

Chapter 6

Examining Performance of Phase I Designs: 3+3 Versus Bayesian Optimal Interval (BOIN)



Kenneth R. Hess and Bryan M. Fellman

Abstract In Phase I oncology trials, the primary goal is to assess dose limiting toxicities (DLT) and estimate the maximum tolerated dose (MTD). The classical 3+3 design is still used in the vast majority of studies. In this chapter, we review the 3+3 design and the new Bayesian Optimal Interval (BOIN) design. BOIN is easy to implement, similar to the 3+3, using a simple table to guide dose escalation/de-escalation. As opposed to the 3+3 design, BOIN can target a DLT rate well above or below the usual 25–33% target. We explain how computer simulations can be used to evaluate phase I designs and present results comparing the designs under a large number of true dose-toxicity scenarios. We show that BOIN has better performance than 3+3. BOIN selects the true MTD at a much higher rate and treats a higher percentage of patients at the MTD. BOIN allocates fewer patients to low toxicity doses. Unlike older Bayesian designs (e.g., modified continual reassessment method), BOIN does not require a statistician to be available during the trial. Readily-available, free software makes BOIN simple to implement. We recommend the use of BOIN over the 3+3 design.

Keywords 3+3 · Bayesian optimal interval · Statistical properties and performance Novel phase I design

Key Points

1. A key goal of phase I oncology trials is to estimate the maximum tolerated dose (MTD)
2. The vast majority of phase I oncology trials use the simple 3+3 design
3. The Bayesian Optimal Interval (BOIN) design is one of a new class of model-assisted designs

K. R. Hess (✉) · B. M. Fellman

Department of Biostatistics, UT MD Anderson Cancer Center, Houston, TX, USA

e-mail: khess@mdanderson.org; bmfellman@mdanderson.org

© Springer Nature Switzerland AG 2020

T. A. Yap et al. (eds.), *Phase I Oncology Drug Development*,

https://doi.org/10.1007/978-3-030-47682-3_6

4. BOIN has superior statistical properties as shown by numerous computer simulation studies
5. Free BOIN software is available on a wide range of platforms.

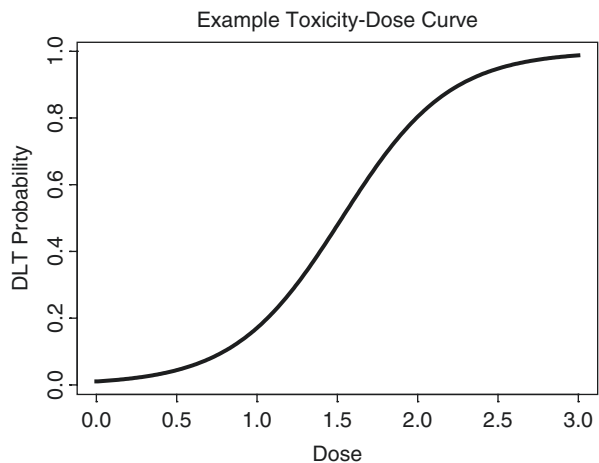
6.1 Introduction

The key goals of a phase I dose-escalation study in oncology are to assess the types, severities and incidences of dose-limiting toxicities (DLTs) and to estimate the maximum tolerated dose (MTD) of experimental therapy. In designing a phase I study, the following must be specified: the patient population, the types of toxicities of interest, the route and schedule of administration of the experimental therapy, and a set of possible doses to be studied (see Chap. 1). We assume that the probability of DLT increases with dose (Fig. 6.1).

In a typical phase I dose-escalation study, at a given dose-level, small cohorts of patients are treated and DLT outcomes are observed. Based on the DLT outcomes the dose is escalated, de-escalated or retained at the current level. This process is repeated until either the maximum dose level is studied or the MTD is reached.

An ideal phase I study design is intuitive both to clinical investigators and statisticians, painless to implement and has good statistical properties including reliably and accurately estimating the MTD.

Fig. 6.1 Example toxicity-dose curve



6.2 Classic 3+3 Dose Escalation Design

Historically, the 3+3 design has been used in the vast majority of oncology studies. This design treats patients in cohorts of 3 following a strict set of rules (Fig. 6.2). At a given dose level, the dose for the next cohort is escalated if 0 of 3 patients experience DLT, the dose is retained if 1 of 3 patients experience DLT, and the dose is de-escalated if >1 of 3 patients experience DLT. The maximum number of patients for the 3+3 design is six times the number of dose levels. However, if no DLTs are encountered, the expected number of patients is three times the number of dose levels plus three additional patients to have six treated at the MTD. So, for 5 dose levels, the maximum number of patients would be 30 and if no DLTs are observed, the expected number of patients would be 18.

For example, a 3+3 design would proceed (given hypothetical DLT results) as:

- Step 1: 3 patients are treated at dose level 1 and 0 experience DLT;
- Step 2: 3 patients are treated at dose level 2 and 0 experience DLT;
- Step 3: 3 patients are treated at dose level 3 and 1 patient develops DLT;
- Step 4: 3 more patients are treated at dose level 3 and 0 of these patients experience DLT (so 1/6 patients experience DLT at dose level 3);
- Step 5: 3 patients are treated at dose level 4 and 1 patient develops DLT;

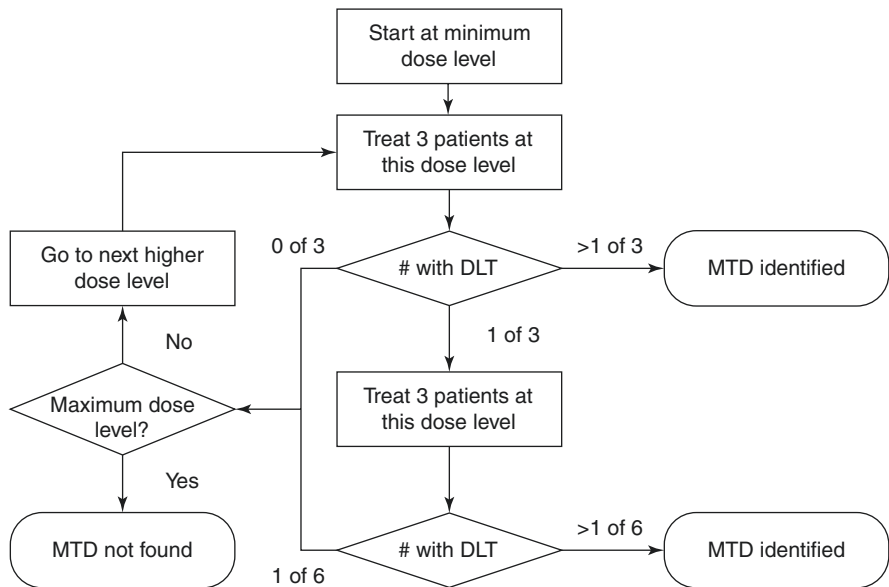


Fig. 6.2 3+3 Flow Diagram

- Step 6: 3 more patients are treated at dose level 4 and 1 of these patients experiences DLT (so 2/6 patients experience DLT at dose level 4);
- Step 7: stop dose escalation and establish dose level 3 as the MTD. Thus a total of 18 patients were treated and the DLT rate at the MTD was $1/6 = 17\%$.

6.3 Newer Designs: Bayesian Statistics Applied to Dose Escalation Designs

Over the years, statisticians have developed several dose escalation designs with superior statistical properties compared to the 3+3 design [1]. Designs have also been developed specifically for targeted and immunotherapy ([2]; see more in Chap. 10).

They are typically model-based or model-assisted. A model-based design like the modified Continual Reassessment Method (mCRM) establishes a statistical model that relates DLT probability to dose level. Other more recent designs are termed model-assisted because, while they are based on probability models, updating the model parameters based on accruing data is not necessary. The decision rule for dose escalation and de-escalation can be predetermined and included in the trial protocol. This greatly simplifies their implementation and means that a statistician does not need to be available to update the model during the trial. The BOIN design described below is an example of such a design.

BOIN Design The Bayesian Optimal Interval, or BOIN design [3] is a model-assisted Bayesian design which is straight-forward to implement and has superior performance to the 3+3 design [1]. The BOIN design is very flexible in that any DLT rate can be used as the target rate for estimating the MTD; any reasonable cohort size (e.g., 1, 2, 3, 4) can be used; and the maximum number of patients to be studied can be pre-specified. The design is easy to implement because the underlying probability model and the design parameters are used to generate *a priori*, a single table that guides dose finding. This table is easily generated using freely available software which can also be used to estimate the statistical properties of the design for a wide range of hypothetical dose-toxicity scenarios using computer simulations.

The goal of the BOIN design [4] is to minimize decision errors of escalating (or deescalating) the dose when the current dose actually is above (or below) the MTD. The design creates three distinct probability regions for the observed DLT rate: “escalate”, “retain” and “de-escalate” (Fig. 6.3). The boundaries between regions are derived to satisfy statistical optimization properties. The boundaries for the oft-chosen 30% target toxicity rate are 0.236 and 0.358 [3]. Thus, the probability regions are: *escalate* = 0 to 0.236, *retain* (i.e., stay at current dose level): 0.237 to 0.357, and *de-escalate* = 0.358 to 1. The design monitors which probability region the observed DLT rate of the current dose level falls and makes decisions

accordingly (Fig. 6.4). For example, if the observed DLT rate at the current dose (e.g., $1/6 = 0.167$) is less than 0.236, the design escalates the dose; if the observed DLT rate at the current dose (e.g., $3/6 = 0.5$) is greater than 0.358, the design de-escalates the dose. BOIN allows for the maximum number of patients to be specified as well as the maximum number of patients to be treated at any given dose level.

An example of a BOIN dose-finding decision table is shown in Table 6.1 (design parameters: target DLT rate = 30%, maximum number of patients treated at single dose = 15, cohort size = 3). For example, if 3 patients have been treated at a given dose level, the decisions are to escalate if 0 patients experience DLT, stay at current dose (i.e., retain) if 1 patient experiences DLT (determined by process of elimination because other options are explicitly ruled out); de-escalate if 2 or more patients experience DLT; and to eliminate the dose level from future consideration if all 3 of the patients experience DLT.

Fig. 6.3 Three distinct probability regions for the observed DLT rate

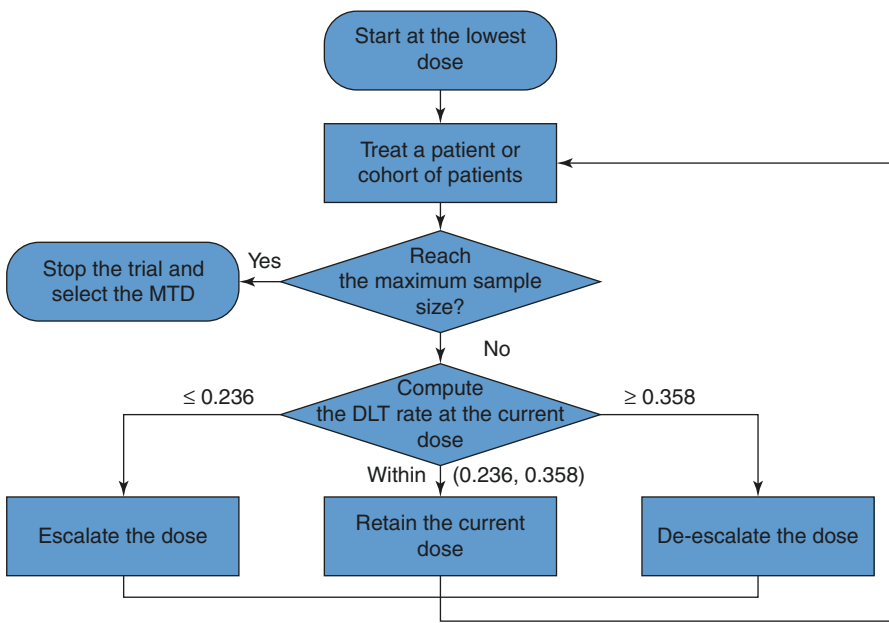
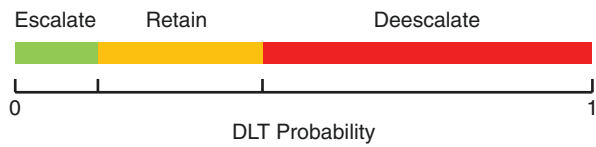


Fig. 6.4 BOIN Flow Diagram

Table 6.1 Dose escalation/de-escalation rule for the BOIN design

Actions	The number of patients treated at the current dose														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Escalate if # of DLT \leq	0	0	0	0	1	1	1	1	2	2	2	2	3	3	3
De-escalate if # of DLT \geq	1	1	2	2	2	3	3	3	4	4	4	5	5	6	6
Eliminate if # of DLT \geq	NA	NA	3	3	4	4	5	5	5	6	6	7	7	8	8

6.4 Making Decisions with Small Cohorts: The Role of Random Variation in DLT Results

Because only a relatively small number of patients are typically studied at a given dose level in a phase I study, it is important to understand the effect of random variation on the observed DLT results. DLT events are binary (yes/no) in nature (i.e., did patient experience DLT during pre-specified observation period). Binary events can be viewed as coin flips (i.e., two well-defined, mutually exclusive outcomes). Observing the number of patients experiencing DLTs out three patients total is analogous to observing the number heads in 3 coin flips. With 3 flips, there are 4 possibilities: 0, 1, 2, or 3 heads. If we assume that the number of heads follows the binomial distribution, then we can compute the probability of observing 0, 1, 2, or 3 heads in 3 flips if we know the true probability of getting a head on a given flip of the coin.

Figure 6.5 shows the probability distribution for a 0.5 probability of heads. It shows that even though the true probability of heads is 50%, the probability of observing zero heads in 3 flips or 3 heads in 3 flips is 12.5%. Of course the probability of observing 1 head in 3 flips or 2 heads in 1 flip is much higher at 27.5%. So given the small number of flips, while there is a 75% chance of observing either 1/3 or 2/3 heads, there is a 25% chance of observing either 0/3 or 3/3 heads. The lesson for phase I trials is that basing decisions on the observed DLT outcomes of 3 patients is fraught with uncertainty. If the true DLT rate at a given dose level is 0.5, and we treat 3 patients at this level, there is a 25% probability that the observed DLT rate will be extreme (0/3 = 0% or 3/3 = 100%). Figure 6.6 shows the probability distribution for a 0.1 probability of heads. It shows that the probability of observing zero heads in 3 flips is 72.9%, probability of 1 head is 24.3%, the probability of 2 heads is 2.7% and the probability of 3 heads is 0.1%. So in the phase I setting, even though the true DLT rate is 0.1, there is a 27% chance that the observed DLT rate will be >0.1 .

We can also think of the precision in our estimation of the DLT rate at the dose level selected as the MTD. If $1/6 = 17\%$ patients develops DLT, the exact 95% confidence interval for this estimate ranges from 0% to 64%. However, if we treat 12 patients at the selected MTD and observe 2 patients with DLT, then $2/12 = 17\%$ and the exact 95% confidence interval extends from 2% to 48% and if we treat 24 patients and observe 4 patients with DLT, then $4/24 = 17\%$ with an exact 95% confidence interval extending from 5% to 37%. Clearly the estimate based on 24

Fig. 6.5 Probability of 0–3 heads in 3 flips

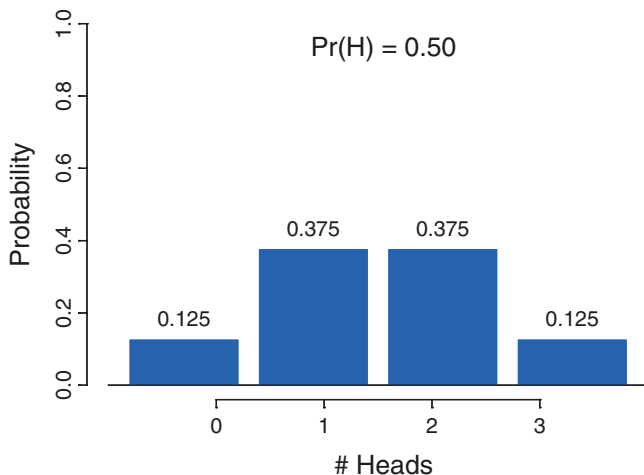
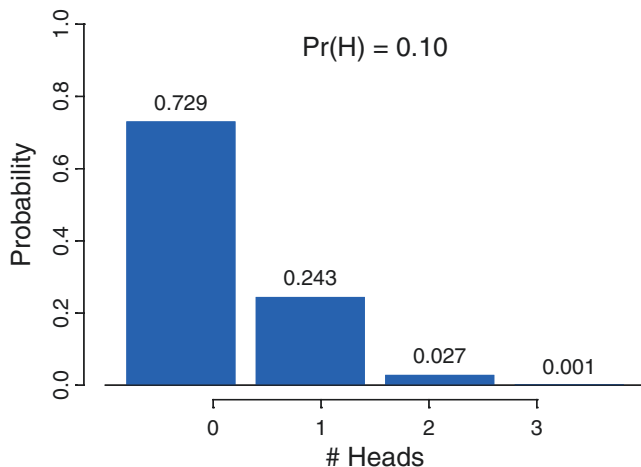


Fig. 6.6 Probability of 0–3 heads in 3 flips



patients has much more precision (i.e., less uncertainty). Knowing that the observed data at the presumed MTD are consistent with a range of DLT rates from 5% to 37% is much more informative than one that ranges from 0% to 64%.

6.5 Simulating Trial Conduct to Assess Design Performance

In order to assess the statistical properties of how a phase I dose-finding method performs, investigators design computer simulations. Given a number of dose levels, they specify a series of scenarios of true DLT rates at each dose level (i.e., dose-toxicity relationships). Given a cohort size (typically 3 patients), the idea is to

generate random data to represent the number of patients experiencing DLTs given the cohort size and true DLT rate (e.g., 3 patients total with DLT rate = 0.5 might yield 2 patients with DLT). The computer uses a random number generator to randomly select a number representing the number of patients experiencing DLT given the cohort size and true DLT rate and assuming the results follow the binomial distribution. A computer program implementing the dose-finding method (e.g., 3+3 or BOIN) is used to determine for each set of generated data how the method would act in terms of escalating, de-escalating, or retaining the dose level. The simulation continues for a given trial until the MTD is found, the maximum number of allowable patients is treated at the highest dose level or, in the case of BOIN, the maximum total number of patients is reached. This process is repeated many, many times (typically 10,000) and the results averaged over these simulations.

Generally a wide range of scenarios is specified to “stress test” the design to show how it operates under a wide range of scenarios, including ones where the true DLT rates are very low for all dose levels (well below the targeted DLT rate); ones where the targeted DLT rate is associated with the lowest dose level; ones where the targeted DLT rate is associated with a middle dose level; ones where the targeted DLT rate is associated with the highest dose level; and ones where all dose levels have high true DLT rates (well above the targeted DLT rate).

Various performance metrics are reported as “operating characteristics” of a phase I design. These are the statistical properties of the design that demonstrate how it performs over a wide range of hypothetical dose-toxicity scenarios. These metrics include: the probability of selecting the dose level with DLT rate closest to the target as the MTD; the percentage of patients treated at the dose level with DLT rate closest to the target as the MTD; percentage of patients treated at dose levels with DLT rates well above the target; and the percentage of patients treated at dose levels with DLT rates well below the target.

6.5.1 Example Simulation

Table 6.2 shows results from a very limited simulation study of the 3+3 design with 3 dose levels with true DLT rates of 0.10, 0.25 (target), and 0.45. For each simulated cohort, random data are generated using the true DLT rates. In the first experiment, first cohort, 0 simulated patients experience DLT and the dose is escalated to level 2. In the second cohort of 3 patients now being treated at dose level 2, 1 patient develops DLT so 3 more patients are treated at this dose level. In the third cohort of 3 patients (2nd cohort at dose level 2), 1 patient develops DLT (for a total of 2/6 DLTs at dose level 2). Given the rules of 3+3 design this means that we conclude that dose level 2 exceeds the MTD and that dose level 1 is chosen as the MTD. For the 6 experiments (simulated trials) shown, dose level 1 is selected as the MTD 2

times (33%) while dose level 2 (with target DLT rate of 25%) is selected as the MTD 3 times (50%), dose level 3 is selected 0 times (0%), and no dose is selected 1 time (17%).

6.5.2 BOIN Versus 3+3 Comparison, Example Simulation

For the second experiment in Table 6.2 and Fig. 6.7 illustrates the results graphically for both 3+3 and BOIN. For 3+3 (Fig. 6.7, top), in the first cohort of 3 patients, 0 patients experience DLT and the dose is escalated to level 2. In the second cohort of 3 patients now being treated at dose level 2, 1 patient develops DLT so 3 more patients are treated at this dose level. In the third cohort of 3 patients (2nd cohort at dose level 2), 0 patients develops DLT (for a total of 1/6 DLTs at dose level 2). Given the rules of 3+3 design this means that we escalate to dose level 3. In the fourth cohort of 3 patients, 2 patients develop DLT and we conclude that dose level 3 exceeds the MTD and that dose level 2 is chosen as the MTD. The DLT rate at the selected MTD is $1/6 = 17\%$ with 95% exact confidence interval from 0% to 64%.

For BOIN (Fig. 6.7, bottom), following the decision rules shown in Table 6.1, the path is the same as for 3+3 (for ease of comparison we are using the same random data for the first 4 cohorts). After cohort #4, Table 6.1 indicates we should de-escalate from dose level 3 to dose level 2. For cohort #5 (now the third cohort treated at dose level 2), 0 patients develop DLT and Table 6.1 indicates we should escalate back to dose level 3. For cohort #5, (now the second cohort treated at dose level 3), 2 patients develop DLT and Table 6.1 indicates we should de-escalate back to dose level 2 (and eliminate dose level 3 from future consideration). For cohort #6, (now the fourth cohort treated at dose level 2), 1 patient develops DLT and Table 6.1 indicates we should stay at dose level 2. For cohort #7 (now the fifth cohort treated at dose level 2), 2 patients experience DLT and Table 6.1 indicates we should remain at dose level 2. However, we have treated 15 patients at dose level 2 which is the maximum allowed. Dose level 2 is chosen as the MTD and the corresponding DLT

Table 6.2 Six simulated 3+3 trials with true DLT rates of 10%, 25%, 45%

EXP #	Results	MTD	# Pts
1	0/3, 1/3 +1/3	1	9
2	0/3, 1/3 + 0/3, 2/3	2	12
3	0/3, 3/3	1	6
4	2/3	Exceeded	3
5	1/3 + 0/3, 0/3, 2/3	2	12
6	0/3, 0/3, 1/3 + 2/3	2	12

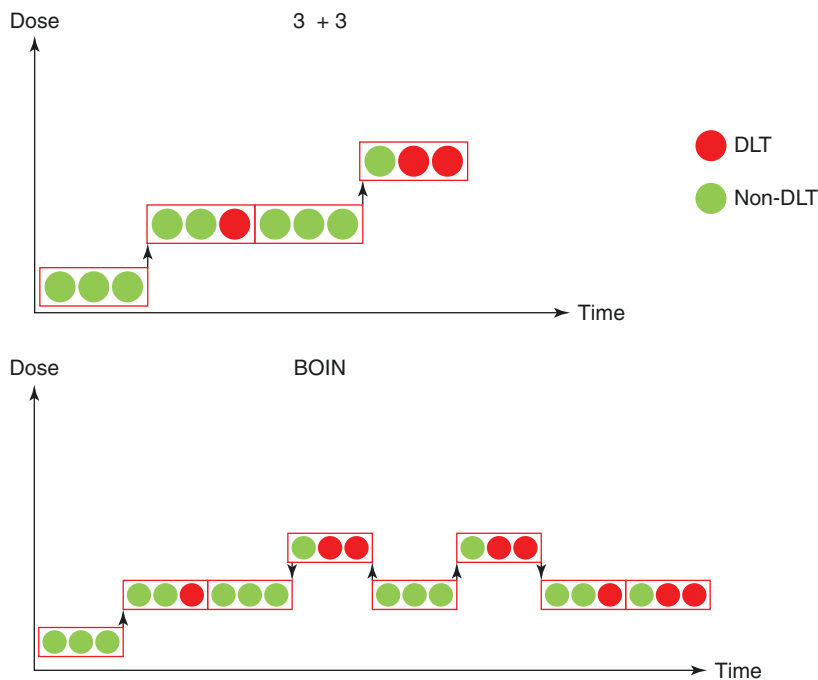


Fig. 6.7 Graphical illustration of phase I trial conduct

rate = $4/15 = 27\%$ with exact 95% confidence interval from 8% to 55%. Thus, the BOIN estimate of the DLT rate at the MTD is both more accurate (i.e., closer to the target of 25%) and more precise (narrower confidence interval) than that from 3+3. Key differences between the 3+3 and BOIN designs are the ability of BOIN to visit a dose level multiple times, the ability of BOIN to treat more than 6 patients at a dose level and the ability of BOIN to repeatedly escalate and de-escalate across dose levels. By doing so, BOIN allows us to collect more information to learn the true DLT rate of a dose and adaptively corrects incorrect decisions possibly made at earlier stages of the trial caused by the random variation of small amounts of data described previously.

6.5.3 BOIN Versus 3+3 Comparison, Simulations

An increasingly common approach to computer simulations for phase I studies is to generate a large number of dose-toxicity scenarios using an algorithm that creates a wide range of scenarios such that the DLT probabilities are a non-decreasing function of dose level [1]. Based on 10,000 generated scenarios with 2000 trials simulated trials each, we can compare the performance of the 3+3 design and the BOIN design (assuming 6 dose levels, a DLT target of 25%, and a maximum sample size

Table 6.3 Comparison of simulation results for 3+3 and BOIN designs

	3+3	BOIN
Probability of correctly selecting MTD	33%	49%
Proportion of patients treated at MTD	26%	31%
Probability of selecting doses with DLT rate $\geq 33\%$	8%	12%
Proportion of patients treated at doses with DLT rate $\geq 33\%$	10%	14%
% of selecting doses with DLT rate $\leq 16\%$	40%	25%
% of patients treated at doses with DLT rate $\leq 16\%$	45%	42%

Note: These metrics are more completely defined in [1]

for 36 patients for BOIN, Table 6.3) using the data from simulations reported in Zhou et al. [1]. The probability of correctly selecting the target dose level (i.e., dose level with DLT rate closest to target) is 33% for 3+3 compared to 49% for BOIN. The proportion of patients treated at the target dose level is 26% for 3+3 and 31% for BOIN. The probability of selecting as the MTD doses with DLT rate $\leq 16\%$ is 40% for 3+3 and 25% for BOIN. Thus, importantly, BOIN is more likely to correctly select the MTD (49% vs 33%) and less likely to select as the MTD doses with low DLT rates (25% vs 40%).

6.6 Expansion Cohorts

Increasingly, phase I studies include expansion cohorts to confirm safety and develop preliminary efficacy data for patients treated at the MTD. Multiple cohorts with different molecular defects or different histologies are sometimes specified. Expansion cohorts typically include 10–30 patients. The number of patients included is usually not given statistical justification but can be specified to achieve a given level of precision in estimation. E.g., to have 95% confidence interval with half-width not more than 0.2 for a targeted proportion of 0.3, requires at least 21 patients while targeting a 0.15 half-width would require at least 36 patients.

Because the toxicity data for patients treated at the MTD at the time the expansion phase begins is generally limited (often based on just 6 patients), it is prudent to monitor toxicity in the expansion phase. For the 3+3 design, we can add Bayesian monitoring rules [5] based on the beta distribution which stop enrollment if the probability of excessive toxicity exceeds some pre-specified threshold. Such rules are superior to deterministic rules such as stopping enrollment if DLT rate in the expansion cohort exceeds some threshold (e.g., 0.30) because their statistical properties can be calibrated by changing the threshold. Also, when excessive toxicity is observed in the expansion phase after 3+3, it is often not clear how to proceed. Since 6 is a very small sample, as more patients are treated at the MTD in an expansion cohort, the additional toxicity data may easily contradict the earlier conclusion that the selected dose is the MTD. For example, what should we do if the first three patients in an expansion cohort all have toxicity? Should we stop the trial according

to Bayesian monitoring rule or de-escalate? If we de-escalate, what sort of rule or algorithm should be applied to choose a dose? On the other hand, if the toxicity rate in the cohort expansion is very low, should we escalate the dose? If so, what sort of rule or algorithm should be applied to choose a dose? The point is that the idea of treating a fixed expansion cohort at a chosen MTD may seem sensible, but in practice can be very problematic. BOIN does not have this issue as it allows dose escalation/de-escalation continuously in light of accruing data. With the BOIN design, the expansion toxicity can be monitored using an expanded decision table like the one used for dose escalation. This allows the dose level declared to be the MTD to be updated based on accruing toxicity data.

6.7 Conclusion/Discussion

The classical 3+3 design is simple in concept and implementation and transparent during trial conduct. However, this design suffers from poor statistical properties. The BOIN design is somewhat complicated in concept due to its statistical underpinnings but is easy to implement and has clearly superior statistical properties compared to the 3+3 design. The BOIN design is also considerably more flexible than the 3+3 design. The target DLT rate for the MTD can be set at any value. Extensions for the BOIN design have been developed for drug combination studies [6] and for phase I studies with prolonged observation periods [7]. Free software can be downloaded to develop BOIN designs and compute operating characteristics. A web-based version, a desktop version, and versions for R and Stata are available for download. The web-based and desktop versions also provide templates of text describing the design giving necessary parameters and tables of decision rules and operating characteristics. These templates can be inserted into appropriate sections in phase I protocols. The web links for the software are provided at the end of the chapter.

Key Expert Opinion Points

1. The 3+3 design in phase I oncology studies should be replaced by newer designs with better statistical properties
2. Phase I designs with better statistical properties have been available for decades but have not been widely used
3. This is partly related to the difficulty in creating and implementing model-based designs
4. Another difficulty is that the comparisons between designs are based on results of computer simulations which many clinical trialists find inaccessible
5. BOIN is easy to implement; very flexible; has superior statistical properties and is widely available as free software in several platforms.

Acknowledgements Authors would like to thank Ying Yuan, PhD for advice; Heng Zhou, PhD for assistance with computer simulations; Rachel Hess for assistance with graphics and editing;

and the clinical investigators in the Phase I program at MDACC for their continued collegial collaboration.

References

1. Zhou H, Yuan Y, Nie L. Accuracy, safety, and reliability of novel phase I trial designs. *Clin Cancer Res.* 2018;24:4357–64.
2. Zhou Y, Lee JJ, Yuan Y. A utility-based Bayesian optimal interval (U-BOIN) phase I/II design to identify the optimal biological dose for targeted and immune therapies. *Statistics in Medicine.* 2019;38(28):5299–316.
3. 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.* 2016;22:4291–301.
4. Liu S, Yuan Y. Bayesian optimal interval designs for phase I clinical trials. *Appl Stat.* 2015;64:507–23.
5. Thall PF, Simon RM, Estey EH. Bayesian sequential monitoring designs for single-arm clinical trials with multiple outcomes. *Stat Med.* 1995;14:357–79.
6. Lin R, Yin G. Bayesian optimal interval designs for dose finding in drug-combination trials. *Stat Methods Med Res.* 2017;26:2155–67.
7. 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;24:4921–30.

Web Links

Web-based version: <http://www.trialdesign.org/one-page-shell.html#BOIN>

Desktop version: https://biostatistics.mdanderson.org/SoftwareDownload/SingleSoftware.aspx?Software_Id=99

R version: <https://cran.r-project.org/web/packages/BOIN/index.html>

Stata version: <https://www.stata-journal.com/article.html?article=st0372>