Received: 13 September 2016 | Revised: 3 March 2018 | Accepted: 5 March 2018

DOI: 10.1002/sim.7674

RESEARCH ARTICLE

WILEY Statistics in Medicine

Comparative review of novel model-assisted designs for phase I clinical trials

Heng Zhou[1] | Thomas A. Murray[2] | Haitao Pan[3] | Ying Yuan[1]

[1]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

[2]Department of Biostatistics, The University of Minnesota, Minneapolis, MN, USA

[3]Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, TN, USA

Correspondence
Ying Yuan, Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA.
Email: yyuan@mdanderson.org

A number of novel phase I trial designs have been proposed that aim to combine the simplicity of algorithm-based designs with the superior performance of model-based designs, including the modified toxicity probability interval, Bayesian optimal interval, and Keyboard designs. In this article, we review these "model-assisted" designs, contrast their statistical foundations and pros and cons, and compare their operating characteristics with the continual reassessment method. To provide unbiased and reliable results, our comparison is based on 10 000 dose-toxicity scenarios randomly generated using the pseudo-uniform algorithm recently proposed in the literature. The results showed that the continual reassessment method, Bayesian optimal interval, and Keyboard designs provide comparable, superior operating characteristics, and each outperforms the modified toxicity probability interval design. These designs are more likely to correctly select the maximum tolerated dose and less likely to overdose patients.

KEYWORDS
dose finding, interval design, maximum tolerated dose, model-assisted design, toxicity probability interval

1 | INTRODUCTION

Phase I clinical trial designs aim to identify the maximum tolerated dose (MTD) of a new drug, which is defined as the dose with a dose-limiting toxicity (DLT) probability that is closest to the target probability. Traditionally, phase I dose-finding designs can generally be classified as algorithm based and model based.[1] Algorithm-based designs use simple, prespecified rules to govern dose escalation and de-escalation. Examples include the 3+3 design, the "rolling-six" design,[2] the biased-coin design,[3] and its variations.[4,5] Despite that the 3+3 design poorly identifies the MTD,[1] it is the most common phase I trial design used in practice, mainly because of its transparency and simplicity.

Model-based designs have been proposed that improve upon the performance of algorithm-based designs. The most well-known model-based design is the continual reassessment method (CRM).[6] The CRM begins with a prior dose-toxicity curve and continuously updates this curve based on the accruing toxicity outcomes from patients in the trial. Each new patient or cohort of patients is assigned to the dose that corresponds to the estimated DLT probability closest to the prespecified target, where the estimated DLT probabilities are derived from the updated dose-toxicity curve. Various extensions of the CRM have been proposed, including dose escalation with overdose control (EWOC),[7] time-to-event CRM,[8] Bayesian model averaging CRM,[9] Bayesian data-augmentation CRM,[10] partial order CRM,[11] and bivariate CRM.[12] Cheung provides a comprehensive review of the CRM and its related methods.[13] Compared with algorithm-based designs, model-based designs typically have superior operating characteristics.

However, because model-based designs require repeated model fitting and estimation, many practitioners view them as conceptually and computationally complex, as though the decisions are coming from a "black box." This perception likely has limited the use of model-based designs, such as the CRM, in practice.

Recently, a new class of designs, known as model-assisted designs,[14] have been proposed to combine the simplicity of algorithm-based designs with the superior performance of model-based designs. Model-assisted designs use a model for efficient decision making like model-based designs, whereas their dose escalation and de-escalation rules can be tabulated before the onset of a trial as with algorithm-based designs. Unlike the model-based design, such as the CRM, which assumes a dose-toxicity curve across all doses, the model-assisted design often models only local data (ie, the data observed at the current dose), typically using a binomial model, which renders it possible to enumerate the dose escalation and de-escalation rules before the trial begins. Examples of model-assisted designs include the modified toxicity probability interval (mTPI) design[15] and its variation mTPI-2,[16] Bayesian optimal interval (BOIN) design,[17,18] Keyboard design,[14] BOIN combination design[19] and phase I/II design, [20,21] and Keyboard combination design.[22] Recently, Mu et al[23] proposed a generalized BOIN design that handles toxicity grades, binary or continuous toxicity endpoints under a unified framework.

In this manuscript, we review several model-assisted designs, including the mTPI, BOIN, and Keyboard designs and compare their operating characteristics to the CRM. The mTPI-2 ends up with the same design as the Keyboard design but is less transparent and relies on perplexing statistical concepts and methods (eg, Ockham's razor). Thus, here we only present the Keyboard design, whose results apply to the mTPI-2 as well. Several simulation studies were carried out to compare the operating characteristics of novel phase I designs but based on a limited number of dose-toxicity scenarios (or curves). For example, Horton, Wages, and Conaway[24] compared the CRM, mTPI, and BOIN designs in a simulation study with 16 dose-toxicity scenarios, and Ananthakrishnan et al[25] considered only 3 dose-toxicity scenarios with the MTD at the same dose level. As a result, these simulation studies may produce the results that do not represent the general performance of the designs. In this article, we conduct a large scale simulation study that includes 10 000 dose-toxicity scenarios. These 10 000 dose-toxicity scenarios are randomly generated using a new pseudo-uniform algorithm, recently proposed by Clertant and O'Quigley.[26] Because a priori that algorithm does not favor any particular dose as the MTD or a particular shape of the dose-toxicity curve, it provides a neutral and objective basis for comparison.

The remainder of the manuscript is as follows. In Section 2, we review the the CRM, mTPI, BOIN, and Keyboard designs. In Section 3, we describe and report the results of our simulation study. In Section 4, we conclude with some further discussion of our findings.

2 | METHODS

Before reviewing the designs, we establish notation. We use $d_1 < \cdots < d_J$ to denote the J prespecified doses of the new drug that is under investigation in the trial, p_j to denote the DLT probability that corresponds to d_j, and ϕ to denote the target DLT probability for the MTD. We use n_j to denote the number of patients who have been assigned to d_j, and y_j to denote the number of DLTs observed at d_j, $j = 1, ..., J$. Therefore, at a particular point during the trial, the observed data are $D = \{D_j, j = 1, ..., J\}$, where $D_j = (n_j, y_j)$ are the "local" data observed at dose level j. For completeness and illustrating the differences between model-based and model-assisted designs, we first briefly describe the CRM, followed by the mTPI, Keyboard, and BOIN designs.

2.1 | Continual reassessment method

The CRM is a model-based dose-finding approach that assumes a parametric model for the dose-toxicity curve. As information accrues during the trial, the dose-toxicity curve is reevaluated by updating the estimates of the unknown model parameters and the corresponding DLT probability at each investigational dose. The current estimates for the DLT probabilities are used to determine the dose allocation for the next patient or cohort of patients. One commonly used model for the CRM is the power model (also known as the empiric model) that assumes

$$p_j(\alpha) = a_j^{\exp(\alpha)}, \quad \text{for} \quad j = 1, \cdots, J, \tag{1}$$

where α is the unknown parameter and $0 < a_1 < \cdots < a_J < 1$ are prior guesses for the DLT probability at each dose. The $\{a_j, j = 1, ..., J\}$ often are called the "skeleton" of CRM.

Under the power model in (1), the likelihood function for α

$$L(\alpha|D) = \prod_{j=1}^{J} \left\{ a_j^{\exp(\alpha)} \right\}^{y_j} \left\{ 1 - a_j^{\exp(\alpha)} \right\}^{n_j - y_j},$$

and thus, the posterior mean estimate for p_j is calculated as

$$\hat{p}_j = \int a_j^{\exp(\alpha)} \frac{L(\alpha|D) f(\alpha)}{\int L(\alpha|D) f(\alpha)\, d\alpha} d\alpha,$$

where $f(\alpha)$ denotes the prior distribution for α, eg, $N(0, 2)$. Upon updating the posterior mean estimate of the DLT probability at each dose, the next patient or cohort of patients is assigned to the dose with an estimated DLT probability closest to the target ϕ. That is, the next patient or cohort of patients is assigned to dose level j^* such that

$$j^* = \underset{j \in (1, \cdots, J)}{\operatorname{argmin}} |\hat{p}_j - \phi|.$$

The trial continues in this manner until the prespecified sample size is exhausted. At that point, the MTD is selected as the dose with an estimated DLT probability closest to the target ϕ. In practice, typically dose escalation and de-escalation is restricted to one level at a time, and a safety stopping rule is included such that the trial is terminated if $\Pr(p_1 > \phi \mid D) > 0.9$ (ie, the lowest dose d_1 has more than 90% chance of being above the MTD). We imposed these practical rules for the CRM in Section 4.

2.2 | mTPI design

The mTPI design requires the investigator to prespecify 3 intervals, the underdosing interval $(0, \delta_1)$, the proper dosing interval (δ_1, δ_2), and the overdosing interval $(\delta_2, 1)$. For example, given a target rate of $\phi = 0.20$, the 3 intervals may be defined as (0, 0.15), (0.15, 0.25), and (0.25,1), respectively. The mTPI design assumes

$$\begin{aligned} y_j|n_j, p_j &\sim \text{Binom}\,(n_j, p_j) \\ p_j &\sim \text{Beta}\,(1,1) \equiv \text{Unif}\,(0,1), \end{aligned} \tag{2}$$

ie, a beta-binomial model, and thus, the posterior distribution arises as

$$p_j|D_j \sim \text{Beta}\,(y_j + 1, n_j - y_j + 1), \text{ for } j = 1, \cdots, J. \tag{3}$$

Unlike the CRM, which models the toxicity across doses using the power model (1), the mTPI models toxicity only at the current dose d_j. To determine the next dose, based on D_j, the mTPI design uses the unit probability mass (UPM) corresponding to each of the 3 intervals, which are defined as

$$\begin{aligned} \text{UPM1} &= \Pr(p_j \in (0, \delta_1)|D_j)/\delta_1, \\ \text{UPM2} &= \Pr(p_j \in (\delta_1, \delta_2)|D_j)/(\delta_2 - \delta_1), \\ \text{UPM3} &= \Pr(p_j \in (\delta_2, 1)|D_j)/(1 - \delta_2). \end{aligned} \tag{4}$$

That is, the UPM is the posterior probability that p_j lies in the corresponding interval divided by the length of that interval.

Suppose j is the current dose level. The mTPI design determines the next dose as follows:

- If $\text{UPM1} = \max\{\text{UPM1}, \text{UPM2}, \text{UPM3}\}$, then escalate the dose to level $j + 1$.
- If $\text{UPM2} = \max\{\text{UPM1}, \text{UPM2}, \text{UPM3}\}$, then stay at the current dose level j.
- If $\text{UPM3} = \max\{\text{UPM1}, \text{UPM2}, \text{UPM3}\}$, then de-escalate the dose to level $j - 1$.

Because the 3 UPMs can be determined for all possible outcomes $D_j = (n_j, y_j)$, the dose escalation and de-escalation rules can be tabulated before the trial begins, which makes the mTPI design easy to implement in practice. The trial

continues until the prespecified sample size is exhausted. At that point, the MTD is selected based on isotonic estimates of the p_j that are calculated using the pooled adjacent violators algorithm.[27] As the decision of dose escalation and de-escalation is based only on the local data at the current dose, it is possible that the dose transition oscillates between a safe dose and the next higher dose that is toxic. To avoid that issue, the mTPI design includes a dose exclusion/safety stopping rule: if $\Pr(p_j > \phi \mid n_j, y_j) > 0.95$, dose level $j$and higher are excluded from the trial. If the lowest dose is excluded, the trial is stopped for safety.

One drawback of using the UPM to guide dose escalation is that it lacks clear interpretation and leads to a high risk of overdosing patients.[14] To see the problem, consider a trial with a target toxicity rate of 0.20 and underdosing, proper dosing, and overdosing intervals of (0, 0.17), (0.17, 0.23), and (0.23, 1), respectively. Suppose 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. That is, there is a 90% chance that the current dose is overdosing patients and only a 9% chance that the current dose is properly dosing patients. Despite such dominant evidence of overdosing, the mTPI design stays the same dose for treating the next patient or patient cohort, since the UPM that corresponds to the proper dosing interval is the largest. In particular, the UPM that corresponds to the proper dosing interval is $0.09/(0.23-0.17) = 1.5$, whereas the UPM that corresponds to the overdosing interval is $0.9/(1 - 0.23) = 1.17$.

2.3 | Keyboard design

The Keyboard design[14] resolves the overdosing issue of the mTPI by defining a series of equal-width dosing intervals (or keys) that correspond to the potential locations of the true DLT probability of a particular dose and using the interval (or key) with the highest posterior probability to guide dose escalation and de-escalation; see Figure 1B. Specifically, the Keyboard design starts by specifying a proper dosing interval $\mathcal{I}^* = (\delta_1, \delta_2)$, referred to as the "target key," and then populates this interval toward both sides of the target key, forming a series of keys of equal width that span the range of 0 to 1. For example, given the proper dosing interval or target key of (0.25, 0.35), on its left side, we form 2 keys of width 0.1, ie, (0.15, 0.25) and (0.05, 0.15), and on its right side, we form 6 keys of width 0.1, ie, (0.35, 0.45), (0.45, 0.55), (0.55, 0.65), (0.65, 0.75), (0.75, 0.85), and (0.85, 0.95). We denote the resulting intervals/keys as $\mathcal{I}_1, \cdots, \mathcal{I}_K$. As all keys have the equal width and must be within [0, 1], some DLT probability values at the two ends (eg, < 0.05 or > 0.95 in the example) may not be covered by keys because they are not long enough to form a key. As explained in Yan et al,[14] ignoring these "residual" DLT probabilities at the two ends does not pose any issue for decision making of dose escalation and de-escalation.

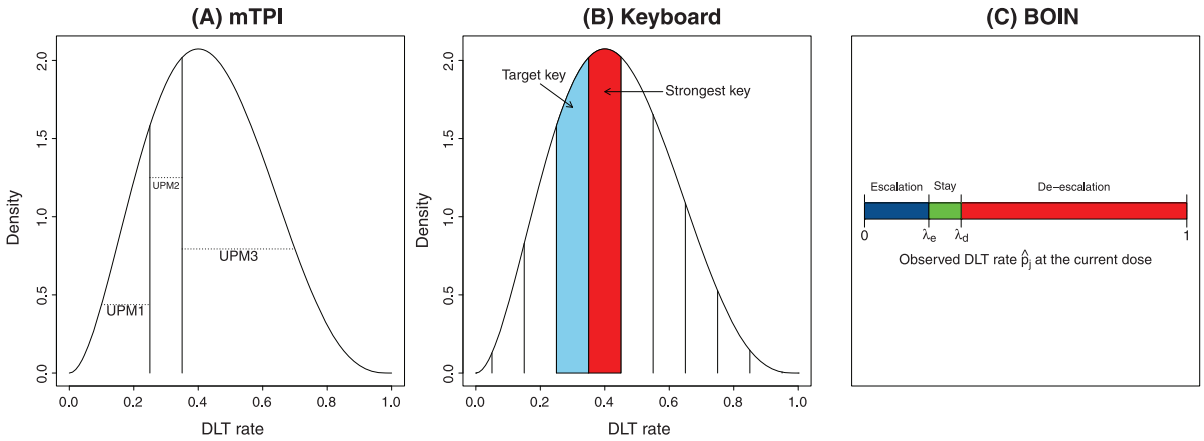

FIGURE 1 Illustration of (A) the modified toxicity probability interval (mTPI) design, (B) the Keyboard design, and (C) the Bayesian optimal interval (BOIN) design. The curves are the posterior distributions of p_j. To determine the next dose, the mTPI design compares the values of the 3 unit probability masses (UPMs), whereas the Keyboard design compares the location of the strongest key with respect to the target key. Bayesian optimal interval compares the observed dose-limiting toxicity (DLT) rate at the current dose with the prespecified dose escalation boundary λ_e and de-escalation boundary λ_d [Colour figure can be viewed at wileyonlinelibrary.com]

To make the decision of dose escalation and de-escalation, given the observed data $D_j = (n_j, y_j)$ at the current dose level j, the Keyboard design identifies the interval $\mathcal{I}_{\max}$ that has the largest posterior probability, ie,

$$\mathcal{I}_{\max} = \underset{\mathcal{I}_1, \cdots, \mathcal{I}_K}{\operatorname{argmax}} \{\Pr(p_j \in \mathcal{I}_k | D_j);\ k = 1, \cdots, K\},$$

which can easily be evaluated based on p_j's posterior distribution given by Equation 3, assuming that p_j follows a beta-binomial model (2). $\mathcal{I}_{\max}$ represents the interval that the true value of p_j is most likely located, referred to as the "strongest" key by Yan et al.[14] Graphically, the strongest key is the one with the largest area under the posterior distribution curve of p_j (see Figure 1B). If the strongest key is on the left (or right) side of the target key, that means that the observed data suggest that the current dose is most likely underdosing (or overdosing), and thus, dose escalation (or de-escalation) is needed. If the strongest key is the target key, the observed data support that the current dose is most likely to be in the proper dosing interval, and thus, it is desirable to retain the current dose for treating the next patient. In contrast, the UPM used by the mTPI design does not have such an intuitive interpretation and tends to distort the evidence for overdosing, as described previously.

Suppose j is the current dose level. The Keyboard design determines the next dose as follows:

- If the strongest key is on the left side of the target key, then escalate the dose to level $j + 1$.
- If the strongest key is the target key, then stay the current dose level j.
- If the strongest key is on the right side of the target key, then de-escalate the dose to level $j - 1$.

The trial continues until the prespecified sample size is exhausted, and the MTD is selected based on isotonic estimates of p_j as described previously. During the trial conduct, the Keyboard design imposes the dose exclusion/early stopping rule such that if $\Pr(p_j > \phi \mid n_j, y_j) > 0.95$ and $n_j \geq 3$, dose level j and higher are eliminated from the trial, and the trial is terminated if the lowest dose is eliminated, where $\Pr(p_j > \phi \mid n_j, y_j)$ is evaluated based on the posterior distribution (3).

Similar to the mTPI design, the dose escalation and de-escalation rules of the Keyboard design can be tabulated before the trial begins, making it easy to implement in practice. As the location of the strongest key approximately indicates the mode of the posterior distribution of p_j, the Keyboard design can be approximately viewed as a posterior-mode-based Bayesian dose-finding method. This makes the Keyboard design a new method different from the UPM-based mTPI design, despite some structural similarities between 2 designs (eg, partitioning the toxicity probability into intervals and the dose escalation and de-escalation rules can be pretabulated). Pan et al[22] showed that the Keyboard design is optimal under the 0-1 loss, long-memory coherent and extended it to drug-combination trials.

2.4 | BOIN design

Compared with the mTPI and Keyboard designs, the BOIN design is more straightforward and transparent. The dose escalation and de-escalation in the BOIN design is determined simply by comparing the observed DLT rate at the current dose with a pair of fixed dose escalation and de-escalation boundaries. Specifically, let $\hat{p}_j = y_j/n_j$ denote the observed DLT rate at the current dose, and λ_e and λ_d denote the predetermined dose escalation and de-escalation boundaries. Suppose j is the current dose level. The BOIN design determines the next dose as follows (see Figure 1C):

- If $\hat{p}_j \leq \lambda_e$, then escalate the dose to level $j + 1$.
- If $\hat{p}_j \geq \lambda_d$, then de-escalate the dose to level $j - 1$.
- Otherwise (ie, $\lambda_e < \hat{p}_j < \lambda_d$), stay at the current dose level j.

The trial continues until the prespecified sample size is exhausted. At that point, select the MTD based on the isotonic estimates of DLT probabilities as described previously. During the trial conduct, the BOIN design imposes a dose elimination (or overdose control) rule as follows: if $\Pr(p_j > \phi \mid n_j, y_j) > 0.95$ and $n_j \geq 3$, dose level $j$and higher are eliminated from the trial, and the trial is terminated if the lowest dose is eliminated, where $\Pr(p_j > \phi \mid n_j, y_j)$ is evaluated based on the posterior distribution (3).

To determine the dose escalation and de-escalation boundaries (λ_e, λ_d), the BOIN design requires the investigator(s) to specify ϕ_1, which is the highest DLT probability that is deemed to be underdosing such that dose escalation is required, and ϕ_2, which is the lowest DLT probability that is deemed to be overdosing such that dose de-escalation is required. Liu and Yuan[17] provided general guidance to specify ϕ_1 and ϕ_2 and recommended default values of $\phi_1 = 0.6\phi$ and $\phi_2 = 1.4\phi$ for general use. When needed, the values of ϕ_1 and ϕ_2 can be calibrated to achieve a particular requirement of the trial at hand. For example, if more conservative dose escalation is required, setting $\phi_2 = 1.2\phi$ may be appropriate. Given ϕ_1 and ϕ_2 and assuming a noninformative prior (ie, a priori the current dose is equally likely to be below, equal to, or above the MTD), the optimal escalation and de-escalation boundaries (λ_e, λ_d) that minimizes the decision error of dose escalation and de-escalation arise as

$$\lambda_e = \frac{\log\left(\frac{1-\phi_1}{1-\phi}\right)}{\log\left\{\frac{\phi(1-\phi_1)}{\phi_1(1-\phi)}\right\}}, \qquad \lambda_d = \frac{\log\left(\frac{1-\phi}{1-\phi_2}\right)}{\log\left\{\frac{\phi_2(1-\phi)}{\phi(1-\phi_2)}\right\}}. \tag{5}$$

The following table provides the dose escalation and de-escalation boundaries (λ_e, λ_d) for commonly used target DLT rate ϕ using the recommended default values $\phi_1 = 0.6\phi$ and $\phi_2 = 1.4\phi$. For example, given the target DLT rate $\phi = 0.25$, the corresponding escalation boundary $\lambda_e = 0.197$ and the de-escalation boundary $\lambda_d = 0.298$, that is, escalate the dose if the observed DLT rate at the current dose $\hat{p}_j \leq 0.197$ and de-escalate the dose if $\hat{p}_j \geq 0.298$. It has been shown that λ_e and λ_d are the boundaries corresponding to the Bayes factors, and thus, the resulting BOIN design is optimal with desirable finite-sample and large-sample properties, ie, long-memory coherence and consistency.[17]

	Target DLT Rate ϕ					
Boundaries	**0.15**	**0.2**	**0.25**	**0.3**	**0.35**	**0.4**
λ_e	0.118	0.157	0.197	0.236	0.276	0.316
λ_d	0.179	0.238	0.298	0.358	0.419	0.479

One interesting note is that the decision rule of the BOIN (with the noninformative prior) has an appearance of the classical frequentist design and only involves the observed DLT rate. This is common in Bayesian statistics. Many well-established Bayesian methods (eg, estimation for normal linear regression models) result in the same estimators as the frequentist approach when noninformative priors are used. Actually, the BOIN can also be derived as a frequentist design, and its decision rule is equivalent to using the likelihood ratio test to determine dose escalation/de-escalation,[17] providing another way to prove its optimality. Having both Bayesian and frequentist interpretations is a strength of the BOIN, making it appealing to wider audiences. In contrast, the mTPI and Keyboard designs only have a Bayesian interpretation and require specifying priors and calculating posterior distributions.

As the observed DLT rate $\hat{p}_j$ is the most natural and intuitive estimate of p_j that is accessible by nonstatisticians, the use of $\hat{p}_j$ to determine the dose escalation and de-escalation makes the BOIN design simpler and more transparent than the mTPI/mTPI-2 and Keyboard designs. It is particularly easy for clinicians and regulatory agents to assess the safety of a trial using the BOIN design, thanks to the feature that the BOIN design guarantees de-escalating the dose when $\hat{p}_j \geq \lambda_d$. For example, given a target DLT rate $\phi = 0.25$, we know a priori that a phase I trial using the BOIN design guarantees de-escalating the dose if the observed DLT rate is higher than $\lambda_d = 0.298$(ie, the default value). Accordingly, the BOIN design also allows users to easily calibrate the design to satisfy a specific safety requirement mandated by regulatory agents through choosing an appropriate target DLT rate ϕ or ϕ_2. For example, supposing for a phase I trial with a new compound, the regulatory agent mandates that if the observed toxicity rate is higher than 0.25, the dose must be de-escalated. We can easily fulfill that requirement by setting the target DLT rate $\phi = 0.21$, under which the BOIN automatically guarantees de-escalating the dose if the observed toxicity rate $\hat{p}_j > \lambda_d = 0.250$. Such flexibility and transparency renders the BOIN design an important advantage in practice.

As a side note, ϕ_1 and ϕ_2 used in the BOIN design have different interpretations than the proper dosing interval (δ_1, δ_2) used in the mTPI/mTPI 2 and Keyboard designs. Specifically, ϕ_1 and ϕ_2 represent the DLT rates that should be regarded as unacceptable (more precisely, underdosing and overdosing, respectively), whereas δ_1 and δ_2 represent the range of DLT probabilities that are acceptable. For example, given that the target DLT probability $\phi = 0.25$, setting

$\phi_1 = 0.15$ and $\phi_2 = 0.35$ mean that the doses with the DLT rates of 0.15 and 0.35 are respectively regarded as unacceptably underdosing and overdosing, whereas setting $\delta_1 = 0.15$ and $\delta_2 = 0.35$ means that the dose with a DLT rate between 0.15 and 0.35 is regarded as acceptable. Thus, in general, the value of ϕ_1 should be smaller than δ_1 and the value of ϕ_2 should be greater than δ_2.

3 | SOFTWARE

The software for implementing the CRM is freely available at the MD Anderson software download website https://biostatistics.mdanderson.org/softwaredownload/SingleSoftware.aspx?Software_Id=81. The R code for implementing the mTPI design is available at https://biostatistics.mdanderson.org/softwaredownload/SingleSoftware.aspx?Software_Id=72. The software for the BOIN design is available in 3 forms, including a standalone graphical user interface–based Windows desktop program freely available from MD Anderson software download website https://biostatistics.mdanderson.org/softwaredownload/SingleSoftware.aspx?Software_Id=99, Shiny online apps freely available at http://www.trialdesign.org, and R package "BOIN" available from the CRAN. The Keyboard design can be implemented using the Shiny online app freely available at http://www.trialdesign.org.

4 | SIMULATION STUDY

4.1 | Generating dose-toxicity scenarios

We generated true dose-toxicity scenarios using the pseudo-uniform algorithm proposed by Clertant and O'Quigley.[26] Given a target DLT probability ϕ and J dose levels, we generated scenarios as follows:

a. Select one of the J dose levels as the MTD with equal probabilities.
b. Sample $M \sim \text{Beta}(\max\{J-j, 0.5\}, 1)$, where j denotes the selected dose level, and set an upper bound $B = \phi + (1-\phi) \times M$ for the toxicity probabilities.
c. Repeatedly sample J toxicity probabilities uniformly on [0, B] until these correspond to a scenario in which dose level j is the MTD.

In these scenarios, the MTD is the dose with the DLT probability closest to the target ϕ but not necessarily equal to the target ϕ. Consequently, it is possible to obtain scenarios in which all the doses have DLT probabilities below or above the target ϕ, as could happen in practice. If the DLT probability at the lowest dose level is greater than $\phi + 0.1$, we will claim the scenario does not have MTD and define the selection percentage of the MTD as the percentage of simulated trials that terminate early. This is one of the strength of the algorithm, which provides extensive coverage on possible dose-toxicity scenarios that we may encounter in practice.

We considered target toxicity rates of $\phi = 0.20$ and $\phi = 0.30$, each with $J = 6$ or $J = 8$ dose levels. Under each setting, we generated 10 000 scenarios, and under each scenario, we simulated 2000 trials. Figure 2 displays 50 randomly selected scenarios with $\phi = 0.20$ and $J = 6$. These exhibit a variety of dose-toxicity curve shapes and spacings. The complete set of 10 000 scenarios are provided in Supporting Information.

For the CRM, we used the `getprior`(·) function in `R` to obtain the skeleton. We set the middle dose level (ie, dose level 3 for $J = 6$ doses and dose level 4 for $J = 8$ doses) as the prior MTD and the halfwidth of the indifference interval equal to 0.06. Specifically, when $\phi = 0.20$, the skeleton is (0.032, 0.095, 0.200, 0.332, 0.470, 0.596) for $J = 6$ doses and (0.007, 0.032, 0.095, 0.200, 0.332, 0.470, 0.596, 0.701) for $J = 8$ doses; when $\phi = 0.30$, the skeleton is (0.095, 0.186, 0.300, 0.422, 0.540, 0.643) for $J = 6$ doses and (0.038, 0.095, 0.186, 0.300, 0.422, 0.540, 0.643, 0.729) for $J = 8$ doses. For the mTPI and Keyboard designs, we set $\delta_1 = \phi - 0.05$ and $\delta_2 = \phi + 0.05$, which are the recommended default values. For the BOIN design, we set $\phi_1 = 0.6\phi$ and $\phi_2 = 1.4\phi$, which are the recommended default values. We set the maximum sample size equal to 36 and 48 for $J = 6$ and $J = 8$, respectively, and considered cohort size = 1 or 3.

4.2 | Performance metrics

For each of the 10 000 scenarios, we calculated the following metrics:

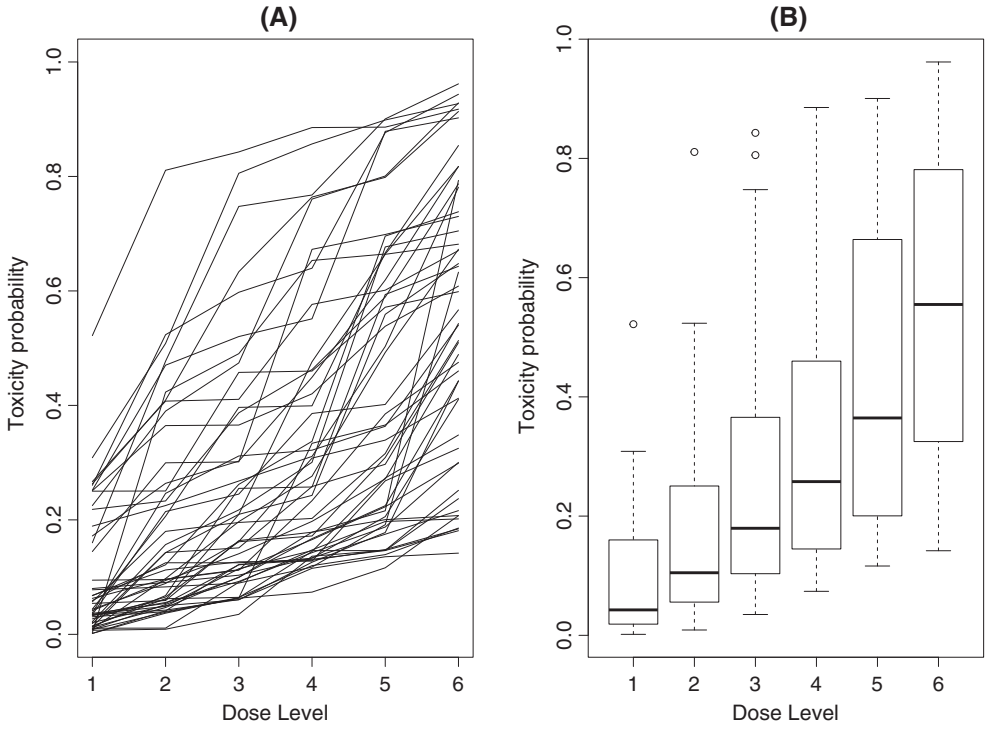

FIGURE 2 A, Fifty randomly selected dose-toxicity curves. B, The distribution of the toxicity probabilities by dose level from the 10 000 scenarios with 6 dose levels

A. MTD selection
   A1. The percentage of correct selection (PCS), which we defined as the percentage of simulated trials in which the correct dose is selected as the MTD.
   A2. The PCS within a 5% acceptable region, which we defined as the percentage of simulated trials in which the dose selected as the MTD has a toxicity probability that lies in the interval $[\phi - 0.05, \phi + 0.05]$.

B. Patient allocation
   B1. The average percentage of patients in the simulated trials who are assigned to the MTD.
   B2. The average percentage of patients in the simulated trials who are assigned to a dose with a toxicity probability that lies in the interval $[\phi - 0.05, \phi + 0.05]$.

C. Overdose control
   C1. The average percentage of patients in the simulated trials who are assigned to a dose that is above the MTD.
   C2. The risk of overdosing, which we define as the percentage of simulated trials in which a large percentage of patients (eg, 70%) are assigned to a dose that is above the MTD. This metric quantifies how likely a particular design is to overdose a large percentage of patients.

One might think that for overdose control, metric C2 is redundant and has been covered by metric C1. That is not true. These 2 metrics actually measure 2 different important aspects of the trial design. Specifically, C1 measures the mean of the number of patients overdosed, whereas C2 measures the tail probability of the number of patients overdosed. As shown later, 2 designs can have a similar average number of patients overdosed but rather different risks of overdosing 70% of patients. Although largely overlooked in the existing literature, metrics C2 is of great practical importance because it measures the likelihood of a design demonstrating extreme problematic behaviors, eg, treating 70% or more patients at toxic doses, ie, the reliability of the design.[17] The incidence of such extreme behavior in a trial design may be low, but is of serious practical concern when it occurs. For ethical reasons, we certainly want to minimize the occurrence of such extreme behavior in the trial and choose a more reliable design.

To compare the relative performance of the designs, we used the CRM design as a benchmark and report the difference between each of the model-assisted designs and the CRM design for each metric. For example, the PCS for the BOIN design is reported as (the PCS of BOIN) – (the PCS of CRM).

4.3 | Results

4.3.1 | MTD selection

Figure 3 shows the results for PCS and PCS within a 5% acceptable region for the mTPI, BOIN, and Keyboard designs, with respect to the CRM. As each dose-toxicity scenario generates a value of the performance metric (eg, PCS), we obtained a total of 10 000 values for each of the metrics across the 10 000 scenarios. Each boxplot reflects the distribution of the metrics across the 10 000 scenarios. As an example, the top-left panel of Figure 3 shows a boxplot of the PCS difference between mTPI and CRM, between BOIN and CRM, and between Keyboard and CRM when $\phi = 0.20$ with $J = 6$ doses. For mTPI versus CRM, when the cohort size equals to 1, most of the data points are negative, which indicates that the CRM tends to outperform mTPI. For BOIN and Keyboard versus CRM, respectively, most of the data points are close to zero, which indicates that the BOIN and Keyboard tend to perform similarly to the CRM. When the cohort size equals to 3, these designs have comparable PCS. For PCS within 5% (bottom-left panel), we see a similar pattern. As evidenced by the right panels of Figure 3, when $\phi = 0.30$,the CRM, mTPI, BOIN, and Keyboard have comparable PCS and PCS within 5%. Table 1 reports the average PCS for each design, along with the averages of the other metrics.

4.3.2 | Patient allocation

Figure 4 shows the results for the average percentage of patients who are assigned to the MTD and the average percentage of patients treated at the doses within 5% acceptable region of target toxicity probability, respectively, for $\phi = 0.20$

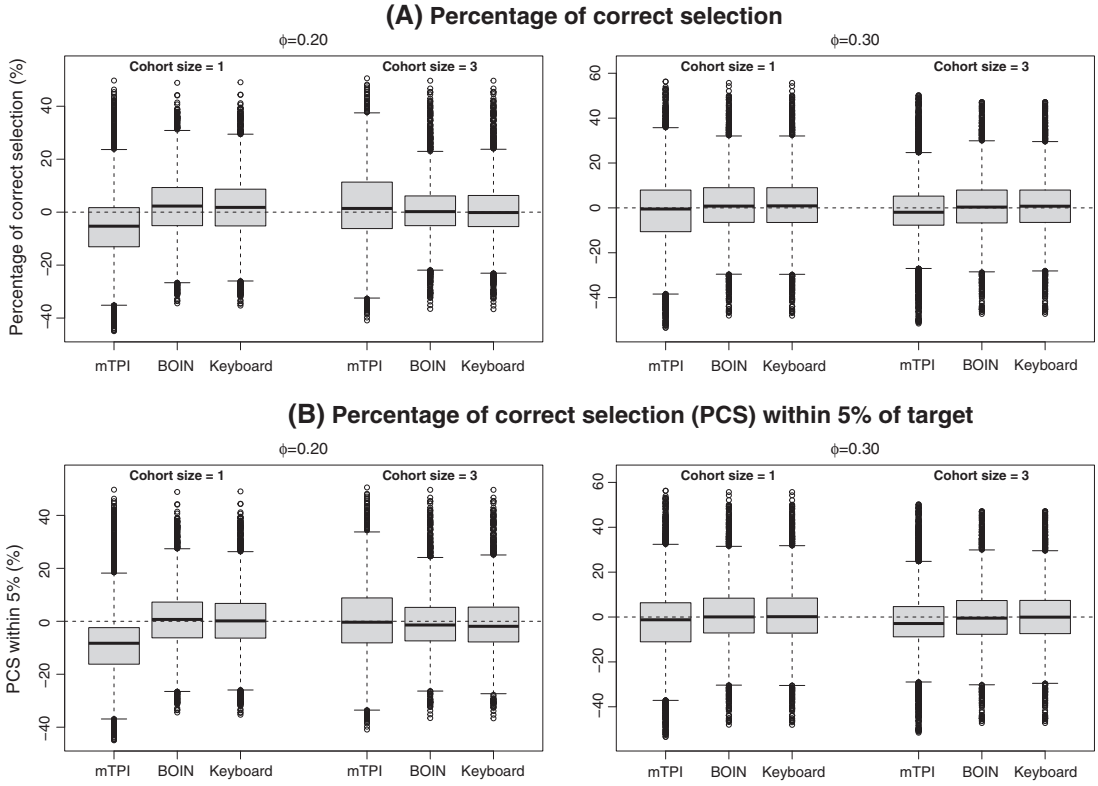

FIGURE 3 Boxplot of the difference in the percentage of correct selection (PCS) of the maximum tolerated dose (MTD) and the PCS of the doses within 5% of the target for mTPI vs CRM, BOIN vs CRM, and Keyboard vs CRM under 10 000 scenarios with 6 dose levels

TABLE 1 Average performance of the CRM, mTPI, BOIN, and Keyboard designs across 10 000 scenarios with 6 dose levels

	Target $\phi = 0.20$							
	Cohort Size = 1				Cohort Size = 3			
Performance Metric	**CRM**	**mTPI**	**BOIN**	**Keyboard**	**CRM**	**mTPI**	**BOIN**	**Keyboard**
PCS, %	49.1	43.0	50.9	50.5	45.8	47.4	46.1	46.0
PCS within 5%, %	60.5	51.5	61.0	60.7	57.2	57.3	56.3	56.1
Patients treated at MTD, %	36.9	34.3	37.3	37.3	31.4	32.6	30.2	30.2
Patients treated within 5%, %	46.7	42.0	46.2	46.3	40.4	41.2	38.4	38.4
Patients treated above MTD, %	18.6	21.6	22.2	21.5	14.5	16.8	12.7	12.6
Risk of overdosing 70%, %	8.8	16.8	9.6	9.8	5.3	8.7	3.6	3.8
	Target $\phi = 0.30$							
	Cohort Size = 1				Cohort Size = 3			
	CRM	**mTPI**	**BOIN**	**Keyboard**	**CRM**	**mTPI**	**BOIN**	**Keyboard**
PCS, %	50.5	49.0	51.7	51.7	49.5	48.2	50.3	50.5
PCS within 5%, %	58.1	55.8	58.8	58.8	58.6	56.6	58.7	58.8
Patients treated at MTD, %	38.0	38.3	38.1	38.1	34.1	33.6	33.5	33.4
Patients treated within 5%, %	44.5	44.3	44.2	44.2	40.1	39.2	39.0	39.0
Patients treated above MTD, %	20.4	24.3	21.6	21.7	17.7	15.3	15.7	15.8
Risk of overdosing 70%, %	9.9	17.4	9.8	9.8	6.8	7.1	5.0	5.0

Abbreviations: CRM, continual reassessment method; BOIN, Bayesian optimal interval; mTPI, modified toxicity probability interval; PCS, percentage of correct selection.

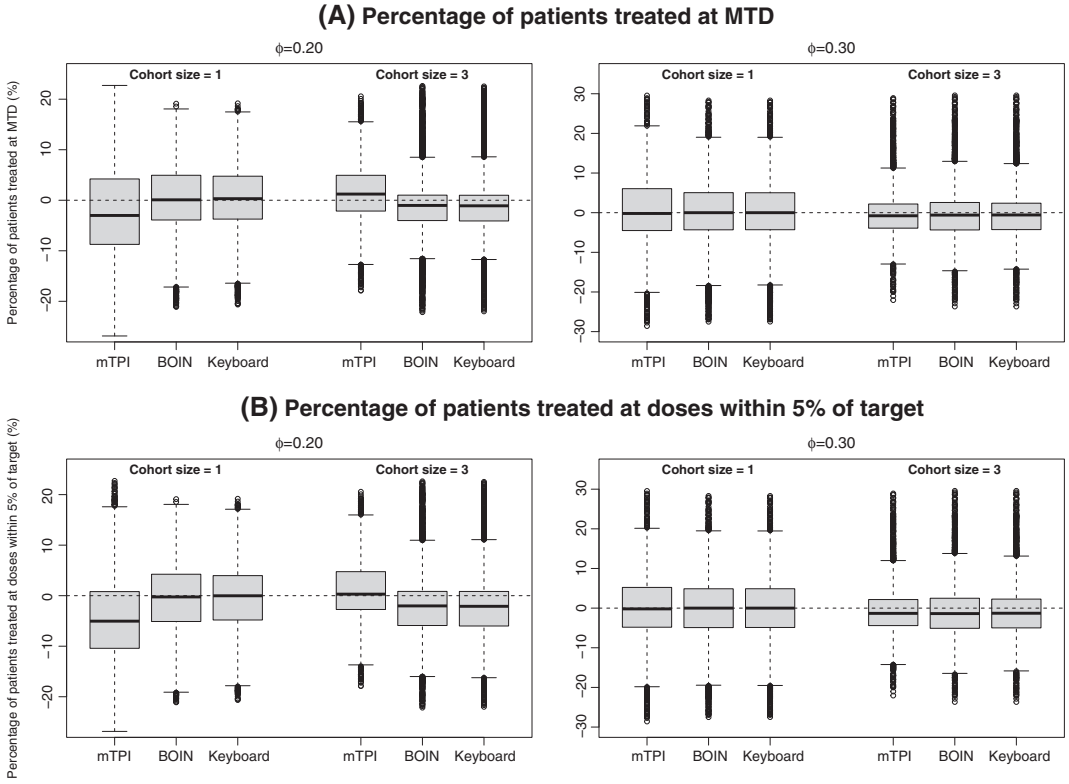

FIGURE 4 Boxplot of the difference in the percentage of patients treated at the maximum tolerated dose (MTD) and the percentage of patients treated at doses within 5% of the target for mTPI vs CRM, BOIN vs CRM, and Keyboard vs. CRM under 10 000 scenarios with 6 dose levels

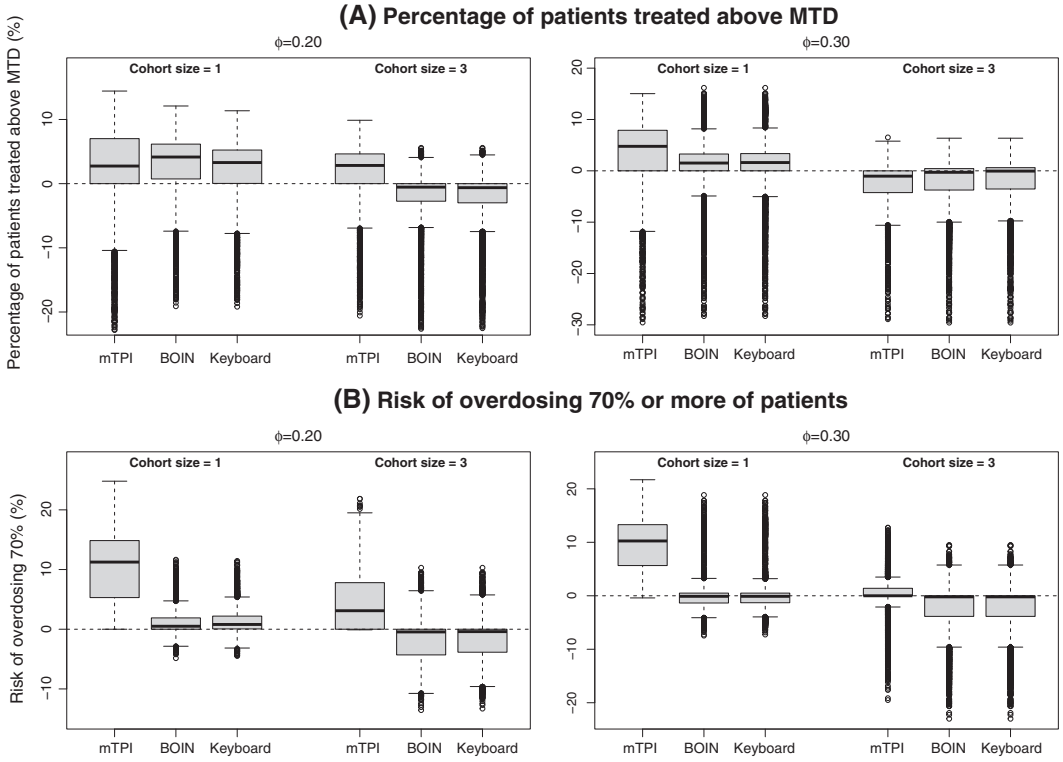

FIGURE 5 Boxplot of the difference in the number of patients treated above maximum tolerated dose (MTD) and the risk of overdosing at least 70% of patients for mTPI vs CRM, BOIN vs CRM, and Keyboard vs CRM under 10 000 scenarios with 6 dose levels

and $\phi = 0.30$ with $J = 6$ doses. When $\phi = 0.20$, CRM, BOIN, and Keyboard are comparable and all outperform mTPI when the cohort size is 1, whereas these 4 designs are comparable when the cohort size is 3. When $\phi = 0.30$, all 4 designs are comparable with respect to these metrics.

4.3.3 | Overdose control

The upper panel of Figure 5 shows the results for the average percentage of patients treated above MTD across simulated trials under 10 000 scenarios. When $\phi = 0.20$, compared with the CRM, mTPI, BOIN, and Keyboard tend to treat slightly more patients to the doses that are above the MTD when the cohort size is 1. When the cohort size is 3, compared with the CRM, the BOIN and Keyboard designs tend to treat fewer patients above the MTD, whereas mTPI treats more patients above the MTD. The pattern for $\phi = 0.30$ is generally similar to that for $\phi = 0.20$.

The lower panel of Figure 5 shows the results for the risk of overdosing at least 70% of patients. The mTPI has substantially higher (ie, approximately doubled when the cohort size = 1) risks of overdosing than the CRM, BOIN, and Keyboard designs. Table 1 shows the average value for this metric. This result shows that 2 designs (eg, mTPI versus CRM or BOIN) can have similar percentages of patients treated above the MTD but dramatically different risk of overdosing a large percentage of patients (ie, design reliability), demonstrating the importance of considering both metrics (ie, C1 and C2) when evaluating a design. The CRM, BOIN, and Keyboard are comparable in the risk of overdosing 70% or more of patients. When $\phi = 0.20$, the CRM is slightly safer than BOIN and Keyboard when the cohort size is 1, whereas BOIN and Keyboard are slightly safer than the CRM when the cohort size is 3.

The reason mTPI is more likely than the other designs to overdose at least 70% of the patients is explained previously (eg, the UPM cannot appropriately measure the evidence of the toxicity of a dose) and can also be seen through the dose escalation and de-escalation rules for the 3 model-assisted designs reported in Table 2. When the target is $\phi = 0.20$, the default BOIN, mTPI, and Keyboard designs use different thresholds for dose escalation and de-escalation. In particular, compared with the BOIN and Keyboard designs, the mTPI design is less likely to de-escalate the dose when a high rate of toxicity is observed. For example, suppose 6 patients have been treated at the current dose, the BOIN and Keyboard designs will de-escalate the dose if 2 DLTs are observed, whereas the mTPI requires observing 3 DLTs before

TABLE 2 Escalation and de-escalation rules for the mTPI, BOIN, and Keyboard designs under their default settings for a target toxicity rate of ϕ=0.2

	Number of Patients Treated at the Current Dose															
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
mTPI Design																
Escalate if number of DLTs ≤	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
De-escalate if number of DLTs ≥	1	2	2	2	3	3	4	4	4	5	5	5	5	6	6	6
BOIN Design																
Escalate if number of DLTs ≤	0	0	0	0	0	0	1	1	1	1	1	1	2	2	2	2
De-escalate if number of DLTs ≥	1	1	1	1	2	2	2	2	3	3	3	3	4	4	4	4
Keyboard Design																
Escalate if number of DLTs ≤	0	0	0	0	0	0	0	1	1	1	1	1	1	1	2	2
De-escalate if number of DLTs ≥	1	1	1	1	2	2	2	2	3	3	3	3	4	4	4	4

Abbreviations: BOIN, Bayesian optimal interval; DLTs, dose-limiting toxicities; mTPI, modified toxicity probability interval.

de-escalating the dose. Consequently, the mTPI design tends to stay long (ie, get stuck) at a particular dose. If that particular dose is above the MTD, a large percentage of patients are overdosed.

4.4 | Analysis of simulation results

Our simulation results show that the CRM, BOIN, and Keyboard designs have comparable, good operating characteristics, especially in terms of PCS and the risk of overdosing a large percentage of patients. However, when we examine each scenario individually, we find that in certain scenarios, the CRM has much higher PCS than the BOIN and Keyboard designs, whereas in other scenarios, the reverse is true. In this section, we aim to characterize the scenarios in which the CRM outperforms the BOIN design and vice versa. Because the Keyboard design has very similar performance as the BOIN, in what follows, we focus on the CRM and BOIN. In the trial conduct of CRM, the parameter

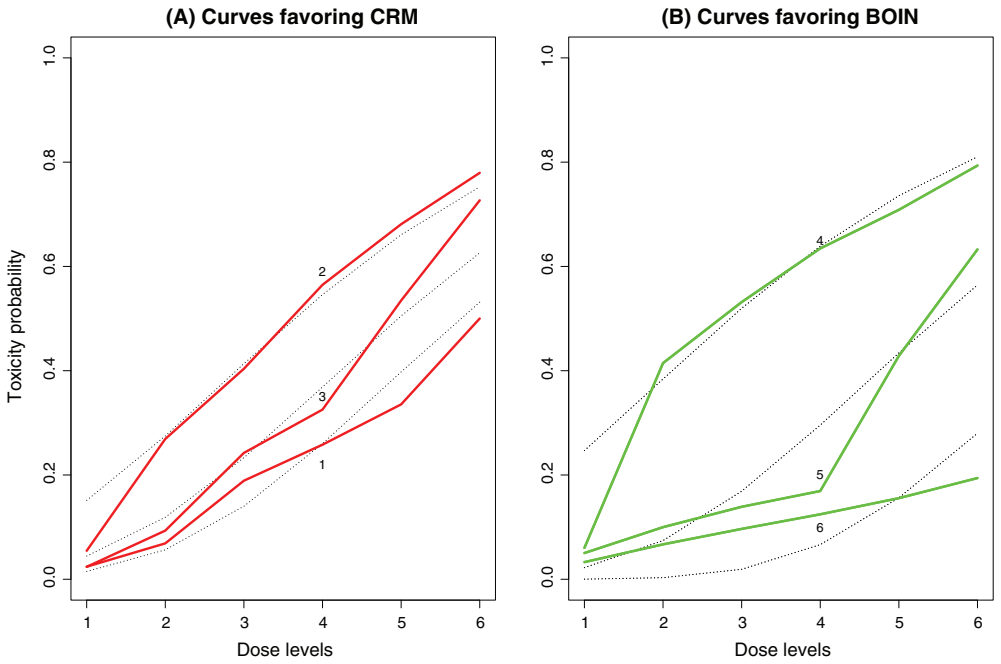

FIGURE 6 Medians of clustered dose-toxicity curves favoring continual reassessment method (CRM) and Bayesian optimal interval (BOIN) designs. Dotted lines are the best model-fitted curves from the CRM design [Colour figure can be viewed at wileyonlinelibrary.com]

α in (1) is continuously updated to reflect the accruing data. We hypothesized that if there exists an α_0 such that the fitted toxicity probabilities $\boldsymbol{\pi}(\alpha_0) = (a_1^{\exp(\alpha_0)}, \cdots, a_J^{\exp(\alpha_0)})$ from the power model are close to the true toxicity probabilities (ie, if the power model provides a good fit to the true dose-toxicity curve), then the CRM will outperform the BOIN design and vice versa. To verify our hypothesis, first, given a specific dose-toxicity scenario with true toxicity rates $(p_1^*, \cdots, p_J^*)$, we defined a goodness-of-fit (GOF) index for the toxicity probabilities as follows

$$GOF = \min_{\alpha} \sqrt{\sum_{j=1}^{J} \left(a_j^{\exp(\alpha)} - p_j^*\right)^2}.$$

The GOF index summarises the difference or distance between the best-fitted CRM model and the true dose-toxicity curve, in terms of the mean square error. A smaller value indicates that the CRM model can provide a better fit to the true dose-toxicity curve. The value of GOF is determined through the grid search over α. Second, we selected the scenarios in which CRM had a PCS that was at least 10% higher than BOIN, and the scenarios in which BOIN has a PCS that was at least 10% higher than CRM. Third, using k-means clustering,[28] we partitioned these 2 sets of scenarios into 3 clusters. Figure 6 shows the median scenario in each cluster, as well as the best model-fitted curve (dotted lines). Figure 6 shows that, compared with the dose-toxicity curves that favor the BOIN design, the dose-toxicity curves that favor the CRM are closer to the corresponding best model-fitted curve. The first 3 scenarios—in which the CRM has better PCS than BOIN—correspond to smaller GOF indexes than the latter 3 scenarios—in which BOIN has better PCS than the CRM.

5 | CONCLUSION

Using 10 000 randomly generated scenarios, we provided a more complete and reliable comparison of the CRM with the mTPI, BOIN, and Keyboard designs. Our results showed that the CRM, BOIN, and Keyboard design provide comparable, excellent operating characteristics, and each outperforms the mTPI design. In particular, these design are more likely to correctly select the MTD and less likely to overdose a large percentage of patients. Based on our analysis of our simulation results, the performance of the CRM is affected by the specification of the skeleton. When the model-fitted curve is close to the true toxicity curve, the CRM tends to provide better operating characteristics than the BOIN design, whereas in scenarios where it is impossible for the model-fitted curve to provide a close approximation for the true toxicity curve, the performance of CRM tends to suffer relative to BOIN. The implication of this result is that in the case that we have good prior knowledge on the true dose-toxicity curve, the CRM may be a better choice, whereas in the case that we are lack of good prior knowledge on the true dose-toxicity curve, the BOIN and Keyboard designs may be a better choice.

The BOIN and Keyboard designs have extremely similar performance in all performance metrics, although they are based on different statistical approaches. As described previously, because of directly using the observed DLT rate for decision making, the BOIN is more transparent and accessible to nonstatisticians. It allows clinicians and regulatory agents to easily assess the safety of the trial design and also allows users to easily calibrate their trials to satisfy the safety requirement mandated by the regulatory agents. In addition, the BOIN design is a versatile method that has been extended to find the MTD or MTD contour for drug combination trials,[19,29] account for toxicity grades,[23] and conduct phase I-II trials.[20,21] Nevertheless, for practitioners who have used and are comfortable with the mTPI design, the Keyboard design provides a useful, seamless upgrade of the mTPI design.

We have compared the model-assisted designs with the CRM. One may wonder how are the model-assisted designs compared with other model-based designs, such as the EWOC and Bayesian logistic regression model design (BLRM).[30] Zhou et al[31] performed a comprehensive comparison of model-assisted designs with these model-based designs in terms of design accuracy, safety, and reliability. Their results show that these model-assisted designs, in particular the BOIN and Keyboard designs, yield comparable or better performance than these model-based designs. Recently, Clertant and O'Quigley[26] propose a flexible semiparametric dose-finding methods that reduces to the CRM under some added parametric conditions and is equivalent to the mTPI or BOIN design under some relaxation of the underlying structure. The semiparametric dose-finding method shows competitive performance. Comprehensively investigating the existing phase I designs under such a unified framework is of interest and warrants further research.

ORCID

Heng Zhou http://orcid.org/0000-0003-3611-028X
Thomas A. Murray http://orcid.org/0000-0003-2769-4957
Haitao Pan http://orcid.org/0000-0003-4457-7349
Ying Yuan http://orcid.org/0000-0003-3163-480X

REFERENCES

1. Le Tourneau C, Lee JJ, Siu LL. Dose escalation methods in phase I cancer clinical trials. *J Nat Cancer Inst*. 2009;101:708-720.
2. 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(2):190-195.
3. Durham SD, Flournoy N, Rosenberger WF. A random walk rule for phase I clinical trials. *Biometrics*. 1997;53:745-760.
4. Ivanova A, Montazer-Haghighi A, Mohanty SG, D Durham S. Improved up-and-down designs for phase I trials. *Stat Med*. 2003;22(1):69-82.
5. Stylianou M, Follmann DA. The accelerated biased coin up-and-down design in phase I trials. *J Biopharm Stat*. 2004;14(1):249-260.
6. O'Quigley J, Pepe M, Fisher L. Continual reassessment method: a practical design for phase 1 clinical trials in cancer. *Biometrics*. 1990;46(1):33-48.
7. Babb J, Rogatko A, Zacks S. Cancer phase I clinical trials: efficient dose escalation with overdose control. *Stat Med*. 1998;17(10):1103-1120.
8. Cheung YK, Chappell R. Sequential designs for phase I clinical trials with late-onset toxicities. *Biometrics*. 2000;56(4):1177-1182.
9. Yin G, Yuan Y. Bayesian model averaging continual reassessment method in phase I clinical trials. *J Am Stat Assoc*. 2009;104(487):954-968.
10. Liu S, Yin G, Yuan Y. Bayesian data augmentation dose finding with continual reassessment method and delayed toxicity. *Ann Appl Stat*. 2013;4:2138-2156.
11. Wages NA, Conaway MR, O'Quigley J. Continual reassessment method for partial ordering. *Biometrics*. 2011;67(4):1555-1563.
12. Braun TM. The bivariate continual reassessment method: extending the CRM to phase I trials of two competing outcomes. *Controlled Clin Trials*. 2002;23(3):240-256.
13. Cheung YK. *Dose Finding by the Continual Reassessment Method*. Boca Raton, FL: CRC Press; 2011.
14. Yan F, Mandrekar SJ, Yuan Y. Keyboard: a novel bayesian toxicity probability interval design for phase I clinical trials. *Clin Cancer Res*. 2017;23:3994-4003.
15. Ji Y, Liu P, Li Y, Bekele BN. A modified toxicity probability interval method for dose-finding trials. *Clin Trials*. 2010;7(6):235-244.
16. Guo W, Wang SJ, Yang S, Lynn H, Ji Y. A Bayesian interval dose-finding design addressing Ockham's razor: mTPI-2. *Contemp Clin Trials*. 2017;58:23-33.
17. Liu S, Yuan Y. Bayesian optimal interval designs for phase I clinical trials. *J R Stat Soc Ser C (Appl Stat)*. 2015;64(3):507-523.
18. Yuan Y, Hess KR, Hilsenbeck SG, Gilbert MR. Bayesian optimal interval design: a simple and well-performing design for phase I oncology trials. *Clin Cancer Res*. 2016;22(17):4291-4301.
19. Lin R, Yin G. Bayesian optimal interval design for dose finding in drug-combination trials. *Stat Methods Med Res*. 2017;26:2155-2167.
20. Lin R, Yin G. STEIN: a simple toxicity and efficacy interval design for seamless phase I/II clinical trials. *Stat Med*. 2017;36:4106-4120.
21. Takeda K, Taguri M, Morita S. Bayesian optimal interval design for dose finding based on both efficacy and toxicity outcomes. *Pharm Stat*. 2018. in press.
22. Pan H, Lin R, Yuan Y. Statistical properties of the keyboard design with extension to drug-combination trials. http://arxiv.org/abs/1712.06718; 2018
23. Mu R, Yuan Y, Xu J, Mandrekar SJ, Yin JY. gBOIN: a unified model-assisted phase I trial design accounting for toxicity grades, binary or continuous end points. *J R Stat Soc Series C (Appl Stat)*. 2018;1-20. https://doi.org/10.1111/rssc.12263.
24. Horton BJ, Wages NA, Conaway MR. Performance of toxicity probability interval based designs in contrast to the continual reassessment method. *Stat Med*. 2017;36(2):291-300.
25. Ananthakrishnan R, Green S, Chang M, Doros G, Massaro J, LaValley M. Systematic comparison of the statistical operating characteristics of various phase I oncology designs. *Contemp Clin Trials Commun*. 2017;5:34-48.
26. Clertant Matthieu, O'Quigley John. Semiparametric dose finding methods. *J R Stat Soc Ser B (Stat Method)*. 2017;79:1487-1508.
27. Barlow RE, Bartholomew DJ, Bremner J, Brunk HD. *Statistical Inference under Order Restrictions: The Theory and Application of Isotonic Regression*. London, New York: Wiley; 1972.
28. Hartigan JA, Wong MA. Algorithm as 136: A k-means clustering algorithm. *J R Stat Soc Ser C (Appl Stat)*. 1979;28(1):100-108.
29. Zhang L, Yuan Y. A practical Bayesian design to identify the maximum tolerated dose contour for drug combination trials. *Stat Med*. 2016;35(27):4924-4936.

30. Neuenschwander B, Branson M, Gsponer T. Critical aspects of the Bayesian approach to phase I cancer trials. *Stat Med*. 2008;27(13):2420-2439.

31. Zhou H, Yuan Y, Nie L. Accuracy, safety and reliability of novel phase I trial designs. *Clin Cancer Res*. 2018. in press.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Zhou H, Murray TA, Pan H, Yuan Y. Comparative review of novel model-assisted designs for phase I clinical trials. *Statistics in Medicine*. 2018;37:2208–2222. https://doi.org/10.1002/sim.7674