Contemporary Clinical Trials Communications 12 (2018) 40–50

Contents lists available at ScienceDirect

Contemporary Clinical Trials Communications

journal homepage: www.elsevier.com/locate/conctc

Overall success rate of a safe and efficacious drug: Results using six phase 1 designs, each followed by standard phase 2 and 3 designs

Amy S. Ruppert[a,b,*], Abigail B. Shoben[b]

[a] Division of Hematology, College of Medicine, The Ohio State University, Columbus, OH, 43210, USA
[b] Division of Biostatistics, College of Public Health, The Ohio State University, Columbus, OH, 43210, USA

ARTICLE INFO

Keywords:
Dose-finding
Phase 1 designs
Overall success

ABSTRACT

To evaluate the overall success rate of a new drug, phase 1, 2, and 3 trials were simulated using eight toxicity and two non-decreasing efficacy profiles. Six phase 1 designs including the standard 3 + 3, CCD, BOIN, mTPI, mTPI-2, and CRM were considered with standard phase 2 and 3 designs.

Based on our results, phase 1 design recommendations are provided when data informing the general shape of the dose-toxicity curve exist. If a large jump in toxicity between dose levels is expected, the standard 3 + 3 design is recommended; it more often recognized when the MTD was exceeded and had the highest overall success rates. If gradually increasing toxicity is expected, a nonstandard design other than the CRM is recommended. Nonstandard designs were more aggressive in dosing and MTD estimation than the standard 3 + 3 and had higher overall success rates, but the CRM was too aggressive and most frequently overestimated the true MTD. If fairly constant, safe toxicity is expected across dose levels, the BOIN or CRM designs are recommended; they escalated to the highest dose most frequently with superior overall success rates.

Without data informing the shape of the dose-toxicity curve, nonstandard phase 1 designs with a modified excessive toxicity rule more easily eliminating unsafe dose levels are recommended. With this modification, MTD overestimation error decreased and overall success rates were similar or higher with nonstandard designs. Among nonstandard designs, the modified CCD and BOIN perform well and are as transparent and simple to implement as the standard 3 + 3 design.

1. Introduction

The primary objective of a phase 1 clinical trial is to determine a safe and tolerable dose level to recommend for further study of efficacy in subsequent phase 2 and 3 trials. Under the assumption that both efficacy and toxicity increase with increasing dose levels, the recommended phase 2 dose is generally the maximum tolerated dose (MTD), defined as the highest dose level where the percentage of patients experiencing predefined dose limiting toxicity (DLT) is below a specified acceptable level. Selection of a dose level that is at or closely below the true MTD is most desirable.

For the past 25 years, the most common dose-finding phase 1 design has been the rule-based standard 3 + 3 design [1–3]. Many have advocated for the use of the model-based continual reassessment method (CRM) for dose-finding [4], but the CRM has been met with resistance due to its unfamiliarity, assumptions that must be made on the shape of the dose-toxicity curve, statistical complexity, need for specialized software, and increased communication required during trial design and implementation [3]. A new type of phase 1 design, the interval design has emerged, and includes the cumulative cohort design (CCD) [5], the modified toxicity probability interval design (mTPI) [6], the Bayesian optimal interval design (BOIN) [7] and the mTPI-2 design [8].

All of these designs except for mTPI-2 have been directly compared to the standard 3 + 3 design and better estimated the true MTD in most scenarios [9]. The CRM was superior in scenarios with six or eight dose levels, followed by the BOIN and then mTPI [9]. However, the ranking of design performance was less clear for smaller dose-finding studies with fewer dose levels. Further, phase 1 design performance has been primarily measured by estimating the percentage of simulations that correctly identify the true MTD and by estimating the average number of simulated patients treated above the true MTD during phase 1. Evaluations from simulation studies rarely measure the downstream effects due to selecting dose levels above or below the true MTD.

Thus, we herein evaluate the performance of all six phase 1 designs (rule-based standard 3 + 3, CCD, BOIN, mTPI, and mTPI-2 interval designs, and model-based CRM design), in the context of a moderately

* Corresponding author. Division of Hematology, College of Medicine, The Ohio State University, Columbus, OH, 43210, USA.
E-mail address: amy.stark@osumc.edu (A.S. Ruppert).

https://doi.org/10.1016/j.conctc.2018.08.010
Received 27 March 2018; Received in revised form 9 August 2018; Accepted 23 August 2018
Available online 24 August 2018

sized phase 1 trial with four escalation dose levels, using overall success rate as a performance measure. Each phase 1 design is followed by Simon's optimal two-stage phase 2 design [10] and a randomized group sequential phase 3 design [11], with overall success rate defined as the percentage of simulations spanning phase 1, 2, and 3 that identify a new drug as safe and efficacious when it actually is safe and efficacious.

Overall success rates are compared by phase 1 design and clinical scenario defined by different dose-toxicity, dose-response, and dose-survival profiles. The impact of excessive toxicity rules and sample size on overall success rates are investigated. Guidelines for phase 1 statistical design choice in different clinical settings are presented, considering the trade-offs between measures of performance with design complexity and ease of implementation.

2. Materials and methods

2.1. Clinical scenarios

In phase 1, five dose levels of a new drug were considered, including four escalation and one de-escalation dose level. Eight toxicity profiles were evaluated: three with the MTD at dose level 2, three with the MTD at dose level 3, and two with the MTD at dose level 4 (Table 1). All but one of the toxicity profiles were monotonically increasing and mirrored shapes that have been commonly included in other phase 1 simulation studies [12–15]. Linear profiles had toxicity probabilities that increased fairly linearly with increasing dose levels, a typical assumption with standard chemotherapy. Jump profiles had a sharp increase in toxicity probability between dose levels 2 and 3, and represented an increase in dose outside the therapeutic window or target saturation. The Plateau profile had increasing toxicity, with smaller increases in toxicity at higher dose levels, which has been described with orally administered, molecularly targeted agents [16]. Constant toxicity profiles had acceptable toxicity with equal probability across dose levels, and have been described with molecularly targeted agents administered within the therapeutic window [17,18]. One nonmonotonic toxicity profile was included and was Tub-shaped. The Tub-shaped toxicity profile had acceptable but moderately high toxicity probabilities at dose levels −1 and 1, lower toxicity probabilities at dose levels 2 and 3, and a sudden increase in toxicity probability above the acceptable level at dose level 4. This toxicity profile represented a scenario in which disease-related adverse events are observed at low inactive dose levels and called DLTs [19]. As the drug becomes more active at higher dose levels and disease-related adverse events are no longer observed, the DLT rate then decreases. Eventually the drug is delivered at a dose level outside the therapeutic window and the DLT rate increases once again.

Each of the eight toxicity profiles was mapped to a Continuous response/survival profile and a Step response/survival profile (Table 1). Continuous response profiles occurred with Continuous survival profiles and represented therapy that had steadily increasing efficacy with increasing dose levels. Step response profiles occurred with Step survival profiles and represented agents that remained inactive until critical mass was reached between dose levels 1 and 2. In the efficacy profiles evaluated, response rate was not lower than 5% and median survival was not shorter than 6 months at any dose level, the response rate and median survival assumed for the standard of care. Scenarios in which a safe and efficacious drug existed were of primary interest, and so all profiles included the optimal or target response rate and median survival at the true MTD. In this study, the target response rate was 20% and the target median survival was 9 months, corresponding to a hazard ratio of 0.67 when compared to standard of care and assuming exponential survival times. Scenarios with suboptimal response or suboptimal survival at the true MTD were not explored.

Collectively, eight toxicity profiles and two efficacy profiles were simulated, resulting in 16 total scenarios. Six phase 1 designs (i.e., standard 3 + 3, CCD, BOIN, mTPI, mTPI-2, CRM), each followed by Simon's optimal two-stage phase 2 design [10] and a two-arm randomized group sequential phase 3 design [11], were applied to each clinical scenario.

Table 1
Assumed toxicity, response, and survival profiles across dose levels. Toxicity, response, and survival, respectively, are the true proportion of DLT, true response proportion, and true median survival in months at each dose level.

MTD at Dose Level 2

Dose Level	Toxicity Profiles			Continuous Efficacy		Step Efficacy	
	Linear A	Jump A	Jump B	Response	Survival	Response	Survival
−1	0.10	0.05	0.20	0.10	7	0.05	6
1	0.20	0.05	0.20	0.15	8	0.05	6
2	0.30	0.05	0.20	0.20	9	0.20	9
3	0.40	0.60	0.40	0.25	10	0.20	9
4	0.50	0.60	0.40	0.30	11	0.20	9

MTD at Dose Level 3

Dose Level	Toxicity Profiles			Continuous Efficacy		Step Efficacy	
	Linear B	Plateau	Tub	Response	Survival	Response	Survival
−1	0.05	0.05	0.20	0.05	6	0.05	6
1	0.10	0.15	0.20	0.10	7	0.05	6
2	0.20	0.25	0.10	0.15	8	0.20	9
3	0.30	0.30	0.10	0.20	9	0.20	9
4	0.40	0.35	0.35	0.25	10	0.20	9

MTD at Dose Level 4

Dose Level	Toxicity Profiles		Continuous Efficacy		Step Efficacy	
	Constant A	Constant B	Response	Survival	Response	Survival
−1	0.05	0.20	0.05	6	0.05	6
1	0.05	0.20	0.05	6	0.05	6
2	0.05	0.20	0.10	7	0.20	9
3	0.05	0.20	0.15	8	0.20	9
4	0.05	0.20	0.20	9	0.20	9

2.2. Description of phase 1 designs

The standard 3 + 3 design is a rule-based design in which patients are enrolled in cohorts of three, beginning at the starting dose level [1]. If there are no DLTs in the first cohort of three patients treated at a dose level, the dose is escalated. If one DLT is observed in the first cohort of three patients, a second cohort of three patients is treated at the same dose level. If at most one DLT is observed in six patients at a dose level, then escalation to the next highest dose level is permitted. At a dose level with two or more DLTs, the MTD has been exceeded and the dose is de-escalated until at most one DLT is observed in a total of six patients.

The CCD is an interval design in which a target DLT rate (p_t) and small fractions of error (e_1 and e_2) about p_t are specified to form a proper-dosing interval (p_t - e_1, p_t + e_2) [5]. Throughout the trial, the observed DLT rate at a dose level is compared to the proper-dosing interval to make dosing decisions. The decision to escalate, stay at the same dose level, or de-escalate corresponds respectively with whether the observed DLT rate at the current dose level is below, within, or above the proper-dosing interval. As in all interval designs, the MTD is estimated at the end of the trial after applying isotonic regression to estimated DLT probabilities at each dose level and selecting the dose level with estimated DLT probability closest to p_t.

The BOIN is an interval design similar to the CCD [7]. Dosing decisions are based on the observed DLT rate as compared to the proper-dosing interval. However, the recommended proper-dosing interval for a given p_t is different between the CCD and BOIN designs.

The mTPI design is the Bayesian analog of the CCD design [6]. With

the mTPI design, the posterior probability of DLT at each dose level is calculated according to a hierarchical *Beta-Binomial* distribution. Dosing decisions are based on the unit probability mass (UPM) of an under-dosing (0, p_t - e_1), proper-dosing (p_t - e_1, p_t + e_2), or over-dosing (p_t + e_2, 1) interval, and is defined as the posterior probability of DLT for an interval divided by the length of the interval. The decision to escalate, stay at the same dose level, or de-escalate corresponds with whether the under-dosing, proper-dosing, or over-dosing interval, respectively, has the largest UPM.

In mTPI-2, the single under-dosing interval used with mTPI is divided into multiple under-dosing intervals with lengths equal to e_1 + e_2 [8]. Likewise, the single over-dosing interval used with mTPI is divided into multiple over-dosing intervals with lengths equal to e_1 + e_2. UPMs are calculated for all intervals. The decision to escalate, stay at the same dose level, or de-escalate corresponds with whether one of the under-dosing intervals, the proper-dosing interval, or one of the over-dosing intervals, respectively, has the largest UPM.

The CRM is a model-based design, in which the best guess of the shape of the dose-toxicity curve across a range of dose levels is made via a statistical model prior to any data collection [4]. As patients are treated and data are observed, the dose-toxicity curve is updated. From the dose-toxicity curve, DLT probabilities at each dose level are estimated; the dose level with estimated DLT probability closest to p_t informs the dose level at which to treat the next cohort of patients. To address safety concerns, dosing is constrained such that the starting dose level is below the dose level suggested by the model, and escalation is restricted to one dose level at a time. The MTD is estimated as the dose level with a model estimate of DLT probability closest to p_t at the end of the trial.

2.3. Simulations

Letting i = dose level for $i = 1,\ldots,d$ where d was the total number of candidate doses proposed in the phase 1 design, the true DLT probability at each dose level was denoted as p_i. During trial implementation, n_i patients were simulated at each dose level and x_i patients experienced DLT. The estimated DLT probability at dose level i was denoted by $\hat{p}_i$, which was either the observed DLT proportion or the posterior mean probability, depending on design.

The clinical trial process was simulated 4000 times across phases, with appropriate decision-making both within and at the conclusion of each phase. The binomial distribution was used to generate the number of DLTs observed at each dose level of the phase 1 trial, with true probabilities of DLT at each dose level specified in Table 1. All phase 1 trials started at dose level 1 and proceeded using cohorts of 3 simulated observations. Dosing decisions continued according to the standard 3 + 3 design, or until a fixed total number of patients specified *a priori* was reached when using nonstandard designs (CCD, BOIN, mTPI, mTPI-2, CRM), or when dose level −1 was identified as excessively toxic.

When implementing the standard 3 + 3 design, the MTD was defined as the highest dose level with 1 or fewer DLTs in 6 patients. For all nonstandard phase 1 designs, the target toxicity probability was at $p_t = 0.30$.

For all interval designs, e_1 and e_2 were selected to form proper dosing intervals indicated in previous publications. When using the CCD, $e_1 = e_2 = 0.10$ to form a proper dosing interval (0.20, 0.40) [5]. When using the BOIN, $e_1 = 0.064$ and $e_2 = 0.058$ to form a proper dosing interval (0.236, 0.358) [7]. When using the mTPI and mTPI-2 designs, $e_1 = 0.05$ and $e_2 = 0.03$ to form the proper dosing interval (0.25, 0.33), and the *Beta-Binomial*(x_i + 1, n_i - x_i + 1) distribution was used in the unit probability mass calculations to guide dosing decisions [6,8]. For the mTPI and mTPI-2, $\hat{p}_i$ were estimated at the end of the study using the *Beta-Binomial*(x_i + 0.005, n_i - x_i + 0.005) distribution [6]. For all interval designs, the Iso package in R applied isotonic regression with the pool adjacent violator algorithm (PAVA) to the $\hat{p}_i$ and obtained $\tilde{p}_i$ at the end of the study [20,21]. The MTD was selected as the dose level with $\tilde{p}_i$ closest to p_t among dose levels considered safe. If the value of $\tilde{p}_i$ closest to p_t mapped to multiple dose levels, and the value of $\tilde{p}_i$ was less than p_t, then the highest dose level was selected; otherwise the lowest dose level was selected for further study in the phase 2 setting.

When implementing the CRM design, the bcrm package in R was used to specify the one-parameter hyperbolic-tangent model for the dose-toxicity curve [20,22]. The prior distribution on the unknown parameter was assumed to be *Gamma*(1,1). The prior probabilities specified for dose levels −1 to 4 were 0.05, 0.10, 0.20, 0.30, and 0.40, a fairly linear and generic dose-toxicity relationship. To guide dosing decisions, the posterior mean estimates of the model parameter were evaluated in the dose-toxicity function to obtain $\hat{p}_i$. At the end of the study, the MTD was selected as the dose level with $\hat{p}_i$ closest to p_t among dose levels considered safe.

In all interval designs and the CRM, a rule excluding dose levels due to excessive toxicity (e.g., $\Pr(p_i > p_t \mid \text{data}) > 0.95$) was implemented throughout the trial. This rule used posterior probabilities from a Beta-Binomial distribution with *Beta*(1, 1) as the prior on p_i. For the primary simulation study, the posterior probability cutoff for excessive toxicity was 0.95, cohort size was set equal to 3, and the fixed total number of patients was 24.

In phase 2, Simon's optimal two-stage design was used to test H_0: $\pi_E \leq 5\%$ versus H_a: $\pi_E > 5\%$, where π_E was the true response rate for the experimental drug. Type I error was constrained to 0.10 and there was at least 90% power to detect a response rate of 20%. This design required a total of 37 patients, allowing for an interim analysis when 12 patients were evaluated. Responses were generated for each simulated patient assuming a binomial distribution, with the true probability of response equal to the assumed response probability of the dose level selected in phase 1.

In phase 3, a group sequential design was used to test the null hypothesis that median survival with the experimental drug was the same as with the standard of care and equal to 6 months versus the alternative hypothesis that median survival was greater than 6 months with the experimental drug. With 1:1 randomization, this design corresponded to a one-sided test with type I error constrained to 0.025 and 90% power to detect a median survival of 9 months under the alternative hypothesis if the test was performed after 267 events had been observed. Assuming uniform accrual over 36 months and a minimum follow-up of 12 months, a total of 296 patients were expected to yield the necessary number of events. To anticipate an attrition rate of 5–10%, study accrual was planned for a total of 320 patients (n = 160 per arm). Survival times and survival status indicator variables were generated for each simulated individual assuming exponential distributions for failure and censoring times. Median failure time equaled 6 months for the standard of care arm. Median failure time for the experimental arm equaled the assumed median survival of the dose level selected in phase 1.

Simulation results from the 16 scenarios were summarized and organized by phase and across phases 1 to 3. Overall success rate, defined as the percentage of simulations spanning phases 1, 2, and 3 that identified a new drug as safe and efficacious when it actually was safe and efficacious, was calculated. The percentage of simulations in which a dose level was selected as the estimated MTD was also calculated. These results were used to compare and contrast the phase 1 designs in their ability to correctly estimate the true MTD but also used to compare and contrast MTD overestimation error, defined as selecting a dose level for further study above the true MTD, and underestimation error, defined as selecting a dose level for further study below the true MTD.

3. Results

3.1. Overall success rate

In most, but not all clinical scenarios considered, overall success

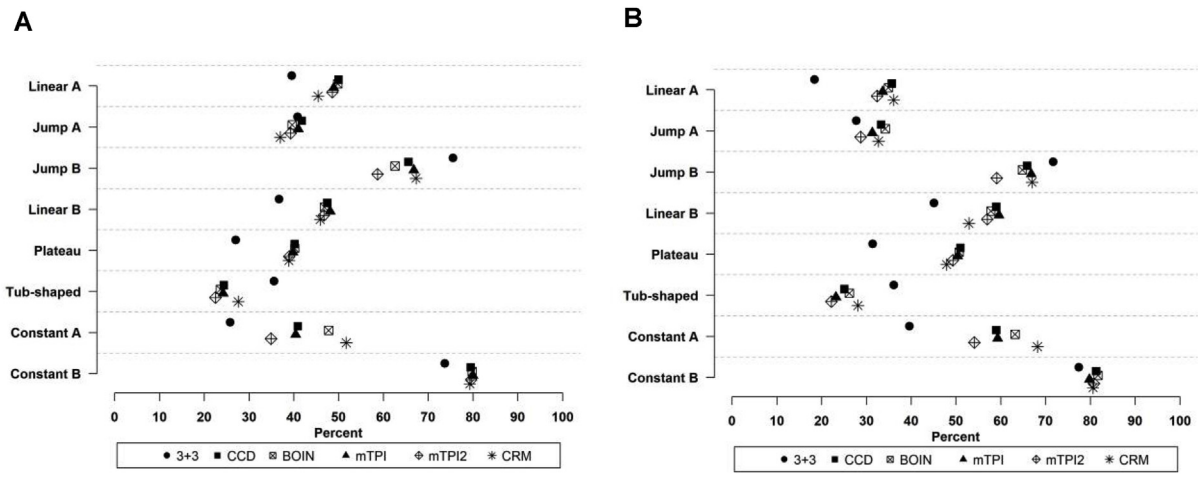

Fig. 1. Overall success rates of a favorable drug by phase 1 design, toxicity, and efficacy profile. A) Continuous efficacy profile and B) Step efficacy profile.

Table 2
Percentage of simulations with dose levels selected as the estimated MTD by phase 1 design and toxicity profile. Next to each dose level, the true DLT probability is listed in parentheses. Results for the true MTD, which is at dose level 2 for the Linear A, Jump A, and Jump B toxicity profiles, are in bold.

Dose Level (Tox)	Standard 3 + 3	CCD	BOIN	mTPI	mTPI-2	CRM
Linear A Toxicity Profile						
−1 (0.10)	28.0	4.3	4.5	4.4	4.8	4.1
1 (0.20)	36.9	32.6	29.0	33.0	35.7	21.8
2 (0.30)	**23.0**	**42.9**	**42.7**	**40.9**	**38.9**	**43.4**
3 (0.40)	7.3	16.6	19.7	18.0	16.7	24.7
4 (0.50)	1.2	3.5	4.3	3.8	3.9	6.0
Too Toxic	3.7	< 0.1	0.0	< 0.1	0.1	< 0.1
Jump A Toxicity Profile						
−1 (0.20)	20.2	7.0	6.5	6.5	11.1	5.6
1 (0.20)	21.6	18.7	13.8	19.6	20.3	10.6
2 (0.20)	**34.8**	**41.1**	**40.8**	**39.2**	**35.5**	**39.8**
3 (0.40)	8.9	24.8	27.9	25.0	23.9	29.7
4 (0.40)	3.7	8.1	10.6	9.5	7.5	13.9
Too Toxic	10.9	0.5	0.5	0.4	1.9	0.6
Jump B Toxicity Profile						
−1 (0.05)	2.7	< 0.1	0.0	0.0	0.0	0.0
1 (0.05)	3.2	0.4	0.2	0.4	0.7	0.0
2 (0.05)	**90.1**	**80.3**	**78.9**	**82.3**	**72.9**	**83.5**
3 (0.60)	3.6	18.7	20.1	16.4	25.8	14.8
4 (0.60)	0.4	0.6	0.9	0.9	0.8	1.8
Too Toxic	< 0.1	0.0	0.0	0.0	0.0	0.0

rates were improved when using a phase 1 design other than the standard 3 + 3 design. Overall success rates ranged from 18.4% to 81.7%, with variability in overall success rates attributed mostly to underlying toxicity profile and the phase 1 design. Since the patterns in overall success rates according to toxicity profile and phase 1 design were similar between the Continuous and Step efficacy profiles (Fig. 1), the results that follow are discussed in detail only for the Continuous efficacy profile.

With the Linear A, Linear B, and Plateau toxicity profiles, characterized by gradually increasing toxicity across dose levels, overall success rates ranged from 27.0% to 50.0% and the overall success rate was on average 10.7 percentage points higher with the CCD, BOIN, mTPI, mTPI-2, or CRM designs compared to the standard 3 + 3 design. Among the nonstandard phase 1 designs, overall success rates were not largely different. Overall success rates tended to be lower with the CRM but always within 5 percentage points of the highest overall success rate of the other nonstandard designs (Fig. 1A).

With the Constant A toxicity profile, where all dose levels had an acceptable but moderately high toxicity probability equal to 0.20, the overall success rates were lowest with the standard 3 + 3 design at 25.8%, higher with the CCD, mTPI, or mTPI-2 designs at 40.9%, 40.4% and 34.9% respectively, and highest with the BOIN or CRM designs at 47.8% and 51.7%, respectively. With the Constant B toxicity profile, where all dose levels had minimal toxicity with toxicity probability equal to 0.05, overall success rates were high across designs and ranged from 73.7% with the standard 3 + 3 design to approximately 80% with all five nonstandard phase 1 designs.

With the Jump A toxicity profile, characterized by a jump in toxicity probability from 0.20 to 0.40 between dose levels 2 and 3, overall success rates hovered around 40%; the overall success rate using the standard 3 + 3 design was just as good as the nonstandard phase 1 designs. With the Jump B toxicity profile, with a large jump in toxicity probability from 0.05 to 0.60 between dose levels 2 and 3, the overall success rate with the standard 3 + 3 design was highest at 75.5%, on average 9.9 percentage points higher compared with nonstandard phase 1 designs. The overall success rate was also highest with the standard 3 + 3 design under the nonmonotonic Tub-shaped toxicity profile at 35.6%, on average 11.2 percentage points higher compared with nonstandard phase 1 designs.

3.2. MTD selection rate

The most influential factor driving overall success rates was correct selection of the MTD. Thus, with the Linear A, Linear B, and Plateau toxicity profiles that had higher overall success rates using nonstandard phase 1 designs, the correct MTD selection rate was roughly 10–20% higher with nonstandard designs compared with the standard 3 + 3 design (Fig. 1 and Tables 2 and 3). In alignment with published literature, the standard 3 + 3 design was more likely to underestimate the true MTD and to underestimate the true MTD to a greater degree than the nonstandard designs [23]. Among the nonstandard designs, MTD selection rates were within 4–6% of one another, with the CRM always having the highest MTD selection rate and the mTPI-2 design always having the lowest MTD selection rate; the CCD, BOIN, and mTPI had MTD selection rates in between the CRM and mTPI-2 and were quite similar to one another. Interestingly, the CRM had the highest MTD selection rate but the lowest overall success rate among the nonstandard designs due to its propensity to overestimate the MTD.

The greatest variability in overall success rates and MTD selection rates across phase 1 designs was observed with the Constant A toxicity profile, where all dose levels had an acceptable but moderately high toxicity probability equal to 0.20 (Fig. 1A and Table 4). Correct MTD selection rates ranged from 23.7% using the standard 3 + 3 design to 51.6% using the CRM design. This was the only scenario where the CRM and the BOIN clearly had higher overall success rates compared with the CCD, mTPI, and mTPI-2 designs, and can be traced back to the

Table 3
Percentage of simulations with dose levels selected as the estimated MTD by phase 1 design and toxicity profile. Next to each dose level, the true DLT probability is listed in parentheses. Results for the true MTD, which is at dose level 3 for the Linear B, Plateau, and Tub-shaped toxicity profiles, are in bold.

Dose Level (Tox)	Standard 3 + 3	CCD	BOIN	mTPI	mTPI-2	CRM
Linear B Toxicity Profile						
−1 (0.05)	9.6	0.3	0.2	0.3	0.4	0.3
1 (0.10)	27.6	7.2	5.0	7.9	8.5	3.5
2 (0.20)	34.3	35.9	32.4	34.4	35.0	24.3
3 (0.30)	**20.8**	**37.4**	**39.1**	**39.2**	**36.5**	**42.1**
4 (0.40)	7.4	19.3	23.4	18.5	19.6	29.9
Too Toxic	0.4	0.0	0.0	0.0	0.0	0.0
Plateau Toxicity Profile						
−1 (0.05)	19.2	1.2	1.1	1.0	1.3	1.4
1 (0.15)	33.4	18.5	14.3	18.6	20.9	10.6
2 (0.25)	24.5	37.8	34.7	36.3	36.1	28.2
3 (0.30)	**14.2**	**25.8**	**27.7**	**25.9**	**25.0**	**31.2**
4 (0.35)	8.1	16.8	22.3	18.2	16.9	28.8
Too Toxic	0.8	0.0	0.0	0.0	0.0	0.0
Tub-shaped Toxicity Profile						
−1 (0.20)	19.3	6.7	6.6	6.6	11.0	5.6
1 (0.20)	5.8	13.7	9.7	16.6	15.4	5.7
2 (0.05)	6.9	6.2	3.7	4.2	5.9	6.5
3 (0.05)	**38.4**	**25.0**	**26.9**	**24.9**	**22.2**	**29.3**
4 (0.35)	19.2	47.9	52.8	47.2	44.0	52.4
Too Toxic	10.4	0.7	0.5	0.5	1.7	0.5

Table 4
Percentage of simulations with dose levels selected as the estimated MTD by phase 1 design and toxicity profile. Next to each dose level, the true DLT probability is listed in parentheses. Results for the true MTD, which is at dose level 4 for the Constant A and Constant B toxicity profiles, are in bold.

Dose Level (Tox)	Standard 3 + 3	CCD	BOIN	mTPI	mTPI-2	CRM
Constant A Toxicity Profile						
−1 (0.20)	19.5	7.4	6.5	6.1	10.9	5.2
1 (0.20)	20.6	18.6	13.9	20.7	20.8	9.4
2 (0.20)	14.4	19.3	15.8	17.6	19.1	15.2
3 (0.20)	11.4	16.0	16.3	16.2	15.9	18.3
4 (0.20)	**23.7**	**38.2**	**47.2**	**39.1**	**31.7**	**51.6**
Too Toxic	10.7	0.6	0.5	0.4	1.8	0.4
Constant B Toxicity Profile						
−1 (0.05)	2.5	0.1	0.0	0.0	0.0	0.0
1 (0.05)	2.5	0.3	0.2	0.5	0.7	0.0
2 (0.05)	2.5	1.0	0.4	0.9	1.1	0.3
3 (0.05)	2.9	1.9	0.9	2.2	2.4	1.0
4 (0.05)	**89.0**	**96.8**	**98.6**	**96.5**	**95.9**	**98.8**
Too Toxic	0.1	0.0	0.0	0.0	0.0	0.0

distribution of dose levels selected as the MTD. With the CRM and the BOIN, the highest dose level was correctly reached in 51.6% and 47.2% of simulations, respectively, compared with 38.2%, 39.1% and 31.7% of simulations for the CCD, mTPI, and mTPI-2 designs, respectively. Instead of escalating to the highest dose level, the CCD, mTPI, and mTPI-2 were more likely to suggest staying at the starting dose level, which had a toxicity probability relatively close to the target toxicity probability (i.e., 0.20 versus the target of 0.30). In addition, the mTPI-2 was the most conservative of the nonstandard designs, with the highest likelihood of de-escalating to dose level −1.

With the Constant B toxicity profile, where the toxicity probability was low and equal to 0.05 across dose levels, the CRM also reached the highest dose level most frequently (Table 4). However, under this toxicity profile, all phase 1 designs resulted in correctly calling the highest dose level the MTD a large percentage of time: 89.0%, 96.8%, 98.6%, 96.5%, 95.9%, and 98.8%, for the standard 3 + 3, CCD, BOIN, mTPI, mTPI-2, and CRM designs, respectively. Even though the nonstandard designs had higher MTD selection rates than the standard 3 + 3 design, this translated into only slightly higher overall success rates observed across all nonstandard designs (Fig. 1A and Table 4).

With the Jump A toxicity profile that had a jump in toxicity probability from 0.20 to 0.40 between dose levels 2 and 3, the overall success rate for the standard 3 + 3 design clustered with the overall success rates of the nonstandard designs and the MTD selection rate was roughly 5 percentage points lower (Fig. 1A and Table 2). Based on these two measures alone, it appeared that the nonstandard designs had an advantage over the standard 3 + 3 design. However, when looking more closely at the distribution of the dose levels selected as the MTD, it was clear that the paths to similar correct MTD selection and overall success rates were not the same.

When using the standard 3 + 3 design, the phase 1 study stopped prematurely for excessive toxicity in 10.9% of simulations, the MTD was underestimated in 41.8% of simulations, and the MTD was overestimated in 12.6% of simulations. In contrast, with most nonstandard designs, a drug was discontinued for excessive toxicity in < 1% of simulations, the MTD was underestimated in approximately 15–25% of simulations and usually not by more than one dose level, and the MTD was overestimated in approximately 30–45% of simulations.

Differences in underestimation and overestimation errors between designs were more apparent with the Jump B toxicity profile, which was characterized by a large jump in toxicity probability from 0.05 to 0.60 between dose levels 2 and 3 (Table 2). With the standard 3 + 3 design, a drug was discontinued for excessive toxicity in < 1% of simulations, the MTD was underestimated in 5.9% of simulations, and the MTD was overestimated in only 4.0% of simulations. With the nonstandard designs and under this particular toxicity profile, the drug was never discontinued for excessive toxicity, the MTD was underestimated in < 1% of simulations, and the MTD was overestimated in approximately 15–25% of simulations. Whereas the conservative standard 3 + 3 design quickly recognized when the MTD had been exceeded, nonstandard designs did not.

In an attempt to better estimate the MTD, the nonstandard designs had more aggressive rules for dose escalation, which resulted in more frequent selection of dose levels above the true MTD. Unacceptable overestimation errors occurred even in the presence of toxicity probabilities grossly higher than the target toxicity probability.

Lastly, with the nonmonotonic Tub-shaped profile that also had a jump in toxicity probabilities, the MTD selection and overall success rates were highest using the standard 3 + 3 design compared with the nonstandard designs (Fig. 1A and Table 3). In this scenario, selection of a dose level above the true MTD in a larger fraction of simulations when using nonstandard designs outweighed the selection of a dose level below the true MTD when using the standard 3 + 3 design.

Irrespective of the toxicity profile, the average number of patients treated at the true MTD during phase 1 was higher using nonstandard designs compared to the standard 3 + 3 design (Table 5). With the exception of mTPI-2 under the Plateau, Tub-shaped, and Constant A toxicity profiles, the percentage of patients treated at the true MTD during phase 1 was also higher using nonstandard designs compared to the standard 3 + 3 design.

3.3. Safety measures

Common measures of safety reported for phase 1 designs include the average percentage of patients who experience a DLT and the average percentage of patients treated above the true MTD during the phase 1 trial. These measures were summarized for the six different phase 1 designs and eight toxicity profiles included in this simulation study (Table 6). The CRM consistently resulted in the highest percentages of patients with DLT and treated above the true MTD in phase 1. Compared to the CRM, the BOIN, CCD, and mTPI designs had similar but lower percentages of patients with DLT and patients treated above the MTD in phase 1. The lowest percentages of patients with DLT and patients treated above the MTD in phase 1 were observed with the standard 3 + 3 and mTPI-2 designs. These results do not support a previous claim that the mTPI design is safer than the standard 3 + 3 design [15].

Table 5
The average number (no.) of patients treated at the true MTD and the average percentage of patients treated at the true MTD during the phase 1 study by phase 1 design.

Measure	Standard 3 + 3	CCD	BOIN	mTPI	mTPI-2	CRM
Linear A Toxicity Profile						
No. at true MTD	3.7	8.2	8.4	8.5	6.8	7.5
% at true MTD	27.9	34.4	35.0	35.4	28.1	31.4
Jump A Toxicity Profile						
No. at true MTD	3.8	7.7	7.8	7.6	6.7	7.2
% at true MTD	27.4	32.2	32.4	31.7	28.1	30.1
Jump B Toxicity Profile						
No. at true MTD	5.7	12.9	12.9	12.6	13.0	11.9
% at true MTD	43.3	53.7	53.7	52.4	54.0	49.7
Linear B Toxicity Profile						
No. at true MTD	3.2	6.3	6.7	6.4	4.8	6.7
% at true MTD	19.8	26.1	27.7	26.7	20.0	28.1
Plateau Toxicity Profile						
No. at true MTD	2.5	4.6	5.0	4.7	3.4	5.0
% at true MTD	15.0	19.1	20.9	19.6	14.2	20.6
Tub-shaped Toxicity Profile						
No. at true MTD	3.3	4.8	5.0	4.6	4.5	5.3
% at true MTD	18.9	19.9	20.9	19.2	18.8	21.9
Constant A Toxicity Profile						
No. at true MTD	2.1	4.4	5.2	4.4	2.8	5.0
% at true MTD	12.0	18.4	21.7	18.5	11.8	20.9
Constant B Toxicity Profile						
No. at true MTD	5.5	13.0	13.2	13.0	11.4	13.4
% at true MTD	34.3	54.1	55.1	54.1	47.4	55.8

Table 6
The average percentage of patients with DLT and average percentage of patients treated above the true MTD during the phase 1 study by phase 1 design.

Measure	Standard 3 + 3	CCD	BOIN	mTPI	mTPI-2	CRM
Linear A Toxicity Profile						
% with DLT	25.6	26.6	27.5	27.0	23.5	29.0
% above true MTD	13.4	17.2	20.0	18.0	11.9	28.1
Jump A Toxicity Profile						
% with DLT	25.8	25.2	26.1	25.5	23.9	27.2
% above true MTD	19.0	25.7	29.8	27.1	17.1	35.2
Jump B Toxicity Profile						
% with DLT	21.7	22.1	22.4	22.8	19.9	25.1
% above true MTD	29.1	31.0	31.9	32.6	27.1	36.6
Linear B Toxicity Profile						
% with DLT	21.1	22.6	23.3	22.8	19.8	25.3
% above true MTD	8.7	12.5	14.8	12.5	8.6	21.7
Plateau Toxicity Profile						
% with DLT	22.1	23.1	23.7	23.2	20.5	25.2
% above true MTD	7.0	10.0	12.5	10.5	6.8	18.8
Tub-shaped Toxicity Profile						
% with DLT	20.7	20.6	20.9	20.9	19.6	21.2
% above true MTD	17.1	26.5	28.6	26.8	17.4	31.4
Constant A Toxicity Profile						
% with DLT	22.1	20.1	20.1	20.0	20.5	20.0
% above true MTD	0.0	0.0	0.0	0.0	0.0	0.0
Constant B Toxicity Profile						
% with DLT	5.0	5.0	5.0	5.0	5.0	4.9
% above true MTD	0.0	0.0	0.0	0.0	0.0	0.0

However, results do support that dose escalation decision modifications used in the mTPI-2 design have made the mTPI safer for patients treated in the phase 1 study, where mTPI-2 is as safe or more safe than the conservative standard 3 + 3 design. The increased safety of mTPI-2 affects one or two patients enrolled in phase 1, but has the same risk of selecting a dose level above the true MTD as all of the other nonstandard designs and can result in a large number of patients treated at unsafe dose levels in subsequent phase 2 and 3 trials.

3.4. Overall success rate according to excessive toxicity rule

Since all nonstandard phase 1 designs overestimated the MTD more often than the standard 3 + 3 design, sometimes negating the benefit of higher MTD selection rates, simulations were repeated using different rules for declaring a dose level excessively toxic. The rule suggested across the literature [7,8,15,24,25] and the rule used in all previous simulations of this study was to declare a dose level too toxic if at any time the $\Pr(p_i > 0.30 \mid \text{data}) > 0.95$, where p_i was the probability of DLT at dose level *i*. As the posterior probability cutoff decreased, a dose level was more easily eliminated for excessive toxicity. Subsequently, there was less frequent selection of dose levels above the true MTD. Decreasing the posterior probability cutoff from 0.95 to 0.90 only led to a change in the excessive toxicity rule in the first cohort of 3 patients treated at a dose level or when at least 15 patients had been treated at a dose level (Table 7). For example, when 3 patients were treated at a dose level and the cutoff was 0.95, all 3 patients needed to have DLT to declare excessive toxicity, as opposed to only 2 patients if the cutoff was 0.90. When 6, 9, or 12 patients were treated at the same dose level, common sample sizes in a phase 1 trial, the rule for excessive toxicity was exactly the same. Hence, differences in simulation results were only apparent when using cutoff values of 0.95, 0.85, and 0.80.

Table 7
Number of patients with DLT required to claim excessive toxicity at a particular dose level when the target toxicity probability is 0.30. Results are presented for different posterior probability cutoffs.

Posterior Probability	Total Number of Patients at Current Dose Level							
	3	6	9	12	15	18	21	24
> 0.95	3	4	5	7	8	9	10	11
> 0.90	2	4	5	6	7	8	9	10
> 0.85	2	3	5	6	7	8	9	10
> 0.80	2	3	4	5	6	7	8	9

As the posterior probability cutoff in the excessive toxicity rule decreased, the overall success rates when using the nonstandard designs became more similar to the overall success rates when using the standard 3 + 3 design (Fig. 2). In scenarios where the nonstandard designs resulted in higher overall success rates compared with the standard 3 + 3 design (i.e., the Linear A, Linear B, Plateau, Constant A, and Constant B toxicity profiles), the favorable gap in the overall success rates decreased, although never to the point where benefit of the nonstandard designs was no longer observed. In the scenarios where the nonstandard designs resulted in lower or similar overall success rates compared with the standard 3 + 3 design (i.e., the Jump A, Jump B and Tub-shaped toxicity profiles), the overall success rates increased to the point where the nonstandard designs performed similarly to the standard 3 + 3 design. As the rule for excessive toxicity tightened, correct MTD selection rates and the distribution of dose levels selected as the MTD also became more similar between the nonstandard designs and the standard 3 + 3 design. This pattern is illustrated in Fig. 3, using the mTPI design as a representative nonstandard phase 1 design and toxicity profiles with the true MTD at dose level 2.

The redistribution in dose levels selected as the MTD when using nonstandard designs can also be achieved by changing the target toxicity probability. All nonstandard phase 1 designs implemented in the primary simulation study used a target toxicity probability of 0.30, just below the unacceptable toxicity probability of 0.33 defined when using the standard 3 + 3 design. However, the implicit target toxicity probability when using the standard 3 + 3 design is closer to 0.25 [23,24], and simulations were repeated using lower target toxicity probabilities equal to 0.25 and 0.20. Lowering the target toxicity probability resulted in more similar distributions of dose levels selected as the MTD between the nonstandard designs and the standard 3 + 3 design, but there remained higher selection rates of dose levels at and closely below the true MTD for most toxicity profiles when using nonstandard designs with the target toxicity probability as low as 0.25 (Fig. 4). A target toxicity probability equal to 0.20 often resulted in

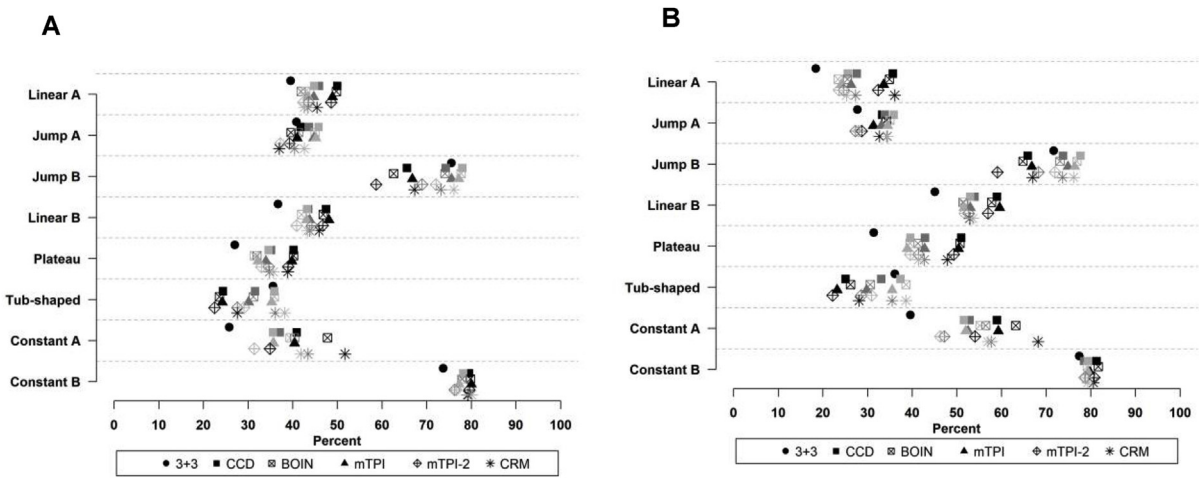

Fig. 2. Overall success rates for each phase 1 design, by toxicity profile, efficacy profile, and excessive toxicity rule. As the posterior probability cutoff used in the excessive toxicity rule for nonstandard designs decreases from 0.95 to 0.85 to 0.80 (darkest to lightest symbols), the safety control increases. As reference, overall success rates when using the standard 3 + 3 design are provided. A) Continuous efficacy profile and B) Step efficacy profile.

inferior correct MTD selection rates for nonstandard designs compared to the standard 3 + 3 design, since lowering the target toxicity probability not only impacted the excessive toxicity rule but also impacted the dosing decisions. For example, under the Constant A toxicity profile (where all dose levels had acceptable toxicity probabilities equal to the 0.20 target) the ability to escalate to the highest dose level was hindered by increased decisions to stay at the same, lower dose (Fig. 4C and F). The distribution of dose levels selected as the MTD for different target toxicity probabilities are shown for one of the interval designs, the mTPI, and the CRM in comparison to the standard 3 + 3.

3.5. Overall success rate according to fixed total sample size

All nonstandard phase 1 designs implemented in the primary simulation study used a fixed total sample size of 24 patients. This number corresponds to the maximum sample size across four escalation dose levels using the standard 3 + 3 design. In practice, when using a standard 3 + 3 design, the MTD is often estimated prior to when this number of patients is enrolled. In fact, in this simulation study, the MTD was estimated using the standard 3 + 3 design when the average number of patients was roughly 15. For this reason, simulation results are presented for the nonstandard designs using fixed total sample sizes of 24, 18, and 15 patients (Fig. 5).

In almost all scenarios, overall success rates for nonstandard designs decreased with decreasing sample size (Fig. 5). With the Linear A, Linear B, and Plateau toxicity profiles, which had gradually increasing toxicity across dose levels, and large gains in the overall success rates had been made with the nonstandard designs versus the standard 3 + 3 design, some benefit was retained with the smaller sample sizes.

With the Constant A and Constant B toxicity profiles, overall success rates decreased with smaller sample sizes, particularly when using the mTPI-2 design. With the mTPI-2 design, dose escalation with nearly no DLTs was required to reach the highest dose level. Any decision to incorrectly stay or de-escalate at a dose level became more difficult to overcome with increasingly fewer number of patients treated during the phase 1 study. The mTPI-2 design was most impacted by the decreasing sample size since it was developed to implement more cautious dose escalation decisions than the mTPI design. In fact, with a sample size of 15, the mTPI-2 design performed worse than the standard 3 + 3 design, even when the toxicity probability was as safe and low as 0.05 across dose levels as with the Constant B toxicity profile.

With the Jump A and Jump B toxicity profiles, characterized by a

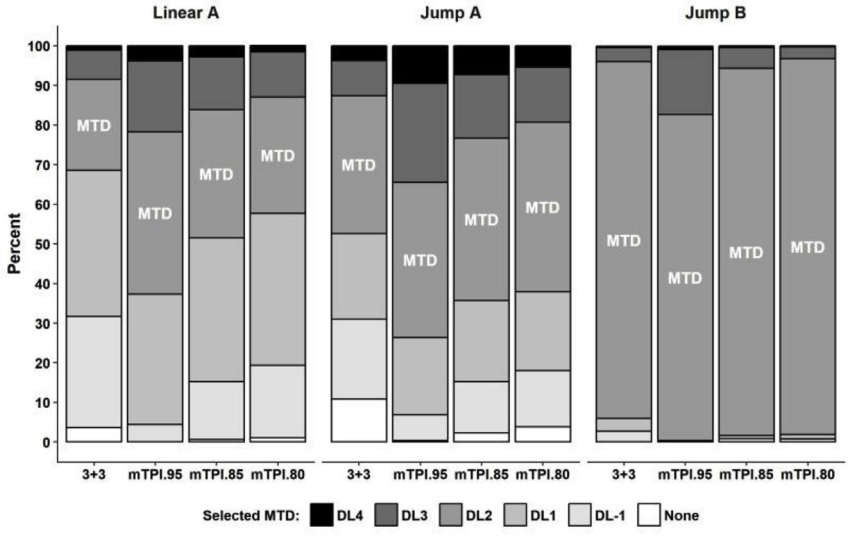

Fig. 3. Distribution of dose levels selected as the MTD when using the mTPI phase 1 design with posterior probability cutoff values of 0.95, 0.85, and 0.80 in the excessive toxicity rule. For reference, distributions of dose levels selected as the MTD when using the standard 3 + 3 design are provided.

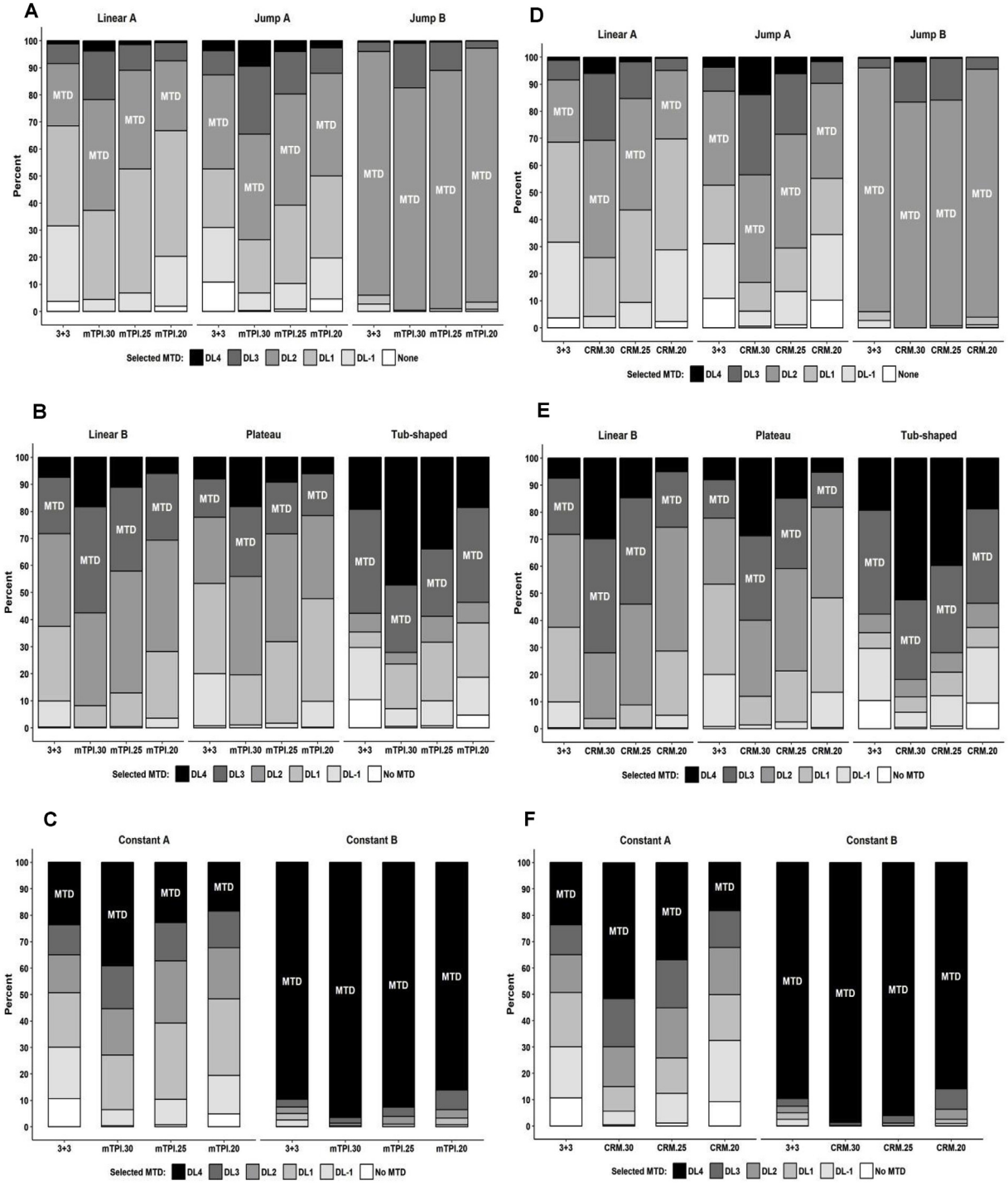

Fig. 4. Distribution of dose levels selected as the MTD when using the mTPI or CRM phase 1 designs with target toxicity probability values of 0.30, 0.25, and 0.20. For reference, distributions of dose levels selected as the MTD when using the standard 3 + 3 design are provided. A-C) mTPI and D-F) CRM.

jump in toxicity between dose levels 2 and 3, nonstandard designs performed increasingly worse than the standard 3 + 3 design with smaller sample sizes. Selection of dose levels above the true MTD was exacerbated by the smaller sample size, with too few patients remaining in the study to inform whether the highest dose levels were excessively toxic and led to lower overall success rates.

Only with the nonmonotonic Tub-shaped profile, where the MTD was exceeded at the highest dose level, was it beneficial to have too few patients to reach the highest dose level.

4. Discussion

Many factors affect the likelihood that a safe and effective drug will successfully progress from phase 1 through phase 3 of the clinical trial process. In this study, correct selection of the MTD in phase 1 drove the overall success rate, but underestimation and overestimation errors of the true MTD contribute to the overall success rate in a disproportionate manner. Underestimation error may or may not lead to discontinuation of drug development, depending on how grossly the true MTD is underestimated and the width of the therapeutic window. Overestimation error leads to additional patients treated at a dose level with

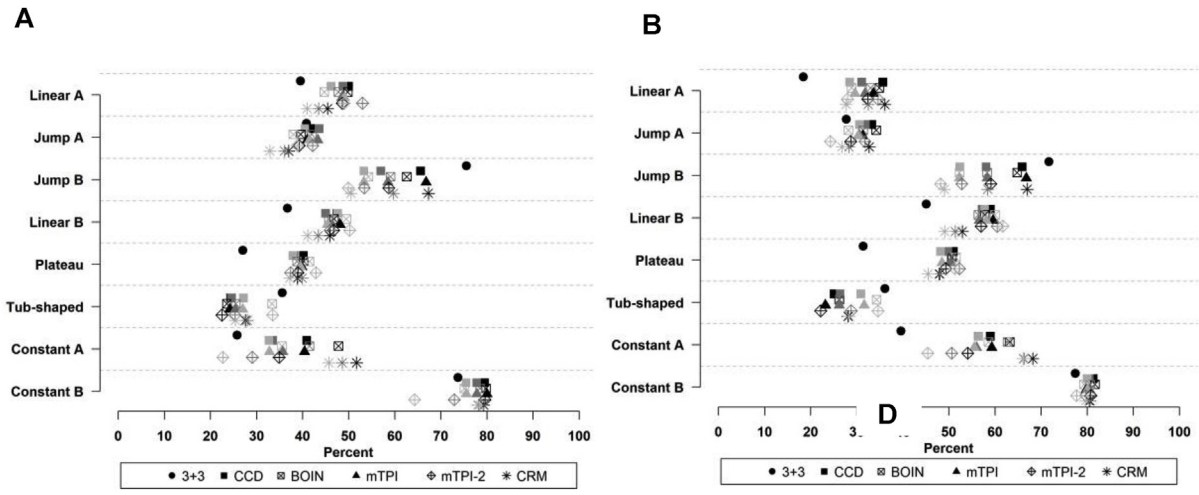

Fig. 5. Overall success rates for each phase 1 design, by toxicity profile, efficacy profile, and fixed total sample size. As the sample size decreases from 24 to 18 to 15 for nonstandard designs, the symbols go from darkest to lightest. As reference, overall success rates when using the standard 3 + 3 design are provided. A) Continuous efficacy profile and B) Step efficacy profile.

unacceptable toxicity, and eventually, drug development is terminated due to safety concerns. Thus, overestimation error is more grievous.

This simulation study showed that no one phase 1 design performed best for all toxicity profiles. However, specific phase 1 design features including the aggressiveness of dose escalation decisions, aggressiveness in estimating the MTD, and complexity related to the design and implementation of the phase 1 trial can be used to select among them when considering different clinical scenarios.

4.1. Phase 1 design selection guidelines

The following guidelines may be used to select a phase 1 design if there is confidence in the general shape of the dose-toxicity curve *a priori*, either from extensive preclinical studies or from other phase 1 studies performed with the same drug or class of drugs in different patient populations.

1) If gradually increasing toxicity among dose levels is anticipated, then use a nonstandard design other than the CRM.

The standard 3 + 3 design is conservative in its dosing decisions and in estimation of the MTD, leading to underestimation of the MTD in too many instances and inferior overall success rates. Among nonstandard designs, the CRM is most aggressive in its dosing decisions and MTD estimation, leading to the highest occurrences of selection of dose levels above the true MTD, percentage of patients with DLT, and percentage of patients treated above the MTD during phase 1. Among the CCD, BOIN, mTPI, and mTPI-2 designs, the mTPI-2 results in lower percentages of patients with DLT or treated above the true MTD during phase 1. However the mTPI-2 design, and by extension the mTPI, is logistically more complex and dosing decisions are not as intuitive when compared with the CCD or BOIN designs. Both the mTPI and mTPI-2 designs require specialized software to generate dosing decisions, whereas the CCD and BOIN designs do not. In addition, the dosing decisions using the mTPI and mTPI-2 designs are directly linked to the UPM as opposed to the observed DLT rate at a dose level. The UPM and the way in which it is calculated is unfamiliar to many statisticians and practitioners, whereas an estimated DLT rate is less abstract and more tangible. Collectively, the use of the CCD and BOIN designs is strongly encouraged.

2) If a jump in toxicity between adjacent dose levels is anticipated, then use the standard 3 + 3 phase 1 design.

The standard 3 + 3 design quickly recognizes when there is escalation from a dose level with toxicity below the target probability to a dose level with toxicity significantly higher than the target probability. In this context, using a nonstandard design is discouraged.

3) If a fairly constant and safe toxicity profile is anticipated, then use nonstandard BOIN or CRM designs.

The highest dose level is reached most often using the BOIN and CRM designs, and the BOIN achieves almost as high overall success rates as the CRM across scenarios. However, unlike the CRM, the BOIN design is simple and easy to implement, with rules similar to the standard 3 + 3 design, and does not require specialized software to generate dosing decisions.

Among interval designs, the CCD, mTPI, and mTPI-2 have less aggressive dose escalation rules than the BOIN and fails to reach the highest dose level as frequently. This problem is exacerbated the closer the true toxicity is to the target toxicity probability. If the true toxicity is above the lower limit of the proper dosing interval, then all interval designs will have difficulty escalating to the highest dose level. The standard 3 + 3 design reaches the highest dose level least frequently and has the lowest overall success rates. The standard 3 + 3 design also has the highest discontinuation rates when the true toxicity is closer to the target toxicity probability.

4.2. Impact of sample size on phase 1 design selection

Sample size impacts the MTD selection and overall success rates, sometimes to a great degree. The primary simulation study used a fixed sample size of 24 for all nonstandard designs, corresponding to the maximum sample size across four escalation dose levels using the standard 3 + 3 design. Ji and Wang [15] recommended using a sample size equal to the number of patients needed to escalate to the highest dose level in the absence of DLT plus one additional cohort of patients. Using this recommendation, which translates to 15 patients in this simulation study, leads to substantial decreases in the MTD selection and overall success rates for all toxicity profiles except those with gradually increasing toxicity. The nonstandard design most impacted by the smaller sample size is the mTPI-2, sometimes with decreases in correct selection of the MTD that the overall success rates are lower than when using the standard 3 + 3 design.

In scenarios where the standard 3 + 3 design already outperformed the nonstandard designs, a decrease in sample size for nonstandard designs only results in even lower MTD selection and overall success

rates. Thus, when using a nonstandard design, it is recommended to use a fixed sample size equal to the maximum sample size across escalation dose levels when using a standard 3 + 3 design. If a smaller sample size is used, simulation studies justifying the choice of sample size should be presented.

4.3. Impact of safety rule on phase 1 design selection

The safety rule used to eliminate dose levels with excessive toxicity also impacts the MTD selection and overall success rates. The dose elimination rule is based on the posterior probability that toxicity at a dose level is greater than the target probability. A posterior probability cutoff suggested across the literature is 0.95 [7,8,15,24,25]. However, this high cutoff value allows for selection of dose levels above the true MTD too frequently. When there is a jump in toxicity between adjacent dose levels about the true MTD, the high cutoff value leads to lower MTD selection and overall success rates for nonstandard designs compared to the standard 3 + 3 design.

By lowering the cutoff value, less evidence is required to exclude a dose level for excessive toxicity. With the Jump A and Jump B toxicity profiles, when the cutoff is lowered, the MTD selection and overall success rates increase when using the nonstandard designs and become more similar to the favorable rates observed when using the standard 3 + 3 design. With other toxicity profiles where nonstandard designs clearly outperform the standard 3 + 3 design, the MTD selection and overall success rates decrease when the cutoff value is as low as 0.80, but the MTD selection and overall success rates do not become as low as when using the standard 3 + 3 design.

Collectively, greater balance in errors are obtained across toxicity profiles when using nonstandard designs if the cutoff value for excessive toxicity is less than 0.95. To provide a safeguard against selecting a dose level above the true MTD while still maintaining better operating characteristics for nonstandard designs, a cutoff value of 0.85 is recommended.

The excessive toxicity rule can also be modified by changing the target toxicity probability. If the target toxicity probability is lowered from 0.30 to 0.25 or 0.20, then less evidence is required to exclude a dose level for excessive toxicity. However, dosing decisions are also impacted and there is less ability to escalate when the true toxicity is at or close to the target toxicity probability.

In this simulation study, the true MTD was defined as the highest dose level with DLT probability below 0.33. Selection of a dose level with corresponding DLT probability at or above 0.33 was therefore considered an overestimation error, even in the setting of interval designs in which the proper dosing interval included the value 0.33; the proper dosing interval was used as a tool to guide dosing decisions, but not to define multiple dose levels that could be considered the true MTD. If determined *a priori* that dose levels with a DLT probability within a certain distance above the target toxicity probability would be acceptable, then the MTD overestimation error described using the Tub-shaped or Constant A toxicity profiles with the nonstandard designs may not be highly concerning. However, the MTD overestimation error described using the Jump B toxicity profile with the nonstandard designs persists unless an adjustment is made to the excessive toxicity rule.

5. Conclusion

This study elucidated the downstream impact of design decisions typically made during phase 1 dose-finding trials when the number of dose levels considered is moderate in size, a setting in which design choice remained unclear. We have shown that in an attempt to better estimate the MTD, nonstandard phase 1 designs incorporate more aggressive dose escalation decisions and more aggressive estimation of the MTD, to varying degrees, compared to the standard 3 + 3 design. Under some toxicity profiles, this aggressiveness results in higher overall success rates of a safe and favorable drug. However, under toxicity profiles with a jump in toxicity between adjacent dose levels, this aggressiveness results in frequent selection of dose levels above the true MTD and outweighs the benefit of using nonstandard designs.

If there is a strong *a priori* belief in the general shape of the dose-toxicity curve, a phase 1 design can be selected among the standard 3 + 3 or nonstandard designs that will have a high likelihood of performing well. If there is not a strong *a priori* belief in the shape of the dose-toxicity curve, adjustments can be made to the excessive toxicity rule used in nonstandard designs to limit the negative downstream effects of MTD overestimation error, while retaining the positive benefit of better estimating the true MTD.

The increased sample size required to conduct a nonstandard phase 1 design compared to the standard 3 + 3 design is small. Although all designs require specialized software to generate operating characteristics for various scenarios, and the nonstandard designs may require a probability calculator to define and implement an excessive toxicity rule, only the CCD and BOIN do not require specialized software to generate dosing decisions among the nonstandard designs evaluated. Due to the ease of implementation and good operating characteristics, the CCD and BOIN should be strongly considered when designing a phase 1 trial in oncology.

Disclosure of conflicts of interest

The authors declare no competing financial interests.

Acknowledgement

The authors would like to thank Dr. John Byrd for providing input on the dose-toxicity profiles included in the clinical scenarios.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.conctc.2018.08.010.

References

[1] B.E. Storer, Design and analysis of phase 1 clinical trials, Biometrics 45 (3) (1989) 925–937.
[2] J. Shtaynberger, S. Lee, J. Duong, C. Chiuzan, What Is the Most Popular Dose-finding Design for Phase 1 Oncology Trials? Annual Society for Clinical Trials Meeting Abstract, Arlington, Virginia, 2015.
[3] Y. Yuan, K.R. Hess, S.G. Hilsenbeck, M.R. Gilbert, Bayesian optimal interval design: a simple and well-performing design for phase 1 oncology trials, Clin. Canc. Res. 22 (17) (2016) 4291–4301.
[4] J. O'Quigley, M. Pepe, L. Fisher, Continual reassessment method: a practical design for phase 1 clinical trials in cancer, Biometrics 46 (1) (1990) 33–48.
[5] A. Ivanova, N. Flournoy, Y. Chung, Cumulative cohort design for dose-finding, J. Stat. Plann. Inference 137 (7) (2007) 2316–2327.
[6] Y. Ji, P. Liu, Y. Li, B.N. Bekele, A modified toxicity probability interval method for dose-finding trials, Clin. Trials 7 (6) (2010) 653–663.
[7] S. Liu, Y. Yuan, Bayesian optimal interval designs for phase 1 clinical trials, J. Roy. Stat. Soc.: Series C (Applied Statistics) 64 (3) (2015) 507–523.
[8] W. Guo, S.-J. Wang, S. Yang, H. Lynn, Y. Ji, A Bayesian interval dose-finding design addressing Ockham's razor: mTPI-2, Contemp. Clin. Trials 58 (2017) 23–33.
[9] B.J. Horton, N.A. Wages, M.R. Conaway, Performance of toxicity probability interval based designs in contrast to the continual reassessment method, Stat. Med. 36 (2) (2017) 291–300.
[10] R. Simon, Optimal two-stage designs for phase 2 clinical trials, Contr. Clin. Trials 10 (1) (1989) 1–10.
[11] S.S. Emerson, D.L. Gillen, J.K. Kittelson, S.C. Emerson, G.P. Levin, RCTdesign: Group Sequential Trial Design, (2012).
[12] A. Hoering, M. LeBlanc, J. Crowley, Seamless phase 1-2 trial design for assessing toxicity and efficacy for targeted agents, Clin. Canc. Res. 17 (4) (2011) 640–646.
[13] X. Huang, S. Biswas, Y. Oki, J.-P. Issa, D.A. Berry, A parallel phase 1/2 clinical trial design for combination therapies, Biometrics 63 (2) (2007) 429–436.
[14] Y. Ji, L. Feng, P. Liu, E.J. Shpall, P. Kebriaei, R. Champlin, et al., Bayesian continual reassessment method for dose-finding trials infusing T cells with limited sample size, J. Biopharm. Stat. 22 (6) (2012) 1206–1219.
[15] Y. Ji, S.-J. Wang, Modified toxicity probability interval design: a safer and more reliable method than the 3 + 3 design for practical phase 1 trials, J. Clin. Oncol. 31 (14) (2013) 1785–1791.

[16] J.R. Brown, J.C. Byrd, S.E. Coutre, D.M. Benson, I.W. Flinn, N.D. Wagner-Johnston, et al., Idelalisib, an inhibitor of phosphatidylinositol 3-kinase p110, for relapsed/refractory chronic lymphocytic leukemia, Blood 123 (22) (2014) 3390–3397.
[17] J. Menis, S. Litiere, K. Tryfonidis, V. Golfinopoulos, The European Organization for Research and Treatment of Cancer perspective on designing clinical trials with immune therapeutics, Ann. Transl. Med. 4 (14) (2016) 267–278.
[18] Y. Zang, J.J. Lee, Y. Yuan, Adaptive designs for identifying optimal biological dose for molecularly targeted agents, Clin. Trials 11 (3) (2014) 319–327.
[19] P.F. Thall, J.D. Cook, E.H. Estey, Adaptive dose selection using efficacy-toxicity trade-offs: illustrations and practical considerations, J. Biopharm. Stat. 16 (5) (2006) 623–638.
[20] R Core Team, R: a Language and Environment for Statistical Computing. Vienna, Austria, (2015) Retrieved from http://www.R-project.org/ , Accessed date: 13 February 2018.
[21] R. Turner, Iso: Functions to Perform Isotonic Regression [Computer Software Manual], (2015) Retrieved from http://CRAN.R-project.org/package=Iso , Accessed date: 13 February 2018.
[22] M. Sweeting, A. Mander, T. Sabin, bcrm: Bayesian continual reassessment method designs for phase 1 dose-finding trials, J. Stat. Software 54 (13) (2013) 1–26.
[23] E. Reiner, X. Paoletti, J. O'Quigley, Operating characteristics of the standard phase 1 clinical trial design, Comput. Stat. Data Anal. 30 (3) (1999) 303–315.
[24] S. Liu, C. Cai, J. Ning, Up-and-down designs for phase 1 clinical trials, Contemp. Clin. Trials 36 (1) (2013) 218–227.
[25] H. Pan, F. Xie, P. Liu, J. Xia, Y. Ji, A phase 1/2 seamless dose escalation/expansion with adaptive randomization scheme (SEARS), Clin. Trials 11 (1) (2014) 49–59.