Full title: The 3+3 Design in Dose-Finding Studies with Small Sample Sizes: Pitfalls and Possible Remedies

Running title: 3+3 Design: Pitfalls and Remedies

Word count: 3,885

Authors: Cody Chiuzan[1], Hakim-Moulay Dehbi[2]

[1] Institute of Health System Science, The Feinstein Institutes for Medical Research, New York, NY, 10022, USA

[2] Comprehensive Clinical Trials Unit, University College London, London, United Kingdom

Author for correspondence:

Cody Chiuzan

Institute of Health System Science

Feinstein Institutes for Medical Research

Northwell Health

130 East 59th St, Suite 14C

New York, NY 10022

Tel: 919-793-6266

Email: cchiuzan@northwell.edu

Abstract

Background: In the last few years, numerous novel designs have been proposed to improve the efficiency and accuracy of phase I trials to identify the maximum-tolerated dose (MTD) or the optimal biological dose (OBD) for non-cytotoxic agents. However, the conventional 3+3 approach, known for its limited sample size and poor performance continues to be an attractive choice for many trials, despite these alternative suggestions.

Objective: The article seeks to underscore the importance of moving beyond the 3+3 design by highlighting a different key element in trial design: the estimation of sample size and its crucial role in predicting toxicity and determining the MTD. We use simulation studies to compare the performance of the most used phase I approaches: 3+3, CRM, Keyboard and BOIN designs regarding three key operating characteristics: the percentage of correct selection (PCS) of the true MTD, the average number of patients allocated per dose level, and the average total sample size.

Results: The simulation results consistently show that the 3+3 algorithm underperforms in comparison to model-based and model-assisted designs across all scenarios and metrics. The 3+3 method yields significantly lower (up to three times) probabilities in identifying the correct MTD, often selecting doses one or even two levels below the actual MTD. The 3+3 design allocates significantly fewer patients at the true MTD, assigns higher numbers to lower dose levels, and rarely explores doses above the target DLT rate.

Conclusions: The overall performance of the 3+3 method is suboptimal, with a high level of unexplained uncertainty and significant implications for accurately determining the MTD. While

the primary focus of the paper is to demonstrate the limitations of the 3+3 algorithm, the question remains about the preferred alternative approach. The intention is not to definitively recommend one model-based or model-assisted method over others, as their performance can vary based on parameters and model specifications. However, the presented results indicate that the CRM, Keyboard, and BOIN designs consistently outperform the 3+3 and offer improved efficiency and precision in determining the MTD, which is crucial in early-phase clinical trials.

Keywords: phase I dose-finding studies, small sample size, model-based designs, model-assisted designs, 3+3 algorithm.

Introduction

Sample size determination is a critical aspect of phase I dose-finding designs in clinical trials. "How many patients do I need?" is a challenging question that does not always have a simple answer and usually requires a collaborative effort between clinicians and statisticians to translate clinical goals into a study that generates reliable and meaningful results. In phase I studies, sample size estimation is a fundamental process that balances the need to ensure patient safety with the goal of selecting the optimal dose of a new drug (single agent or combination) to move forward into later phases of drug development. Traditionally, for cytotoxic oncology agents, the primary objective of phase I trials is the identification of the maximum tolerated dose (MTD) based on assessment of toxicity, with the underlying assumption that both toxicity and efficacy increase with the dose level. The primary toxicity endpoint is usually a binary dose-limiting toxicity (DLT, yes/no) based on protocol-specific adverse event definitions. Upon completion of dose escalation, the MTD is identified as the highest dose that can be administered with an acceptable level of toxicity.

In the last decades, the rapid therapeutic development of non-cytotoxic agents (e.g., molecularly targeted agents or immunotherapy agents) motivated the inclusion of additional endpoints in dose-escalation such as pharmacokinetics (PK), pharmacodynamics (PD) or markers of clinical efficacy. These novel agents can have lower toxicity profiles and efficacy may occur at doses that do not induce clinically significant toxicity (1–3). On these premises, recent dose-finding trials have incorporated a proof-of-principle of biologic effect, that is, evidence of antitumor activity or other immunologic parameters, and targeted the identification of the optimal biological dose (OBD) rather than the MTD. Several sophisticated dose escalation designs (e.g., bivariate models versus joint models, binary versus ordinal endpoints) have been proposed in the context of OBD in an attempt to consider both efficacy and toxicity during the

course of the study (4–8). Recent publications have suggested to guide the escalation decisions based on toxicity only and incorporate the efficacy and PK/PD markers at study end to choose the OBD dose (9,10). However, the majority of phase I oncology trials are still driven by a single binary toxicity endpoint with the MTD being the primary trial objective (11). Various dose-finding methods have been proposed for designing phase I trials. Comparisons of these methods have been covered thoroughly in the literature including in recent reviews (12), and guidelines on key aspects of trial design have been published for the clinical community (13,14). The three main approaches currently used in phase I studies are: rule-based, model-based, and model-assisted designs.

Rule-based or algorithmic designs such as the 3+3 continue to be popular approaches for dose-finding mainly due to the practical simplicity (15–17). The 3+3 assigns patients sequentially starting at the lowest dose and escalating after every three to six patients per dose; the recommended dose is defined as the largest dose with fewer than two patients experiencing a DLT during the observation window (i.e., first treatment cycle). With no underlying dose-toxicity model, dose escalation relies more on empirical reasoning and several limitations have been consistently raised over time. The 3+3 algorithm lacks the capability to specify a target DLT rate, is slow in escalation, has high error rates that lead to inaccurate dose recommendations, and enrolls a significant proportion of patients at subtherapeutic doses (16,18). Since cohorts are limited to six patients per dose, in order to obtain additional evidence for further characterizing safety and preliminary efficacy signals, trials have used prespecified expansion cohorts after the dose-escalation component of the trial is completed (19,20).

Model-based designs have been developed to target a pre-specified toxicity rate and improve precision in estimating the MTD and/or recommended phase 2 dose (RP2D) as well as

efficacy during the dose escalation. Implemented in both Bayesian and Frequentist frameworks, model-based designs use a pre-specified dose-toxicity curve that is updated with observed toxicity data as the trial proceeds. With increasing awareness of the limitation and poor operating characteristics of the algorithmic design, investigators (especially from the pharmaceutical industry) have been more open to using model-based designs that achieve a better estimation of the target DLT rate at the MTD/RP2D while minimizing suboptimal dosing (21–24). One of the most commonly applied model-based approaches is the Continual Reassessment Method (CRM) (25) and its variants that can incorporate both toxicity (including late-onset) and efficacy outcomes and can be implemented for testing single and multiple-agent combinations (5,26–28). Despite several advantages of model-based designs over traditional 3+3 designs, the high requirement of statistical and computational expertise coupled with the lack of predetermined algorithms to be followed limited their adoption in phase I trials.

Model-assisted designs can be viewed as a hybrid between rule- and model-based designs. This category also relies on a statistical model for decision making, but like rule-based designs, the escalation/de-escalation rules can be tabulated before the trial starts and easily followed for dose allocation. Examples of model-assisted designs include the modified toxicity probability interval designs: mTPI (29), Keyboard/mTPI-2 (30,31) and the Bayesian Optimal Interval (BOIN) design (32). Simulations results showed that the model-assisted designs significantly outperform the 3 + 3 design and have comparable characteristics to model-based designs on several metrics, including the accuracy of identification of the MTD, allocation of patients to the MTD and reduction of overdosing risk (33,34).

In a recent article published in the American Society of Clinical Oncology Educational Book, Kurzrock et al., 2021, provide a comprehensive review of the *status quo* in phase I trial

design methodology highlighting the superiority of model-based or model-assisted designs compared with traditional 3+3 design in efficiency, safety, and flexibility. They also emphasize that "there is no reason that the 'standard of trial design' has to be fixed in the traditional 3+3 design, ignoring the demonstrated advantages of novel designs" (12). On the same note, this article aims to reinforce the message of stepping away from the 3+3 by focusing upon another critical aspect of trial design: sample size estimation and understanding its critical role in estimating toxicity and the MTD.

So, what constitutes an adequate sample size? In phase I trials, the sample size is generally determined considering two metrics: the percentage of correct selection (PCS) of the true MTD and the average number of patients allocated per dose level. We are using these measures in simulation studies to compare the performance of the most used phase I approaches: 3+3, CRM, Keyboard and BOIN designs. The objective of this study is not to recommend the 'best' model-based or method-assisted method, but to demonstrate the poor behavior and emphasize the severe limitations of the 3+3 under different toxicity scenarios.

Overview of Selected Dose-Escalation Designs

3+3 Algorithm

This rule-based design guides 'up-and-down' decisions, using the modified Fibonacci mathematical series to determine the amount of dose increase for cohorts of sequentially enrolled patients. In a 3 + 3 design, three patients are initially enrolled into a given dose cohort. If no DLT is observed in any of these subjects, the trial proceeds to enroll additional patients into the next higher dose cohort. If one subject develops a DLT at a specific dose, an additional three subjects are enrolled into that same dose cohort. The dose escalation continues until at least two patients

among a cohort of three to six patients experience DLTs (i.e., ≥33% of patients with a DLT at that dose level). The MTD is then defined as one dose level just below this toxic level.

As previously mentioned, the 3+3 has no specific target DLT rate. Using only data at the current dose to choose the next dose and MTD, the algorithm results in uncertainty surrounding the estimated DLT at each dose, with values ranging between 17% and 33% (35, 38).

Continual Reassessment Method (CRM)

The original CRM was implemented in a Bayesian framework relying on the use of a working dose-toxicity model with a prior distribution to sequentially update the dose-toxicity curve and estimate the dose level at which to treat the next available cohort of patients. The dose-toxicity model describes the probability of a patient experiencing a DLT at a given dose. The most common implementation of the CRM uses the 'empiric' model $F(d_k, a) = d_k^{\exp(a)}$, with a mean zero normal prior $N(0, \sigma_a^2)$ (25). The model is updated with accumulating toxicity data from the trial and allocates the next patient cohort to the dose level with an estimated DLT rate closest to the prespecified target rate (varying between 20 to 35%). Simulations showed that the CRM design achieves the recommended MTD after a median of three to four patients fewer than the '3+3' design (36). This not only reduces the cost and time required for the trial but also reduces the risk to patients by minimizing their exposure to potentially toxic doses.

Modified Toxicity Probability Interval (mTPI) and Keyboard Designs

The mTPI design starts with a definition of three toxicity probability intervals: underdosing, proper dosing, and overdosing intervals. A Bayesian framework, taking into consideration the relative distance between toxicity rate at each dose level and target probability for a fixed sample

size, is used to calculate the posterior probabilities of intervals. Patients in the first cohort are treated at the lowest dose level and the next assignments (escalation/de-escalation) are made based on prespecified algorithmic-like rules until the maximum sample size is reached or a certain number of patients is treated at a single dose (e.g., 6 or 9). In mTPI the overdosing interval is typically wider than the proper dosing interval; this can lead to a high risk of overdose of patients (at doses greater than the MTD). The Keyboard design (also known as mTPI-2) addresses the overdose issue by using a dose escalation determined by the location of the strongest *key* relative to the target dosing interval that includes target toxicity θ. The strongest *key* is defined to be the dosing interval that most likely contains the current dose's true toxicity rate, which is determined based on the posterior probability that each interval includes the target toxicity. With substantially lower risk of overdosing patients and better accuracy in identifying the MTD, the Keyboard design has been shown to outperform the mTPI (31).

Bayesian Optimal Interval (BOIN) Design

The BOIN design mimics the 3+3 simplicity and makes the dose escalation/de-escalation recommendations by comparing $\hat{p}$, the observed DLT rate at the current dose, with two boundaries (λ_e, λ_d) that depend on a target toxicity rate. Specifically, $\hat{p}$ is defined as the number of patients experiencing DLT at the current dose divided by the total number of DLT-evaluable patients treated at the current dose. The design starts by treating a cohort of patients at the lowest dose and subsequent dose assignments are based on the following rules: if $\hat{p} \leq \lambda_e$, escalate the dose to the next higher level; if $\hat{p} \geq \lambda_d$, de-escalate the dose to the next lower level; otherwise, stay at the current dose. These steps repeat until the maximum sample size is reached or a certain number of patients is treated at a single dose (e.g., 9 or 12). More details of determining the two

boundaries for commonly used target toxicity rates are provided in Yuan et al, 2016 (32). Simulations studies showed that the BOIN design is more likely to correctly select the MTD and allocate more patients to the MTD than the 3+3 design and it has a lower risk of overdosing patients than the mTPI design. Comparisons of BOIN and CRM suggested that these two designs have comparable performance (37).

Phase I Trial Example

For illustration we consider a hypothetical phase I trial that aims to identify the MTD of a single agent defined as the dose with a DLT rate $\theta = 0.33$. Five prespecified doses are considered with a cohort size of three patients. The trial starts with the first cohort of patients receiving dose level 1. For the CRM, Keyboard and BOIN designs, we chose $\theta = 0.33$ as target toxicity rate. For the 3+3 design, the target DLT is not actually specified, thus the MTD is determined according to the rules described above. The prespecified maximum sample size is of 30 patients (i.e., six times the number of dose levels). The number of doses investigated in a trial is often driven by clinical considerations. Previous reviews of phase I trials have shown that the median number of dose levels explored was 5 (range 2–12) (39). Moreover, Wheeler et al., 2019, show that different dose range choices affect MTD selection; too few doses may lead to inaccurate MTD estimation, while too many doses may have a slow dose escalation towards the MTD (40).

Design Specifications and Simulations Setting

Simulations were performed using the following software/apps: 3+3 - R function *ssim3p3* (UBCRM package), CRM web application (41) available at https://uvatrapps.shinyapps.io/crmb/, and web applications “Keyboard” and “BOIN” (BOIN V2.7.6.0) available at

http://www.trialdesign.org. For the CRM, Keyboard and BOIN designs we implemented the default and/or recommended parameters specified in the web applications.

For the Bayesian CRM, we employed the 'empiric' model with the least non-informative normal prior distribution $N(0, \sigma_a^2)$, with a standard deviation $\sigma_a = 0.94$ on the model parameter a (25). The skeleton values (i.e., initial DLT probability estimates) were generated by setting the prior MTD (ν) to be the median dose (42, 43), and a spacing measure (δ) of 0.05 (44), which produce reasonable skeletons for many scenarios. The trial stops for safety if the lower limit of the 90% probability interval for the lowest study dose level exceeds the target DLT rate.

For the Keyboard (i.e., mTPI-2) design the acceptable toxicity probability interval corresponding to the target DLT rate is (0.28, 0.38). Patients in the first cohort are treated at the lowest dose level, and the next dose escalation/de-escalation assignments are conducted according to the rules displayed in Table S1.

The BOIN design uses the following rules to guide dose escalation/de-escalation: if the observed DLT rate at the current dose is ≤ 0.26 (λ_e), escalate the dose to the next higher dose level; if the observed DLT rate at the current dose is > 0.395 (λ_d), de-escalate the dose to the next lower dose level; otherwise, stay at the current dose. These boundaries are dependent on target DLT (in our example 0.33) and they use a binomial model to minimize the incorrect decisions of escalation/de-escalation the dose when it is actually greater/lower than the MTD (37). If there is a 95% probability that toxicity at a certain dose exceeds 0.33, then the current dose and all higher levels are eliminated. Dose escalation/de-escalation assignments are conducted according to the rules displayed in Table S2.

In the case of CRM, Keyboard, and BOIN simulations, the trial ends either upon reaching the maximum sample size of 30 or when the recommendation is to assign the next cohort to a dose that has already been assigned to 9 patients (maximum). This additional stopping rule proves advantageous in saving sample size and reducing the trial duration, particularly in scenarios where there is prior knowledge that the first dose is safe.

As shown in Figure1, four dose-toxicity scenarios are considered with five dose levels. All scenarios, inspired by clinical applications, assume a monotonic increasing relationship, with various slopes, locations and spacing of the true MTD. We ran 1,000 simulations per scenario and computed the following operating characteristics for each method: the percentage of correct selection (PCS) of the true MTD, the percentage of early stopping, the average number of patients allocated per dose, and the average total sample size.

Results

Tables 1 and 2 show the operating characteristics of the four dose escalation methods under different dose-toxicity scenarios, with the true MTD (target DLT rate of 0.33) varying from dose level 1 to dose level 5. Toxicity probabilities are based on clinical applications and with the objective of observing the methods' more extreme behavior under high or low toxicity profiles, as seen in scenarios 1 and 4, respectively.

Simulation results indicate that under all scenarios and metrics, the 3+3 considerably underperforms compared to model-based and model-assisted designs. Table 1 presents the percentages of correctly selecting (PCS) the MTD and early stopping generated by the four approaches. In three out of four scenarios, the algorithm fails to recommend the true MTD generating the highest percentage of selection at one or even two dose levels below it.

This conservative nature and risk of underestimation are on par with previous theoretical results showing that in fact 3+3 targets DLT rates between 16% to 27% (33). In scenario 1, where the true MTD is dose level 1, the 3+3 correctly identifies it in only 30.7% of the simulated trials and stops with no recommendation for the MTD 63.9% of the time. Overall, with a high percentage of unexplained uncertainty and severe implications on accurately determining the MTD, the algorithm never exceeds 44% in (correct or incorrect) MTD selection.

An attractive feature of the 3+3 is the small sample size. Enrolling a small number of patients is usually faster and less logistically challenging than recruiting larger cohorts. Results in Table 2 show that with five doses the 3+3 produces average sample sizes ranging from 7.4 to 18.5 saving an average of five patients in total compared to CRM and up to nine patients compared to BOIN and Keyboard designs. However, in terms of allocation per dose, the algorithm treats up to 50% less patients at the true MTD, generates higher allocations at lower dose levels and barely visits dose levels above the target DLT rate. As a result, the identified MTD may not be optimal and there may be a higher risk of underestimating or overestimating toxicity. From an ethical point of view, if assuming an underlying increasing dose-efficacy relationship, this inefficient patient allocation exposes patients to subtherapeutic doses (as seen in our simulations) raising concerns about patient welfare. Notably, our simulations consider a fixed sample size of 30, chosen to equal the maximum number of patients that could be enrolled under the 3+3 setting. In practice though, expanding phase I trials to include dose-expansion cohorts has become common in the last decade. These expansion cohorts usually follow a 3+3 dose escalation and are used to collect more information at the estimated MTD to further define safety and collect preliminary signs of efficacy. A recent analysis found that dose-expansion cohorts enrolled up to 271 patients and were less likely used for testing cytotoxic agents (45).

Alternatively, the CRM and BOIN have the advantage of allowing for flexibility in sample size and previous studies have shown that these models can have a sizable impact on reducing the sample size, while maintaining the ability to identify the MTD (46).

While the focus of this paper is to demonstrate the poor performance of the 3+3 algorithm, the inevitable question remains: what method should be used instead? Our goal is not to ultimately and decisively recommend *one* of the model-based or mode-assisted approaches, as their operating characteristics can vary depending on several parameters and model specifications (and across scenarios, though that is not the scope here). Tables 1 and 2 show that the CRM, Keyboard and BOIN designs have similar levels of efficiency for identifying the MTD and generate PCS values two or three times (scenario 4) higher than the 3+3. In some situations, the non-algorithmic designs select the true MTD in less than 50% of the trials, but still show a significant improvement from the 3+3. Overall, the CRM correctly selects the true MTD in all toxicity scenarios with the smallest average number of patients treated per dose and total sample size. The BOIN and Keyboard designs have almost identical results, but that is expected under these simulations setting since both follow the same escalation/de-escalation rules (see Tables S1 and S2).

Discussion

Several factors can affect the efficiency of a phase I trial (e.g., number of patients enrolled and the proportion of those treated at subtherapeutic doses, number of dose levels tested, target DLT rate, trial duration). Phase I trials are usually small with a limited number of patients (i.e., 12 to 30) and the maximum sample size being set as six times the number of dose levels considered. This estimation is often inspired from the 3+3 setting, with no other clear justification. In this

paper, we seek to compare the relative performances of several phase I clinical trial designs under such a fixed sample size setting and point out several limitations with using the 3+3 algorithm and its simplistic sample size approach. We have ultimately demonstrated that, regardless of the underlying true DLT scenario, the 3+3 has a poor performance with a low ability of identifying the MTD and high risk of underdosing. The performances of the model-assisted designs (i.e., BOIN and Keyboard/mTPI-2) are comparable to that of the model-based CRM design and overall superior to the 3+3. However, in most scenarios, a fair amount of uncertainty remains even with model-based/assisted approaches. This is to be expected with any study that uses a fixed small sample size which is often the case in early phase dose-finding trials. The adequate sample size varies with different parameters such as target toxicities, different dose-toxicity relationships, and range of doses. Our goal was not to extensively study the design characteristics and propose an optimal sample size, but to show that the 'standard' 3+3 setting consistently fails and that model alternatives are to be preferred. We should not necessarily assume that a larger sample size leads to improved accuracy. The answer is not straightforward and should be addressed in the planning stage of the trial by performing simulations with input from the clinical team. Another misconception regards the number of doses and novel designs' inability to provide much benefit in the event of fewer doses. Zhu et al., 2019 show that even with three dose levels, the 3 + 3 design still performs much worse than the CRM, BOIN, and Keyboard designs (47). The choice of cohort size is another important factor impacting sample size calculations. Three patients per cohort is, again, not based on any statistical justification, but rather on practicality and simplicity. Model-based approaches can be implemented for one or more patients per cohort and thereby allow for a more frequent updating

of dose-toxicity curves, a more accurate MTD estimation and skewing of patients' allocation to more promising/effective doses.

In the era of molecularly targeted agents (MTA) and biological agents, the assumption of monotonic toxicity is not necessarily met. The novel agents have different toxicity profiles and may show non-monotonic dose-efficacy curve (e.g., plateau of antitumor effect). Under these circumstances, toxicity is no longer the main endpoint of interest and the process of selecting the optimal biological dose (OBD) usually incorporates both measures of toxicity and efficacy. The 3+3 design's use of solely toxicity to guide dose escalation makes it a very reductionist approach, unlikely to be optimized to ascertain the OBD.

Lastly, model-based and model-assisted designs can adeptly handle a broader range of endpoints beyond binary DLT including ordinal (48) or time-to-event outcomes (26,49). These designs are also better suited for evaluating combination therapies, which pose unique challenges due to the larger dimensions of the dose search space and the partial ordering among drug combinations.

Therefore, all evidence from simulations or real trials suggests that the 3+3 no longer has a place in the design of phase I trials, especially nowadays, when implementation of model-based/-assisted designs is increasingly facilitated by freely available, user-friendly software.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

1. Corbaux P, El-Madani M, Tod M, Péron J, Maillet D, Lopez J, et al. Clinical efficacy of the optimal biological dose in early-phase trials of anti-cancer targeted therapies. Eur J Cancer. 2019 Oct;120:40–6.

2. Patil V, Noronha V, Joshi A, Abhyankar A, Menon N, Banavali S, et al. Low doses in immunotherapy: Are they effective? Cancer Res Stat Treat. 2019;2(1):54.

3. Sachs JR, Mayawala K, Gadamsetty S, Kang SP, De Alwis DP. Optimal Dosing for Targeted Therapies in Oncology: Drug Development Cases Leading by Example. Clin Cancer Res. 2016 Mar 15;22(6):1318–24.

4. Piantadosi S, Liu G. Improved designs for dose escalation studies using pharmacokinetic measurements. Stat Med. 1996 Aug 15;15(15):1605–18.

5. Braun TM. The bivariate continual reassessment method. extending the CRM to phase I trials of two competing outcomes. Control Clin Trials. 2002 Jun;23(3):240–56.

6. Bekele BN, Shen Y. A Bayesian approach to jointly modeling toxicity and biomarker expression in a phase I/II dose-finding trial. Biometrics. 2005 Jun;61(2):343–54.

7. Dragalin V, Fedorov V. Adaptive designs for dose-finding based on efficacy–toxicity response. J Stat Plan Inference. 2006 Jun;136(6):1800–23.

8. Houede N, Thall PF, Nguyen H, Paoletti X, Kramar A. Utility-based optimization of combination therapy using ordinal toxicity and efficacy in phase I/II trials. Biometrics. 2010 Jun;66(2):532–40.

9. Dehbi HM, O'Quigley J, Iasonos A. Controlled backfill in oncology dose-finding trials. Contemp Clin Trials. 2021 Dec;111:106605.

10. Dehbi HM, O'Quigley J, Iasonos A. Controlled amplification in oncology dose-finding trials. Contemp Clin Trials. 2023 Feb;125:107021.

11. Fraisse J, Dinart D, Tosi D, Bellera C, Mollevi C. Optimal biological dose: a systematic review in cancer phase I clinical trials. BMC Cancer. 2021 Dec;21(1):60.

12. Kurzrock R, Lin CC, Wu TC, Hobbs BP, Pestana RC, Hong DS. Moving Beyond 3+3: The Future of Clinical Trial Design. Am Soc Clin Oncol Educ Book Am Soc Clin Oncol Annu Meet. 2021 Jun;41:e133–44.

13. Lee SM, Wages NA, Goodman KA, Lockhart AC. Designing Dose-Finding Phase I Clinical Trials: Top 10 Questions That Should Be Discussed With Your Statistician. JCO Precis Oncol. 2021 Jan;5:317–24.

14. Araujo D, Greystoke A, Bates S, Bayle A, Calvo E, Castelo-Branco L, et al. Oncology phase I trial design and conduct: time for a change - MDICT Guidelines 2022. Ann Oncol. 2023 Jan;34(1):48–60.

15. Storer BE. Design and analysis of phase I clinical trials. Biometrics. 1989 Sep;45(3):925–37.

16. Hansen AR, Graham DM, Pond GR, Siu LL. Phase 1 Trial Design: Is 3 + 3 the Best? Cancer Control. 2014 Jul;21(3):200–8.

17. Chiuzan C, Shtaynberger J, Manji GA, Duong JK, Schwartz GK, Ivanova A, et al. Dose-finding designs for trials of molecularly targeted agents and immunotherapies. J Biopharm Stat. 2017 May 4;27(3):477–94.

18. Reiner E, Paoletti X, O'Quigley J. Operating characteristics of the standard phase I clinical trial design. Comput Stat Data Anal. 1999 May;30(3):303–15.

19. Manji A, Brana I, Amir E, Tomlinson G, Tannock IF, Bedard PL, et al. Evolution of clinical trial design in early drug development: systematic review of expansion cohort use in single-agent phase I cancer trials. J Clin Oncol Off J Am Soc Clin Oncol. 2013 Nov 20;31(33):4260–7.

20. Iasonos A, O'Quigley J. Design considerations for dose-expansion cohorts in phase I trials. J Clin Oncol Off J Am Soc Clin Oncol. 2013 Nov 1;31(31):4014–21.

21. Mick R, Ratain MJ. Model-guided determination of maximum tolerated dose in phase I clinical trials: evidence for increased precision. J Natl Cancer Inst. 1993 Feb 3;85(3):217–23.

22. Thall PF, Lee SJ. Practical model-based dose-finding in phase I clinical trials: methods based on toxicity. Int J Gynecol Cancer Off J Int Gynecol Cancer Soc. 2003;13(3):251–61.

23. Le Tourneau C, Gan HK, Razak ARA, Paoletti X. Efficiency of new dose escalation designs in dose-finding phase I trials of molecularly targeted agents. PloS One. 2012;7(12):e51039.

24. Love SB, Brown S, Weir CJ, Harbron C, Yap C, Gaschler-Markefski B, et al. Embracing model-based designs for dose-finding trials. Br J Cancer. 2017 Jul 25;117(3):332–9.

25. O'Quigley J, Shen LZ. Continual reassessment method: a likelihood approach. Biometrics. 1996 Jun;52(2):673–84.

26. Cheung YK, Chappell R. Sequential designs for phase I clinical trials with late-onset toxicities. Biometrics. 2000 Dec;56(4):1177–82.

27. Yuan Z, Chappell R, Bailey H. The continual reassessment method for multiple toxicity grades: a Bayesian quasi-likelihood approach. Biometrics. 2007 Mar;63(1):173–9.

28. Wages NA, Conaway MR, O'Quigley J. Continual reassessment method for partial ordering. Biometrics. 2011 Dec;67(4):1555–63.

29. Ji Y, Liu P, Li Y, Bekele BN. A modified toxicity probability interval method for dose-finding trials. Clin Trials Lond Engl. 2010 Dec;7(6):653–63.

30. Guo W, Wang SJ, Yang S, Lynn H, Ji Y. A Bayesian interval dose-finding design addressingOckham's razor: mTPI-2. Contemp Clin Trials. 2017 Jul;58:23–33.

31. Yan F, Mandrekar SJ, Yuan Y. Keyboard: A Novel Bayesian Toxicity Probability Interval Design for Phase I Clinical Trials. Clin Cancer Res Off J Am Assoc Cancer Res. 2017 Aug 1;23(15):3994–4003.

32. Yuan Y, Hess KR, Hilsenbeck SG, Gilbert MR. Bayesian Optimal Interval Design: A Simple and Well-Performing Design for Phase I Oncology Trials. Clin Cancer Res Off J Am Assoc Cancer Res. 2016 Sep 1;22(17):4291–301.

33. Yuan Y, Lee JJ, Hilsenbeck SG. Model-Assisted Designs for Early-Phase Clinical Trials: Simplicity Meets Superiority. JCO Precis Oncol. 2019 Dec;(3):1–12.

34. Zhou H, Yuan Y, Nie L. Accuracy, Safety, and Reliability of Novel Phase I Trial Designs. Clin Cancer Res. 2018 Sep 15;24(18):4357–64.

35. Ivanova A. Escalation, group and A + B designs for dose-finding trials. Stat Med. 2006 Nov 15;25(21):3668–78.

36. Onar A, Kocak M, Boyett JM. Continual reassessment method vs. traditional empirically based design: modifications motivated by Phase I trials in pediatric oncology by the Pediatric Brain Tumor Consortium. J Biopharm Stat. 2009;19(3):437–55.

37. Liu S, Yuan Y. Bayesian optimal interval designs for phase I clinical trials. J R Stat Soc Ser C Appl Stat. 2015;64(3):507–23.

38. Dent SF, Eisenhauer EA. Phase I trial design: Are new methodologies being put into practice? Ann Oncol. 1996 Aug;7(6):561–6.

39. Penel N, Kramar A. What does a modified-Fibonacci dose-escalation actually correspond to? BMC Med Res Methodol. 2012 Jul 23;12:103.

40. Wheeler GM, Mander AP, Bedding A, Brock K, Cornelius V, Grieve AP, et al. How to design a dose-finding study using the continual reassessment method. BMC Med Res Methodol. 2019 Dec;19(1):18.

41. Wages NA, Petroni GR. A web tool for designing and conducting phase I trials using the continual reassessment method. BMC Cancer. 2018 Feb 5;18(1):133.

42. Lee SM, Ying Kuen Cheung null. Model calibration in the continual reassessment method. Clin Trials Lond Engl. 2009 Jun;6(3):227–38.

43. Lee SM, Cheung YK. Calibration of prior variance in the Bayesian continual reassessment method. Stat Med. 2011 Jul 30;30(17):2081–9.

44. Cheung YK. Dose Finding by the Continual Reassessment Method [Internet]. 0 ed. Chapman and Hall/CRC; 2011 [cited 2024 Jan 27]. Available from: https://www.taylorfrancis.com/books/9781420091526

45. Bugano DDG, Hess K, Jardim DLF, Zer A, Meric-Bernstam F, Siu LL, et al. Use of Expansion Cohorts in Phase I Trials and Probability of Success in Phase II for 381 Anticancer Drugs. Clin Cancer Res Off J Am Assoc Cancer Res. 2017 Aug 1;23(15):4020–6.

46. Horton BJ, O'Quigley J, Conaway MR. Consequences of Performing Parallel Dose Finding Trials in Heterogeneous Groups of Patients. JNCI Cancer Spectr. 2019 Jun 1;3(2):pkz013.

47. Zhu Y, Hwang WT, Li Y. Evaluating the effects of design parameters on the performances of phase I trial designs. Contemp Clin Trials Commun. 2019 Sep;15:100379.

48. Van Meter EM, Garrett-Mayer E, Bandyopadhyay D. Dose-finding clinical trial design for ordinal toxicity grades using the continuation ratio model: an extension of the continual reassessment method. Clin Trials. 2012 Jun;9(3):303–13.

49. Yuan Y, Lin R, Li D, Nie L, Warren KE. Time-to-Event Bayesian Optimal Interval Design to Accelerate Phase I Trials. Clin Cancer Res. 2018 Oct 15;24(20):4921–30.

Tables

Table 1. Percentage of correctly selecting the MTD and early stopping generated by the four approaches

		Dose 1	Dose 2	Dose 3	Dose 4	Dose 5	% Early Stopping
Scenario 1		**0.33**	**0.50**	**0.56**	**0.61**	**0.67**	
% MTD Selection	3+3	30.7	4.9	0.4	0.1	0	63.9
	CRM	71.9	13.1	2.3	0.2	0	12.5
	Keyboard	74.2	13.6	1.0	0	0	11.2
	BOIN	74.7	13.4	0.9	0.1	0	10.9
Scenario 2		**0.16**	**0.33**	**0.45**	**0.52**	**0.60**	
% MTD Selection	3+3	44.3	25.8	5.0	0.6	0.1	24.2
	CRM	32.9	41.6	20.5	3.8	0.2	1.0
	Keyboard	28.1	50.3	16.6	3.5	0.1	1.4
	BOIN	28.4	49.7	17.1	3.4	0.1	1.3
Scenario 3		**0.05**	**0.15**	**0.25**	**0.33**	**0.45**	
% MTD Selection	3+3	19.5	34.1	27.8	12.8	3.5	2.3
	CRM	1.4	13.8	34.3	35.6	14.9	0
	Keyboard	1.4	15.7	35.0	32.7	15.2	0
	BOIN	1.3	15.9	34.6	33.3	14.8	0
Scenario 4		**0.02**	**0.08**	**0.12**	**0.18**	**0.33**	
% MTD Selection	3+3	7.5	12.7	23.3	36.8	19.4	0.3
	CRM	0	1.5	5.1	29.5	63.9	0
	Keyboard	0.1	1.7	6.9	30.4	60.9	0
	BOIN	0.1	1.7	6.9	30.4	60.9	0

Operating characteristics are averaged across 1,000 simulated trials.
CRM: Continual Reassessment Method; Keyboard Design; BOIN: Bayesian Optimal Interval Design
% Early Stopping refers to early stopping due to excessive DLTs at the lowest dose.

Table 2. Average sample size and number of patients treated per dose generated by the four approaches

		Dose 1	Dose 2	Dose 3	Dose 4	Dose 5	# Total Patients
Scenario 1		**0.33**	**0.50**	**0.56**	**0.61**	**0.67**	
# Patients treated	3+3	5.2	1.9	0.3	0.03	0	7.4
	CRM	7.9	2.9	0.6	0.04	0	11.4
	Keyboard	8.9	4.3	0.6	0.05	0	13.9
	BOIN	8.9	4.3	0.5	0.05	0	13.8
Scenario 2		**0.16**	**0.33**	**0.45**	**0.52**	**0.60**	
# Patients treated	3+3	5.1	4.1	1.6	0.3	0.06	11.2
	CRM	6.4	6.1	3.3	0.7	0.1	16.6
	Keyboard	7.6	8.4	3.7	0.9	0.1	20.7
	BOIN	7.5	8.3	3.7	0.9	0.1	20.5
Scenario 3		**0.05**	**0.15**	**0.25**	**0.33**	**0.45**	
# Patients treated	3+3	3.9	4.7	4	2.4	0.9	15.9
	CRM	3.9	5.1	5.9	4.6	1.9	21.4
	Keyboard	3.8	6.2	7.1	5	2.2	24.3
	BOIN	3.8	6.2	7	5	2.2	24.2
Scenario 4		**0.02**	**0.08**	**0.12**	**0.18**	**0.33**	
# Patients treated	3+3	3.4	3.9	4.1	4.2	2.9	18.5
	CRM	3.3	3.9	4.4	5.8	6.3	23.7
	Keyboard	3.2	4.2	4.9	6.5	6.6	25.4
	BOIN	3.1	4.2	4.9	6.5	6.5	25.2

Operating characteristics are averaged across 1,000 simulated trials.
CRM: Continual Reassessment Method; Keyboard Design; BOIN: Bayesian Optimal Interval Design

Figure

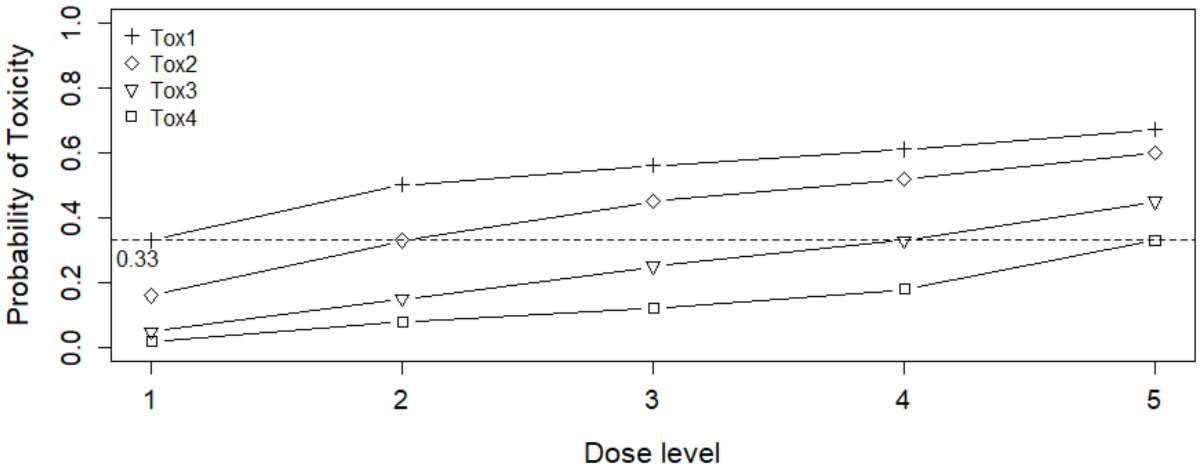

Figure 1 legend: Dose-toxicity scenarios as a function of dose level. Dashed line marks the target dose limiting toxicity (DLT) rate θ=0.33

Supplementary Tables

Table S1. Dose escalation/de-escalation rules for the keyboard design

Number of patients treated at the current dose	1	2	3	4	5	6	7	8	9
Escalate if # of DLT ≤	0	0	0	1	1	1	1	2	2
Deescalate if # of DLT ≥	1	1	2	2	2	3	3	4	4
Eliminate if # of DLT ≥	NA	NA	3	3	4	4	5	5	6

Note. "of DLT" is the number of patients with at least one DLT. When none of the actions (i.e., escalate, de-escalate or eliminate) is triggered, stay at the current dose for treating the next cohort of patients. "NA" means that a dose cannot be eliminated before treating three patients.

Table S2. Dose escalation/de-escalation rule for the BOIN design

Number of patients treated at the current dose	1	2	3	4	5	6	7	8	9
Escalate if # of DLT ≤	0	0	0	1	1	1	1	2	2
Deescalate if # of DLT ≥	1	1	2	2	2	3	3	4	4
Eliminate if # of DLT ≥	NA	NA	3	3	4	4	5	5	6

Note. "# of DLT" is the number of patients with at least one DLT. When none of the actions (i.e., escalate, de-escalate or eliminate) is triggered, stay at the current dose for treating the next cohort of patients. "NA" means that a dose cannot be eliminated before treating three patients.