Clinical Cancer Research

Bayesian Optimal Interval Design: A Simple and Well-Performing Design for Phase I Oncology Trials

Ying Yuan[1], Kenneth R. Hess[1], Susan G. Hilsenbeck[2], and Mark R. Gilbert[3]

Abstract

Despite more than two decades of publications that offer more innovative model-based designs, the classical 3 + 3 design remains the most dominant phase I trial design in practice. In this article, we introduce a new trial design, the Bayesian optimal interval (BOIN) design. The BOIN design is easy to implement in a way similar to the 3 + 3 design, but is more flexible for choosing the target toxicity rate and cohort size and yields a substantially better performance that is comparable with that of more complex model-based designs. The BOIN design contains the 3 + 3 design and the accelerated titration design as special cases, thus linking it to established phase I approaches. A numerical study shows that the BOIN design generally outperforms the 3 + 3 design and the modified toxicity probability interval (mTPI) design. The BOIN design is more likely than the 3 + 3 design to correctly select the MTD and allocate more patients to the MTD. Compared with the mTPI design, the BOIN design has a substantially lower risk of overdosing patients and generally a higher probability of correctly selecting the MTD. User-friendly software is freely available to facilitate the application of the BOIN design. *Clin Cancer Res; 22(17); 4291–301. ©2016 AACR.*

Disclosure of Potential Conflicts of Interest

Y. Yuan is a consultant/advisory board member for Agenus. K.R. Hess is an uncompensated consultant/advisory board member for Angiochem. No potential conflicts of interest were disclosed by the other authors.

Editor's Disclosures

The following editor(s) reported relevant financial relationships: W.E. Barlow—None.

CME Staff Planners' Disclosures

The members of the planning committee have no real or apparent conflicts of interest to disclose.

Learning Objectives

Upon completion of this activity, the participant should have a better understanding of using the Bayesian optimal interval (BOIN) design for phase I clinical trials. BOIN is a novel phase I design that is as simple to implement as the 3 + 3 design, but yields significantly better performance comparable to more complicated model-based designs.

Acknowledgment of Financial or Other Support

This activity does not receive commercial support.

Introduction

Despite more than 20 years of publications with innovative model-based clinical trial designs that offer widely acknowledged improvements in efficiency, such designs are implemented in only a very small fraction of phase I trials. The 3 + 3 design (1–3), although widely criticized for its poor operating characteristics (i.e., poor performance in computer simulations of a wide variety of dose–toxicity scenarios; refs. 4–7), remains the dominant phase I trial design used in practice. As evidence, of 34 phase I trials published in *Clinical Cancer Research* in 2015, 32 used classical 3 + 3 designs or a related design with a minor variation. Most phase I trials conducted with the Children's Oncology Group have used the rolling 6 design (8), which trades a larger cohort and sample size in the face of rapid accrual for a faster completion of the trial, but shares similar operating characteristics with the 3 + 3 design for identifying the MTD.

The major reason for the dominance of the 3 + 3 design is its simplicity and transparency. The decision rule for dose escalation and de-escalation is determined before trial conduct, and

[1]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, Texas. [2]Duncan Cancer Center, Baylor College of Medicine, Houston, Texas. [3]Center for Cancer Research, National Cancer Institute, Bethesda, Maryland.

Note: Supplementary data for this article are available at Clinical Cancer Research Online (http://clincancerres.aacrjournals.org/).

Corresponding Author: Ying Yuan, The University of Texas MD Anderson Cancer Center, Unit 1411, 1400 Pressler Street, Houston, TX 77030. Phone: 713-563-4271; Fax: 712-563-4243; E-mail: yyuan@mdanderson.org

doi: 10.1158/1078-0432.CCR-16-0592

physicians can easily inspect the rules to judge whether they fit with clinical practice. In contrast, the well-performing and innovative model-based designs, for example, the continual reassessment method (CRM; refs. 9–13), are considered by many to be statistically and computationally complex, leading practitioners to perceive dose allocations as coming from a "black box," which has hindered their use in practice. It would be ideal to have a phase I trial design that is as simple as the 3 + 3 design, but yields a performance that is comparable with that of the model-based designs.

The goal of this article is to introduce a novel Bayesian optimal interval (BOIN) design (14) that is simple to implement, similar to the 3 + 3 design, but is much more flexible and possesses superior operating characteristics that are comparable with those of the more complex model-based methods. On the basis of our experience, the underlying idea of the BOIN design has been well received by oncologists and has been used to design a number of phase I trials at The University of Texas MD Anderson Cancer Center and Baylor College of Medicine. The statistical methodology of the BOIN design was provided by Liu and Yuan (14). Here, we focus on delineating the links and differences between the BOIN and the 3 + 3 and related designs from a practical standpoint, paired with comprehensive numerical studies. Our goal is to change the current practice in which the vast majority of phase I trials use the 3 + 3 design, and expedite the adoption of novel clinical trial designs, leading to improved efficacy and ethics of phase I trials.

Improved algorithm-based designs have been proposed to obtain better operating characteristics than the 3 + 3 design. Durham and Flournoy (15) proposed the biased coin design that uses a biased coin to determine dose escalation and de-escalation. Lin and Shih (16) studied statistical properties of general "A+B" designs. Ivanova and colleagues (17) developed and compared several improved up-and-down designs. Ji and colleagues (18) proposed the modified toxicity probability interval (mTPI) design that performs better than the 3 + 3 design.

BOIN Design

The BOIN design takes a very simple form, rendering it easy to implement in practice. The decision of dose escalation and de-escalation involves only a simple comparison of the observed dose-limiting toxicity (DLT) rate at the current dose with a pair of fixed, prespecified dose escalation, and de-escalation boundaries. Specifically, let $\hat{p}$ denote the observed DLT rate at the current dose, defined as

$$\hat{p} = \frac{\text{the number of patients experiencing DLT at the current dose}}{\text{the total number of patients treated at the current dose,}}$$

and λ_e and λ_d denote prespecified dose escalation and de-escalation boundaries. The BOIN design can be described as follows (see also Fig. 1).

1. Treat the first patient or cohort of patients at the lowest dose. (In some trials, another dose, such as the second lowest dose, may be used as the starting dose.)
2. To assign a dose to the next patient or cohort of patients,
   a. if $\hat{p} \le \lambda_e$, escalate the dose;
   b. if $\hat{p} \ge \lambda_d$, de-escalate the dose;
   c. otherwise, retain the current dose.
3. Repeat step 2 until the maximum sample size is reached.

The BOIN design shares the simplicity of the 3 + 3 design, which makes the decision of dose escalation/de-escalation by comparing the observed DLT rate $\hat{p}$ with 0/3, 1/3, 2/3, 0/6, 1/6, and 2/6. The BOIN design makes the decision by comparing $\hat{p}$ with two fixed boundaries, λ_e and λ_d, which is arguably even simpler. The statistical rationale behind the BOIN design and the technical details of determining λ_e and λ_d are outlined in the Supplementary Data. Table 1 provides the values of λ_e and λ_d for commonly used target toxicity rates. For example, given the target DLT rate of 30%, the corresponding escalation boundary $\lambda_e = 0.236$ and the de-escalation boundary $\lambda_d = 0.358$. A BOIN design with cohorts of 3 patients will escalate the dose if 0 of 3 patients has DLT because the observed DLT rate $\hat{p} = 0/3 < 0.236$; de-escalate the dose if 2 of 3 patients have DLTs because the observed toxicity rate $\hat{p} = 2/3 > 0.358$; and retain the current dose if 1 of 3 patients has DLT because $0.236 < 1/3 < 0.358$. This example demonstrates that the 3 + 3 rule is actually nested within the BOIN design when the target DLT rate is 30% and the cohort size is 3. Because the 3 + 3 design requires that the number of patients treated at a dose cannot exceed 6 patients, whereas the BOIN design does not impose that requirement, the dose escalation/deescalation rule for 6 patients may be different between the two designs.

The BOIN design, however, is much more flexible than the 3 + 3 design. It can target any prespecified DLT rate. Such flexibility is of great clinical utility. For instance, for some cancer populations for whom there is no effective treatment, a target DLT rate higher than 30% may be an acceptable trade-off to achieve higher treatment efficacy, while for other cancer populations, a lower target DLT rate, for example, 15% or 20%, may be more appropriate.

In addition, unlike the 3 + 3 design, for which the dose escalation and de-escalation decisions can be made only when we have 3 or 6 evaluable patients, the BOIN design does not require a fixed cohort size and allows for decision making at any time during the trial by comparing the observed DLT rate at the current dose with the escalation and de-escalation boundaries. Decisions regarding dose escalation and de-escalation can be made at any time as long as we can calculate the DLT rate at the current dose. Given the target DLT rate of 30%, the escalation boundary $\lambda_e = 0.236$ and the de-escalation boundary $\lambda_d = 0.358$ are equivalent to the dose escalation and de-escalation rules shown in Table 2. Similar dose escalation and de-escalation rules for the target DLT rates of 15%, 20%, and 25% are provided in Supplementary Table S1 in the Supplementary Data. Such flexibility has important practical utility and implications. First, it allows clinicians to "adaptively" change the cohort size during the course of the trial to achieve certain design goals. For example, to shorten the trial duration and reduce the sample size, clinicians often prefer to use a cohort size of 1 for the initial dose escalation, and then switch to a cohort size of 3 after observing the first DLT, as with the commonly used accelerated titration design (ATD; ref. 19). Such an accelerated titration can be easily and seamlessly performed using the BOIN design by simply switching the cohort size from 1 to 3 when the first DLT is observed. Unlike the ATD, which combines two independent empirical rules, the accelerated titration rule and the 3 + 3 rule, in an *ad hoc* way, the BOIN design achieves the same design goal under a single, coherent framework with assured statistical properties. In addition, the BOIN design includes the rolling 6 design as a special case. By allowing for the accrual of 2 to 6 patients concurrently, the BOIN design can mimic

the rolling 6 design to achieve the same goal of trading a larger cohort and sample size for a faster completion of the trial. Therefore, in a sense, the BOIN design provides a generalization of the $3 + 3$, ATD, and rolling 6 designs.

The BOIN design also offers clinicians the flexibility to handle a "passive" change in the cohort size. Often, the actual number of patients available for decision making deviates from the planned cohort size. In many phase I trials that use the $3 + 3$ design, the actual number of patients treated at a dose often deviates from 3 or 6 for various logistic reasons; for example, some patients are not evaluable or have not received adequate treatment to be eligible (or many eligible patients become available in a short period). When that occurs, the decision of dose assignment is difficult for the next new patient because the $3 + 3$ design does not tell us how to assign the dose if the number of (evaluable) patients is not 3 or 6. In contrast, the BOIN design can easily handle that issue because its decision of dose escalation/de-escalation only involves assessing the observed toxicity rate, which is calculable as long as at least one patient has been treated at the current dose and is evaluable, with escalation and de-escalation boundaries λ_e and λ_d. For example, if only 4 of 6 patients enrolled at a dose level were evaluable and provided toxicity data, assuming the target DLT rate of 30%, the dose would be escalated if 0 of 4 patients has DLT (because the observed toxicity rate < 0.236), and de-escalated if ≥ 2 of 4 patients have DLTs (because the observed toxicity rate > 0.359), or the current dose would be retained if 1 of 4 patients has DLT.

The $3 + 3$ and BOIN designs take different approaches to select the MTD at the end of the trial. The $3 + 3$ design directly chooses the MTD as the dose that is one level below the dose that yields 2 or more DLTs, ignoring the data observed at other doses, whereas the BOIN design uses a statistical technique called isotonic regression to pool information across doses to obtain a more efficient statistical estimate of the MTD. The BOIN design offers some desirable statistical properties that the $3 + 3$ design lacks, such as coherence and consistency (see Supplementary Data for details).

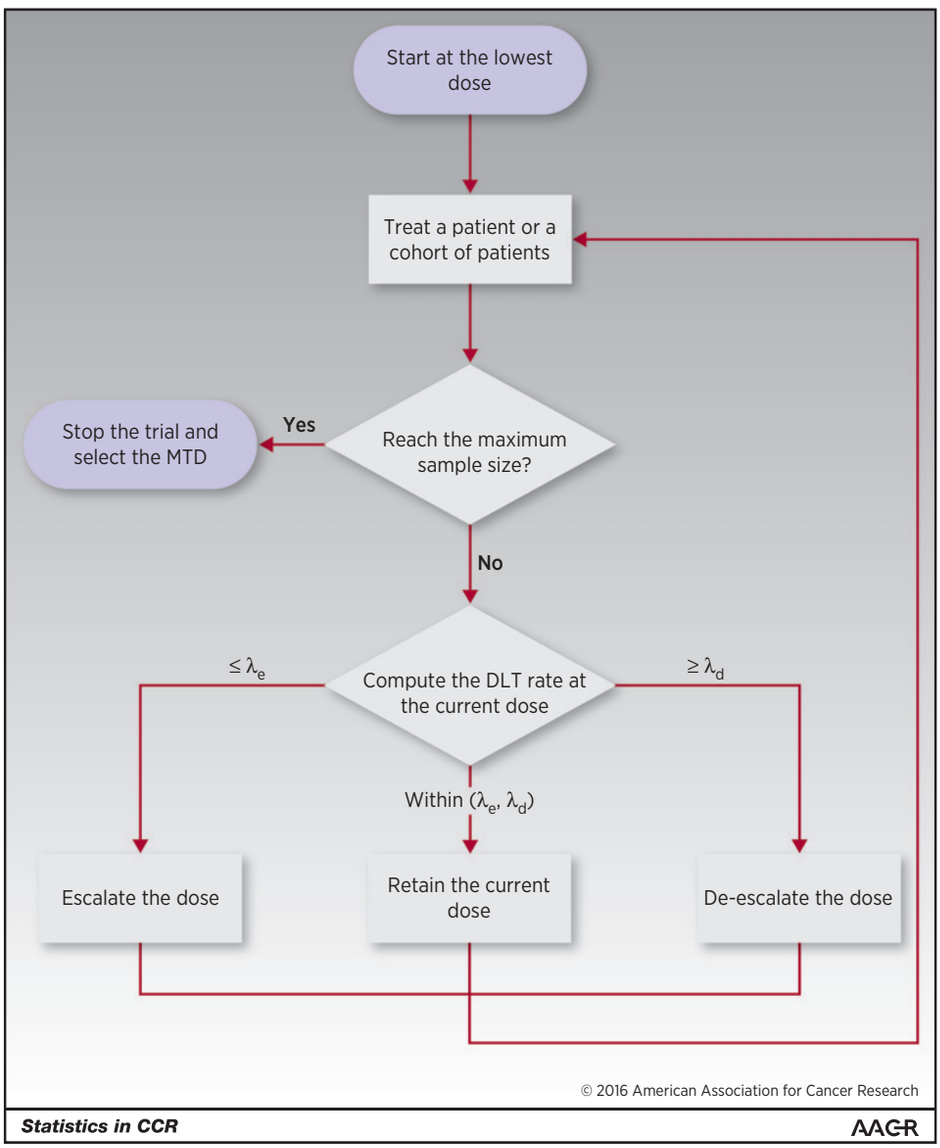


Figure 1.
Flowchart of the BOIN design.

Table 1. Dose escalation and de-escalation boundaries

	Target toxicity rate for the MTD						
Boundary	**0.1**	**0.15**	**0.2**	**0.25**	**0.3**	**0.35**	**0.4**
$\lambda_{e\ (escalation)}$	0.078	0.118	0.157	0.197	0.236	0.276	0.316
$\lambda_{d\ (de\text{-}escalation)}$	0.119	0.179	0.238	0.298	0.358	0.419	0.479

Another feature of the BOIN design is that the sample size is prespecified, which allows researchers to calibrate and choose appropriate sample sizes to achieve the desirable probability of correctly estimating and selecting the MTD. In contrast, with the 3 + 3 design, the sample size actually used in a clinical trial is random because the trial stops whenever 2 or more DLTs are observed at a dose. Because of such a stopping rule, the sample size of a 3 + 3 design tends to be excessively small. One might regard that as an advantage. However, it is actually one of the major drawbacks of the 3 + 3 design. The excessively small and random sample size means that the 3 + 3 design has a low chance of correctly identifying the MTD (see "Numerical Study" below), and precludes the possibility of calibrating the sample size to obtain good operating characteristics. Under the 3 + 3 design, the number of patients treated at any dose cannot be more than 6, which provides too little information to reliably estimate the true toxicity rate. For example, if 1 out of 6 patients experiences DLT, the estimate of the toxicity rate, 1/6 = 16.7%, seems low, but the 95% exact confidence interval (CI) for that estimate is (0.004–0.641), indicating that the true toxicity rate can be as high as 64.1%. Conversely, if 3 out of 6 patients experience DLTs, the estimate of the toxicity rate, 3/6 = 50%, seems very high, but the 95% CI for that estimate is (0.118–0.88); and the true toxicity rate can be as low as 11.8%. In practice, this deficiency is often remedied by expanding the cohort at the "MTD" selected by the 3 + 3 trial. Thus, the final sample size of a realized 3 + 3 trial and a BOIN trial without an expansion cohort might be similar. However, the difference is that under the approach of the 3 + 3 design plus cohort expansion, we lose the flexibility to continuously update our best estimate of the MTD based on the data accumulating during cohort expansion. Were the cohort expansion data to indicate that the MTD selected from the 3 + 3 trial was overdosing or underdosing patients, we would have to manually modify the dose decision in an *ad hoc* way. In contrast, the BOIN design does not require *post hoc* cohort expansion, and the dose escalation/de-escalation explicitly continues by treating each patient at a dose near the evolving estimate of the MTD.

An Example Trial

To illustrate the application of the BOIN design, we construct a hypothetical phase I trial that aims to find the MTD with a target DLT rate of 30%, 5 prespecified doses, and 30 patients. Figure 2 shows the process of the trial conduct. To accelerate dose escalation, the trial starts with a cohort size of 1, and then expands to a cohort size of 3 after the first DLT is observed, as in the ATD design. The trial starts with the first patient receiving dose level 1 without experiencing DLT. On the basis of the dose escalation rules given in Table 2, the dose is then escalated to level 2 for the second patient, who also does not experience DLT. The dose escalation continues until the third patient experiences DLT at dose level 3, at which time the cohort is expanded to 3 by adding 2 more patients (i.e., patients 4 and 5) at dose level 3. Patients 4 and 5 do not experience DLTs. Retaining that dose, patients 6–8 are treated at dose level 3. Patients 6 and 7 do not experience DLTs and patient 8 is not evaluable. At that point, 5 evaluable patients have been treated at dose level 3 and one has experienced DLT. If the 3 + 3 design were used, it would be difficult to make the decision of dose assignment for the subsequent cohort because the number of patients at the current dose is not 3 or 6. In contrast, according to Table 2, the BOIN design escalates the dose to level 4 to treat patients 9–11, among whom patients 10 and 11 experience DLTs. If the 3 + 3 design were used, the trial would stop because 2 of the 3 most recently enrolled patients experience toxicity, precluding the chance to further learn the toxicity profile of the doses and "claim" dose level 3 as the MTD. In contrast, the BOIN design allows us to continue to learn the toxicity of the doses by de-escalating the dose to level 3 to treat patients 12–14, none of whom experiences DLT.

Then, at dose level 3, among a total of 9 treated patients, 8 are evaluable and only one patient has experienced DLT. According to the rules in Table 2, the dose is then escalated to level 4 to treat patients 15 to 17, none of whom experiences DLT. Of the 6 patients treated at dose level 4, only 2 of them have experienced DLTs. Thus, that dose is retained and patients 18 to 20 are treated at dose level 4. Similarly, based on the dose escalation/de-escalation rule of the BOIN design, patients 21–30 are all treated at dose level 4. Although patients 19 and 23 are not evaluable and the last patient (patient 30) does not form a complete cohort (of 3 patients), there is no issue under the BOIN design because it allows for decision making with any number of patients. At the end of the trial, a total of 17 evaluable patients have been treated at dose level 4, and 5 patients have experienced DLTs. Thus, dose level 4 is chosen as the MTD, with the estimated DLT rate = 29.4% and the 95% CI, 0.10–0.56. In contrast, using the 3 + 3 design, dose level 3 would have been chosen as the MTD, with an estimated DLT rate of 20% and a much wider 95% CI, 0.005–0.72.

Numerical Study

Simulation setting

We used computer simulations to evaluate the operating characteristics of the BOIN design. We considered a dose-finding trial with 5 dose levels and a maximum sample size of 30 patients (i.e.,

Table 2. Dose escalation and de-escalation boundaries for target toxicity rate = 30%

	The number of patients treated at the current dose																	
Action	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
Escalate if no. of DLTs ≤	0	0	0	0	1	1	1	1	2	2	2	2	3	3	3	3	4	4
De-escalate if no. of DLTs ≥	1	1	2	2	2	3	3	3	4	4	4	5	5	6	6	6	7	7

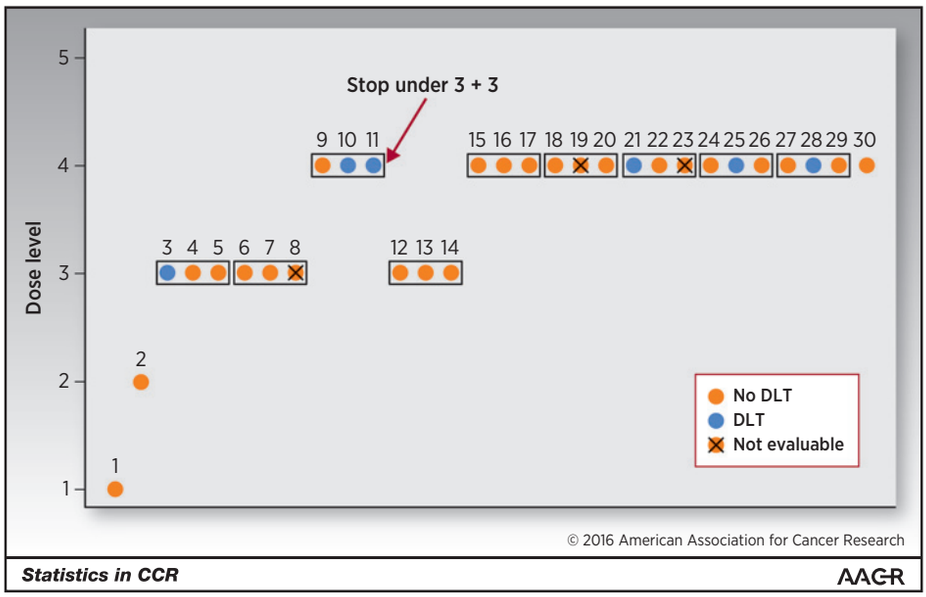


Figure 2.
A hypothetical phase I clinical trial using the BOIN design. The numbers indicate the patient's identification. Three patients in each box form a cohort.

the maximum sample size of the 3 + 3 design). We investigated four commonly used target DLT rates: 15%, 20%, 25%, and 30%. For each of the target DLT rates, we examined 16 representative toxicity scenarios (i.e., the true toxicity rates of the five investigational doses), which varied in the location of the MTD and the gaps around the MTD. Under the standard assumption that toxicity monotonically increases with the dose, the gap (i.e., difference) between the MTD and its two adjacent doses controls the difficulty of dose finding because these adjacent doses are the most difficult ones to distinguish from the MTD. Table 3 shows the 16 true toxicity scenarios with target DLT rates of 20% and 25%. The scenarios for target DLT rates of 15% and 30% are given in Supplementary Table S2 in the Supplementary Data. Similar toxicity scenarios have been used to compare different phase I trial designs (20). Under each scenario, we simulated 10,000 trials to compare the BOIN design with the 3 + 3 and mTPI designs. Because the 3 + 3 design often stopped the trial early (e.g., when 2 out of 3 patients experienced DLTs) before reaching 30 patients, in these cases, the remaining patients were treated at the selected "MTD" as the cohort expansion, such that the three designs had comparable sample sizes. An alternative approach to match the average sample size of three designs is to use the average sample size of the 3 + 3 design as the sample size for the mTPI and BOIN designs. However, as explained in the Supplementary Data, that approach yields severely biased results. There are many variations of the 3 + 3 design. The 3 + 3 design that we used for the comparison is described in the Supplementary Data. We implemented the BOIN design using the R package "BOIN" with its default design parameters (21), the mTPI design using the Web application with the interval width epsilon1 = epsilon2 = 0.03 (22). The mTPI and BOIN designs were implemented in a more efficient, fully sequential way (i.e., patients were treated one by one) because that is one important advantage of these two designs.

Performance metrics

We considered four metrics to measure the performance of the designs:

Table 3. Sixteen true toxicity scenarios with the target DLT rates of 0.2 and 0.25

	Dose level						Dose level				
Scenario	**1**	**2**	**3**	**4**	**5**	**Scenario**	**1**	**2**	**3**	**4**	**5**
1	**0.20**[a]	0.25	0.35	0.45	0.50	1	**0.25**[a]	0.35	0.45	0.60	0.70
2	**0.20**	0.30	0.40	0.50	0.60	2	**0.25**	0.32	0.40	0.50	0.60
3	0.15	**0.20**	0.25	0.35	0.45	3	0.20	**0.25**	0.35	0.45	0.60
4	0.15	**0.20**	0.30	0.45	0.55	4	0.18	**0.25**	0.32	0.40	0.50
5	0.10	**0.20**	0.25	0.35	0.45	5	0.15	**0.25**	0.35	0.50	0.65
6	0.10	**0.20**	0.30	0.40	0.55	6	0.13	**0.25**	0.32	0.40	0.50
7	0.08	0.15	**0.20**	0.25	0.35	7	0.10	0.15	**0.25**	0.30	0.40
8	0.08	0.15	**0.20**	0.30	0.45	8	0.10	0.18	**0.25**	0.32	0.40
9	0.05	0.10	**0.20**	0.25	0.40	9	0.10	0.15	**0.25**	0.35	0.50
10	0.05	0.10	**0.20**	0.30	0.45	10	0.06	0.13	**0.25**	0.32	0.40
11	0.05	0.10	0.15	**0.20**	0.25	11	0.02	0.10	0.20	**0.25**	0.30
12	0.05	0.10	0.15	**0.20**	0.30	12	0.02	0.10	0.18	**0.25**	0.32
13	0.02	0.06	0.10	**0.20**	0.25	13	0.02	0.10	0.20	**0.25**	0.35
14	0.02	0.06	0.10	**0.20**	0.30	14	0.01	0.07	0.13	**0.25**	0.32
15	0.02	0.05	0.07	0.10	**0.20**	15	0.01	0.05	0.10	0.15	**0.25**
16	0.01	0.06	0.10	0.15	**0.20**	16	0.01	0.06	0.12	0.18	**0.25**

[a]Bold font indicates the MTD.

i The percentage of correct selection (PCS) of the true MTD in 10,000 simulated trials.

ii The average number of patients allocated to the MTD across 10,000 simulated trials.

iii The risk of overdosing, which is defined as the percentage of simulated trials in which a large percentage (e.g., more than 60% or 80%) of patients are treated at doses above the MTD, that is, how likely the design treats more than 60% or 80% of the patients at doses above the MTD. This risk measure is practically more relevant and useful than the average number of patients treated above the MTD across 10,000 simulated trials because in practice, the trial is typically conducted only once. What concerns the investigator is how likely the current trial overdoses a large percentage of patients, not if the trial was repeated thousands of times, on average how many patients would be overdosed.

iv The risk of underdosing, which is defined as the percentage of simulated trials in which more than 80% of patients are treated at doses below the MTD (i.e., potentially subtherapeutic doses). We chose a higher cut-off value of 80% to define underdosing because in practice, underdosing tends to be of less concern than overdosing.

Results

The percentage of correct selection of the MTD

As shown in Fig. 3, the BOIN design outperforms the 3 + 3 design with a substantially higher percentage of correct selection (PCS) of the MTD. For example, when the target DLT rate is 25%, the PCS of the BOIN design is mostly 12% to 16% higher than that of the 3 + 3 design. In particular, when the MTD is the highest dose (i.e., scenarios 15 and 16), the PCS of the BOIN design almost triples that of the 3 + 3 design. The BOIN design also outperforms the mTPI design, especially when the target DLT rate is low, such as 15% or 20%. In these cases, the PCS of the BOIN design is often about 6% to 10% higher than that of the mTPI design.

Figure 3.
The PCS of the MTD under the 3 + 3, mTPI, and BOIN designs when the target toxicity rates are 15%, 20%, 25%, and 30%. A higher value is better.

Figure 4.
Average number of patients allocated to the MTD under the 3 + 3, mTPI, and BOIN designs when the target toxicity rates are 15%, 20%, 25%, and 30%. A higher value is better.

Average number of patients allocated to the MTD

The performance of the 3 + 3 design depends on the location of the MTD and the target DLT rate (see Fig. 4). When the MTD is located at low doses (e.g., doses 1 and 2, corresponding to scenarios 1–6), the 3 + 3 design performs reasonably well. However, when the MTD is located at high doses (doses 4 and 5, corresponding to scenarios 11–16) or the target DLT rate is 30%, the 3 + 3 design performs substantially worse than the mTPI and BOIN designs. The BOIN design generally outperforms the mTPI design, assigning more patients to the MTD when the target DLT rate is 15% or 20%. The two designs are comparable when the target DLT rate is 25% or 30%.

Risk of overdosing

Among the three designs, the mTPI design has the highest risk of overdosing (i.e., assigning more than 60% or 80% of the patients to doses above the MTD), especially when the target DLT rates are 20%, 25%, and 30% (see Figs. 5 and 6). For example, when the target DLT rate is 25%, the mTPI design often has more than 40% chance of assigning more than 60% of patients to overly toxic doses, and more than 30% chance of assigning more than 80% of patients to overly toxic doses. In the Discussion section, we provide a theoretical explanation why the mTPI design tends to have such an alarmingly high risk of overdosing patients. The 3 + 3 design generally has the lowest risk of overdosing when the target DLT rates are 25% and 30%. This is consistent with previous research that found the 3 + 3 design to be overly conservative (4–7). Although being safe is desirable, being overly conservative is undesirable and results in poor precision for identifying the MTD. Because the dose selected in phase I is used in subsequent phase II trials to treat a much larger number of patients, misidentification of the MTD has the serious consequence of potentially treating a large number of patients at overly toxic or subtherapeutic doses. The BOIN design strikes a good balance in safety (i.e., the risk of overdosing) and identifying the MTD. Compared with the 3 + 3 design, the BOIN design has much higher PCS of the MTD

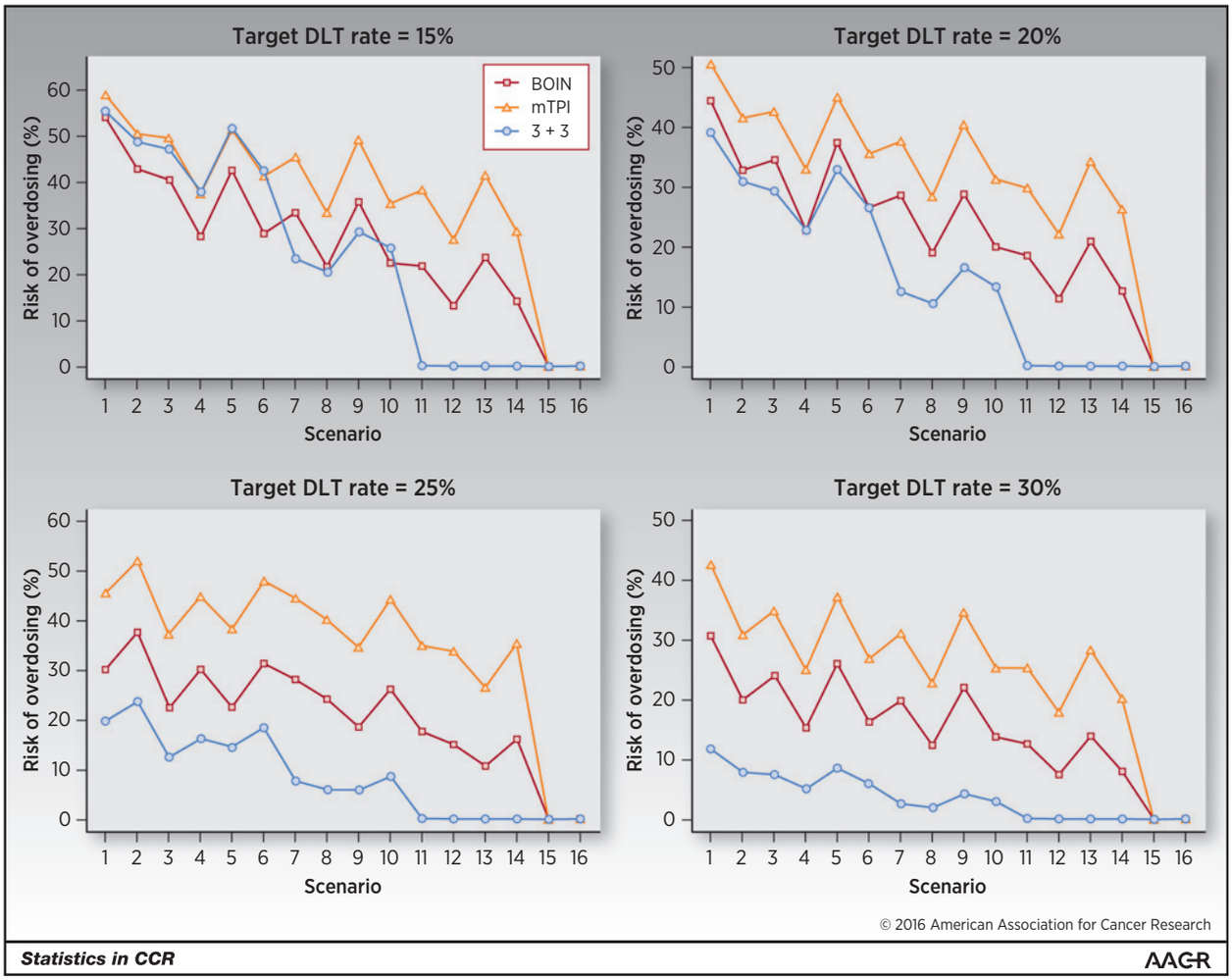


Figure 5.
Risk of overdosing 60% or more of patients under the 3 + 3, mTPI, and BOIN designs when the target toxicity rates are 15%, 20%, 25%, and 30%. A lower value is better.

(see Fig. 3). Compared with the mTPI, the BOIN design has a substantially lower risk of overdosing in almost all scenarios. Specifically, the risk of overdosing 80% or more of patients under the BOIN design is less than half of that under the mTPI design in most scenarios (see Fig. 6).

Risk of underdosing

As the 3 + 3 design is conservative, it is not surprising that it generally has a higher risk of underdosing (i.e., treating more than 80% of patients at doses below the MTD), especially when the target DLT rate is 25% or 30% (see Fig. 7). The mTPI performs well when the target DLT rate is 25% or 30%, but has the highest risk of underdosing when the target DLT rate is 15%. In most scenarios, the BOIN design has the lowest or close to the lowest risk of underdosing.

Software for Practical Implementation

To facilitate the use of the BOIN design, we developed two freely available programs: an R package "BOIN" and a standalone desktop application. The desktop application has an intuitive graphical user interface and is convenient to use for most phase I trials. The R package provides extra flexibility that allows users to modify the code, if needed, to add additional features that have not been included in the package. The R package "BOIN" is available from CRAN (21), and the desktop program is available at the MD Anderson Software Download website (23). A statistical tutorial and protocol template for using the BOIN design are provided at the first author's website (24).

Discussion

This article introduces the BOIN design and compares it with the 3 + 3 and mTPI designs. The BOIN design is built upon rigorous statistical principles and treats each patient at dose levels near the evolving estimate of the MTD. This design is easy to implement in a manner similar to the 3 + 3 design, but provides much more flexibility in choosing the target toxicity rate and cohort size, and yields a substantially better performance. A numerical study showed that the BOIN design is more likely to

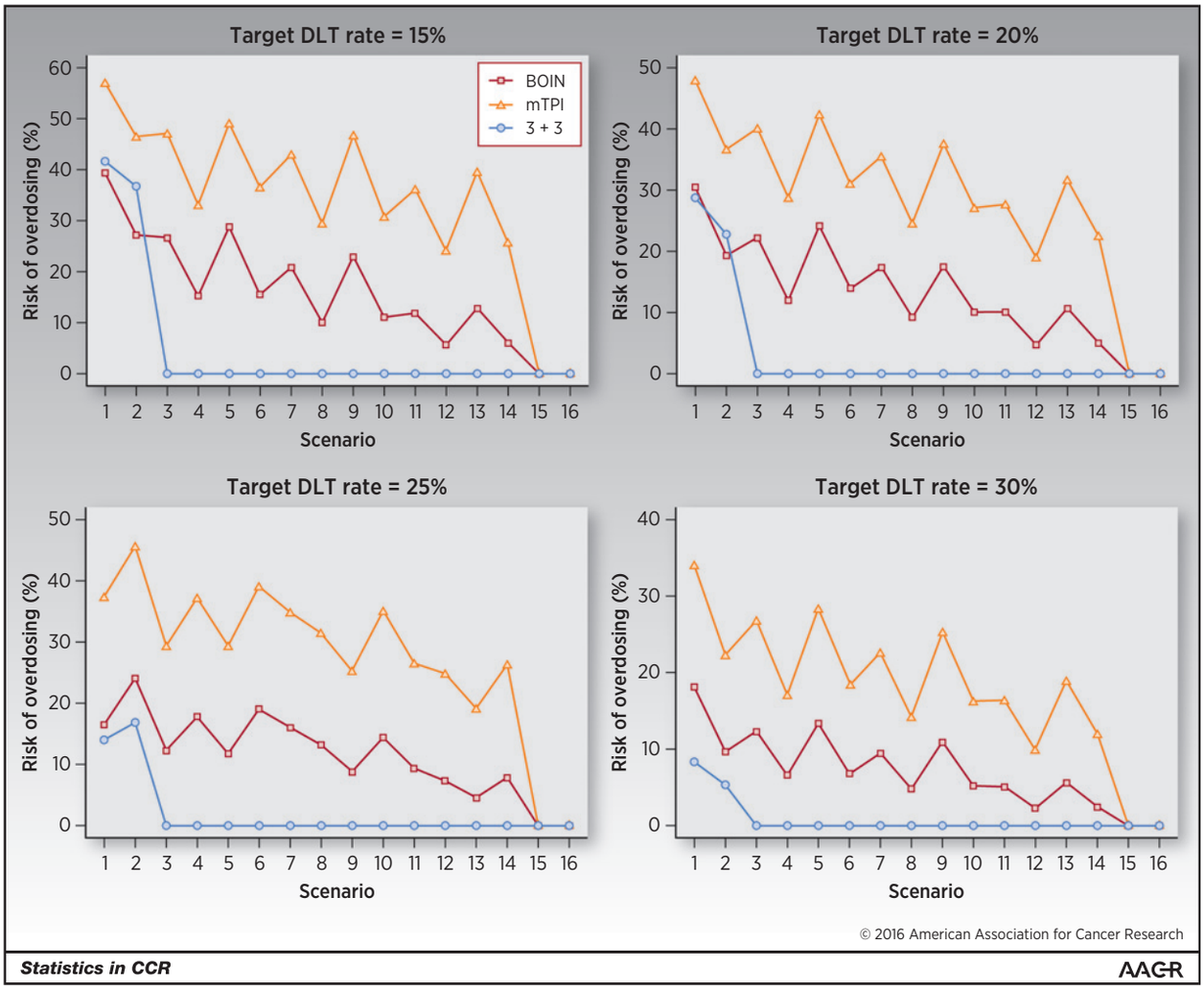


Figure 6.
Risk of overdosing 80% or more of patients under the 3 + 3, mTPI, and BOIN designs when the target toxicity rates are 15%, 20%, 25%, and 30%. A lower value is better.

correctly choose the MTD and allocate more patients to the MTD than the 3 + 3 design, and has substantially lower risk of overdosing patients than the mTPI design.

The reason that the mTPI design has an excessively high risk of overdosing patients lies in the core of that method, that is, using the unit probability mass (UPM) as the criterion to determine dose escalation. Specifically, the mTPI defines three dosing intervals (i.e., the underdosing interval, proper dosing interval, and overdosing interval). Given a dosing interval and the observed toxicity data, the UPM is defined as the posterior probability of the interval divided by the length of the interval. The mTPI makes the decision of dose escalation and de-escalation based on which interval has the largest UPM. If the underdosing (or overdosing or proper dosing) interval has the largest UPM, the dose is escalated (or de-escalated or stays at the same level). Unfortunately, the UPM is not an appropriate indication of the toxicity of a dose, and leads to problematic decisions. To visualize this problem, consider a trial for which the target toxicity rate is 0.2, and the underdosing, proper dosing, and overdosing intervals are (0–0.17), (0.17–0.23), and (0.23–1), respectively. Suppose that at a certain stage of the trial, the observed data indicate that the posterior probabilities of the underdosing interval, proper dosing interval, and overdosing interval are 0.01, 0.09, and 0.9, respectively. In other words, the data indicate that there is a 90% chance that the current dose is overdosing and only a 9% chance that the current dose provides proper dosing. Despite such dominant evidence of overdosing, the mTPI dictates that the design stays at the same dose for treating the next new patient because the UPM for the proper dosing interval is the largest. Specifically, the UPM for the proper dosing interval is $0.09/(0.23–0.17) = 1.5$, and the UPM for the overdosing interval is $0.9/(1–0.23) = 1.17$. This example demonstrates that the UPM is not an appropriate indication of the toxicity of a dose, and as a result, the mTPI tends to keep treating patients at a toxic dose even when there is strong evidence for that dose being overly toxic. Our results seem to contradict those of the previous simulation study by Ji and Wang (25), which claimed that the mTPI is safer than the 3 + 3 design. As detailed in the Supplementary Data (see Supplementary Fig. S1

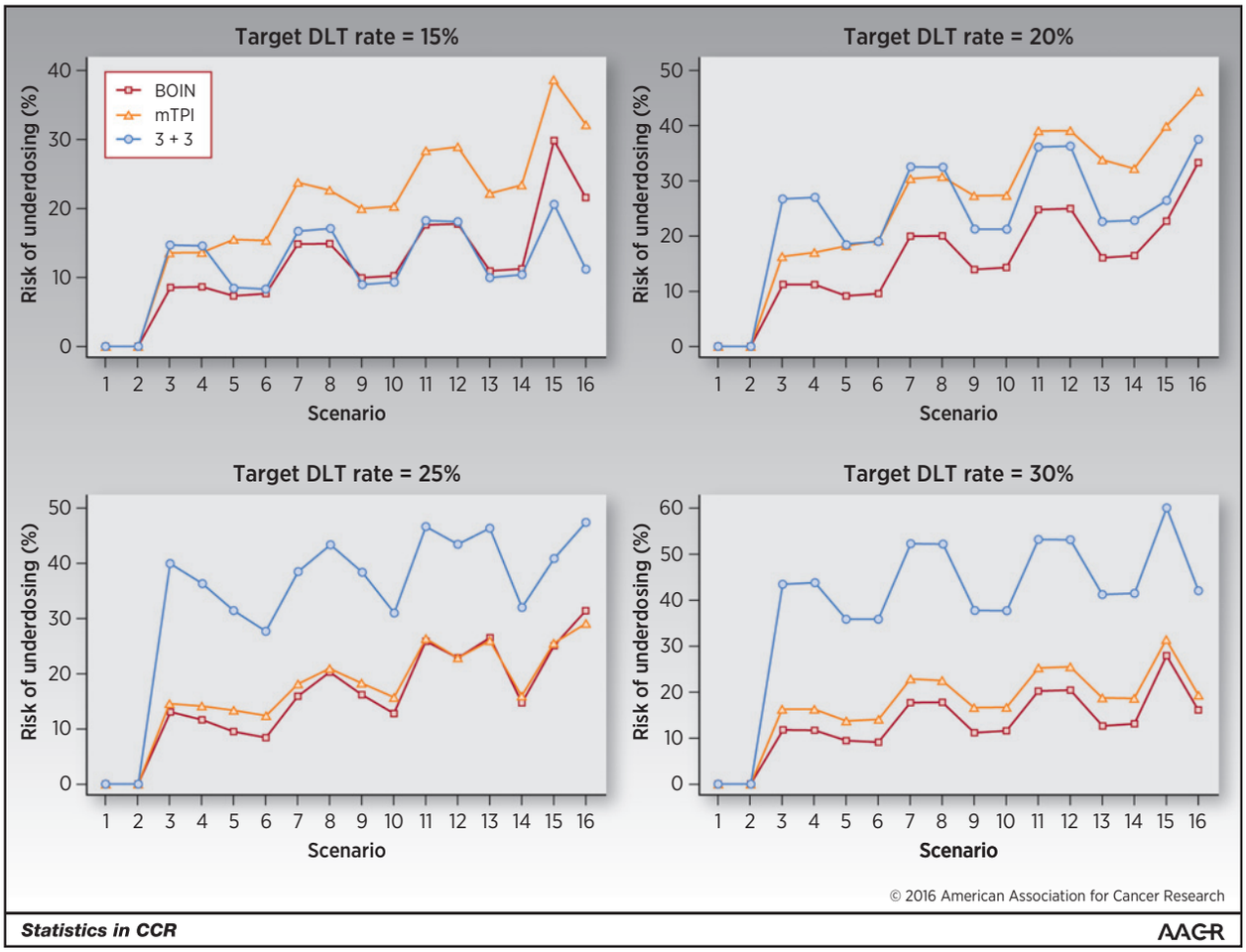


Figure 7.
Risk of underdosing (i.e., assigning more than 80% of patients to doses below the MTD) under the 3 + 3, mTPI, and BOIN designs when the target toxicity rates are 15%, 20%, 25%, and 30%. A lower value is better.

and Supplementary Table S4), the simulation in that study is biased because of the inappropriate way the sample sizes were matched between the designs.

Recently, the BOIN design has been extended to find the MTD for drug–combination trials (26), which may further improve the utility of the BOIN design in practice. Under the BOIN design, many practical considerations are either automatically or easily accommodated. For example, the 3 + 3 design often includes one or more expansion cohorts with no way to monitor toxicity; whereas the BOIN design naturally accommodates ongoing monitoring by continuously treating patients under its dose escalation and de-escalation rules. In addition, the BOIN design allows for starting the trial from any prespecified dose level, and stopping the trial when a dose accumulates a certain number of patients.

The dose escalation and de-escalation boundaries provided in Table 1 are approximately symmetric around the target DLT rate. In some applications, we may prefer a tighter (i.e., lower) de-escalation boundary to impose a higher safety requirement. This can be done by reducing the value of the highest acceptable DLT rate in the BOIN software, which results in a tighter de-escalation boundary. Supplementary Table S3 in the Supplementary Data provides such an example. Using a tighter de-escalation boundary decreases the risk of overdosing, but the tradeoff is that it may reduce the PCS and the number of patients allocated to the MTD. This is because to correctly identify the MTD, it is necessary to experiment at the doses both below and above the MTD. In general, a conservative design tends to yield lower precision to identify the MTD, as exemplified by the 3 + 3 design. Given the fact that the BOIN has a substantially lower risk of overdosing patients than the mTPI, overdosing may not be a particular concern for the BOIN. If the investigator prefers a lower risk of overdosing, we recommend the boundaries in Supplementary Table S3, which generally yield good operating characteristics.

We compared the BOIN, 3 + 3, and mTPI designs because they share similar simplicity and therefore are more likely to be implemented in practice. We did not include the CRM in our comparison because that design is more complicated to implement in practice. In addition to a lack of transparency, the choice of the model and prior in the CRM can be difficult for physicians, and an inappropriate choice can affect the performance of the

design. However, the comparison of BOIN and CRM, which has been investigated elsewhere (14), suggested that these two designs have comparable performance.

As with most existing phase I trial designs, a limitation of the BOIN design is that it requires toxicity to be quickly ascertained with respect to the accrual time. That is, it requires that when the next new cohort of patients is enrolled and ready for dose assignment, the toxicity outcomes of the patients who have been treated in the trial have been fully evaluated. To handle delayed toxicities, some innovative designs have been proposed (27, 28); the extension of the BOIN design to delayed toxicities is a topic for future research.

Authors' Contributions

Conception and design: Y. Yuan, K.R. Hess, M.R. Gilbert
Development of methodology: Y. Yuan
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): Y. Yuan, K.R. Hess, S.G. Hilsenbeck, M.R. Gilbert
Writing, review, and/or revision of the manuscript: Y. Yuan, K.R. Hess, S.G. Hilsenbeck, M.R. Gilbert
Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): Y. Yuan

Acknowledgments

The authors thank Heng Zhou for carrying out the simulation study and preparing the figures.

Grant Support

Y. Yuan was supported in part by the NIH under award numbers P50CA098258 and R01CA154591.

Received March 10, 2016; revised May 11, 2016; accepted June 2, 2016; published OnlineFirst July 12, 2016.

References

1. Storer BE. An evaluation of phase I clinical trials designs in the continuous dose–response setting. Stat Med 2001;20:2399–408.
2. Storer BE. Design and analysis of phase I clinical trials. Biometrics 1989; 45:925–37.
3. Rogatko A, Schoeneck D, Jonas W, Tighiouart M, Khuri FR, Porter A. Translation of innovative designs into phase I trials. J Clin Oncol 2007;25:4982–6.
4. Korn EL, Midthune D, Chen TT, Rubinstein LV, Christian MC, Simon RM. A comparison of two phase I trial designs. Stat Med 1994;13:1799–806.
5. Ahn C. An evaluation of phase I cancer clinical trial designs. Stat Med 1998;17:1537–49.
6. Iasonos A, Wilton AS, Riedel ER, Seshan VE, Spriggs DR. A comprehensive comparison of the continual reassessment method to the standard 3 + 3 dose escalation scheme in phase I dose-finding studies. Clin Trials 2008; 5:465–77.
7. Le Tourneau C, Lee JJ, Siu LL. Dose escalation methods in phase I cancer clinical trials. J Natl Cancer Inst 2009;101:708–20.
8. Skolnik JM, Barrett JS, Jayaraman B, Patel D, Adamson PC. Shortening the timeline of pediatric phase I trials: the rolling six design. J Clin Oncol 2008;26:190–5.
9. O'Quigley J, Pepe M, Fisher L. Continual reassessment method: a practical design for phase I clinical trials in cancer. Biometrics 1990;46:33–48.
10. Cheung YK. Dose finding by the Continual Reassessment Method. Boca Raton, FL: CRC Press; 2011.
11. Goodman SN, Zahurak ML, Piantadosi S. Some practical improvements in the continual reassessment method for phase I studies. Stat Med 1995; 14:1149–61.
12. Iasonos A, O'Quigley J. Continual reassessment and related designs in dose finding studies. Stat Med 2011;30:2057–61.
13. Iasonos A, O'Quigley J. Adaptive dose-finding studies: a review of model-guided phase I clinical trials. J Clin Oncol 2014;32:2505–11.
14. Liu S, Yuan Y. Bayesian optimal interval designs for phase I clinical trials. J R Stat Soc Ser C Appl Stat 2015;64:507–23.
15. Durham SD, Flournoy N. Random walks for quantile estimation. In: Gupta SS, Berger JO, editors. Statistical decision theory and related topics V. New York: Springer-Verlag; 1994. p.467–76.
16. Lin Y, Shih WJ. Statistical properties of the traditional algorithm-based designs for phase I cancer clinical trials. Biostatistics 2001;2: 203–15.
17. Ivanova A, Montazer-Haghighi A, Mohanty SG, Durham SD. Improved up-and-down designs for phase I trials. Stat Med 2003;22: 69–82.
18. Ji Y, Liu P, Li Y, Bekele N. A modified toxicity probability interval method for dose-finding trials. Clin Trials 2010;7:653–63.
19. Simon R, Freidlin B, Rubinstein L, Arbuck SG, Collins J, Christian MC. Accelerated titration designs for phase I clinical trials in oncology. J Natl Cancer Inst 1997;89:1138–47.
20. Liu S, Cai C, Ning J. Up-and-down designs for phase I clinical trials. Contemp Clin Trials 2013;36:218–27.
21. R package BOIN [software on the Internet][cited 2016 May 11]. Available from: https://cran.r-project.org/web/packages/BOIN/index.html.
22. mTPI web application [software on the Internet][cited 2016 May 11]. Available from: http://www.compgenome.org/NGDF/.
23. Bayesian Optimal Interval (BOIN) design desktop program [software on the Internet][cited 2016 May 29]. Available from: https://biostatistics.mdanderson.org/softwaredownload/SingleSoftware.aspx?Software_Id=99.
24. Simulation Codes for My Research Projects [about 5 screens]. [cited 2016 May 29]. Available from: http://odin.mdacc.tmc.edu/~yyuan/index_code.html.
25. Ji Y, Wang S. Modified toxicity probability interval design: a safer and more reliable method than the 3 + 3 design for practical phase I trials. J Clin Oncol 2013;31:1785–91.
26. Lin R, Yin G. Bayesian optimal interval designs for dose finding in Drug-combination trials. Stat Methods Med Res. 2015 Jul 15. [Epub ahead of print].
27. Cheung YK, Chappell R. Sequential designs for phase I clinical trials with late onset toxicities. Biometrics 2000;56:1177–82.
28. Liu S, Yin G, Yuan Y. Bayesian data augmentation dose finding with continual reassessment method and delayed toxicity. Ann Appl Stat 2013; 4:2138–56.

Supplementary Data

Rationale of the BOIN design

In order to understand the rationale behind the BOIN design, we first examine how a phase I cancer trial is conducted in practice. Typically, the trial starts by treating the first cohort of patients at the lowest (or a prespecified intermediate) dose. Based on the toxicity data collected from the first cohort, the most appropriate dose is selected for the second cohort by escalating, de-escalating or retaining the current dose. After we observe the toxicity outcome of the second cohort, the most appropriate dose for the third cohort is selected, based on the cumulative toxicity data from the first two cohorts, and so on until the trial reaches the prespecified maximum sample size. Therefore, the phase I trial is essentially a sequence of decision-making steps of dose assignment for patients who are sequentially enrolled into the trial.

Let ϕ represent the prespecified target toxicity level. If the true toxicity rate of the current dose, say p, was known at each stage of decision making, then it would be straightforward to make the dose assignment. If $p>\phi$, which means that the current dose is above the MTD (i.e., overdosing), the dose should be de-escalated to avoid exposing the next patient to an overly toxic dose; if $p<\phi$, which means that the current dose is below the MTD (i.e., underdosing), the dose should be escalated to avoid treating the next patient at a subtherapeutic dose level; and if $p=\phi$, indicating that the current dose is the MTD, the current dose should be retained to treat the next patient. We refer to such a design as an "oracle" design because (1) it always make correct decisions of dose escalation and de-escalation and thus leads to optimal ethical patient treatment, and (2) it

does not exist in practice because in reality the true toxicity rate of the current dose is never known; otherwise there would be no need to conduct the phase I trial.

In real-world trials, we have to rely on the observed data to make the decision of dose assignment. For example, given the target toxicity rate of ϕ=0.3, if 1 patient out of 5 experiences dose-limiting toxicity (DLT), we might choose to escalate the dose because the observed toxicity rate is only 20%. Because of the randomness of the data observed in the small sample sizes of phase I trials, the decisions regarding dose assignment are often incorrect, leading to erroneous and overly aggressive dose escalation or de-escalation and treating an excessive number of patients at dose levels above or below the MTD. For example, if the true toxicity rate of a dose is 0.4, there is more than 40% chance to see 1 or fewer DLTs among 5 patients (i.e., the actual observed toxicity rate ≤ 0.2), making the dose appear much safer than it actually is. This issue is inherent in small samples and cannot be completely removed. In practice, however, statistical tools can be used to account for such uncertainty and minimize the decision error of dose assignment such that the design approximates the "oracle" design as closely as possible. This is the motivation behind the BOIN design: to optimize patient ethics by minimizing the chance of making incorrect dosing decisions.

Determination of dose escalation and de-escalation boundaries

The basic statistical principles are provided here, and more technical details can be found in the work of Liu and Yuan (14). Under the BOIN design, the dose escalation and de-escalation boundaries λ_e and λ_d are chosen to minimize incorrect decisions of dose assignment. Toward that goal, we first formally define the correct and incorrect

decisions. Toward that goal, let p_j denote the true DLT rate of the current dose j. Three point hypotheses are formulated: H_1: $p_j = \phi$; H_2: $p_j = \phi_1$; H_3: $p_j = \phi_2$, 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 de-escalation is required.

Specifically, H_1 indicates that the current dose is the MTD and we should retain the current dose to treat the next cohort of patients; H_2 indicates that the current dose is subtherapeutic (or below the MTD) and the dose should be escalated; and H_3 indicates that the current dose is overly toxic (or above the MTD) and the dose would be de-escalated. Therefore, the correct decisions under H_1, H_2 and H_3 are retainment, escalation and de-escalation (each based on the current dose level), respectively, while other decisions are incorrect decisions. For example, escalation and de-escalation are incorrect decisions under H_1, de-escalation and retainment are incorrect decisions under H_2, and escalation and retainment are incorrect decisions under H_3.

Our purpose in specifying the three hypotheses, H_1, H_2 and H_3, is not to represent the truth and conduct hypothesis testing, but just to indicate the cases of special interest under which we optimize the performance of our design. In particular, H_2 and H_3 represent the minimal differences (or effect sizes) of practical interest to be distinguished from the target toxicity rate, ϕ (or H_1), under which we want to minimize the average decision error rate for the trial conduct. This approach is analogous to sample size determination, for which we first specify a point alternative hypothesis to represent the minimal effect size of interest and then determine the sample size to ensure a desirable power under that hypothesis. In practice, setting ϕ_1 and ϕ_2 very close to ϕ should be

avoided because the small sample sizes of typical phase I trials prevent us from being able to discriminate the target toxicity rate from the rates close to it. For example, at the significance level of 0.1, there is only 7% power to distinguish 0.25 from 0.35 with a total of 30 patients given just two doses. As default values, we recommend $\phi_1=0.6\phi$ and $\phi_2=1.4\phi$ for most clinical applications.

Under the Bayesian paradigm, we can assign each hypothesis a noninformative equal prior probability of being true and calculate the expected decision error rate, and then minimize it by choosing appropriate values of λ_e and λ_d. Remarkably, the solutions of λ_e and λ_d not only have closed-form expressions, given by

$$\lambda_e = \log\frac{1-\phi_1}{1-\phi}\Big/\log\frac{\phi(1-\phi_1)}{\phi_1(1-\phi)}$$

$$\lambda_d = \log\frac{1-\phi}{1-\phi_2}\Big/\log\frac{\phi_2(1-\phi)}{\phi(1-\phi_2)},$$

but are also independent of the dose level and the number of patients that have been treated. That is, the same boundaries can be used throughout the trial, no matter which dose is currently under consideration and how many patients have been treated.

Because the dose escalation rules (i.e., boundaries λ_e and λ_d) of the BOIN design are chosen on the basis of the formal statistical theory, it offers substantially better operating characteristics than the 3+3 design, as we demonstrate in the numerical study, as well as some desirable statistical properties. Specifically, the BOIN design is (long-memory) coherent and consistent. Being long-memory coherent means that the BOIN design never escalates (or de-escalates) the dose if the observed toxicity rate at the current dose is higher (or lower) than the target toxicity rate. This is a very desirable design property because it automatically satisfies the following (ad hoc) safety

requirement often imposed by clinicians: dose escalation is not allowed if the observed toxicity rate at the current dose is higher than the target toxicity rate. The BOIN design is consistent, which means that it guarantees that the true MTD will be found when the sample size is large.

The 3+3 design in the simulation study

There are many variations of the 3+3 design. The 3+3 design we used for the comparison in the simulation study is described as follows.

- The first cohort of 3 patients is treated at dose level 1.
- If 0 out of 3 patients experiences DLT, the next cohort of 3 patients is treated at the next higher dose level.
- If 1 patient out of 3 develops DLT, 3 more patients are treated at the same dose level. If no more patients experience DLT at that dose, i.e., only 1 out of a total of 6 patients develops DLT, the dose escalation continues to the next higher level for a cohort of 3 patients.
- At any given dose, if more than 1 out of 3 patients or 6 patients experience DLTs, the dose level exceeds the MTD and 3 patients are then treated at the next lower dose level if fewer than 6 patients have already been treated at that dose; otherwise the next lower dose level is claimed as the MTD. If this is the lowest dose level tested, the trial is terminated and the MTD is not found.

Supplementary Table S1. Dose escalation and de-escalation boundaries for the target toxicity rates of 15%, 20%, 25% and 30%.

Target toxicity rate = 15%

Action	The number of patients treated at the current dose																	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Escalate if # of DLTs ≤	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	2	2
De-escalate if # of DLTs ≥	1	1	1	1	1	2	2	2	2	2	2	3	3	3	3	3	4	4

Target toxicity rate = 20%

Action	The number of patients treated at the current dose																	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Escalate if # of DLTs ≤	0	0	0	0	0	0	1	1	1	1	1	1	2	2	2	2	2	2
De-escalate if # of DLTs ≥	1	1	1	1	2	2	2	2	3	3	3	3	4	4	4	4	5	5

Target toxicity rate = 25%

Action	The number of patients treated at the current dose																	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Escalate if # of DLTs ≤	0	0	0	0	0	1	1	1	1	1	2	2	2	2	2	3	3	3
De-escalate if # of DLTs ≥	1	1	1	2	2	2	3	3	3	3	4	4	4	5	5	5	6	6

Target toxicity rate = 30%

Action	The number of patients treated at the current dose																	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Escalate if # of DLTs ≤	0	0	0	0	1	1	1	1	2	2	2	2	3	3	3	3	4	4
De-escalate if # of DLTs ≥	1	1	2	2	2	3	3	3	4	4	4	5	5	6	6	6	7	7

Supplementary Table S2. Sixteen true toxicity scenarios with the target DLT rate of 0.15 and 0.3

Scenario	Dose level					Scenario	Dose level				
	1	2	3	4	5		1	2	3	4	5
1	**0.15***	0.20	0.25	0.30	0.40	1	**0.30***	0.40	0.50	0.60	0.70
2	**0.15**	0.23	0.30	0.40	0.50	2	**0.30**	0.45	0.60	0.70	0.80
3	0.10	**0.15**	0.20	0.30	0.40	3	0.20	**0.30**	0.40	0.50	0.60
4	0.10	**0.15**	0.25	0.35	0.50	4	0.20	**0.30**	0.45	0.60	0.70
5	0.05	**0.15**	0.20	0.30	0.40	5	0.15	**0.30**	0.40	0.50	0.60
6	0.05	**0.15**	0.25	0.35	0.50	6	0.15	**0.30**	0.45	0.60	0.70
7	0.04	0.10	**0.15**	0.20	0.30	7	0.12	0.20	**0.30**	0.40	0.50
8	0.04	0.10	**0.15**	0.25	0.40	8	0.12	0.20	**0.30**	0.45	0.60
9	0.02	0.05	**0.15**	0.20	0.30	9	0.05	0.15	**0.30**	0.40	0.50
10	0.02	0.05	**0.15**	0.25	0.40	10	0.05	0.15	**0.30**	0.45	0.60
11	0.01	0.05	0.10	**0.15**	0.20	11	0.05	0.12	0.20	**0.30**	0.40
12	0.01	0.05	0.10	**0.15**	0.25	12	0.05	0.12	0.20	**0.30**	0.45
13	0.01	0.03	0.05	**0.15**	0.20	13	0.02	0.08	0.15	**0.30**	0.40
14	0.01	0.03	0.05	**0.15**	0.25	14	0.02	0.08	0.15	**0.30**	0.45
15	0.02	0.04	0.06	0.10	**0.15**	15	0.02	0.10	0.15	0.20	**0.30**
16	0.01	0.02	0.04	0.05	**0.15**	16	0.01	0.04	0.08	0.15	**0.30**

* boldface indicates the MTD.

Supplementary Table S3. BOIN design with a tighter de-escalation boundaries*							
	Target toxicity rate for the MTD						
Boundary	0.1	0.15	0.2	0.25	0.3	0.35	0.4
λ_e	0.078	0.118	0.157	0.197	0.236	0.276	0.316
λ_d	0.110	0.165	0.219	0.275	0.330	0.385	0.440

* The dose de-escalation boundary is obtained by setting the upper acceptable toxicity limit $\phi_2=1.2\phi$, where ϕ is the target DLT rate. The default value in the BOIN software is $\phi_2=1.4\phi$.

Problems when matching the sample size of the 3+3 design

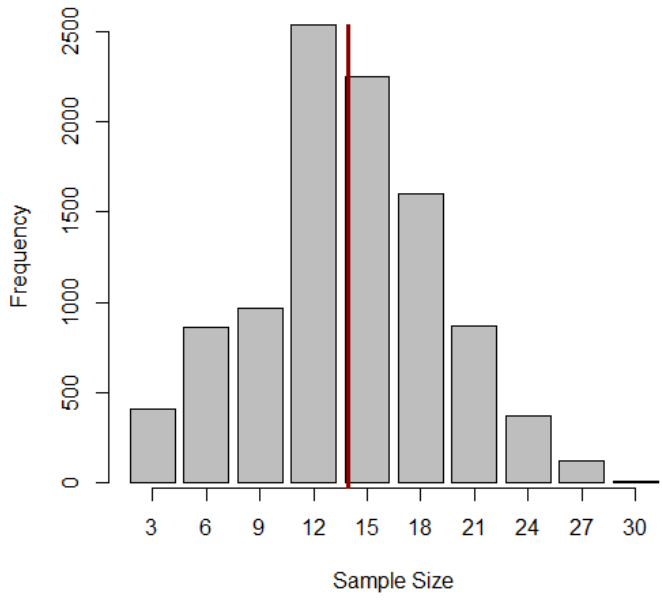
It might seem appealing to use the average sample size of the 3+3 design as the sample size for the designs that are based on fixed sample sizes, such as the mTPI and BOIN designs, to match the average sample size of different designs. However, that approach yields severely biased results because the sample size of the 3+3 design is random and takes a bell-shaped distribution. Supplementary Figure S1 shows the sample size distribution of the 3+3 design based on 10,000 simulated trials when the true DLT rates for 5 dose levels are 0.12, 0.2, 0.3, 0.4 and 0.5 (i.e., scenario 7 with the target DLT rate of 0.3), respectively. Using the mean sample size of the 3+3 design (i.e., 13.9 patients) as the sample size of the mTPI and BOIN designs would truncate all larger sample sizes and thus largely forbid the mTPI and BOIN designs to reach overly toxic doses. In other words, that approach makes the mTPI or BOIN design artificially safer simply because there are not enough patients to reach overly toxic doses. That is the reason why Ji and Wang (25) observed that the mTPI is safer than the 3+3 design in their simulation study. By contrasting the decision rules of the two designs (Supplementary Table S4), it is clear that the mTPI theoretically cannot be safer than the 3+3 design because the dose escalation rule of the 3+3 design is more conservative than that of the mTPI. Following Ji and Wang (25), two versions of the 3+3 design are listed in Supplementary Table S4, with the $3+3^{L}$ design targeting the MTD with the DLT rate $\leq 1/6$, and the $3+3^{H}$ design targeting the MTD with the DLT rate $\leq 2/6$. The details of these two versions of the 3+3 design are provided in Ji and Wang (25). Clearly, the mTPI is more aggressive: when 2/6 patients have DLTs, the $3+3^{L}$ design will de-escalate the dose, whereas the mTPI will continue treating patients at the same dose; and when 3/6 patients have DLTs, the $3+3^{H}$

design will deescalate the dose, whereas the mTPI will continue treating patients at the same dose.

Supplementary Table S4. Dose escalation and de-escalation rule for 3+3 and mTPI								
No. of patients	3			6				
No. of DLTs	0	1	≥2	0	1	2	3	≥4
3+3^{L}	E	S	D	E	Se	**D**	D	D
mTPI (p_T=20%)	E	S	D	E	S	**S**	D	D
3+3^{H}	E	S	D	E	E	Se	**D**	D
mTPI (p_T=30%)	E	S	D	E	E	S	**S**	D

Notation: E, escalation; D, de-escalation; S, stay at same dose; Se, select the MTD; p_T, target DLT rate.

Another issue of using the mean sample size of the 3+3 design as the sample size for the comparative designs is that the average sample size of the 3+3 design somewhat informs the sample size required to reach the MTD, which makes the comparative designs more likely to identify the MTD than the 3+3 design.



Supplementary Figure S1. Sample size distribution of the 3+3 design when the true toxicity rates of 5 dose levels are 0.12, 0.2, 0.3, 0.4 and 0.5, respectively. The red vertical

line indicates the mean of the sample size. Matching mean sample size of the 3+3 design truncates all large sample sizes.