Contemporary Clinical Trials Communications 15 (2019) 100379

Contents lists available at ScienceDirect

Contemporary Clinical Trials Communications

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

Evaluating the effects of design parameters on the performances of phase I trial designs

Yaqian Zhu[*], Wei-Ting Hwang, Yimei Li

Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA, USA

ARTICLE INFO

Keywords:
Phase I trial design
Dose finding
Maximum tolerated dose
Design parameters

ABSTRACT

Numerous designs have been proposed for phase I clinical trials. Although studies have compared their performances, few have considered the effects of changing design parameters. In this article, we review a few popular designs, including the 3 + 3, continuous reassessment method (CRM), Bayesian optimal interval (BOIN) design, and Keyboard design, and evaluate how varying design parameters (such as number of dose levels, target toxicity rate, maximum sample size, and cohort size) could impact the performances of each design through simulations. Excluded from our analysis is the mTPI-2 design, which operates in the same way as the Keyboard. Our results suggest that regardless of the choices of design parameters, the 3 + 3 design performs worse than the other ones, and BOIN and Keyboard have comparable performance to CRM. For any design, the performance varies with the choice of parameters. In particular, it improves as sample sizes increase, but the magnitude of benefit from increasing sample sizes varies substantially across scenarios. The impact of cohort size on design performances seems to have no clear direction. Therefore, BOIN and Keyboard designs are generally recommended due to their simplicity and good performance. With regard to choices of sample size and cohort size in designing a trial, it is recommend that simulations be performed for the particular clinical settings to aid decision making.

1. Introduction

The purpose of phase I clinical trials is to identify an appropriate dose for experimentation in subsequent phase II and phase III studies. This dose is typically the maximum tolerated dose (MTD), defined to be the dose level whose corresponding toxicity probability is closest to the target toxicity probability [1]. Many statistical designs have been proposed for phase I dose finding, and they can be categorized as algorithm-based designs, model-based designs, or model-assisted designs [2,3]. Algorithm-based ones, such as the 3 + 3 design [4], utilize a set of rules for dose escalation and de-escalation. Despite their ease of use and transparency, these designs have been shown to have poor operating characteristics. On the other hand, model-based designs assume a statistical model for the dose-toxicity curve, and model parameters are updated based on observed data accumulated during the trial [2]. The most popular and well-studied design of this kind is the continual reassessment method (CRM) [5,6]. These designs have good operating characteristics but require continuous model updates during the trial, which have limited their implementation. Model-assisted designs (also called interval-based designs) utilize statistical models but their dose escalation rules can be determined prior to the trial and, therefore, offer ease in implementation [3]. Designs of this kind include the modified toxicity probability interval (mTPI) design [7] (as well as its upgrade, the mTPI-2 design) [8], the Bayesian optimal interval (BOIN) design [9], and the Keyboard design [3].

Several studies have compared the performances of the designs described above. Horton et al. [10] compared the CRM, mTPI, and BOIN designs but did not include the Keyboard design in their comparisons and only considered target toxicity rate of 0.20. Zhou et al. [11] added the Keyboard design to their review and evaluated two target toxicity rates, 0.20 and 0.30. However, they focused on scenarios with large number of doses (6 and 8), and we are interested in evaluating if their findings would remain the same under settings with fewer dose levels, as phase I trials sometime have only 3 or 4 dose levels [12,13]. Moreover, both studies considered a few sample sizes and cohort sizes of 1 and 3 and assessed relative performances of various designs given a particular sample size or cohort size. We would like to evaluate from an alternative perspective in the sense that for a given design, we explore how performance varies when the sample size or cohort size changes.

[*] Corresponding author. Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, 423 Guardian Drive, Blockley 108/109, Philadelphia, PA, 19104, USA.

E-mail address: yazhu@pennmedicine.upenn.edu (Y. Zhu).

https://doi.org/10.1016/j.conctc.2019.100379
Received 25 October 2018; Received in revised form 30 April 2019; Accepted 15 May 2019
Available online 17 May 2019

Therefore, this paper has two objectives: First, to extend previous findings about the relative performances of 3 + 3, CRM, BOIN, and Keyboard designs into settings with fewer dose levels, and second, for any given design, to evaluate how the performance changes with varying sample sizes and cohort sizes. The rest of the paper is organized as follows. In the Methods section, we give an overview of the phase I trial designs of interest and describe the simulation procedures. The Results section presents the findings from our simulations. We elaborate on our results in Discussion and offer concluding remarks in Conclusions.

2. Methods

We first summarize the four designs to be evaluated: the most commonly used algorithm-based design (3 + 3) and model-based design (CRM) as well as two interval-based designs (BOIN and Keyboard). We also include an overview of the mTPI-2, which is shown to be the same as the Keyboard design [11]. In all designs, the tested doses d_i, $i = 1, ..., I$, where I is the total number of dose levels, are pre-specified by the clinical investigators. Toxicity is a binary outcome with the target toxicity rate denoted as θ, and the dose toxicity relationship is assumed to be monotonically increasing. All designs start at the lowest dose.

2.1. 3 + 3 design

There are various versions of the 3 + 3 design, and we adopt the most commonly used one that considers the dose to be safe if there is less than one third of patients experiencing toxicity [4,14]. The trial starts with the lowest dose level and treats 3 patients at a time. At the current dose level, if there are no toxicities for the 3 patients, then the procedure escalates to the next dose level. If one toxicity is observed, 3 more patients are treated at the same dose level. If there are 2 or more toxicities, de-escalate to the lower dose, or if the trial is already at the lowest dose, stop the trial and declare all doses are toxic. When an additional 3 patients are treated at the same dose level, if there is 1 toxicity observed out of the total 6 patients, then escalation to next dose occurs. Otherwise, de-escalate to the lower dose level. The dose level that has at most 1 toxicity out of 6 treated patients will be declared as MTD, and 6 patients have to be treated at the dose level to be declared as MTD. We will use this definition to estimate the MTD in our simulations.

It is common to have an expansion cohort after MTD is found that treats additional patients at the MTD—to gather more toxicity data around the MTD and obtain preliminary data about efficacy [15]. The size of the expansion cohort typically ranges from 6 to 15 patients, and it is suggested that 10 to 20 patients in the expansion cohort may significantly increase the probability of selecting the true MTD [16,17]. To make the 3 + 3 design comparable to other designs, after MTD is found, the remaining patients are treated at the estimated MTD until the maximum sample size is reached, but MTD will not be updated during the expansion phase.

2.2. Continual Reassessment Method (CRM) design

The continual reassessment method (CRM) assumes a parametric dose-toxicity relationship and employs a Bayesian approach [5,18]. It specifies a dose-toxicity model and a prior distribution for the model parameters, and posterior distributions are calculated based on the observed toxicity data. Suppose the dose-toxicity model has the general form $p_i = f(d_i, a)$, where p_i is the toxicity probability at dose level i, d_i is the ith dose level, and a is a parameter with the prior distribution denoted as $g(a)$. To improve numerical stability, raw dosage d_i are rarely used; instead we often use the effective doses x_i (also referred to as the skeleton), which are determined as follows. We first elicit a prior guess of toxicity probability at each dose level, denoted as $\tilde{p}_i$, which is either chosen by clinicians, or determined using the algorithm in Lee and Cheung [19]. Given the prior toxicity $\tilde{p}_i$ and based on the dose-toxicity model, we then back-solve the effective dose x_i as $x_i = f^{-1}(\tilde{p}_i,\ a = \tilde{a})$, where $\tilde{a} = 0$ is the prior mean of a. Therefore, the actual model used for model fitting is $p_i = f(x_i, a)$.

Let n_i be the number of patients treated at dose level i and Y_i be the number of toxicities observed at dose level i. Then the likelihood of the observed data is given by $L(x_i,\ Y_i, a) = \prod_{i=1}^{I} f(x_i, a)^{Y_i}(1 - f(x_i, a))^{n_i - Y_i}$, and the posterior distribution of a is $g(a|data\ ) = \frac{g(a)L(x_i,\ Y_i, a)}{\int_{-\infty}^{\infty} g(u)L(x_i,\ Y_i, u)du}$. From this, the estimated probability of toxicity at dose level i given observed data is $\hat{p}_i\ = \int_{0}^{\infty} f(x_i, a)g(a|data\ )da$ and the dose with posterior toxicity probability $\hat{p}_i$ closest to θ is the estimated MTD. The next cohort is treated at the estimated MTD and this process continues until maximum sample size is reached.

We used the "`dfcrm`" package in R to implement the CRM design with the common specification of one-parameter logistic dose-toxicity model and not allowing dose-skipping during dose escalation [6]. We used the `getprior` function to obtain the prior based on the method suggested by Lee and Cheung [19]. Specifically, the desired half width of the indifference interval is set to 0.25 times the target toxicity rate. We provide sample R codes with specifications for argument inputs in the Appendix.

2.3. Bayesian Optimal Interval (BOIN) design

Bayesian optimal interval (BOIN) design was proposed by Liu and Yuan [9]. In this design, dose escalation and de-escalation are determined by where the observed toxicity rate at the current dose $\hat{p}_i$ lies in a prespecified toxicity tolerance interval, which contains the target toxicityθ [20]. Specifically for dose level i, the local BOIN design seeks to identify the interval $[\lambda_{1i}, \lambda_{2i}\ ]$ that contains the target θ, and λ_{1i} and are selected to minimize incorrect decisions about the next dose level assigned under hypothesis testing framework involving three point hypotheses: H_{0i}: $p_i = \theta$, H_{1i}: $p_i = \theta_1$, H_{2i}: $p_i = \theta_2$, where θ_1 is the toxicity probability corresponding to the highest dose below the MTD such that dose escalation should occur and θ_2 is the lowest toxicity probability that is considered toxic and de-escalation should occur. Correct decisions under hypotheses H_0, H_1, and H_2 are staying (S), escalation (E), and de-escalation (D), respectively. Under the Bayesian paradigm, each hypothesis is assigned a prior probability of being true: $\pi_{ki} = P(H_{ki})$, $k = 0,\ 1,\ 2$. The decision error rate is defined as $\alpha(\lambda_{1i},\ \lambda_{2i}) = P(H_{0i})P(S^c|H_{0i}) + P(H_{1i})P(E^c|H_{1i}) + P(H_{2i})P(D^c|H_{2i})$. The values of λ_{1i} and that minimize the decision error rate are the boundaries at which the posterior probabilities of H_1 and H_2, respectively, are greater than that of H_0. Given the observed data, Liu and Yuan showed that the desired values are $\lambda_{1i} = \frac{log\left(\frac{1-\theta_1}{1-\theta}\right) + \frac{1}{n_i}log\left(\frac{\pi_{1i}}{\pi_{0i}}\right)}{log\left\{\frac{\theta(1-\theta_1)}{\theta_1(1-\theta)}\right\}}$ and $\lambda_{2i} = \frac{log\left(\frac{1-\theta}{1-\theta_2}\right) + \frac{1}{n_i}log\left(\frac{\pi_{0i}}{\pi_{2i}}\right)}{log\left\{\frac{\theta_2(1-\theta)}{\theta\ (1-\theta_2)}\right\}}$.

Dose escalation decisions are determined by comparing $\hat{p}_i$ to the boundaries: if $\hat{p}_i \le \lambda_{1i}$, escalate to level $i + 1$, if $\hat{p}_i \ge \lambda_{2i}$, de-escalate to level $i - 1$, and if $\lambda_{1i} \le \hat{p}_i \le \lambda_{2i}$, stay at the same dose level i. The dose also remains at the same level if $\hat{p}_1 \ge \lambda_{2i}$ or if $\hat{p}_I \le \lambda_{1i}$. This process continues until the maximum sample size is reached. Safety rules that override previous dose escalation rules can also be applied. For example, if the current dose is too toxic (say, $P(p_i > \theta|Y_i, n_i) > .95\ and\ n_i \ge 3$), then the current dose and higher ones are eliminated from the remaining portion of the trial. In addition, if the first dose level is too toxic, then the trial is stopped and MTD is deemed not available.

At the completion of the trial, isotonic regression is used to determine an efficient statistical estimate of MTD [21]. This procedure identifies the doses that violate the monotonicity assumption and adjusts their toxicity rates to maintain monotonicity—replacing toxicity estimates of violators with their average.

We implemented the BOIN design using the "`boin`" package in R, and sample R codes with specific choices for arguments are presented in the Appendix. Online software is also available at *trialdesign.org*.

2.4. Keyboard design

The Keyboard design was proposed by Yan et al. [3], in which dose escalation is determined by the location of the strongest key relative to the target dosing interval that includes target toxicity θ. The strongest key is defined to be the dosing interval that most likely contains the true toxicity rate of the current dose, which is determined based on the posterior probability that each interval includes the target toxicity.

More specifically, the Keyboard design partitions the (0,1) interval into a series of equal-width dosing intervals (called keys) and the proper dosing interval is called the "target key," which is the one that includes the target toxicity rate θ. The target key is first specified (e.g. target key is 0.2–0.3) and then other keys are constructed to be of the same length and partition the rest of the (0,1) interval. With the observed data, if the strongest key is below the target key, escalation results; if it is above the target key, then de-escalation should occur; if it is the target key, then stay at the current dose. Like the BOIN design, additional safety rules may also be applied, and once the trial is completed, isotonic regression is used to obtain the MTD.

We implement the Keyboard design using R codes obtained from the authors. Online software is also available at *trialdesign.org*.

2.5. mTPI-2 design

Guo et al. [8] propose the mTPI-2 design as way to overcome problems of the modified toxicity posterior intervals (mTPI) design [7,22]. The mTPI design uses a set of decision rules for dose finding based on toxicity posterior intervals. For each interval, the unit probability mass (UPM), is defined to be "the ratio of the probability of the interval and the length of the interval" [7]. The (0,1) interval is partitioned into three parts: the equivalence interval (ε_1, ε_2) which is the one that includes the target θ, the interval above (called the overdosing interval), and the interval below (called the underdosing interval). The equivalence interval is similar to the target key in the Keyboard design. Based on the observed toxicity data, the posterior probability of each interval reflects the probability that the true toxicity rate is within that interval. If the underdosing interval has the highest UPM, then escalate to the next dose. If the overdosing interval has the highest UPM, then de-escalate to a lower dose. Otherwise, treatment remains at the same dose level.

Due to the mTPI design's high risk of overdosing patients—exposing subjects to doses above the MTD, modifications have been proposed in the mTPI-2 by dividing the interval into subintervals of equal length (the length of the equivalence interval). The optimal rule involves finding the interval with the largest posterior probability—the "winning" interval. Thus, if the winning interval lies above the equivalence interval, escalate to the next dose level; if the winning interval is below, de-escalate; and if the winning interval is the equivalence interval, stay at the same dose.

Although based on different theoretical justifications, it is noted that the mTPI-2 and Keyboard designs are actually identical [11]. Therefore, the subsequent comparisons focus on the Keyboard design and the results apply to mTPI-2 as well. Online software is available for the mTPI-2 design at https://udesign.laiyaconsulting.com/.

2.6. Simulation procedures

We conduct simulations to assess the design performances with varying study design parameters. One thousand trials are generated for all simulations unless specified otherwise. We perform two sets of simulations, one set for 6 dose levels and one set for 3 dose levels. For the 6 dose level setting, we consider five dose toxicity scenarios (Fig. 1a). The first scenario with toxicity probabilities (0.05, 0.12, 0.20, 0.30, 0.45, 0.60) represents a steady linear increasing dose-toxicity curve. The second scenario (0.10, 0.10, 0.11, 0.25, 0.50, 0.70) features a jump in the latter part of the curve. The third scenario (0.01, 0.03, 0.05, 0.15, 0.20, 0.45) represents a situation where toxicities are low at the first few doses but are followed by larger increases. The fourth scenario (0.12, 0.16, 0.20, 0.24, 0.27, 0.30) consists of small increases in toxicity rates between consecutive dose levels, and the range of toxicities is small overall. The fifth scenario (0.05, 0.20, 0.35, 0.50, 0.75, 0.90) consists of a large range of toxicity probabilities. For the 3 dose level setting, we consider three dose-toxicity scenarios (Fig. 1b). Scenario 6 with toxicity rates (0.10, 0.35, 0.60) represents a faster linear increase, Scenario 7 with toxicity rates (0.05, 0.10, 0.35) represents a jump at dose level 3, and Scenario 8 with toxicity rates (0.12, 0.20, 0.28) represents a slower linear increase.

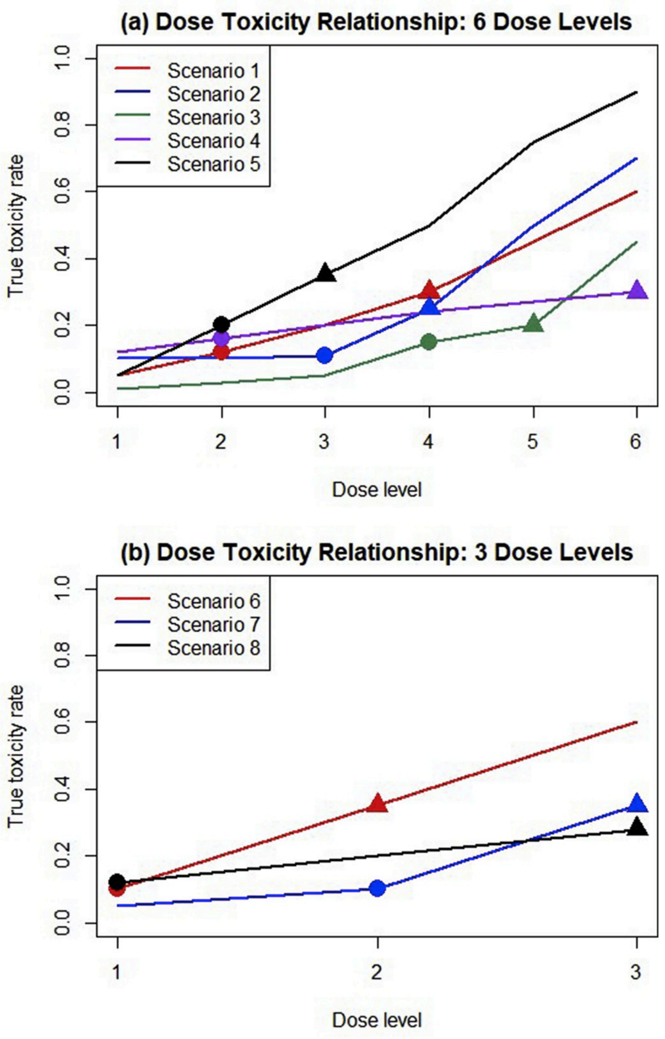

Fig. 1. True toxicity probabilities for simulation scenarios. Circles indicate MTD when target is 0.15, and triangles indicate MTD when target is 0.30.

We evaluated two target toxicity probabilities of 0.15 and 0.30. For our first objective of evaluating relative performances of various designs given a design setting, we assumed the cohort size to be 3 patients and the maximum sample size to be 36 patients for the 6 dose level setting and 18 patients for the 3 dose level setting. As mentioned before for the 3 + 3 design, the expansion cohort was treated at MTD to achieve the maximal sample size. For our second objective of evaluating the impact of sample size and cohort size on a given design's performance, we used varying sample sizes and cohort sizes. More specifically, we considered 15 to 42 patients with increments of 3 patients for the 6 dose level setting and 12 to 33 with increments of 3 patients for the 3 dose level setting. The 3 + 3 design was excluded in these comparisons because it does not update the MTD estimate as more patients are treated (after the MTD is estimated). For cohort size, we considered either 1, 2, or 3 patients per cohort.

2.7. Metrics for performance

In evaluating the performance of a phase I dose-finding trial, factors that come into play include statistical properties and ethical considerations. Statistically, a method that yields accurate estimates of MTD with high precision is desired. Ethically, a design that treats fewer patients at low ineffective doses or at overly toxic doses is preferred. Therefore, we look at common criteria such as percentage of correct selection (PCS)—the probability of selecting the true MTD—and average number of patients treated at the true MTD. In addition, we obtain the probability of selecting each dose level as the MTD to see how likely each design selects the doses near the true MTD. We also evaluate boxplots for the number of patients treated at MTD to better understand its distribution besides the simple summary of an average.

3. Results

Fig. 2 presents PCS and the average number of patients treated at the true MTD, for target toxicity rates of 0.15 and 0.30 and all the dose-toxicity scenarios in the 6 dose level setting. For the target toxicity rate of 0.15, CRM, BOIN and Keyboard designs have similar PCS, while the 3 + 3 has lower PCS in all the dose scenarios (Fig. 2a). For the target toxicity rate of 0.30, the difference between the 3 + 3 design and other designs is more substantial in most scenarios (Fig. 2b). For the average number of patients treated at the true MTD, the performances of all the designs are similar with slightly worse performance by the 3 + 3 design (Fig. 2c and d). We note that in Scenario 4 with target toxicity 0.30 (Fig. 2b), the CRM exhibits lower PCS. This is because the prior toxicity probabilities obtained from the default `getprior` function were substantially different from the true toxicity probabilities. These results suggest that the CRM is sensitive to specification of the prior. Specifically, for target toxicity of 0.30, the prior obtained from the `getprior` function is (0.07, 0.16, 0.30, 0.45, 0.59, 0.69), which differs considerably from the true toxicity probabilities (0.12, 0.15, 0.20, 0.24, 0.27. 0.30). If we instead use the true toxicity probabilities as the prior, then PCS for CRM would increase to 60.6%.

To further explore the patterns in MTD selection, we look at the percent of selecting each dose level as the MTD (Supplementary Fig. 1). In general, the CRM, BOIN, and Keyboard designs have similar proportions of selecting each dose level as the MTD. For these three designs, the percentage selection is highest at the true MTD and tends to be much higher at doses adjacent to the true MTD than those further away. However, the 3 + 3 design is more likely to incorrectly select lower dose levels as the MTD. When looking at the boxplots of number of patients treated at the true MTD (Supplementary Fig. 2), we see that the median number of patients treated at MTD is lower for the 3 + 3 design than the other designs, even though the difference in mean (which is largely influenced by extreme values) is much attenuated. Moreover, the 3 + 3 design either has more between-trial variability as suggested by the larger inter-quartile range or uniformly undertreats patients as suggested by lower values of both the 1st and 3rd quartiles.

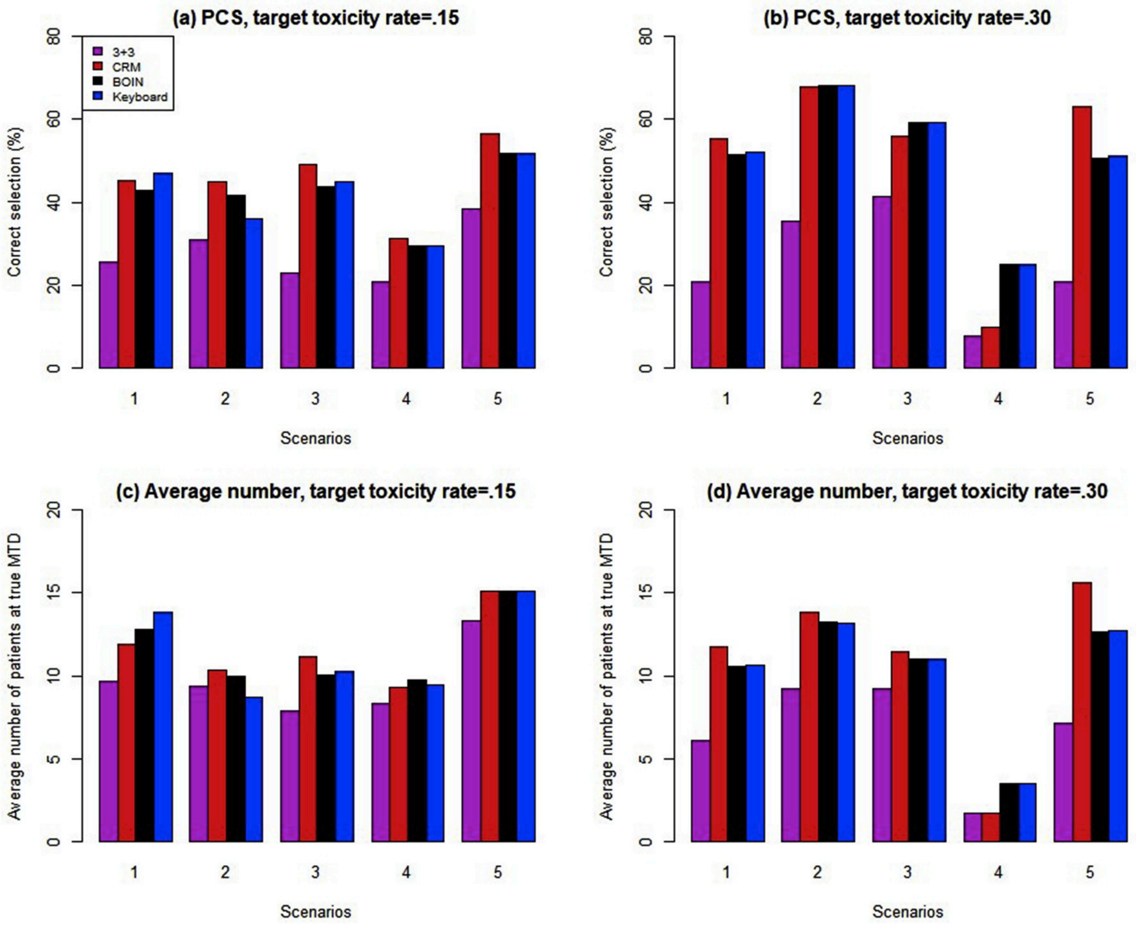

Fig. 2. PCS and average number of patients treated at the true MTD, for 6 dose levels, assuming sample size of 36 and cohort size of 3.

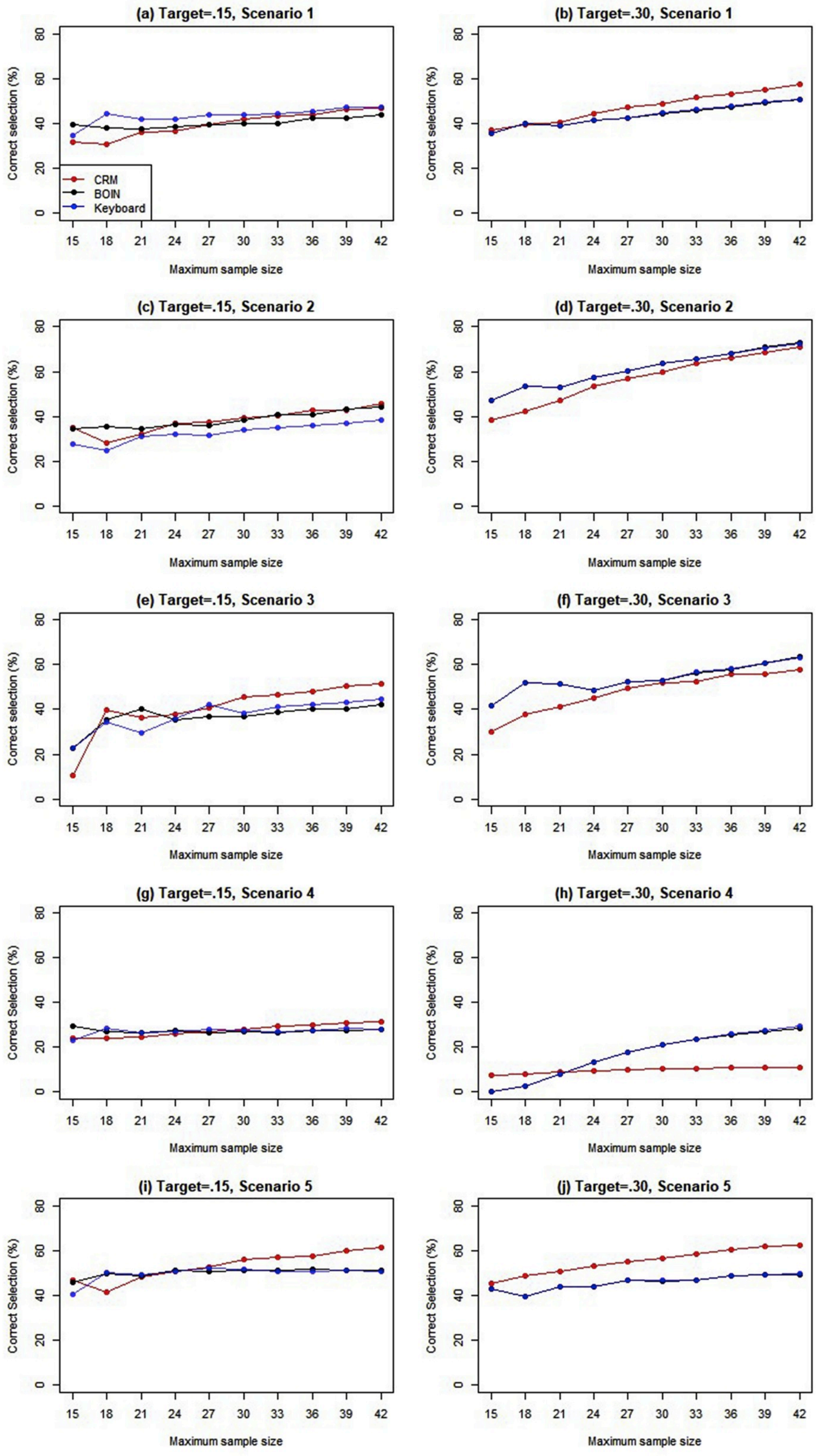

Fig. 3. PCS by maximal sample sizes, for 6 dose levels, assuming cohort size of 3, over 100,000 trials.

Fig. 3 demonstrates the effect of maximal sample size on PCS, for 6 dose levels. In general, increasing the maximum sample size somewhat increases PCS for both target rates of 0.15 and 0.30, for all the three designs, in all the scenarios. However, the magnitude of increases in PCS varies across scenarios and designs. For example, for Scenario 1 with target toxicity of 0.30, increasing sample sizes yields larger increases in PCS for CRM than for BOIN and Keyboard; in contrast, for Scenario 4 with target toxicity of 0.30, increasing sample size yields increases PCS for BOIN and Keyboard but not for CRM. Note again that in Scenario 4 with target toxicity 0.30, CRM has poor performance because the prior toxicity probabilities were misspecified as described before, and this appears to persist even with increasing sample sizes.

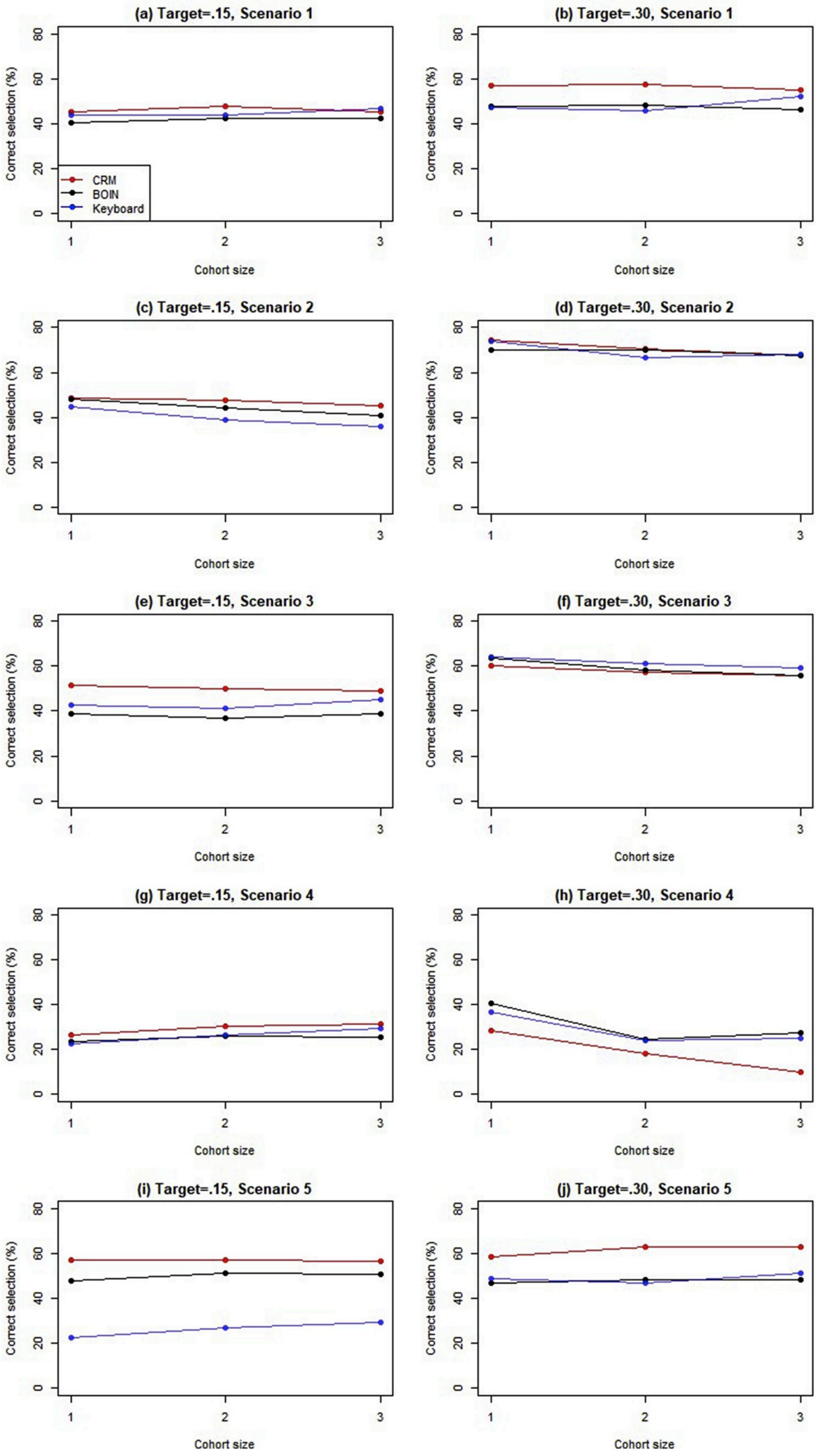

Fig. 4. PCS by cohort sizes, for 6 dose levels, assuming maximum sample size of 36.

The impact of different cohort sizes (1, 2, and 3 patients per cohort) on PCS under 6 dose levels are plotted in Fig. 4. Across all scenarios and designs, PCS can change as much as 20% with changes in cohort size. However, for a given design and given scenario, the effect of cohort size does not appear to follow any one direction—PCS either increases or decreases with increasing cohort size.

For the 3 dose level setting, PCS and average number of patients treated at the true MTD are presented in Fig. 5. The results show that the 3 + 3 design has lower PCS than the other designs with the difference becoming more pronounced for a higher target toxicity rate (Fig. 5a and b).The average number of patients treated at the true MTD is similar across the four designs, with a slightly lower number for the

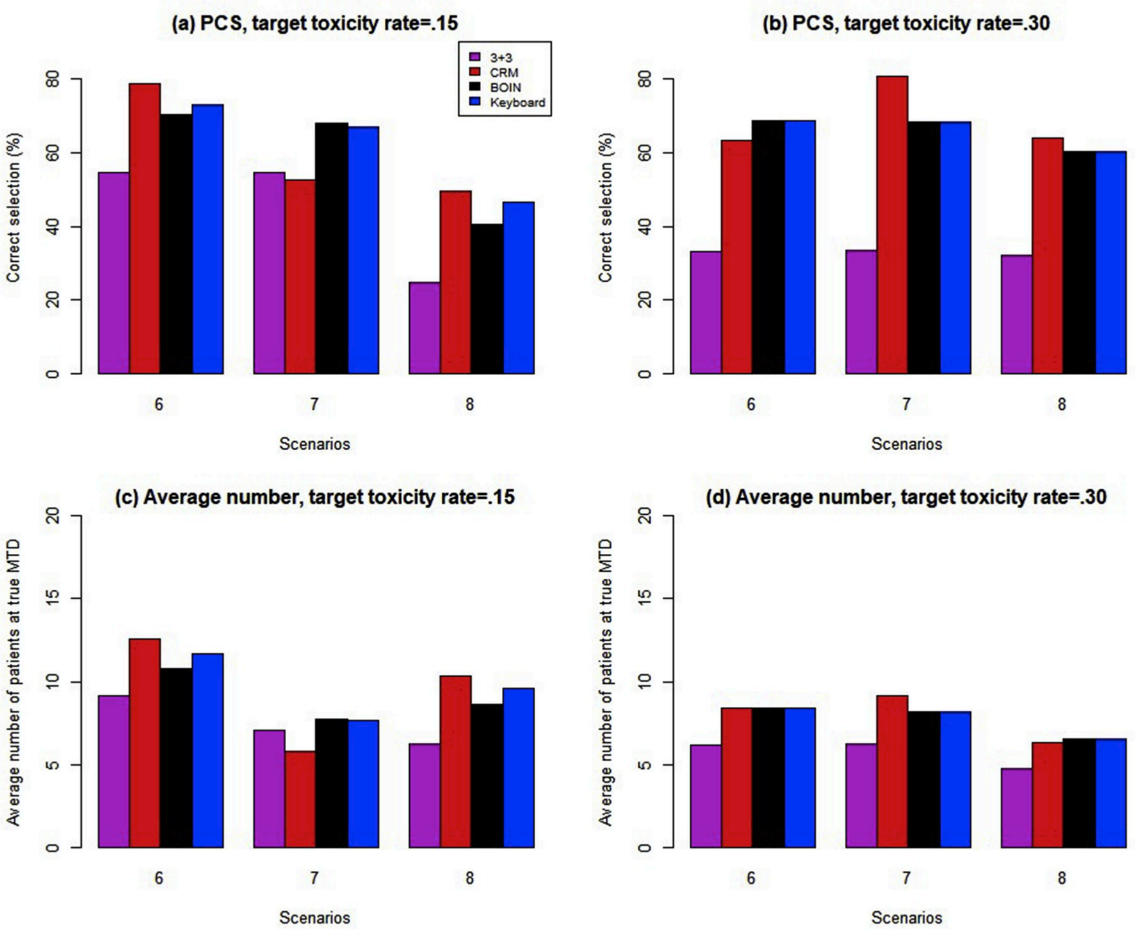

Fig. 5. PCS and average number of patients treated at the true MTD, for 3 dose levels, assuming sample size of 36 and cohort size of 3.

3 + 3 design in some scenarios. As noted in the 6 dose level setting, CRM has lower PCS in the scenario where the prior obtained from the default `getprior` function is substantially different from the truth, which is illustrated again in Scenario 7 with target toxicity 0.15 (Fig. 5a).

When evaluating the effect of maximum sample size for the 3 dose level setting, the results again suggest that the sample size effect varies across scenarios and designs (Fig. 6). Increasing sample sizes yields increased PCS in some but not all scenarios/designs.

For the effect of cohort size, the results are similar to those for 6 dose levels in that there are no directional effects in PCS with increasing cohort size (Fig. 7).

4. Discussion

In this paper, we seek to assess the relative performances of several phase I clinical trial designs under various design parameter settings and also to evaluate how the performance changes with varying design parameters for any given design. We have demonstrated that regardless of the choices of design parameters, the performances of the model-assisted designs (BOIN and Keyboard/mTPI2) are comparable to that of the model-based CRM design while the 3 + 3 design has poorer performance. This confirms previous findings and extends them to the settings with fewer number of dose levels [3,15,22,23].

In particular, our findings under the 3 dose level setting provide evidence to advocate the novel phase I designs over the traditional 3 + 3 design in this setting. As mentioned before, previous studies often assume a large number of dose levels, such as 6 doses [8,10,11,14] or 8 doses [10,11]. However, phase I trials may have fewer number of tested dose levels (such as 3 or 4 dose levels), especially for pediatric trials [24–26] and immunotherapy trials [13,27]. There is the perception that novel phase I trial designs will not provide much benefit and 3 + 3 designs may be adequate if there are fewer doses. However, our simulations suggest that even with three dose levels, the 3 + 3 design still performs much worse than the CRM, BOIN, and Keyboard/mTPI2 designs.

While in many cases CRM performs as well as or better than BOIN and Keyboard designs, CRM is sensitive to the choice of prior toxicity probabilities, which has been noted previously [28]. Moreover, as demonstrated in our simulations, the poor performance of CRM due to a misspecified prior may not be improved by simply increasing the sample size. Therefore, it is strongly recommended that one use the best historical information available to elicit priors (e.g. use adult data to elicit priors for pediatric trials). Another approach is to use the Bayesian Model Averaging Continual Reassessment Method (BMA-CRM) [28], which employs multiple priors and adaptively weights the CRM model associated with a prior based on its consistency with the observed data.

Although phase I trials are often small with a limited number of patients, such as 12 to 18 patients [17], it is unclear how the designs perform under different sample sizes. Most simulation studies assume a maximum sample size that is 6 times the number of dose levels considered [3,10,15], such as 36 patients for 6 dose levels, but there is no clear justification for this choice. Our findings show that an increase in sample size does not always translate to increased PCS, and a sample

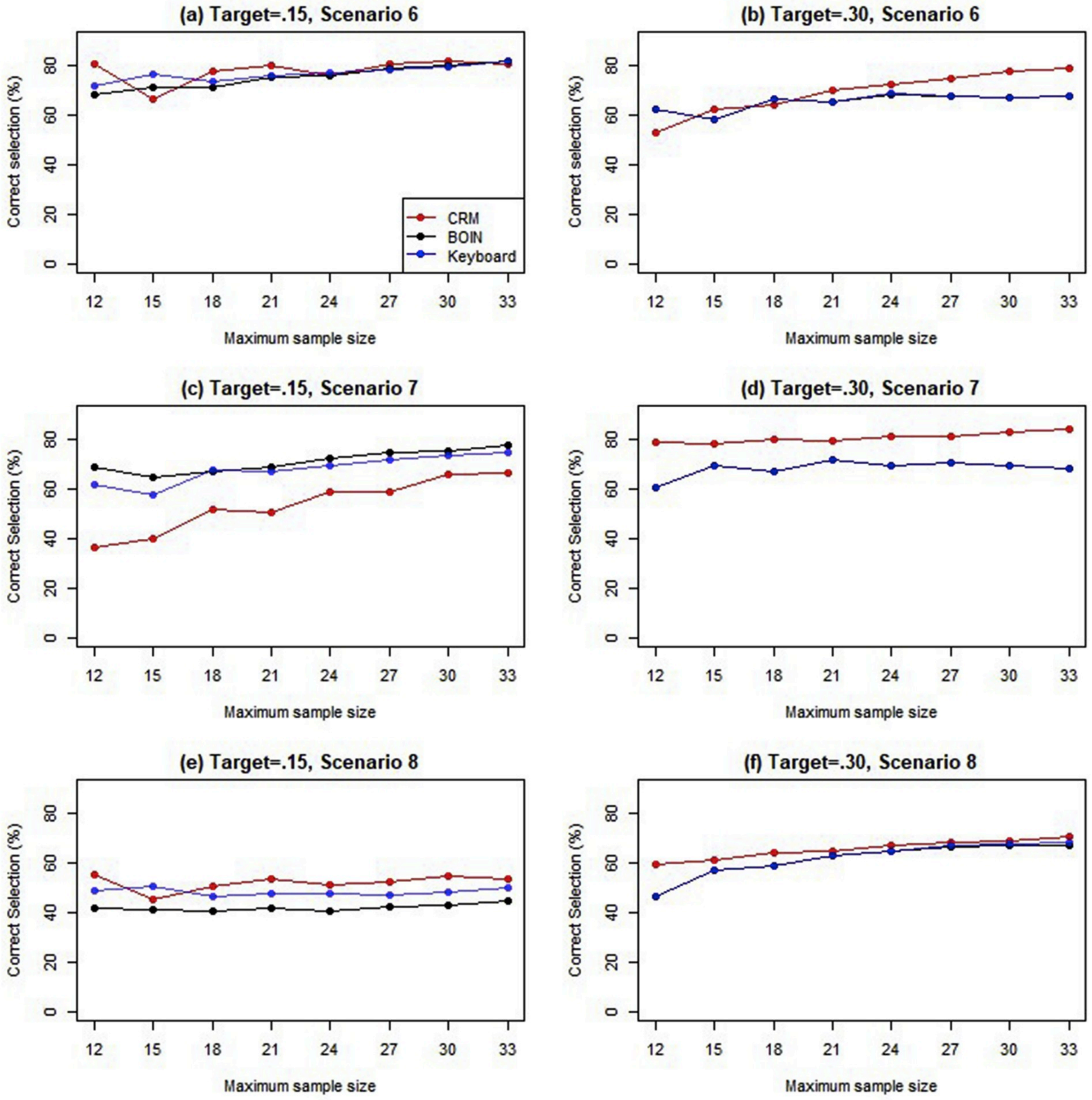

Fig. 6. PCS by maximal sample sizes, for 3 dose levels, assuming cohort size of 3, over 100,000 trials.

size less than 6 times the number of doses may be sufficient. However, the adequate sample size varies with different target toxicities, different dose-toxicity relationships, and different designs; therefore, we cannot assume a larger sample size results in substantial accuracy gain nor that a smaller sample size provides adequate accuracy, without performing simulations to evaluate.

With respect to the choice of cohort size, our results suggest that there is no universal recommendation of a particular size. Ahn [23] considers the impact of having 1, 2, or 3 patients per cohort for the CRM in his simulation study and indicates that a cohort size of 1 patient requires the least number of patients to find the MTD but the largest number of cohorts and thus the longest time to complete the trials. Because patient enrollment often needs to be suspended after each dose cohort to wait for toxicity assessment, for a fixed total sample size, a smaller cohort size or, equivalently, a larger number of cohorts, means that the trial will be suspended more often and thus will take longer time to complete. It may be speculated that having a smaller cohort size allows more frequent updating of the dose-toxicity models and thus more accurate MTD estimation, but our results demonstrate that using a smaller cohort size does not necessarily yield higher PCS. When choosing cohort size in practice, we recommend conducting simulations to understand the differences in PCS and to consider its practical implications such as total length (time) of the trial.

5. Conclusions

In summary, model-assisted designs (BOIN and Keyboard/mTPI2) perform as well as the CRM, and also offer similar simplicity in implementation to the traditional 3 + 3 design. The advantage of the CRM is that it can be more efficient than the BOIN and Keyboard because it utilizes all available information (all toxicities across all doses), but the

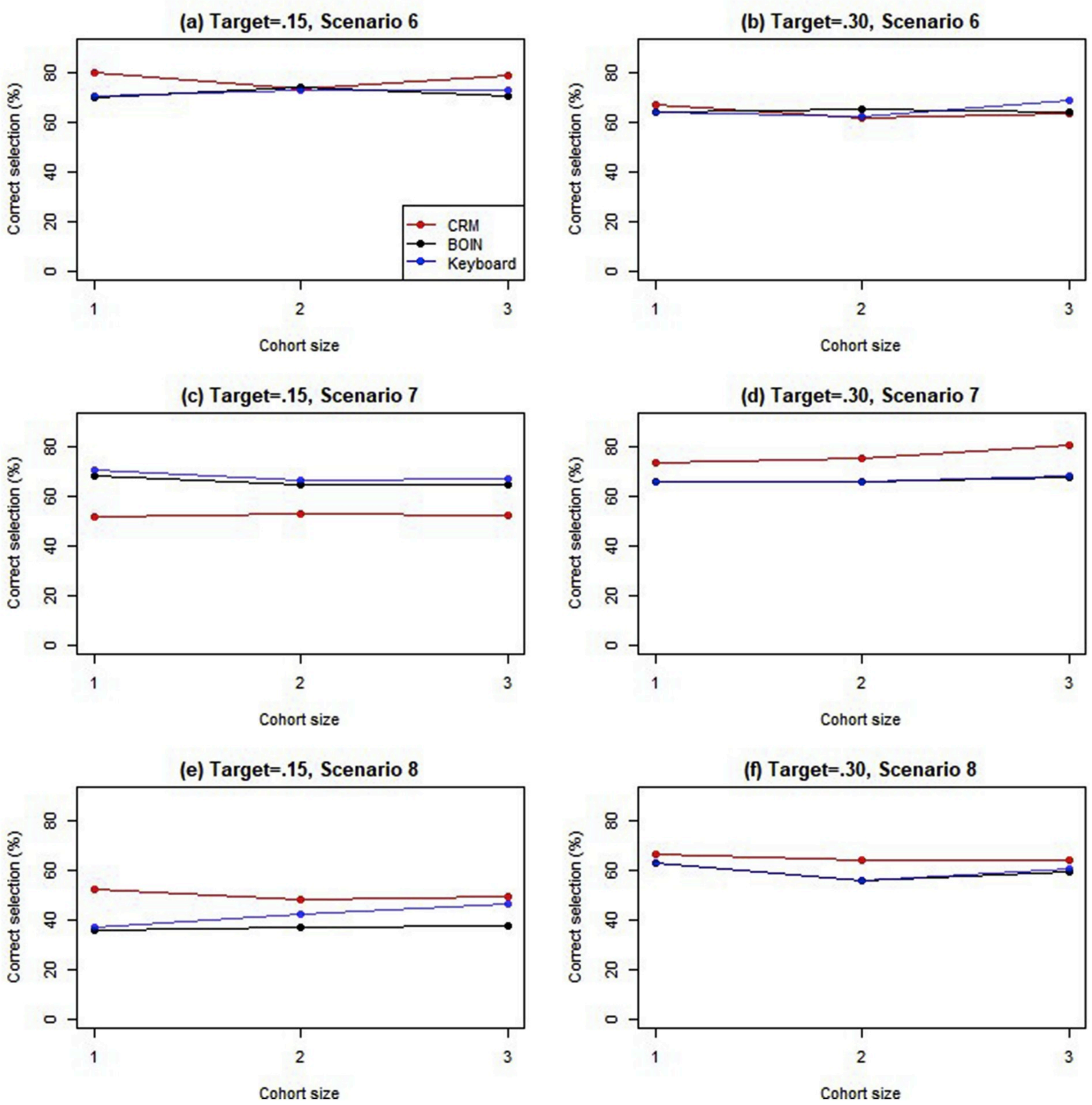

Fig. 7. PCS by cohort sizes, for 3 dose levels, assuming maximum sample size of 18.

performance of the CRM may be affected by how close the prior is to the true toxicity rates. When selecting design parameters such as maximal sample size and cohort size, simulations for the particular clinical settings are necessary to understand the pros and cons of different choices. The decision making should take into account both the statistical performance such as PCS and practical considerations such as feasibility and availability of resources for completing the trial.

Acknowledgements

We would like to acknowledge Ying Yuan for sharing his R codes for the Keyboard design.

Appendix

The following is sample R code for the CRM design for target toxicity of 0.15 using Scenario 1 under the 6 dose level setting, assuming a maximum sample size of 36 and a cohort size of 3. Specifically, we used the default intercept value of 3 for the one-parameter logistic model. The default method is “bayes”, which uses a normal prior with mean 0 and default variance of 1.34.

```
true_tox <-c(0.05, 0.12, 0.20, 0.30, 0.45, 0.60)
theta <-0.15
ndose <-6
library(dfcrm)
prior < -getprior(halfwidth=0.25*theta, target =
theta, nu = round(ndose/2), nlevel = ndose,
model = "logistic")
crmsim(PI = true_tox, prior = prior, target = theta,
n = 36, x0 = 1, nsim = 1000, mcohort = 3, restrict = TRUE,
count = FALSE, model = "logistic”, seed = 134)
```

The following is sample R code for the BOIN design for target toxicity of 0.15 using Scenario 1 under the 6 dose level setting, assuming a maximum sample size of 36 and a cohort size of 3:

```
true_tox <-c(0.05, 0.12, 0.20, 0.30, 0.45, 0.60)
theta <-0.15
library(BOIN)
get.oc(target = theta, p.true = true_tox, ncohort = 12,
cohortsize = 3, seed = 134)
```

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.conctc.2019.100379.

Funding

This research is partly supported by the University of Pennsylvania Abramson Cancer Center Biostatistics Core funded by the NIH/NCI Cancer Center Support Grant P30CA016520 (W-TH and LY).

Declaration of conflicting interests

The authors declare that there is no conflict of interest with respect to the research, authorship, and/or publication of this article.

References

[1] L.V. Rubinstein, R.M. Simon, Phase I clinical trial design, in: D.R. Budman, A.H. Calvert, E.K. Rowinsky (Eds.), Handbook of Anticancer Drug Development, Lippincott Williams & Wilkins, Philadelphia, PA, 2003, pp. 297–308.

[2] T. Jaki, S. Clive, C.J. Weir, Principles of dose finding studies in cancer: a comparison of trial designs, Cancer Chemother. Pharmacol. 71 (2013) 1107–1114, https://doi.org/10.1007/s00280-012-2059-8.

[3] F. Yan, S.J. Mandrekar, Y. Yuan, Keyboard: a novel bayesian toxicity probability interval design for phase I clinical trials, Clin. Cancer Res. 23 (2017) 3994–4003, https://doi.org/10.1158/1078-0432.CCR-17-0220.

[4] B.E. Storer, Design and analysis of phase I clinical trials, Biometrics 45 (1989) 925–937.

[5] J. O'Quigley, M. Pepe, L. Fisher, Continual reassessment method: a practical design for phase 1 clinical trials in cancer, Biometrics 46 (1990) 33–48.

[6] Y.K. Cheung, Dose Finding by the Continual Reassessment Method, Chapman Hall/CRC, Boca Raton, FL, 2011.

[7] Y. Ji, P. Liu, Y. Li, et al., A modified toxicity probability interval method for dose-finding trials, Clin. Trials 7 (2010) 653–663, https://doi.org/10.1177/1740774510382799.

[8] W. Guo, S. Wang, S. Yang, et al., A Bayesian interval dose-finding design addressingOckhams razor: mTPI-2, Contemp. Clin. Trials 58 (2017) 23–33, https://doi.org/10.1016/j.cct.2017.04.006.

[9] S. Liu, Y. Yuan, Bayesian optimal interval designs for phase I clinical trials, J. R. Stat. Soc. Ser. C Appl. Stat. 64 (2014) 507–523, https://doi.org/10.1158/1078-0432.CCR-16-0592.

[10] 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 (2016) 291–300, https://doi.org/10.1002/sim.7043.

[11] H. Zhou, T.A. Murray, H. Pan, et al., Comparative review of novel model-assisted designs for phase I clinical trials, Stat. Med. 37 (2018) 2208–2222, https://doi.org/10.1002/sim.7674.

[12] S.L. Topalian, F.S. Hodi, J.R. Brahmer, et al., Safety, activity, and immune correlates of anti-PD-1 antibody in cancer, N. Engl. J. Med. 366 (2012) 2443–2454, https://doi.org/10.1056/NEJMoa1200690.

[13] C.A. Ramos, B. Ballard, H. Zhang, et al., Clinical and immunological responses after CD30-specific chimeric antigen receptor-redirected lymphocytes, J. Clin. Investig. 127 (2017) 3462–3471, https://doi.org/10.1172/JCI94306.

[14] E.L. Korn, D. Midthune, T.T. Chen, et al., A comparison of phase I trial designs, Stat. Med. 13 (1994) 1799–1806.

[15] Y. Yuan, K.R. Hess, S.G. Hilsenbeck, et al., Bayesian optimal interval design: a simple and well-performing design for phase I oncology trials, Clin. Cancer Res. 22 (2016) 4291–4301, https://doi.org/10.1111/rssc.12089.

[16] S.E. Dahlberg, G.I. Shapiro, J.W. Clark, et al., Evaluation of statistical designs in phase I expansion cohorts: the dana-farber/harvard cancer center experience, JNCI 106 (2014), https://doi.org/10.1093/jnci/dju163.

[17] P.S. Boonstra, J. Shen, J.M. Taylor, et al., A statistical evaluation of dose expansion cohorts in phase I clinical trials, JNCI 107 (2015), https://doi.org/10.1093/jnci/dju429.

[18] E. Garret-Mayer, The continual reassessment method for dose-finding studies: a tutorial, Clin. Trials 3 (2006) 57–71, https://doi.org/10.1191/1740774506cn134oa.

[19] S.M. Lee, Y.K. Cheung, Model calibration in the continual reassessment method, Clin. Trials 6 (2009) 227–238, https://doi.org/10.1177/1740774509105076.

[20] A.P. Oron, D. Azriel, P.D. Hoff, Dose-finding designs: the role of convergence properties, Int. J. Biostat. 7 (2011) 1–17, https://doi.org/10.2202/1557-4679.1298.

[21] R.E. Barlow, Statistical Inference under Order Restrictions: the Theory and Application of Isotonic Regression, Wiley, London, 1972.

[22] Y. Ji, S.J. Wang, Modified toxicity probability interval design: a safer and more reliable method than the 3 + 3 design for practical phase I trials, J. Clin. Oncol. 31 (2013) 1785–1791, https://doi.org/10.1200/JCO.2012.45.7903.

[23] C. Ahn, An evaluation of phase I cancer clinical trial designs, Stat. Med. 17 (1998) 1537–1549.

[24] A. Kim, E. Fox, K. Warren, et al., Characteristics and outcome of pediatric patients enrolled in phase I oncology trials, Oncol. 13 (2008) 679–689, https://doi.org/10.1634/theoncologist.2008-0046.

[25] A. Srinivasan, K.A. Kasow, S. Cross, et al., Phase I study of the tolerability and pharmacokinetics of palifermin in children undergoing allogeneic hematopoietic stem cell transplantation, BB and MT 18 (2012) 1309–1314, https://doi.org/10.1016/j.bbmt.2012.04.013.

[26] M.S. Merchant, M. Wright, K. Baird, et al., Phase I clinical trial of ipilimumab in pediatric patients with advanced solid tumors, Clin. Cancer Res. 22 (2015) 1364–1370, https://doi.org/10.1158/1078-0432.CCR-15-0491.

[27] P. Chevallier, T. Eugene, N. Robillard, et al., 90 Y-labelled anti-CD22 epratuzumab tetraxetan in adults with refractory or relapsed CD22-positive B-cell acute lymphoblastic leukaemia: a phase 1 dose-escalation study, Lancet Haematol. 2 (2015) e108–e117, https://doi.org/10.1016/S2352-3026(15)00020-4.

[28] G. Yin, Y. Yuan, Bayesian model averaging continual reassessment method in phase I clinical trials, J. Am. Stat. Assoc. 104 (2009) 954–968, https://doi.org/10.1198/jasa.2009.ap08425.