 to 5) at ϕ=0.2 and 0.25, assuming both an example and BOIN's default DLT rate matrix

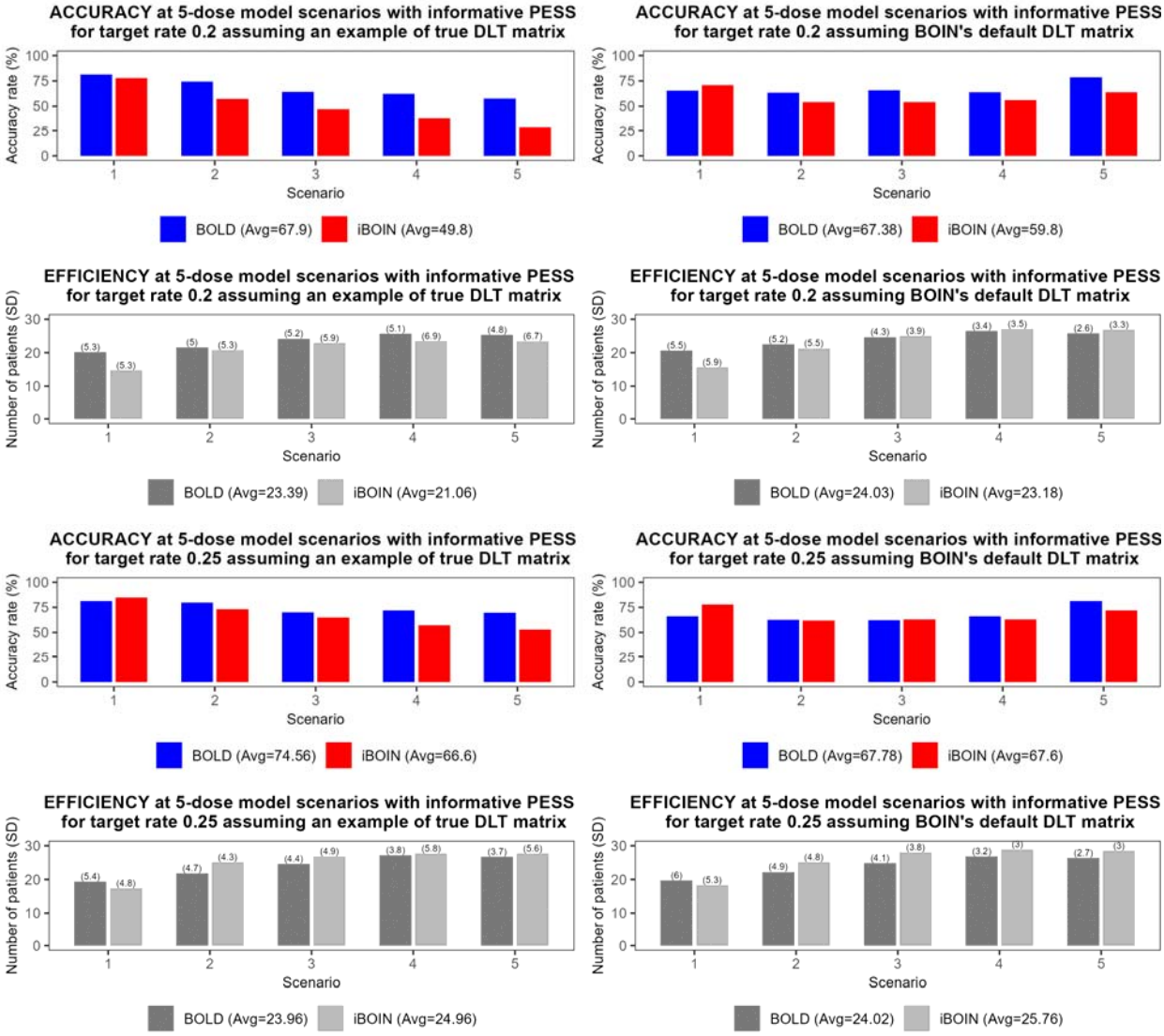
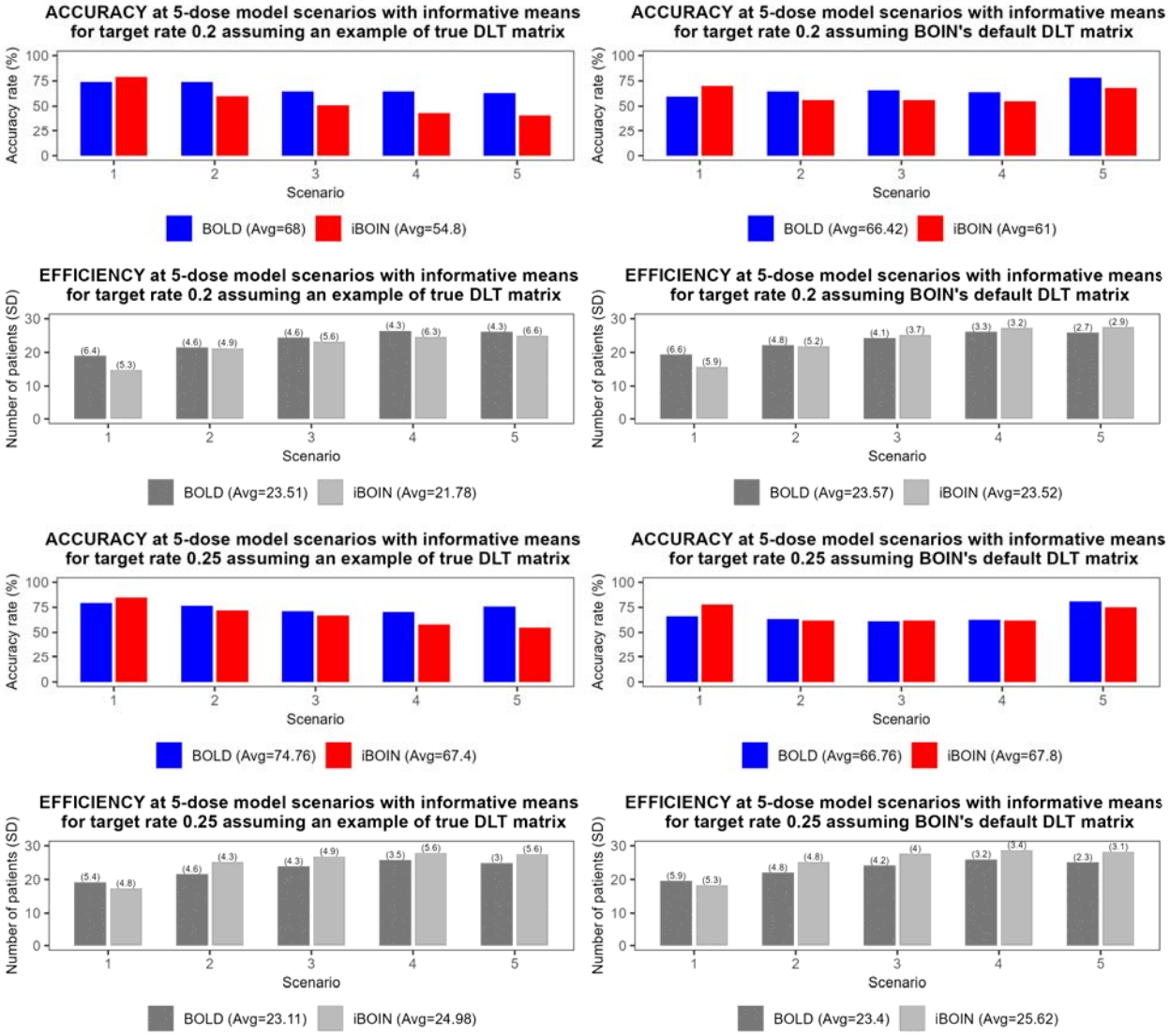
Figure 4. Accuracy and efficiency of BOLD and iBOIN at 5-dose model scenarios with informative prior means/skeletons (preferred dose is true with prior mean/skeleton=ϕ; other doses with prior mean/skeleton=ϕ-0.05 or ϕ+0.05), by scenario (1 to 5) at ϕ=0.2 and 0.25, assuming both an example and BOIN's default DLT rate matrix