



Balancing sample size and accuracy of dose selection in phase 1 oncology trials – designs tiered by cohort size and maximum number of patients treated per dose

Ping Gao , Xun Zhang , Weidong Zhang , Meng Li & Chunming Zhang

To cite this article: Ping Gao , Xun Zhang , Weidong Zhang , Meng Li & Chunming Zhang (26 Jul 2026): Balancing sample size and accuracy of dose selection in phase 1 oncology trials – designs tiered by cohort size and maximum number of patients treated per dose, Journal of Biopharmaceutical Statistics, DOI: [10.1080/10543406.2026.2699850](https://doi.org/10.1080/10543406.2026.2699850)

To link to this article: <https://doi.org/10.1080/10543406.2026.2699850>



© 2026 The Author(s). Published with license by Taylor & Francis Group, LLC.



Published online: 26 Jul 2026.



Submit your article to this journal [↗](#)



View related articles [↗](#)



View Crossmark data [↗](#)



Balancing sample size and accuracy of dose selection in phase 1 oncology trials – designs tiered by cohort size and maximum number of patients treated per dose

Ping Gao^a, Xun Zhang^b, Weidong Zhang^{c*}, Meng Li^d, and Chunming Zhang^e

^aInnovatio Statistics, Inc., Bridgewater, NJ, USA; ^bSchool of Mathematics and Statistics, Changsha University of Science and Technology, Changsha, Hunan, China; ^cSana Biotechnology, Inc, Cambridge, MA, USA; ^dYinchuan University of Science and Technology, Yinchuan, China; ^eDepartment of Statistics, University of Wisconsin-Madison, Madison, WI, USA

ABSTRACT

The objective of phase 1 oncology dose-escalation trials is to identify the dose with a toxicity rate that matches or is the closest to the target dose-limiting toxicity (DLT). There are mainly two categories of design methods for phase 1 trials: the rule-based methods and model-based/assisted methods. All literature has focused on the designs, mostly on how the various model-based/assisted methods are better than the 3 + 3 design, which is the most used rule-based design. The accuracy of dose selection is perhaps the most important operating characteristics (OC). Our investigation shows that the accuracy is more consequentially affected by the maximum number of patients treated per dose (MPTPD) than by the type of designs. Simulation examples show that the model-based BOIN and CRM design has lower accuracy for dose selection than the 3 + 3 design with the same cohort size and the same MPTPD. Larger MPTPD is associated with larger sample sizes. Simulations show that selection accuracy improves with increasing MPTPD (hence sample sizes). Our findings suggest that the design of phase 1 trials should include consideration of the balance between MPTPD, sample size and expectation of selection accuracy. Rule-based designs that are tiered by cohort sizes and MPTPD are proposed. These designs perform similarly with the BOIN and CRM designs with identical cohort sizes and MPTPD, but can be used without any software.

ARTICLE HISTORY

Received 1 May 2025
Accepted 30 June 2026

KEYWORDS

Dose escalation; maximum number of patients treated per dose; dose limiting toxicity; sample size; MTD power

1. Introduction

A phase 1 dose-escalation trial aims to identify the dose with a toxicity rate that is closest to a target DLT. The dose whose toxicity matches the target DLT is commonly referred to as the maximum tolerated dose (MTD). There are mainly two categories of design methods for such trials: the rule-based methods and model-based/assisted methods (Tourneau et al. 2009; LoRusso et al. 2010; Kurzrock et al. 2021). The rule-based methods do not need any software and are easy to understand and execute. The 3 + 3 design (Storer 1989) is the best-known rule-based design and has been the main target of criticism by researchers who prefer model-based/model-assisted designs. There are several model-based methods, including but not limited to, continual reassessment method (CRM) (O'Quigley et al. 1990; O'Quigley and Shen; Cheung 2011), the modified toxicity probability interval (mTPI) design (Ji and Wang 2013), the Bayesian logistic regression model (BLRM) (Neuenschwander

CONTACT Ping Gao ping.gao@innovatiostat.com Innovatio Statistics, Inc., Bridgewater, NJ 08807, USA

*The views expressed in this article do not necessarily represent those of the author's employer.

© 2026 The Author(s). Published with license by Taylor & Francis Group, LLC.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

et al. 2008), and Bayesian Optimal Interval Designs (BOIN) (Liu and Yuan 2015). All literature aimed at exploring new methods to improve upon the 3 + 3 design. With almost no exception, all discussions have been on designs. The impact of MPTPD and sample size, although intuitively obvious, has not been studied in depth. In this article, unlike in all current literature, the focus is not on the type of designs, but on the impact of MPTPD, cohort size and consequently the sample size. In this article,

- The impact of MPTPD on various designs is investigated by systematic isolation of MPTPD under matched design settings.
- Simulations show that small MPTPD leads to uniformly poor performance across designs (including BOIN and CRM).
- We propose tiered rule-based designs with performance comparable to BOIN/CRM with the same MPTPD.

2. Background

2.1. Elements of dose-escalation designs

In a dose-selection trial, several doses (say, K doses) of a new drug are given to patients in sequentially increased dosages. Let $r_1, \dots, r_K$ denote the rate of DLT for each dose, $n_1, \dots, n_K$ the number of treated patients, and $e_1, \dots, e_K$ the number of observed DLT events at each dose. All designs, regardless of rule-based or model-based, include two phases: dose escalation and dose selection.

- Dose escalation and upward push: The MTD will not get selected if the escalation does not reach the MTD. Metaphorically, dose escalation is an upward movement of dose administration. The upward movement happens when there is an upward push. Higher doses are reached with higher probability if there are more and stronger upward pushes.
- Dose selection and pullbacks: Overshooting is the situation when the escalation reaches doses higher than the MTD. In case of overshooting, a pullback to lower doses will be necessary to reduce the likelihood of selecting doses with high toxicity.

Rule-based methods use the information at the current dose, $\hat{r}_i = e_i/n_i$ to determine whether to escalate a dose, or stop the dose-escalation process, and then to determine a selected dose, which may pull back from higher doses. Model-based designs may use the sequence $\hat{r}_1, \dots, \hat{r}_i$ to make similar decisions. Theoretically, model-based designs use all dosing and toxicity information from all dosed patients, instead of just the information at the current dose. Thus model-based designs are expected to be more accurate than rule-based designs. Ideally, as few patients as possible should be treated at doses with low DLT level, so that the trial can escalate quickly. The escalation should end as soon as it reaches the target DLT level so that as few as patients would be exposed to high toxicity. Such OCs are subject to the impact of the variation of $\hat{r}_i$, which is $r_i(1 - r_i)/n_i$. Because of the variations, overshooting will be unavoidable. It is desirable that the high doses be pulled back to doses with lower toxicity level when overshooting happens. A design would be very efficient if all of these ideal features were achieved. On the other hand, the smaller the n_i , the larger the variation, which reduces the precision/reliability of decision-making regarding dose escalation and dose selection in the trial. To increase the reliability, it is necessary to reduce the variation of $\hat{r}_i$, and n_i 's will need to be large enough. If each of $n_1, \dots, n_i$ is small, then the combined information from $\hat{r}_1, \dots, \hat{r}_i$ may not be much more than that from $\hat{r}_i$ alone. Modeling, whether frequentist or Bayesian, will also be subject to large variations, and the theoretical advantage of model-based designs over rule-based designs may be very limited. In addition, the correlations between $\hat{r}_1, \dots, \hat{r}_i$ are often unknown, which could further reduce the usefulness of modeling.

In this article, the last tested dose will be referred to as the last dose, or LD. The last dose that met the escalation criteria will be referred to as the last dose with an escalation, or LDE. Both the LD and LDE may be a dose in the middle of the dosing sequence (if the escalation stopped due to observed DLTs) and may also be the highest dose if the observed DLTs at all doses met the criteria for escalation. The dose that is selected for future efficacy trials will be referred to as the recommended dose, or RD.

2.2. Designs in the literature

2.2.1. The 3 + 3 design

- Rules for dose escalations: In the 3 + 3 design (Storer 1989), patients are enrolled in cohorts of 3. If no DLT is observed in the first 3 patients treated at the dose, the dose will be escalated to the next higher dose. If 1 DLT is observed in the 3 patients, then 3 additional patients will be treated at the current dose. If no more DLT is observed in the next 3 patients, the dose level will be escalated. If at least 2 DLTs are observed in the 6 treated patients (the dose escalation can be stopped if two DLTs were observed before 6 patients have been tested), the dose escalation and the trial stops, and this dose will be the LD. Hence, if the LD is not the planned highest dose, the observed DLT rate for LD will be $2/6 = 0.33$.
- Rules for RD selection: If the LDE is the highest dose, then the LDE will be the RD. Otherwise, the previous lower dose (LD-1) will be the RD. There is a possibility that only 3 patients had been tested at the LD-1 dose. In some practices, a total of 6 patients will be enrolled at the LD-1 dose. The dose will be chosen as the RD if at most 1 DLT was observed among the 6 patients. Hence the observed DLT rate for RD will be either $0/3 = 0$ or $1/6 = 0.167$. This rule for selecting RD results in a high rate that $RD = LD-1$.

2.2.2. Model-based designs

There are many model-based designs, and it is not possible to study all the model-based designs in one article. For this article, the BOIN design (Liu and Yuan 2015; Yuan et al. 2016; Yuan and Liu 2021) and Continual Reassessment Method (CRM) (e.g., O’Quigley et al. 1990; O’Quigley and Shen 1996; Cheung 2011) were used as representatives of model-based designs. The designs were selected mainly for the following reasons: i) CRM and BOIN are the two most famous model-based methods (Li et al. 2024). ii) With both BOIN and CRM, the users can select cohort size, MPTPD, and number of cohorts, or the overall sample size. Thus, both the BOIN and CRM designs are well suited for evaluating the impact of cohort size and MPTPD. iii) BOIN R-package (Yuan and Liu 2021) and CRM R-package *simCrm.R* (CRAN library) are available for simulations. Similar investigations with other model-based designs with flexibility on sample sizes may be conducted in future research.

2.2.2.1. The BOIN design. BOIN, Bayesian Optimal Interval Designs, by its name, is an interval design (Liu and Yuan 2015). In such a design, ϕ indicates a target DLT level. At each dose level i , two boundaries $0 \leq \lambda_{1i}(n_i, \phi) < \lambda_{2i}(n_i, \phi) \leq 1$ are defined. The dose escalates if $\hat{r}_i \leq \lambda_{1i}(n_i, \phi)$, and deescalates if $\hat{r}_i \geq \lambda_{2i}(n_i, \phi)$. The dose is retained if $\lambda_{1i}(n_i, \phi) < \hat{r}_i < \lambda_{2i}(n_i, \phi)$. BOIN defines a decision error rate, $\alpha(\lambda_{1i}, \lambda_{2i})$. There are three hypotheses, $H_{0i} : r_i = \phi$, $H_{1i} : r_i = \phi_1$, $H_{2i} : r_i = \phi_2$, where ϕ_1 denotes the highest toxicity probability that is deemed subtherapeutic (i.e., below the MTD) such that dose escalation should be made, and ϕ_2 denotes the lowest toxicity probability that is deemed overly toxic such that dose deescalation is required. The posterior distributions of H_{1i} and H_{2i} are obtained at each dose level, λ_{1i} and λ_{2i} are chosen to minimize $\alpha(\lambda_{1i}, \lambda_{2i})$. The interval $(\lambda_{1i}, \lambda_{2i})$ is the Bayesian optimal interval.

- Dose escalation rules: the rules escalate, retain, or de-escalate the dose are determined by the interval $(\lambda_{1i}, \lambda_{2i})$. The trial stops when the specified number of cohorts for the trial are reached, or an MPTPD is reached.
- LD: Due to the possibility of de-escalations, the LD may not be the highest administered dose.
- RD selection rules: After the LD has been reached, BOIN determines the RD according to the posterior distributions of the DLT for all tested doses.
- The BOIN design requires the use of the BOIN software.

With the BOIN design, the user must specify the cohort size, the number of cohorts, and the MPTPD.

2.2.2.2. The CRM design. In the CRM model, the dose level of the j -th patient is denoted as X_j , taking value of X_j , from a range of k values, $d_1, \dots, d_K$ (O'Quigley et al. 1990; O'Quigley and Shen 1996). Let Y_j be a binary variable, $Y_j = 1$ indicating observing a DLT event, and $Y_j = 0$ otherwise. The toxicity response is modeled as $Pr(Y_j = 1 | X_j = x_j) = E(Y_j | x_j) = \psi(x_j, a_0)$, where $\psi(x_j, a_0)$ is a function of x_j, a_0 , and a_0 is a parameter. Different versions of CRM may have different functions ψ , and different number of parameters a_i (O'Quigley and Shen 1996). The CRM method includes dose escalation and de-escalation. After each observation, the toxicity level for the doses $d_1, \dots, d_K$ is updated either with Bayesian method or maximum likelihood approach. In this article, simulations on CRM are based on the Bayesian method. The estimated toxicity level is then used to determine the next dose level, which may be an escalation or a deescalation.

With the BOIN design, the user must specify the cohort size, the number of cohorts, the MPTPD, and the total sample size. The use of CRM requires a software.

3. Operating characteristics of the 3 + 3 design

The target DLT level for the 3 + 3 design is often assumed to be 0.33 or 0.3 (e.g., Storer⁴). Higher rate of RD = MTD means higher selection accuracy. This rate depends on the actual toxicity pattern, which is unknown in practice. The accuracy is usually evaluated in literature through simulations (in which the toxicity pattern is assumed). In Figure 1, the distributions of simulated RD rates are presented. In Figure 1a, the toxicity pattern is $p_1 = (0.01, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45)$. In Figure 1b, the toxicity pattern is $p_2 = (0.01, 0.1, 0.2, 0.3, 0.4, 0.5)$.

The dose selection rule results in a high chance of a pullback at dose selection, and the likelihood of RD being equal to LD-1 is very high (0.996 in Figure 1a, and 0.984 in Figure 1b). A main character of the 3 + 3 design is its small MPTPD (at $n = 6$), which is associated with large variance of the $\hat{r}_i$'s, resulting in low precision of escalation decisions. Consequently, the LD and RD curves are spread out and flat. As discussed earlier, the observed DLT rate for RD will be at most $1/6 = 0.167$ unless it is the highest planned dose. The expected peak of the RD curve will be at a dose with DLT no higher than 0.167. This suggests that the target DLT level between 0.167 and 0.2, rather than 0.3 or 0.33, is more reasonable for the 3 + 3 design. The simulated RD curves are mainly between tox levels 0.05 and 0.3, and do not have high concentration at any particular toxicity level, with the highest point of LD and RD curves at 0.21 for tox pattern p_1 (Figure 1a) and the highest point of LD and RD curves at 0.32 for tox pattern p_2 (Figure 1b). As a result, the rate of RD = MTD will be low (e.g., this rate is no more than 0.21 in Figure 1a, and no more than 0.32 in Figure 1b), regardless of what the target DTL level is. In Figure 1a, the toxicity pattern is p_1 , the rate of RD = 0.3 was 0.09, and the rate of RD = 0.2 is 0.203. In Figure 1b, the toxicity pattern is p_2 , the rate of RD = 0.3 was 0.219, and the rate of RD = 0.2 is 0.329. Both Figures 1(a) and 1(b) confirm that 0.2 is a more reasonable choice for MTD than 0.3 for the 3 + 3 design.

In later sections, simulations are conducted to compare 3 + 3, BOIN and CRM at MPTPD = 6. These simulations show that the poor performance of 3 + 3 arises from limited information per dose rather than from the rule-based nature of the design per se.

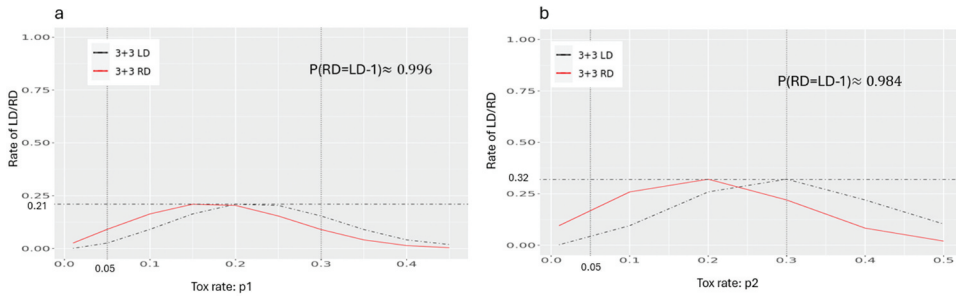


Figure 1. Analysis of the 3 + 3 design.

4. Simulations: the impact of MPTPD, part I

The simulations in this section investigate the impact of MPTPD by comparing the OCs of 3 + 3, BOIN and CRM designs. As discussed earlier, it is more reasonable to use 0.2 as the target DLT for 3 + 3 in the simulations for comparison purposes. Therefore, the simulations will compare 3 + 3, BOIN and CRM with 0.2 as the target DLT level.

Two main characters of the 3 + 3 design are that the cohort size is 3, and the MPTPD is 6. On the other hand, the BOIN and CRM design allow users to select cohort size and the MPTPD. In this section, we present simulations comparing the OC of: i) the 3 + 3 design, cohort size = 3, MPTPD = 6; ii) BOIN design, cohort size = 3, MPTPD = 6; iii) BOIN design, cohort size = 3, MPTPD = 15; iv) CRM design, cohort size = 3, MPTPD = 6; iii) CRM design, cohort size = 3, MPTPD = 15. With these simulations, the 3 + 3 design, BOIN and CRM designs are evaluated under identical restrictions of cohort size = 3 and MPTPD = 6, and the BOIN and CRM designs are also evaluated under identical restrictions of cohort size = 3 and MPTPD = 15.

The most important OC is the percentage of correctly selecting the true MTD, or the rate of RD = MTD. The relatively small sample size limits the accuracy of identifying the MTD. Therefore, the rate of the RD being within a range around the target DLT is also a relevant OC. Such a range could be, say, [Target DLT - 0.05, Target DLT + 0.05], and named as the “MTD ± 0.05 ” interval. The OCs vary, depending on the toxicity rate pattern of the dose sequences. In real trials, the toxicity pattern is unknown. For our simulations, we use two patterns, one with a 5% increase in toxicity rates between the doses, and another with 10% increases. i.e., i) p1 = (0.01, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45); ii) p2 = (0.01, 0.1, 0.2, 0.3, 0.4, 0.5). The simulations with p1 can assess the sensitivity of the escalation strategy, and the rate of selection in the MTD ± 0.05 interval. The simulations on p2 are mainly for comparing similarity with the results on p1.

For the 3 + 3 design, there is no specified limit on the total sample size, although the total sample size cannot exceed 6 times the number of doses. The trial stops according to the escalation rules. For the BOIN design, the users can choose the maximum number of cohorts for the trial. In all of the simulations related to BOIN in this article, the number of cohorts is chosen to be 40. For the CRM method, the maximum sample size is chosen to be 240. For the simulations in Figure 1, the mean sample sizes for the BOIN method are much smaller than 3×40 (cohort size times the maximum number of cohorts allowed), and much smaller than 240 for the CRM, indicating that the maximum number of cohorts was never reached for the BOIN design, and the maximum sample size was never reached for the CRM method, and the trial termination was completely determined by the escalation/deescalation rules of both designs.

Under the restriction of cohort size 3 and MPTPD = 6 for all three methods, for toxicity pattern p1, the rate of correctly selecting the MTD (i.e. the rate of RD = MTD) (Figure 2) for the 3 + 3 design was slightly higher than both BOIN and CRM. For toxicity pattern p2, the rate of correctly selecting the MTD for 3 + 3 was slightly lower than that for BOIN, and slightly higher than that for CRM. The average over p1 and p2 for 3 + 3 was slightly higher than both BOIN and CRM.

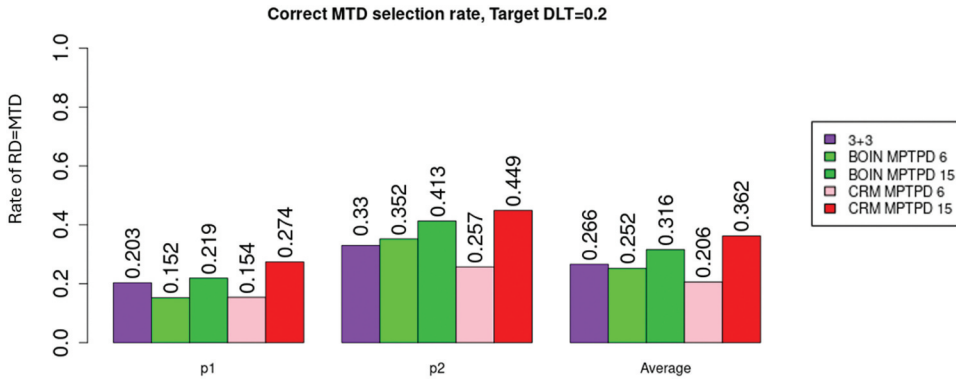


Figure 2. Correct MTD selection: 3 + 3 vs. BOIN MPTPD 6 and BOIN MPTPD 15, CRM MPTPD 6 and CRM MPTPD 15.

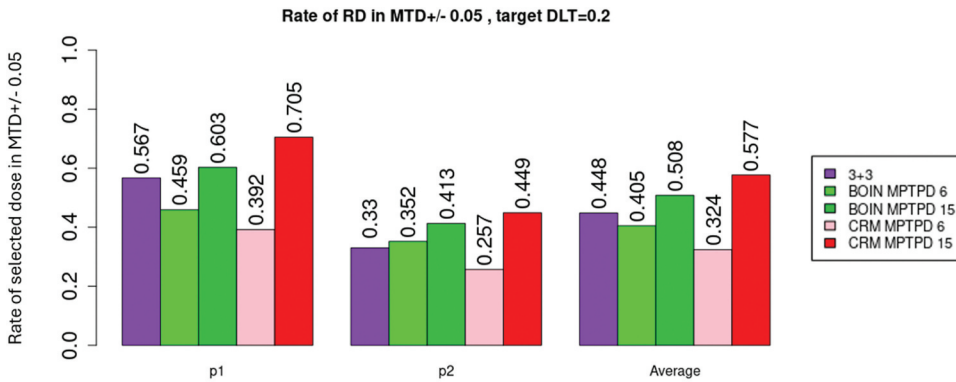


Figure 3. Rate of the selection being in the MTD ± 0.05 interval: 3 + 3 vs. BOIN MPTPD 6 and BOIN MPTPD 15, CRM MPTPD 6 and CRM MPTPD 15.

In [Figure 3](#), this rate increased for both BOIN and CRM and were both higher than 3 + 3 when MPTPD increased to 15.

Under the restriction of cohort size 3 and MPTPD = 6 for all three methods, for toxicity pattern p1, the rate of selection in the MTD ± 0.05 interval ([Figure 2](#)) for the 3 + 3 design was higher than both BOIN and CRM. For pattern p2, the rate of correctly selecting the MTD and the rate of selection in the MTD ± 0.05 interval were the same, because p2 does not include midpoints of $x \pm 0.05$. Hence, the rate for 3 + 3 is slightly lower than those for BOIN and slightly higher than CRM. The average rate over p1 and p2 for 3 + 3 was slightly higher than those for BOIN and CRM.

This rate increased for both BOIN and CRM and were both higher than 3 + 3 when MPTPD increased to 15 ([Figure 4](#)).

When MPTPD increased to 15 for BOIN and CRM, the mean sample sizes for these two methods increased over MPTPD = 6. This shows that the increased precision with MPTPD = 15 is associated with a considerable increase in mean sample size. These findings confirm that MPTPD and mean sample size have fundamental impact on the OCs of the BOIN and CRM designs.

Similar simulations with target DLT = 0.25 showed similar patterns (not presented).

The simulations in this section demonstrate that no design performs well under the severely constrained MPTPD = 6.

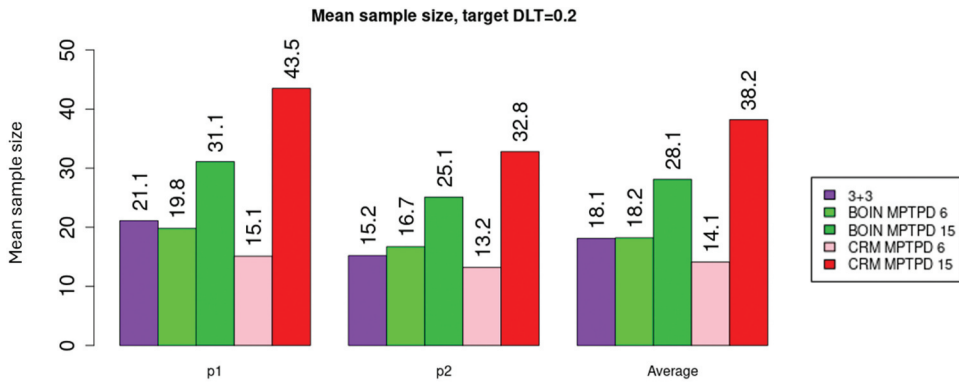


Figure 4. Mean sample size: 3 + 3 vs. BOIN MPTPD 6 and BOIN MPTPD 15, CRM MPTPD 6 and CRM MPTPD 15.

5. Cohort size and MPTPD tiered rule-based designs

5.1. Motivation for rule-based design that performs better than 3 + 3

Despite the almost one-sided criticism of the 3 + 3 design in the literature, the 3 + 3 design is still the most used method (73.85%) in phase 1 oncology dose-escalation trials (Li et al. 2024). The primary reason for this phenomenon is that it is straightforward, being a rule-based approach that does not require complex statistical software or real-time computations. Its predetermined rules for dose escalation or de-escalation are easy for clinicians and researchers to understand and to implement in a clinical setting. The extensive use of 3 + 3 is solid evidence that there is a robust demand for rule-based designs that perform better than the 3 + 3 design and can give effective options to those clinicians who desire designs that are simple to implement and yet perform well (as compared with model-based methods).

This is the motivation for our proposal of the tiered rule-based designs.

5.2. Necessary elements of optimal designs

A desirable design is expected to have the following features:

- Sufficient but not excessive upward push for dose escalation: For doses that are close to the MTD, the rates of them being the LD should be high, and for doses with very low or very high toxicity, the rates of being the LD should be low. The rate of LD should peak at the MTD. Adequate upward push is indicated by a line plot curve of the LD rates (from low dose to high dose) that peaks at the MTD. Higher precision of dose escalation is indicated by higher peak of LD line plot curve at the MTD, and sharper drops of the curve away from the peak.
- Adequate pullbacks: The RD should pull back to lower doses if LD is higher than the MTD. Lower rates of RD above MTD (high toxicity) are desirable.
- The line plot curve of the RD rates (from low dose to high dose) peak at the MTD. Higher precision of dose escalation is indicated by higher peak of RD line plot curve at the MTD, and sharper drops of the curve away from the peak.

These features naturally are dependent on target DLT and the dose toxicity pattern. Desirable minimal optimal criteria for a design are the following:

- The line-plot curve of the LD rates (from low dose to high dose) peaks at the MTD.
- The line-plot curve of the RD rates (from low dose to high dose) peaks at the MTD.

5.3. Applying the learnings from the 3 + 3 design and impact of MPTPD for the proposed designs

In the 3 + 3 design, the MPTPD is 6. Also, the escalation often ends too early, because only two chances of escalations at each dose are possible, after testing 3 and at most 6 patients. In other words, the upward pushes in the 3 + 3 designs are weak. The final dose selection involves at most 6 patients. Cumulatively, the total sample size for 3 + 3 is often smaller than most model-based designs. The small sample size in the 3 + 3 design results in large variation in $\hat{r}_i$ and lower accuracy of selection. These observations suggest that to improve upon the 3 + 3 design, the variance of $\hat{r}_i$ must be reduced, this means that MPTPD must be increased. However, larger MPTPDs are also associated with larger sample sizes which will increase the burden on patient enrollment and cost for conducting the trial. Designs that are tiered based on the size of MPTPD can provide flexibilities and options to trial investigators and clinicians so that they can select different MPTPDs based on the trade-off between the desire for higher accuracy and constraint on feasibility. In addition, there must be more chances for dose escalation. Hence, the proposed rule-based designs include larger MPTPDs and stronger upward pushes with 4 or 5 chances for escalation (see schematic charts in Figures 5–7). The simulations in section 4 suggest that the accuracy of dose selection of any design is highly dependent on the MPTPD. The proposed designs are tiered by cohort size and MPTPD. They can be applied to target DLTs at 0.2, 0.25, and 0.3. Any ratio of MPTPD vs. cohort size in a design may be developed with the help of simulations. For convenience and simplicity, in this article, MPTPD were chosen to be 5 times the cohort size. Therefore, the proposed designs are said to be cohort-size tiered designs, or simply tiered designs. This ratio of 5:1 is obviously somewhat arbitrary, but it is flexible enough and provides sufficient options for trial investigators and clinicians. The designs include cohort sizes from 3 to 10. All the schematic charts are provided in the online supporting material. Designs with all cohorts are supported by the software Design and Analysis for Clinical Trials (DACT) at <https://www.innovatiostat.com/software.html>, and their OCs can be investigated with simulations using DACT.

5.4. The cohort 3, 6, and 9 designs

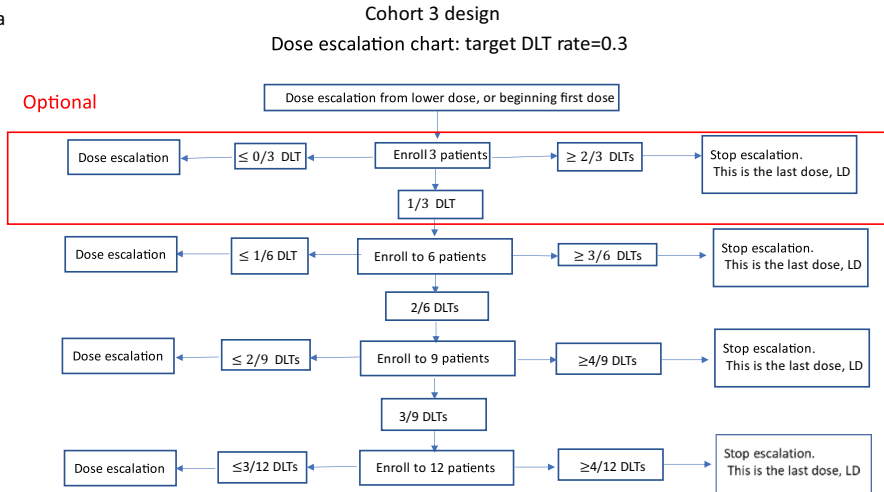
The 3 + 3 design can be clearly described using a flowchart, which is the main reason why it is easy to follow and implement. Likewise, our proposed tiered designs are illustrated by flowcharts. In practice, the clinicians only need to follow that chart without needing to use a computer or software, such as in the case for model-based methods, including BOIN and CRM. The design schematic charts for cohort sizes 3, 6 and 9, and with target DLT = 0.3 are described in this section. The design schematic charts for cohort sizes 3, 6 and 9, and with target DLT = 0.25 and 0.2 are provided in the APPENDIX and the online supporting material.

5.4.1. The cohort 3 design, target DLT = 0.3

In this design, the increment of patients tested at each dose will be in cohorts of 3, and chances of escalation will be given at each dose with accumulatively 3, 6, 9, and 12 patients. At the LD where the dose selection will be made, 15 patients will be treated to increase the precision of $\hat{r}_i$.

The design includes an option such that the user can choose either to or not to check cohort 1 to decide on whether to escalate. The impact of both options will be evaluated with simulations.

a



b

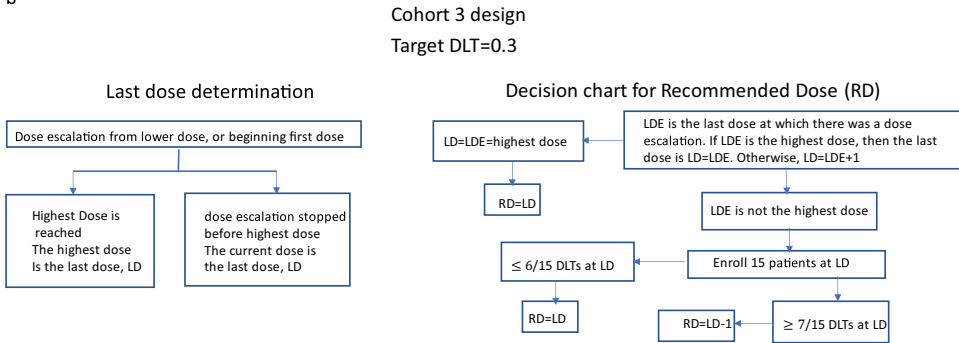


Figure 5. The cohort 3 design. Target DLT = 0.3.

5.4.2. The cohort 6 design, target DLT = 0.3

In this design, the increment of patients tested at each dose will be in cohorts of 6, and chances of escalation will be given at each dose with accumulatively 6, 12, 18 and 24 patients. At the LD where the dose selection will be made, 30 patients will be treated to increase the precision of $\hat{r}_i$ (Figure 6).

5.4.3. The cohort 9 design, target DLT = 0.3

In this design, the increment of patients tested at each dose will be in cohorts of 9, and chances of escalation will be given at each dose with accumulatively 9, 18, 27 and 36 patients. At the LD where the dose selection will be made, 45 patients will be treated to increase the precision of $\hat{r}_i$.

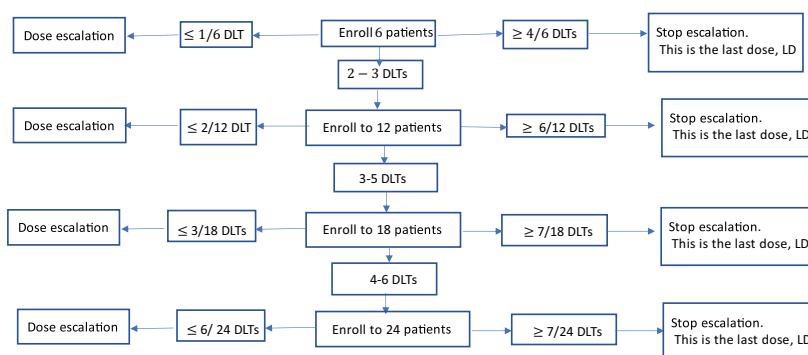
6. Simulations

6.1. Impact of MPTPD, part II

In this section, the impact of MPTPD will be assessed by simulations involving the cohort 3, 6 and 9 designs, and the BOIN and CRM designs with identical cohort size and MPTPD, with target DLT = 0.3. The number of cohorts for the BOIN design is 40, and the maximum sample size for CRM is

a

Cohort 6 design
Dose escalation chart: target DLT rate=0.3



b

Cohort 6 design
Target DLT=0.3

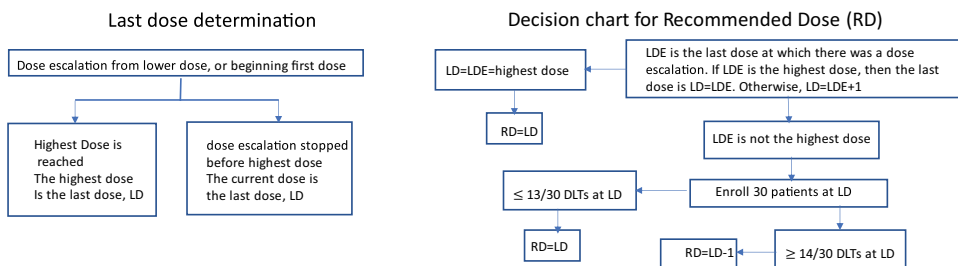


Figure 6. The cohort 6 design. Target DLT = 0.3.

chosen to be 240. Both are unlikely to be reached, so that the termination of the trial is completely determined by the BOIN and CRM escalation rules.

In the simulations,

- Line plots for the last dose (LD) will be presented to assess the effectiveness of escalation.
- Line plots for the recommended dose (RD) will be presented to assess the accuracy of dose selection.
- Mean sample sizes for all designs at all tiers will be presented to demonstrate the association of cohort size, MPTPD and sample size.

In all literature, the operating characteristics of any design are assessed through simulations, using assumed toxicity patterns. It is not possible for the assumed patterns to be exhaustive. Therefore, all such simulation assessments are only partial in nature. Our investigations are no different. As with the simulations comparing BOIN and CRM vs. 3 + 3 in section 4, the same two patterns, i) $p_1 = (0.01, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45)$; ii) $p_2 = (0.01, 0.1, 0.2, 0.3, 0.4, 0.5)$ will be used for the simulations in this section. Both p_1 and p_2 are with equally spaced toxicities. The simulations with p_1 can assess the sensitivity of the escalation strategy, and the rate of selection in the MTD ± 0.05 interval. The simulations on p_2 are mainly for comparing similarity with the results on p_1 . In real trials, the toxicity patterns that are unknown and can be different from either p_1 or p_2 . We note that

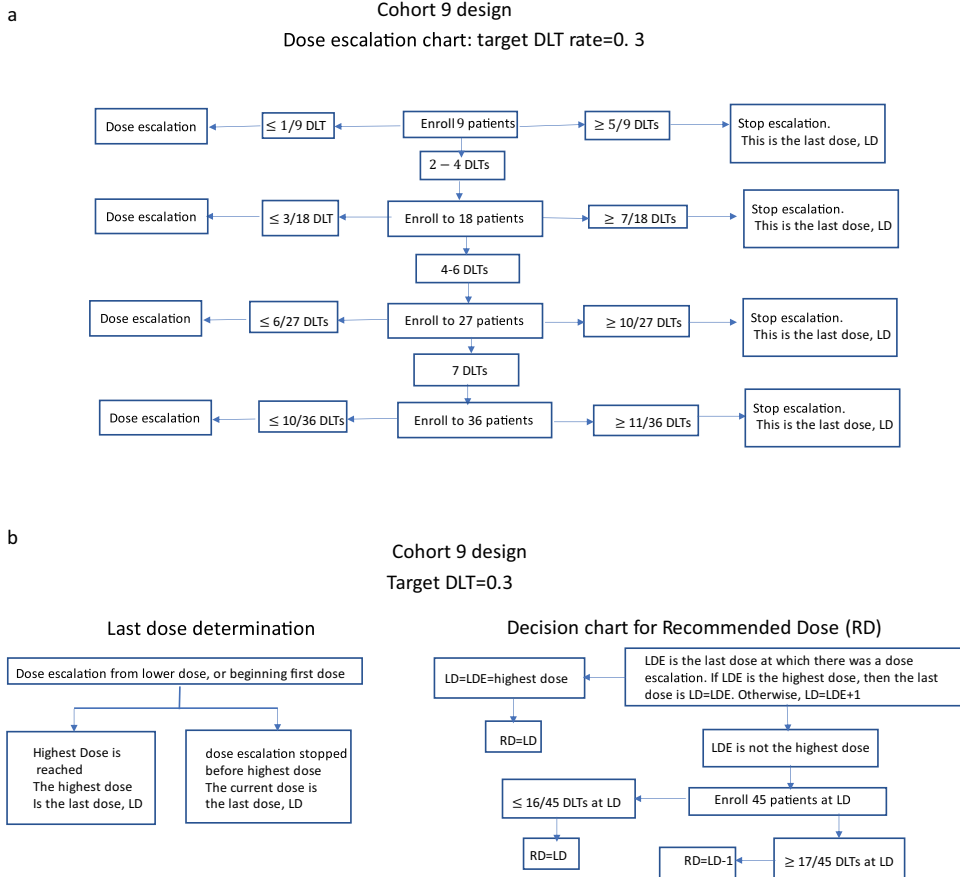


Figure 7. The cohort 9 design. Target DLT = 0.3.

the simulations are not intended as quantitative assessments of the OCs of the proposed designs. Rather, these simulations provide qualitative demonstration that the OCs are consistent across both patterns and are driven primarily by MPTPD rather than the details of the toxicity curves. Trial investigators may conduct their own simulations for various assumed toxicity patterns using the DACT software. However, simulations using the DACT software are not required for the use of the tiered designs in practice.

Simulations for target DLTs at 0.25 and 0.2 have also been conducted with similar results (not presented).

Figure 8a shows that the peak of the LD curves increases with larger MPTP for tiered, BOIN and CRM designs. The curves provide side-by-side comparisons. However, it is too crowded. Figure 8b is a breakdown of Figure 8a by each design.

In tox pattern p1, the increment of tox rate is 0.05. In Figure 8, all of the LD curves (under p1) for the tiered cohort 3, 6 and 9 (with MPTPD 15, 30, and 45) designs and CRM MPTPD 15, 30 and 45 peak at the MTD, while all LD curves for BOIN MPTPD 15, 30 and 45 peak at the dose with toxicity rate of 0.25, just before the MTD, indicating slightly insufficient upward push for escalations for BOIN.

Figure 9a shows that the peak of the RD curves increases with larger MPTP for tiered, BOIN and CRM designs. The curves provide side-by-side comparisons. However, it is too crowded. Figure 9b is a breakdown of Figure 9a by each design.

In Figure 9, all of the RD (under p1) curves for the tiered cohort 3, 6 and 9 (with MPTPD 15, 30, and 45) designs and CRM MPTPD 15, 30 and 45 peak at the MTD, while all RD curves for BOIN MPTPD 15, 30, and 45 peak at the dose with toxicity rate of 0.25, just before the MTD. This resulted in the rates of RD being the MTD (see also Figure 12), as well as the rates of RD being in the MTD ± 0.05 interval (Figure 13) for the tiered designs and CRM design being higher than for the BOIN designs at the same MPTPD levels.

Combining Figures 8 and 9, the tiered designs and the CRM design meet the minimum optimal criteria with tox pattern p1, while none of the BOIN designs do.

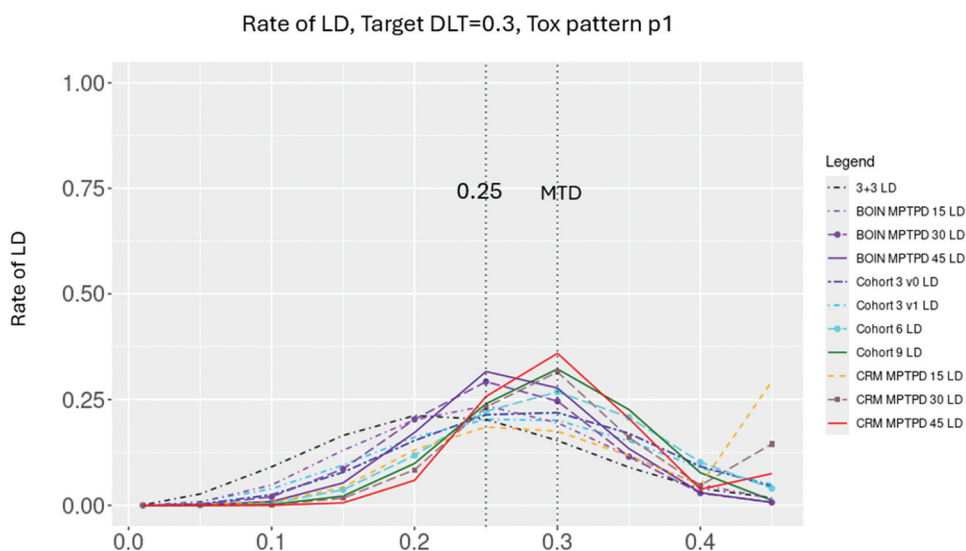


Figure 8a. Line plots for LD, p1, all designs.

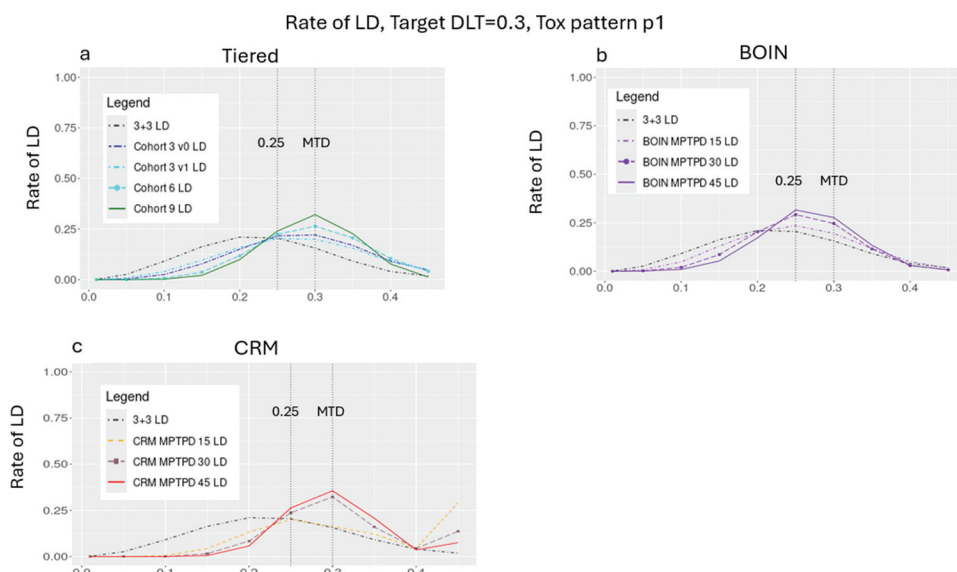


Figure 8b. Line plots for LD, p1. By each design.

Figure 10a shows that the peak of the LD curves increases with larger MPTP for tiered, BOIN and CRM designs. The curves provide side-by-side comparisons. However, it is too crowded. Figure 10b is a breakdown of Figure 10a by each design.

In tox pattern p2, no dose has a toxicity rate of 0.25. Because of the larger increment of tox rate of 0.1, the escalation is faster for all designs. The LD curve for all designs peak at higher doses (compared with the escalations with p1). Hence, in Figure 10, all of the LD curves, including 3 + 3, the tiered

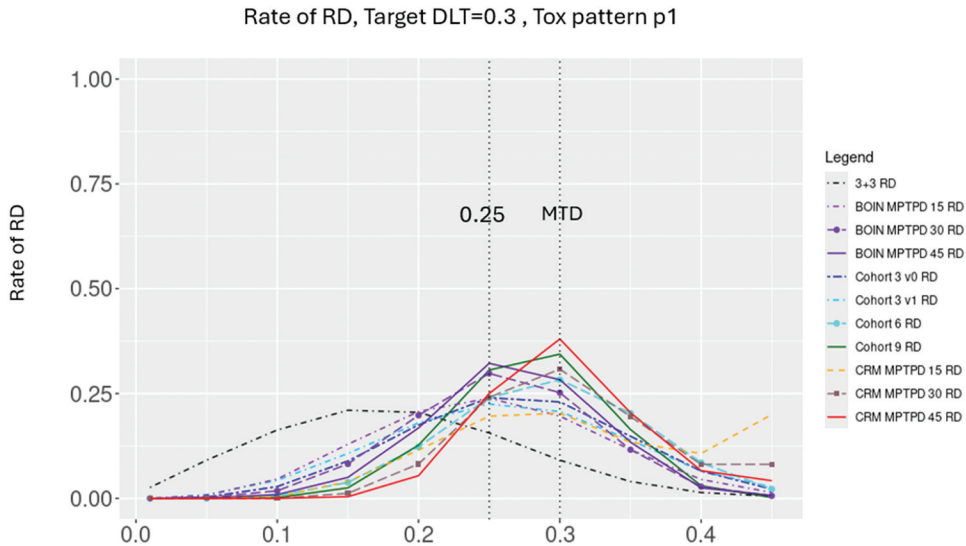


Figure 9a. Line plots for RD, p1. All designs.

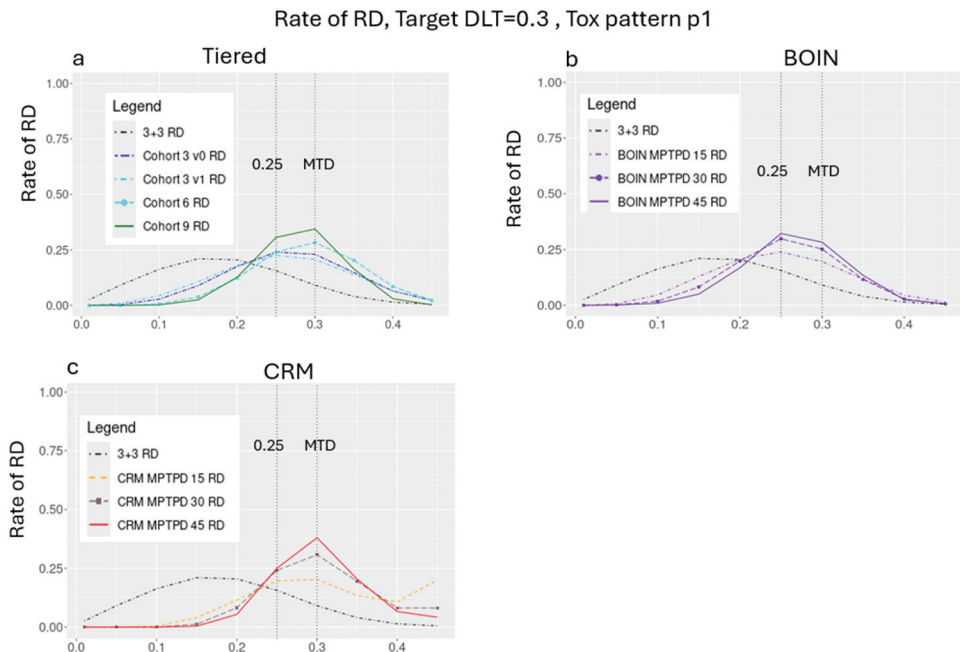


Figure 9b. Line plots for RD, p1. By each design.

cohort 3, 6, and 9 designs, the BOIN MPTPD 15, 30, and 45, and the CRM MPTPD 15, 30, and 45, peak at the MTD.

Figure 11a shows that the peak of the RD curves increases with larger MPTP for tiered, BOIN and CRM designs. The curves provide side-by-side comparisons. However, it is too crowded. Figure 11b is a breakdown of Figure 11a by each design.

In Figure 11, due to faster escalations, all of the RD curves for the tiered cohort 3,6 and 9 (with MPTPD 15, 30, and 45), and BOIN, CRM MPTPD 15, 30, and 45, peak at the MTD. At each of MPTPD levels (MPTPD 45, solid lines; MPTPD 30, dashed lines; MPTPD 15, dashed lines), the peaks of the tiered designs and CRM design are higher than those for BOIN. The peaks are the rate of correctly selecting the MTD and are also plotted as MTD matching rate in Figure 12

In Figure 10, all LD curves peak at the MTD since no dose has a DLT of 0.25. The peaks for the CRM and BOIN designs at all MPTPD were higher than those for the corresponding tiered designs with the same MPTPD. Hence, the escalation for CRM and BOIN is better than the tiered designs for p2. On the other hand, in Figure 11, the peak of RD for BOIN MPTPD 45 design at the MTD is lower than that for the tiered cohort 9 (MPTPD 45) design and CRM design with MPTPD 45, despite the peak of LD curve for BOIN MPTPD 45 design at the MTD being (see Figure 10) higher than the tiered design. In other words, the better LD curves for the BOIN designs did not result in better RD selection. This suggests that the RD selection rule for model-based design (in this case BOIN MPTPD 45) may not always be more optimal than for rule-based designs (in this case, the tiered cohort 9, with MPTPD 45).

In Figures 8–11, all the LD and RD curves have sharper drops from the peaks for all of the tiered, CRM and BOIN designs with larger MPTPD. i.e., the drops from the peaks for tiered with MPTPD 45 (the solid curves) are sharper than those for the tiered designs with MPTPD 30 (the dashed curves), which in turn, are sharper than those for the tiered design with MPTPD 15 (dashed curves). This is because larger MPTPD are associated with smaller variations, thus the LD and RD have higher concentration around the peaks. The patterns of drops are similar for CRM and BOIN designs with MPTPD 45 (the solid curves), 30 (the dashed curves), and 15 (dashed curves), also confirming that larger MPTPD increases accuracy.

In Figure 12, the rates of correctly selecting the MTD increase in stepwise fashion, according to the sizes of MPTPD, for all of the design, the CRM, BOIN designs and the tiered designs. The rates for the

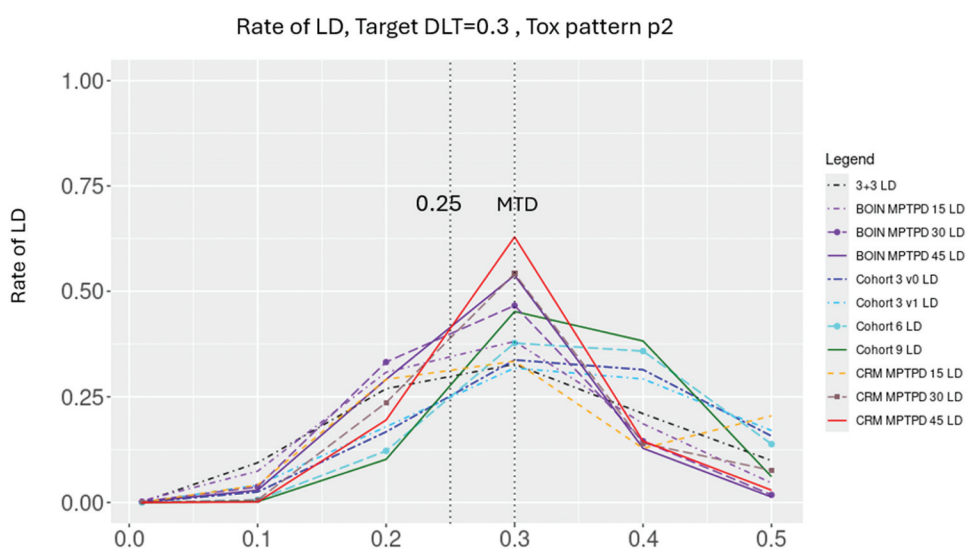


Figure 10a. Line plots for LD, p2. All designs.

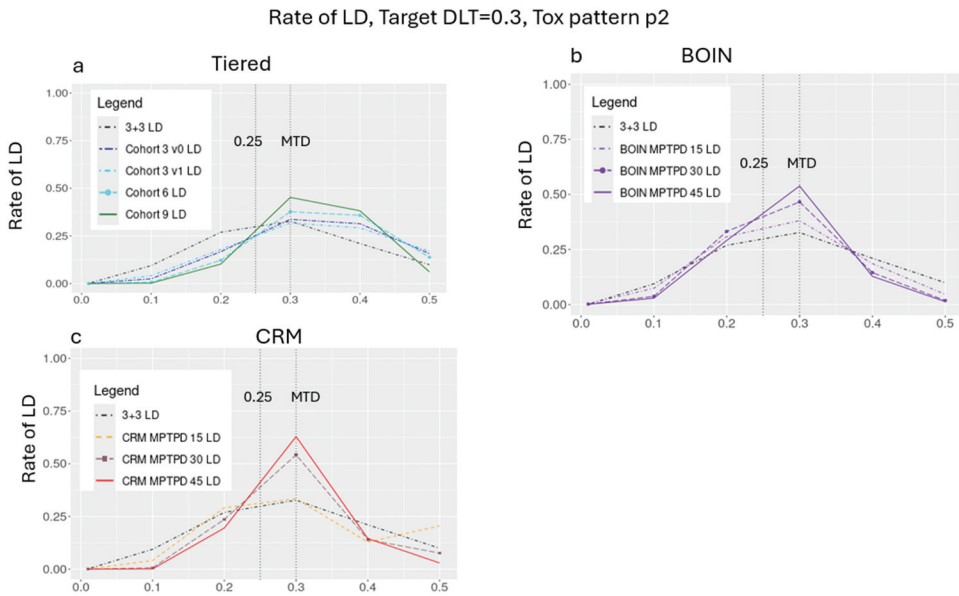


Figure 10b. Line plots for LD, p2. By design.

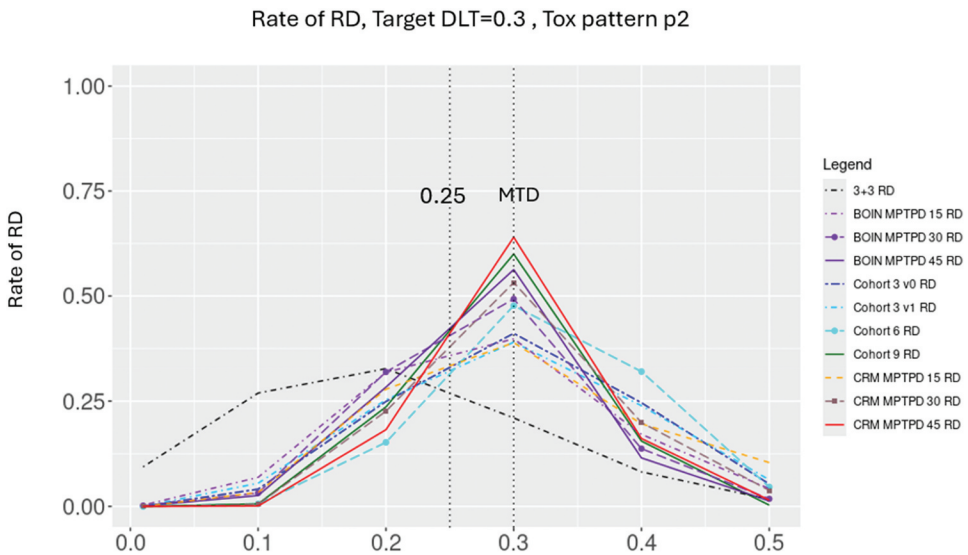


Figure 11a. Line plots for RD, p2. All designs.

tiered designs are similar to those of BOIN and CRM designs at all MPTPD levels. Also, the 3 + 3 design has the lowest rate of correctly selecting the MTD.

In Figure 13, the rate of RD being in the $\text{MTD} \pm 0.05$ interval increase in stepwise fashion, according to the sizes of MPTPD, for all the designs, CRM, the BOIN designs and the tiered designs. The rates for the tiered designs are similar to BOIN and CRM at each corresponding MPTPD size. e.g., the rate for the tiered cohort 9 design (with MPTPD 45) is higher than that for the BOIN MPTPD 45 design, etc. Also, the 3 + 3 design has the lowest rate of RD being at doses with tox rate in the $\text{MTD} \pm 0.05$ interval.

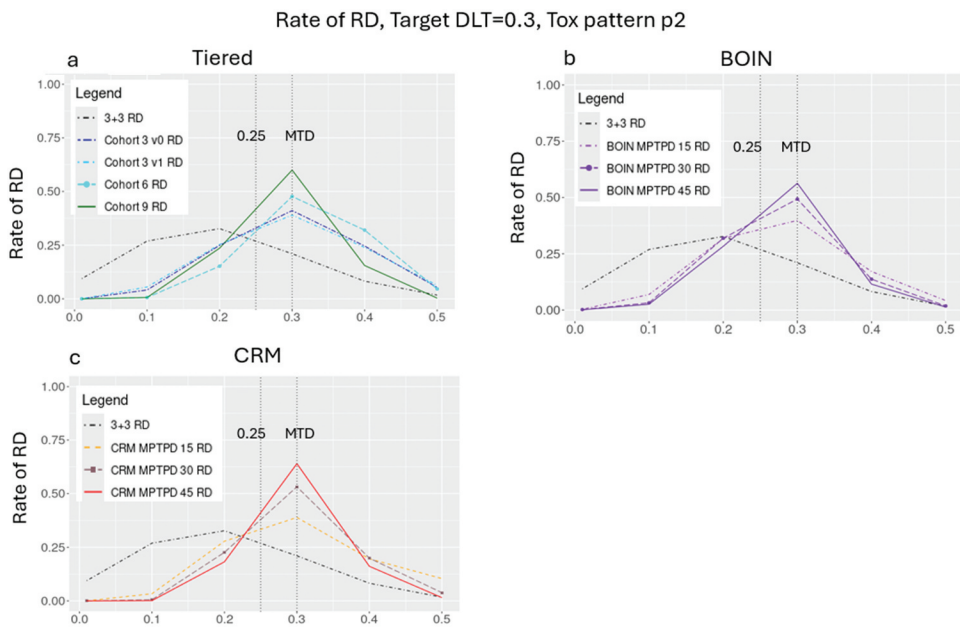


Figure 11b. Line plots for RD, p2. By design.

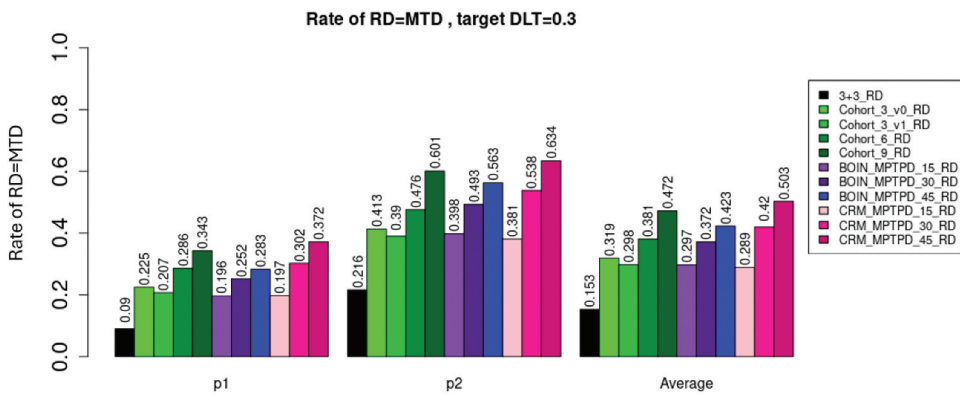


Figure 12. Rates of RD = MTD.

In Figure 14, the mean sample sizes for p1 are larger than those for p2 because the escalations for p1 involve more doses. The mean sample sizes increase in stepwise fashion, for all of the designs, CRM, the BOIN designs and the tiered designs, according to the increases of the sizes of MPTPD. The mean sample sizes for the tiered designs are higher than those for the CRM and BOIN designs at each corresponding MPTPD size. e.g., the sample size for cohort 9 design (with MPTPD 45) is larger than that for the BOIN MPTPD 45 design, etc. This is because the tiered designs have stronger upward push than the CRM and BOIN designs, and the dose escalations reach higher doses.

The tiered cohort 3 design includes an option to skip checking cohort 1 for dose escalation. In Figures 8-14, Cohort_3_v1 indicates that cohort 1 is checked, while Cohort_3_v0 indicates that cohort 1 was not checked. In Figures 12 and 13, the rates for correctly selecting the MTD and the rates of RD being in the $MTD \pm 0.05$ interval were both higher for Cohort_3_v0 than for Cohort_3_v1. This is because the check at cohort 1 with only 3 patients is subject to larger

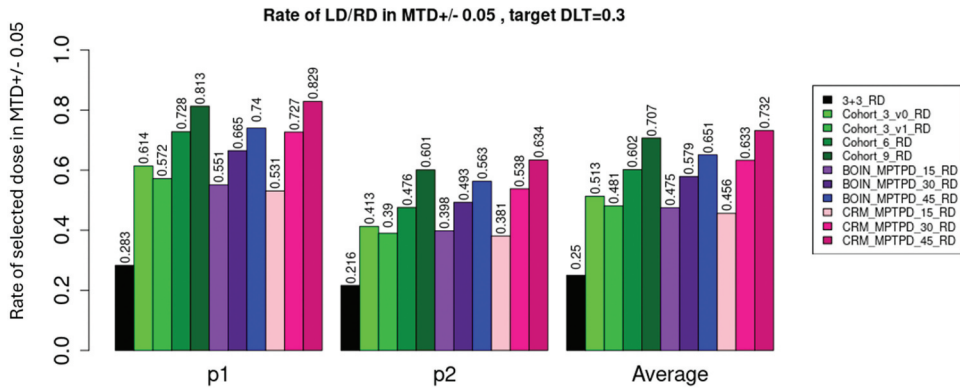


Figure 13. Rates of RD in MTD+/- 0.05.

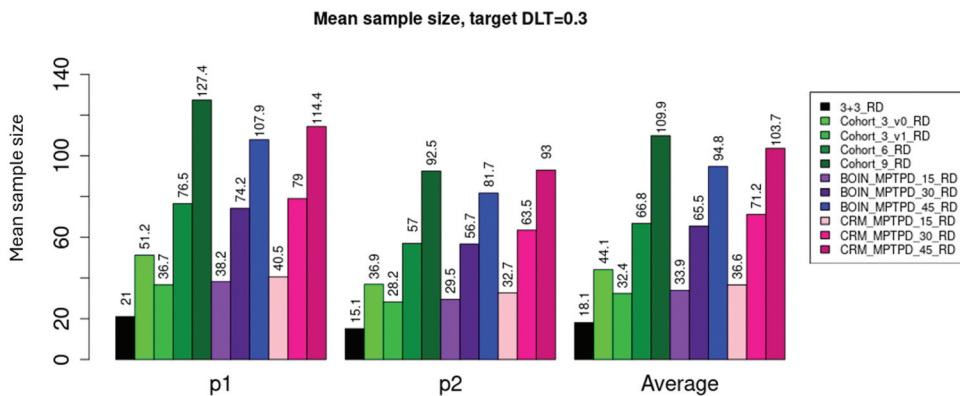


Figure 14. Mean sample size.

variation and higher inaccuracy, and skipping the check actually improves the accuracy. This comparison explains to some extent why the 3 + 3 has low accuracy. However, Figure 14 shows that the mean sample size for Cohort_3_v0 is slightly larger than that for Cohort_3_v1. This demonstrates again that the cost of higher accuracy is larger sample size.

6.2. Summary of simulations

- The tiered designs, the BOIN and CRM were compared under identical cohort size and MPTPD restrictions.
- The patterns of impact of MPTPD are similar on the tiered designs, the BOIN and CRM: larger MPTPDs are associated with higher accuracy of MTD selection
- The accuracy of selection was similar between the tiered designs, the BOIN and CRM for each cohort and MPTPD.

7. Discussions

7.1. Overall summary of simulations

The simulations in section 3:

- The 3 + 3 design is not based on thorough understanding of its OCs and is not an optimal design. It lacks accuracy because of the small MPTPD.

- The simulations were conducted only with two toxicity patterns p1 and p2. Interested readers may conduct their own simulations for any other toxicity patterns to confirm that the 3 + 3 design has very low rate of correctly selecting the MTD with any toxicity pattern and any target DLT level. Hence, the 3 + 3 design is not an advisable design for phase 1 trials.

The simulations in section 4:

- The simulation showed that the BOIN and CRM designs have lower rate of correctly selecting the MTD and lower rate of RD in the $MTD \pm 0.05$ interval than the 3 + 3 design when all designs were under the restriction of $MPTPD = 6$. The simulations only showed that small $MPTPD$ results in low accuracy for any design, and should not be interpreted as supporting that the 3 + 3 design should be the design of choice.
- The accuracy of BOIN and CRM designs increases when $MPTPD$ is increased.

The simulations in section 6:

- The size of $MPTPD$ has fundamental impact on the accuracy. Larger $MPTPD$ is associated with higher accuracy.
- The escalation procedures are more accurate if more patients have been tested at the dose, which means that larger cohort size and larger $MPTPD$ lead to more reliable escalation decisions.
- Larger $MPTPD$ results in larger mean sample sizes and increases the cost of dose-escalation trials.
- For both the p1 and p2 tox patterns, the tiered designs have similar rates of correct dose selection and the rates of RD being in the $MTD \pm 0.05$ interval as those for the BOIN and CRM designs with identical cohort size and $MPTPD$.

Similar to the OC assessments in literature, simulations based on limited assumed toxicity patterns (i.e., p1 and p2) were conducted to compare the OCs of the tiered design, BOIN and CRM. Such simulations should not be interpreted as precise quantitative evaluations of the OCs of the designs. The focus should be on the robustness of the OC patterns. The simulations showed that the rate of correctly selecting the MTD were all very low for 3 + 3, BOIN and CRM designs when $MPTPD = 6$, and thus strongly support the conclusion that no methods would perform well when the $MPTPD$ is small. In addition, in the simulations in section 5 the rates of correctly identifying the MTD for the tiered designs were slightly higher than BOIN design for either p1 or p2 for some $MPTPD$, but slightly lower for some others. Results like these may suggest that the tiered designs are similar to BOIN but are not sufficient to suggest that either design is better than the other. However, the simulations consistently demonstrated the impact of the $MPTPD$ across all designs and for both patterns p1 and p2, and thus strongly support the conclusion that larger $MPTPD$ is associated with higher dose-selection accuracy.

7.2. Sample size and power consideration in the design of dose-escalation trials

The goal of oncology phase 1 dose-escalation trial is to identify a dose that preserves clinical benefit with optimal tolerability. Selecting a dose with low toxicity but insufficient clinical benefit, or a dose with intolerable toxicity can lead to negative outcomes in future efficacy trials, which may undermine the development of a potentially promising drug. It has been reported that about 40–50% of clinical trials failed due to insufficient efficacy (Sacks et al. 2014; Hwang et al. 2016; Sun et al. 2022; Sacks et al. 2014) attributed 15.9% of the failures to uncertainties related to dose selection. Wong et al. (2019) reported that the success (phase 3 to approval by regulatory agencies) rate of phase 3 trials from 2001 to 2015 had been only about 35.5% for phase 3 oncology trials. The accuracy of identifying the correct dose can have far reaching consequences.

Many oncology trials are expensive to conduct and there is strong desire to conduct phase 1 trials as small as possible. We note that a dose-selection trial is only the first step in a potentially very large

clinical program which aims to bring an effective treatment to meet unmet medical needs. From this perspective, correctly identifying an effective dose, even at greater cost, is more important than purely aiming for a small and efficient phase 1 trial. The trade-offs between smaller but less accurate trials vs. larger but more reliable trials require discussions involving strategic vision, feasibility, practicality.

We propose that sample size and power considerations be included in the design of dose-escalation trials to balance the needs for accuracy and the cost of trials. The design of late phase trials (such as phase 3) typically involves sample size and power considerations. A trial is successful if the null hypothesis is rejected. The power of a trial is the probability of rejecting the null hypothesis in the trial, or the probability of the trial being successful. An assumption is made about the effect size of the new drug (although the true effect size is unknown), and a sample size is calculated accordingly to achieve a power of $1 - \beta$. A sample size/power consideration for phase 1 trials may be similarly introduced. Success in the strict sense means that the RD equals the true MTD and termed as MTD success. Success in a broader sense means that the RD is in the $\text{MTD} \pm 0.05$ interval and termed as $\text{MTD} \pm 0.05$ success. Accordingly, the probability of MTD success (i.e., the rate of RD being equal to MTD) is termed as MTD power, and the probability of $\text{MTD} \pm 0.05$ success is termed as $\text{MTD} \pm 0.05$ power. We note that the use of the terminologies of MTD power and $\text{MTD} \pm 0.05$ power is intended as analogies only, and they are based on simulations, rather than the algorithm-based calculated classical power for hypothesis testing. In a phase I trial design, a toxicity pattern is assumed (although the true toxicity pattern is unknown). Then simulations can be done to assess the MTD power and the $\text{MTD} \pm 0.05$ power under the assumed toxicity pattern, and the mean sample size for certain cohort sizes and MPTPDs. Then some combination of cohort size and MPTPD can be selected based on the mean sample size within the feasibility. The following are some hypothetical examples.

Example 1. Suppose that toxicity pattern p1 is assumed for a phase 1 trial in patients with a rare type of cancer and enrollment is expected to be slow and difficult. It is anticipated that a sample size exceeding 40 patients will not be feasible. Four designs have a mean sample size under or about 40 (Table 1):

The sample size of 3 + 3 is the lowest, but the MTD power and the $\text{MTD} \pm 0.05$ power are also the lowest. The Cohort_3_v1 design, the BOIN MPTPD15 the CRM MPTPD15 designs have similar sample sizes at just below or about 40, and similar MTD power and the $\text{MTD} \pm 0.05$ power. Hence any of these designs may be a reasonable candidate.

Example 2. Suppose that toxicity pattern p2 is assumed for a phase 1 trial in patients with a rare type of cancer and enrollment is expected to be slow and difficult. It is anticipated that a sample size exceeding 40 patients will not be feasible. Five designs have a mean sample size under 40 (Table 2):

Table 1. Simulated sample size and power with p1 for rare cancer.

Design	Mean sample size	MTD power	MTD $\pm$ 0.05 power
3 + 3	21.03 ^a	0.09 ^b	0.283 ^c
Cohort_3_v1	36.71 ^a	0.207 ^b	0.572 ^c
BOIN MPTPD 15	38.23 ^a	0.196 ^b	0.551 ^c
CRM MPTPD 15	40.5 ^a	0.197 ^b	0.531 ^c

^aFigure 14; ^bFigure 12; ^cFigure 13.

Table 2. Simulated sample size and power with p2 for rare cancer.

Design	Mean sample size	MTD power	MTD $\pm$ 0.05 power
3 + 3	15.17 ^a	0.216 ^b	0.216 ^c
Cohort_3_v0	36.87 ^a	0.413 ^b	0.413 ^c
Cohort_3_v1	28.16 ^a	0.390 ^b	0.390 ^c
BOIN MPTPD 15	29.51 ^a	0.398 ^b	0.398 ^c
CRM MPTPD 15	32.65 ^a	0.381 ^b	0.381 ^b

^aFigure 14; ^bFigure 12; ^cFigure 13.

Table 3. Simulated sample size and power with p1 for common cancer.

Design	Mean sample size	MTD power	MTD $\pm$ 0.05 power
Cohort_9	127.39 ^a	0.343 ^b	0.813 ^c
BOIN MPTPD 45	107.86 ^a	0.283 ^b	0.74 ^c
CRM MPTPD 45	114.45 ^a	0.372 ^b	0.829 ^c

^aFigure 14; ^bFigure 12; ^cFigure 13.

Table 4. Simulated sample size and power with p2 for common cancer.

Design	Mean sample size	MTD power	MTD $\pm$ 0.05 power
Cohort_9	92.46 ^a	0.601 ^b	0.601 ^c
BOIN MPTPD 45	81.68 ^a	0.563 ^b	0.563 ^c
CRM MPTPD 45	93.0 ^a	0.634 ^b	0.634 ^c

^aFigure 14; ^bFigure 12; ^cFigure 13.

The sample size of the 3 + 3 design is the lowest, but the MTD power and the MTD $\pm$ 0.05 power are also the lowest. The Cohort_3_v0 design has the highest MTD power at 0.413 (MTD $\pm$ 0.05 power is the same). But its sample size is near the maximum. The Cohort_3_v1 design, the BOIN MPTPD 15 and the CRM MPTPD 15 designs have similar sample sizes at below 40, and similar MTD power (MTD $\pm$ 0.05 power is the same) at just below 0.4. Hence any of these designs may be a reasonable candidate.

Example 3. Suppose that toxicity pattern p1 is assumed for a phase 1 trial in patients with a common type of cancer and enrollment is expected to be relatively quick. It is anticipated that a sample size of about 150 patients will be feasible. In Figure 14, the sample sizes of all designs are less than 150. Hence the focus of design selection is on MTD power and MTD $\pm$ 0.05 power (Table 3).

The cohort 9 design (MPTPD = 45) and CRM MPTPD 45 design had MTD power at 0.343 and 0.372, respectively, and the MTD $\pm$ 0.05 power at 0.813 and 0.829, respectively. Hence, the cohort 9 design can be a reasonable choice. The BOIN MPTPD 45 design, with slightly lower MTD power at 0.283 and MTD $\pm$ 0.05 power at 0.74, can be a reasonable choice as well.

Example 4. Suppose that toxicity pattern p2 is assumed for a phase 1 trial in patients with a common type of cancer and enrollment is expected to be relatively quick. It is anticipated that a sample size of about 100 patients will be feasible. In Figure 14, the sample sizes of all designs are less than 100. Hence the focus of design selection is on MTD power and MTD $\pm$ 0.05 power (Table 4).

The cohort 9 design (MPTPD = 45) and CRM MPTPD 45 design had MTD power at 0.601 and 0.634 (MTD $\pm$ 0.05 power is the same), respectively. Hence, either the cohort 9 design and CRM MPTPD 45 design can be a reasonable choice. The BOIN MPTPD 45 design had slightly lower MTD power at 0.563 (MTD $\pm$ 0.05 power is the same). This can be a reasonable choice as well.

7.3. Beyond phase 1: designs for phase 2 and 3 trials

Examples 1. 4 shows that the MTD power (and even the MTD $\pm$ 0.05 power) in phase 1 dose-selection trials can be much lower than 0.8, or that the likelihood of correct dose selection is low. This obviously raises concerns about the chance of success for the entire clinical development of the new drug. Later stage trials will need to be flexible to mitigate the risk of incorrect dose selection and increase the chance of success. Trial designs may include:

- Adaptive single-arm phase 2 trial design. Simon's design (Simon 1989) has been the design of choice in many phase 2 trials. In this design, a response rate is assumed for sample size calculation. Trials with the design will have a power of $1 - \beta$ if the assumed and the true response rates match. The power can be less than $1 - \beta$ if the assumption is incorrect. Gao and Zhang (2024) proposed an adaptive design to mitigate the risk of incorrect assumption, which can provide a power of $1 - \beta$ if the true response rate is in a certain range. However, we note that this adaptive design does not include dose modification. If the selected dose has no or minimal efficacy, this design will not help.
- The FDA guidance (2024), Optimizing the Dosage of Human Prescription Drugs and Biological Products for the Treatment of Oncologic Diseases, called for "Trial Designs to Compare Multiple Dosages" (section B, FDA, 2024). We propose that such trials may include three doses (e.g., RD-1, RD, RD + 1):
 - (a) FDA (2024, section B) suggested that "A recommended trial design to compare multiple dosages in a randomized, parallel dose-response trial". One design that can serve this purpose is the Adaptive Sequential Design for Multiple Comparisons (Gao and Li 2023). The design includes sample size and power calculations, and the determination of critical boundaries for interim analyses for family-wise type I error control. The design also includes final analysis, including p -values, point estimates, and confidence intervals, with or without any adaptive changes. With this design, multiple doses can be tested. Ineffective dose(s) can be dropped at any interim analysis and sample size can be adjusted for the remaining dose(s) to increase power.
 - (b) Another design option is the Adaptive Two-stage Seamless Sequential Design (Gao and Li 2024). The design includes sample size and power calculations, and the determination of critical boundaries for interim analyses for family-wise type I error control. The design also includes final analysis, including p -values, point estimates, and confidence intervals, with or without any adaptive changes. With this design, multiple doses can be tested in the first stage (phase 2). Ineffective doses can be dropped at the end of the first stage and one dose will be carried into the second stage (phase 3). The sample size can be adjusted for the second stage to increase power.
 - (c) Both of these two designs can be used in trials without a placebo control arm.

7.4. Software support

The simulations with the tiered designs and the 3 + 3 design can be conducted with the software DACT at <https://www.innovatiostat.com/software.html>, which is free for academic researchers. R-codes for simulations are available upon request. Phase 2 and 3 design options, such as Gao, Zhang (2024), Gao and Li (2023), and Gao and Li (2024), are also supported by DACT.

7.4.1. R-code free download

All related R-codes are available for free download at <https://github.com/innovatiostat/rcode> and <https://github.com/ChunmingZhangUW>

Author contributions

CRedit: Ping Gao: Methodology; Xun Zhang: Methodology; Weidong Zhang: Methodology; Meng Li: Methodology; Chunming Zhang: Methodology.

Disclosure statement

No potential conflict of interest was reported by the author(s).

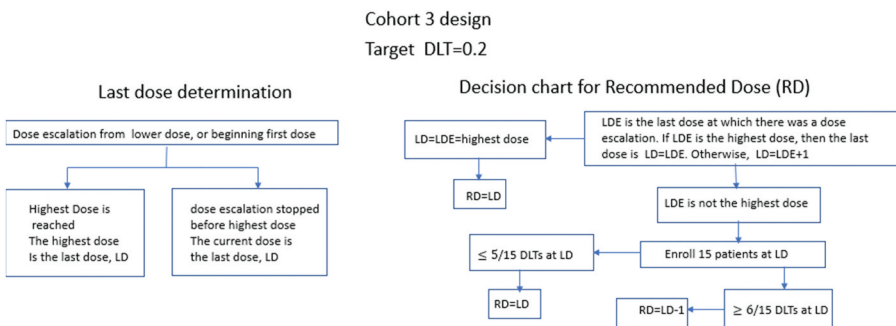
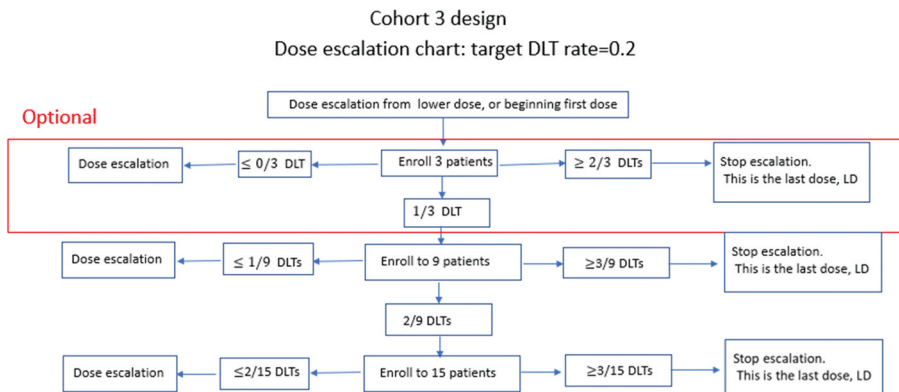
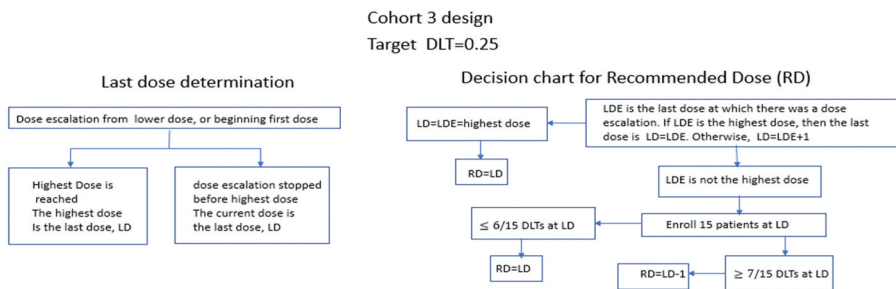
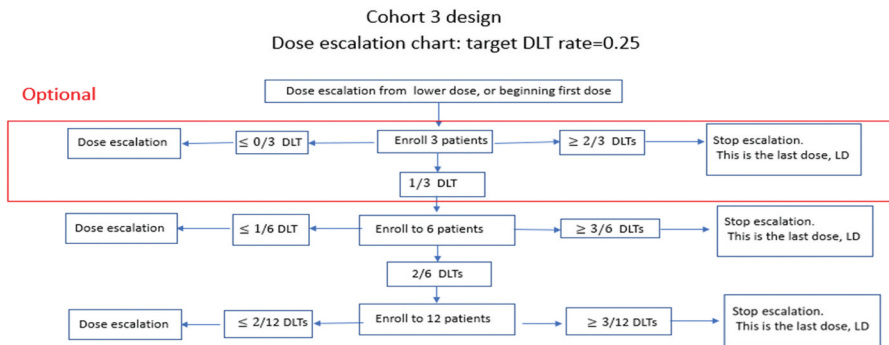
Funding

The author(s) reported there is no funding associated with the work featured in this article.

References

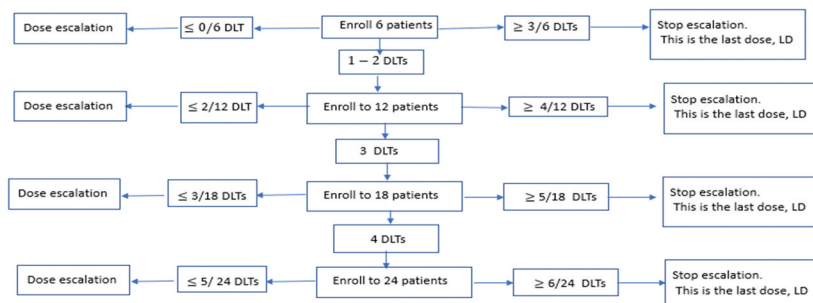
- Cheung YK. 2011. Dose finding by the Continual reassessment method. CRC Press.
- Gao P, Li Y. 2023. Adaptive multiple comparison Sequential design (AMCSD) for clinical trials. *J Biopharm Stat.* 1–17. <https://doi.org/10.1080/10543406.2023.2233590>
- Gao P, Li Y. 2024. Adaptive two-stage seamless Sequential design for clinical trials. *J Biopharm Stat.* 1–23. <https://doi.org/10.1080/10543406.2024.2342518>
- Gao P, Zhang W. 2024. Adaptive sequential design for phase II single-arm oncology trials: an expansion of Simon's design. *J Biopharm Stat.* 35(4):1–15. <https://doi.org/10.1080/10543406.2024.2341673>
- Hwang TJ et al. 2016. Failure of investigational drugs in late-stage clinical development and publication of trial results. *JAMA Intern Med.* 176(12):1826–1833. <https://doi.org/10.1001/jamainternmed.2016.6008>
- Ji Y, Wang S. 2013. Modified toxicity probability interval design: a safer and more reliable method than the 3+3 design for practical phase I trials. *J Clin Oncol.* 31(14):1785–1791. [PubMed: 23569307] <https://doi.org/10.1200/JCO.2012.45.7903>
- Kurzrock R et al. 2021. Moving beyond 3+3: The future of clinical trial design. *Am Soc Clin Oncol Educ Book.* (41): e133–e144. https://doi.org/10.1200/EDBK_319783
- Li N, Zhou X, Yan D. 2024 Sep 24. Phase I clinical trial designs in oncology: a systematic literature review from 2020 to 2022. *J Clin Transl Sci.* 8(1):e134. <https://doi.org/10.1017/cts.2024.599>. PMCI: 39345694; PMCID: PMC11428115.
- Liu S, Yuan Y. 2015. Bayesian optimal interval designs for phase I clinical trials. *J R Stat Soc Ser C Appl Stat.* 64 (3):507–523. <https://doi.org/10.1111/rssc.12089>
- LoRusso PM, Boerner SA, Seymour L. 2010 Mar 15. An overview of the optimal planning, design, and conduct of phase I studies of New therapeutics. *Clin Cancer Res.* 16(6):1710–1718. <https://doi.org/10.1158/1078-0432.CCR-09-1993>
- Neuenschwander B, Branson M, Gsponer T. 2008. Critical aspects of the Bayesian approach to phase I cancer trials. *Stat Med.* 27(13):2420–2439. <https://doi.org/10.1002/sim.3230>
- O'Quigley J, Pepe M, Fisher L. 1990. Continual reassessment method: a practical design for phase I clinical trials in cancer. *Biometrics.* 46(1):33–48. [PubMed: 2350571] <https://doi.org/10.2307/2531628>
- O'Quigley J, Shen LZ. 1996 Jun. Continual reassessment method: a likelihood approach. *Biometrics.* 52(2):673–684. PMCI: 8672707. <https://doi.org/10.2307/2532905>
- Sacks LV et al. 2014 Jan. Scientific and regulatory reasons for delay and denial of FDA approval of initial applications for new drugs. *JAMA.* 311(4):378–384. <https://doi.org/10.1001/jama.2013.282542>
- Simon R. 1989. Optimal two-stage designs for phase II clinical trials. *Control Clin Trials.* 10(1):1–10. [https://doi.org/10.1016/0197-2456\(89\)90015-9](https://doi.org/10.1016/0197-2456(89)90015-9)
- Storer BE. 1989. Design and analysis of phase I clinical trials. *Biometrics.* 45(3):925–937. <https://doi.org/10.2307/2531693>
- Sun D, Gao W, Hu H, Zhou S. 2022. Why 90% of clinical drug development fails and how to improve it? *Acta pharmaceutica Sin B.* 12(7):3049–3062. <https://doi.org/10.1016/j.apsb.2022.02.002>
- Tourneau CL, Lee JJ, Siu LL. 2009 May 20. Dose escalation methods in phase I cancer clinical trials. *J Natl Cancer Inst.* 101(10):708–720. <https://doi.org/10.1093/jnci/djp079>
- U.S. 2024. Food and drug administration: optimizing the dosage of human prescription drugs and biological products for the treatment of oncologic diseases, guidance for industry. <https://www.fda.gov/media/164555/download>
- Wong CH, Siah KW, Lo AW. 2019. Estimation of clinical trial success rates and related parameters. *Biostatistics.* 20 (2):273–286. <https://doi.org/10.1093/biostatistics/kxx069>
- Yuan Y, Hess KR, Hilsenbeck SG, Gilbert MR. 2016 Sep 1. Bayesian optimal interval design: a simple and well-performing design for phase I oncology trials. *Clin Cancer Res.* 22(17):4291–4301. <https://doi.org/10.1158/1078-0432.CCR-16-0592>
- Yuan Y, Liu S. 2021. R package BOIN. <https://cran.rproject.org/web/packages/BOIN/index.html>

Appendix

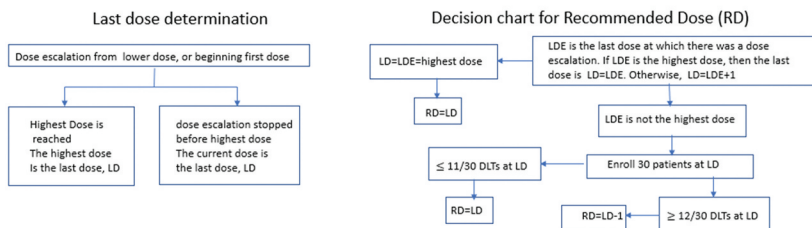


Cohort 6 design

Dose escalation chart: target DLT rate=0.25

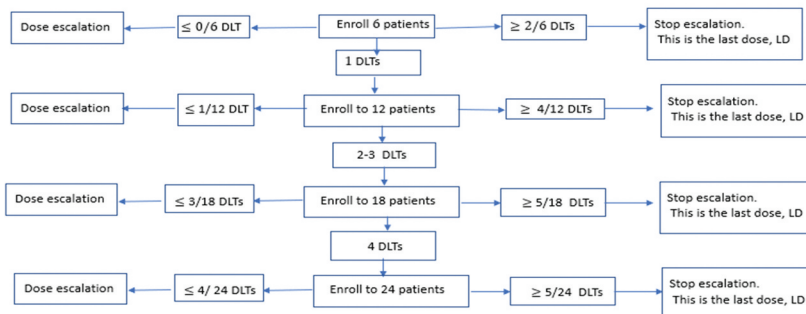


Cohort 6 design
Target DLT=0.25

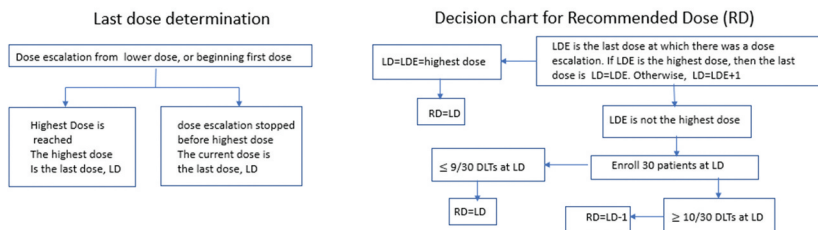


Cohort 6 design

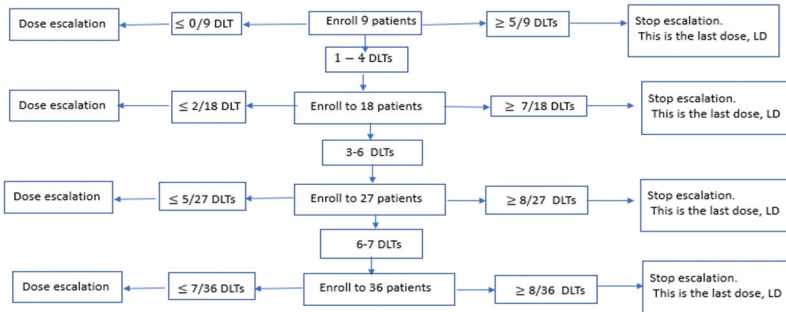
Dose escalation chart: target DLT rate=0.2



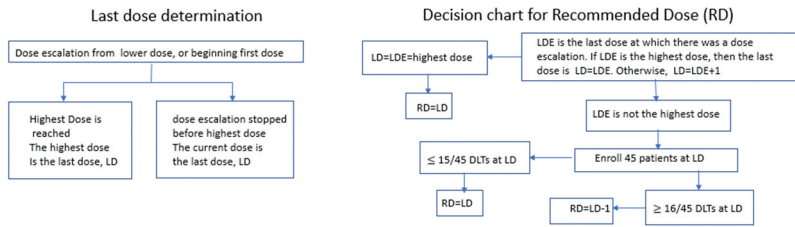
Cohort 6 design
Target DLT=0.2



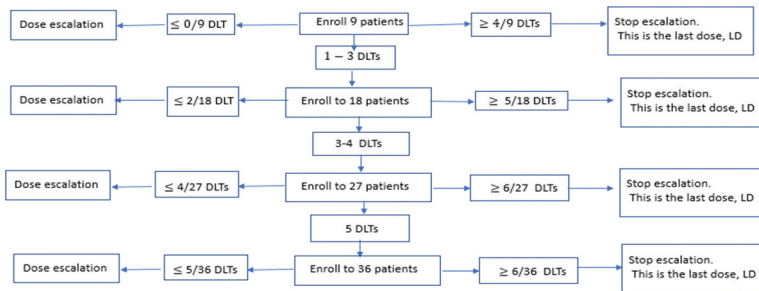
Cohort 9 design
Dose escalation chart: target DLT rate=0. 25



Cohort 9 design
Target DLT=0.25



Cohort 9 design
Dose escalation chart: target DLT rate=0. 2



Cohort 9 design
Target DLT=0.2

