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

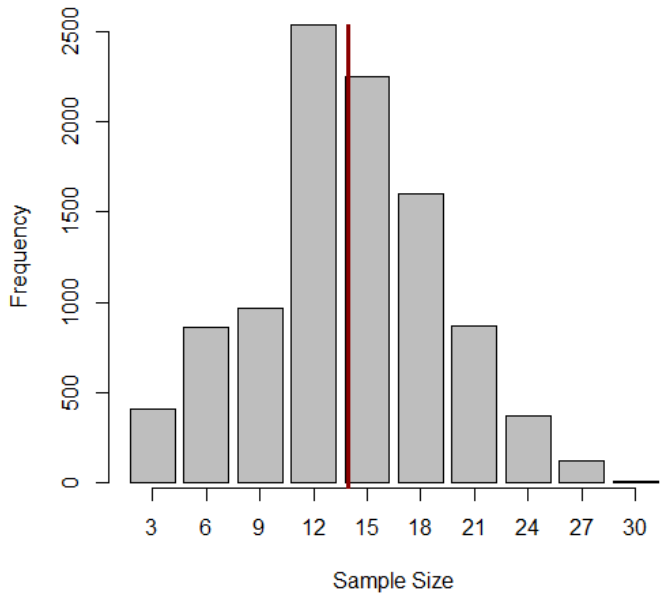
It might seem appealing to use the average sample size of the 3+3 design as the sample size for the designs that are based on fixed sample sizes, such as the mTPI and BOIN designs, to match the average sample size of different designs. However, that approach yields severely biased results because the 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.