Clinical Cancer Research

Accuracy, Safety, and Reliability of Novel Phase I Trial Designs

Heng Zhou[1], Ying Yuan[1], and Lei Nie[2]

Abstract

A number of novel model-based and model-assisted designs have been proposed to find the MTD in phase I clinical trials, but their differences and relative pros and cons are not clear to many practitioners. We review three model-based designs, including the continual reassessment method (CRM), dose escalation with overdose control (EWOC), and Bayesian logistic regression model (BLRM), and three model-assisted designs, including the modified toxicity probability interval (mTPI), Bayesian optimal interval (BOIN), and keyboard (equivalently mTPI-2) designs. We conduct numerical studies to assess their accuracy, safety, and reliability and the practical implications of various empirical rules used in some designs, such as skipping a dose and imposing overdose control. Our results show that the CRM outperforms EWOC and BLRM with higher accuracy of identifying the MTD. For the CRM, skipping a dose is not recommended, as it substantially increases the chance of overdosing patients while providing limited gain for identifying the MTD. EWOC and BLRM appear excessively conservative. They are safe but have relatively poor accuracy of finding the MTD. The BOIN and keyboard (equivalently mTPI-2) designs have similar operating characteristics, outperforming the mTPI, but the BOIN is more intuitive and transparent. The BOIN yields competitive performance comparable with the CRM but is simpler to implement and free of the issue of irrational dose assignment caused by model misspecification, thereby providing an attractive approach for designing phase I trials. *Clin Cancer Res; 24(18); 4357–64. ©2018 AACR.*

Introduction

Most phase I oncology clinical trials are designed to identify the MTD of a new drug, which is defined as the dose with a dose-limiting toxicity (DLT) probability that is closest to the target DLT probability. Due to its simplicity, the 3 + 3 design (1) has been dominant in phase I clinical trials for decades despite its poor ability to identify the MTD and tendency to treat patients at low doses that are potentially subtherapeutic (2). The 3 + 3 design and its variations are called "algorithm-based designs" because they use simple, prespecified rules to guide dose escalation.

Novel phase I trial designs that have been proposed to improve the efficiency of identifying the MTD include model-based designs and model-assisted designs. The continual reassessment method (CRM) is a typical example of a model-based design that assumes a parametric model for the dose–toxicity curve and then based on the accumulating trial data, continuously updates the estimate of the curve to guide the dose assignment and MTD selection (3). Various extensions of the CRM have been proposed, including dose escalation with overdose control (EWOC; ref. 4), the Bayesian logistic regression model (BLRM; ref. 5), the time-to-event CRM (6), and the Bayesian model averaging CRM (7), among others. The CRM has better operating characteristics than algorithm-based designs (2, 8); however, its use in practice has been limited probably due to its requirement of repeated model fitting, its conceptual and computational complexity, and its nontransparent approach to decision-making.

Model-assisted designs are a relatively new class of designs that combine the superior performance of model-based designs with the simplicity of algorithm-based designs (9, 10). Such 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. Examples of model-assisted designs are the modified toxicity probability interval (mTPI) design (11) and its variation, mTPI-2 (12), the Bayesian optimal interval (BOIN) design (13, 14), and the keyboard design (9). Because of their good performance and simplicity, model-assisted designs have been increasingly used in practice.

The development of these novel designs provides practitioners with an array of tools for conducting more flexible and efficient phase I trials. We seek to compare these designs to determine their differences and relative pros and cons. In addition, some designs (e.g., CRM and BLRM) suggest optional empirical rules to regulate dose escalation, such as whether dose skipping should be allowed or an overdose control rule should be applied. On the basis of our experience with phase I trials at the FDA and The University of Texas MD Anderson Cancer Center (Houston, TX), we observe that some protocols impose these empirical rules, whereas others do not. The practical implications of these empirical rules are not clear. To fill these knowledge gaps, we reviewed several novel phase I designs, including the model-based CRM,

[1]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, Texas. [2]DBV/OB/OTS/Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, Maryland.

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

Corresponding Author: Lei Nie, U.S. Food and Drug Administration, 10903 New Hampshire Avenue, Silver Spring, MD 20993. Phone: 301-796-2422; Fax: 301-796-8733; E-mail: Lei.Nie@fda.hhs.gov

doi: 10.1158/1078-0432.CCR-18-0168

©2018 American Association for Cancer Research.

Downloaded from http://aacrjournals.org/clincancerres/article-pdf/24/18/4357/2047511/4357.pdf by guest on 27 August 2022

EWOC, and BLRM designs, and model-assisted mTPI, BOIN, and keyboard designs, and conducted a Monte Carlo experiment (i.e., computer simulations) to compare their operating characteristics. We note that mTPI-2 ends up with the same design as the keyboard design, thus the results for the keyboard design also apply to mTPI-2.

An important issue that we examined that is largely overlooked in the existing literature is the reliability of the design, which is defined as the likelihood of extreme problematic trial behavior occurring under a design (13), for example, the likelihood of a design overdosing more than 50% of the patients and the likelihood of a design failing to de-escalate the dose when 2/3 or $\geq$3/6 patients had DLTs at that dose. The incidence of such extreme behavior in a trial design may be low but is of serious practical concern when it occurs. Our study reveals some new, intriguing design behaviors that have important practical implications. For example, two designs may have similar performance in some commonly used metrics (e.g., the average number of patients treated above the MTD) but rather different likelihood of overdosing more than 50% of the patients and failing to de-escalate the dose when 2/3 or $\geq$3/6 patients had DLTs.

Materials and Methods

Review of Bayesian designs

CRM. The CRM is a model-based design that assumes a parametric model for the dose–toxicity curve. Let p_j denote the true DLT probability of dose level j, and ϕ denote the target DLT probability, for example, $\phi = 0.2$ or 0.3. A commonly used model for the CRM is the power model:

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

where α is the unknown parameter, and $0 < a_1 < \ldots < a_J < 1$ are prior guesses for the DLT probability at each dose (i.e., the initial estimate of the dose–toxicity curve), usually elicited from clinicians. This model links the true dose–toxicity curve (i.e., p_j's) with its prior estimate (i.e., a_j's). After each patient cohort is treated, the CRM updates the estimate of the dose–toxicity curve based on the accumulating DLT data across all dose levels and assigns the next cohort of patients to the "optimal" dose, defined as the dose whose posterior mean estimate of the DLT probability is closest to the target ϕ. Through this continuously updating process, the CRM seamlessly incorporates the newly observed DLT data into the estimation of p_j and the decision of dose escalation and de-escalation. As illustrated in Fig. 1A, the observation of DLTs tends to lift the dose–toxicity curve, leading to dose de-escalation, and the observation of no DLT tends to lower the dose–toxicity curve, leading to dose escalation. 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 ϕ. In the original CRM, new patients are always assigned to the currently estimated "optimal" dose, which may lead to skipping untried doses. In practice, many trials impose a rule that forbids skipping doses and restricts dose escalation and de-escalation to one level at a time. In addition, a safety stopping rule is often used: Stop the trial if the posterior probability that the DLT probability of the lowest dose is greater than ϕ exceeds a certain threshold, such as 0.95, that is, $\Pr(p_1 > \phi | data) > 0.95$.

EWOC. The EWOC is a modification of the CRM (4). The EWOC employs a two-parameter logistic regression model to provide extra flexibility to model the dose–toxicity curve:

$$\text{logit}(p_j) = \beta_0 + \beta_1 d_j,$$

where β_0 and β_1 are the unknown intercept and slope parameters, and d_j is the dosage of dose level j. Similar to the CRM, the EWOC continuously updates the estimate of the dose–toxicity curve based on the accumulating data and assigns the next cohort of patients to the currently estimated "optimal" dose. The difference is that the EWOC uses a different definition of the optimal dose to actively control the risk of overdosing. The EWOC defines the optimal dose as the highest dose whose posterior probability of being higher than the MTD is equal to or less than a prespecified threshold α, with the recommended value of $\alpha = 25\%$. In the EWOC, dose skipping is not allowed, and dose escalation and de-escalation are restricted to one level at a time. A detailed description of EWOC is provided in Supplementary Appendix SA3.

BLRM. The BLRM is another modification of the CRM. The BLRM uses the similar two-parameter logistic regression model as the EWOC (5). Similar to the CRM and EWOC, the BLRM also continuously updates the estimate of the dose–toxicity curve based on the accumulating data and assigns the next cohort of patients to the currently estimated "optimal" dose. The BLRM uses a slightly different definition for the "optimal" dose from the CRM and EWOC. Specifically, let (δ_1, δ_2) denote the proper dosing interval, which means that any dose with the DLT probability within that interval can be approximately accepted as the MTD. For example, given target $\phi = 0.25$, the interval (0.2, 0.3) may be defined as the proper dosing interval. The BLRM defines the "optimal" dose as the dose that has the highest posterior probability of being within (δ_1, δ_2). The BLRM typically imposes an overdose control rule similar to the EWOC, which says that if the observed data suggest that there is 25% or higher (posterior) probability that the DLT rate of a dose is greater than δ_2, that is, $\Pr(p_j > \delta_2 | data) \geq 0.25$, that dose is overdosing and cannot be used to treat patients. This overdose control rule naturally leads to the following safety stopping rule: Stop the trial if the lowest dose is overdosing. In the BLRM, dose skipping is not allowed. A detailed description of the BLRM is provided in Supplementary Appendix SA4.

mTPI design. The mTPI design is a model-assisted design. It starts from the specification of three intervals: the proper dosing interval (δ_1, δ_2), the underdosing interval $(0, \delta_1)$, and the overdosing interval $(\delta_2, 1)$. Unlike the CRM, which assumes a parametric model to specify the whole dose–toxicity curve, the mTPI uses a beta-binomial model locally to describe the toxicities at the current dose only and makes the decision of dose escalation and de-escalation based on the unit probability mass (UPM) of the three intervals. Let p_{cur} denote the DLT probability of the current dose. The UPM of an interval is defined as the posterior probability that p_{cur} is within the interval divided by the length of the interval. Graphically, the UPM of an interval is the area under the posterior distribution curve of p_{cur} within the interval divided by the interval length (see Fig. 1B). Let UPM1, UPM2, and UPM3 denote the UPM for the underdosing, proper dosing, and

Downloaded from http://aacrjournals.org/clincancerres/article-pdf/24/18/4357/2047511/4357.pdf by guest on 27 August 2022

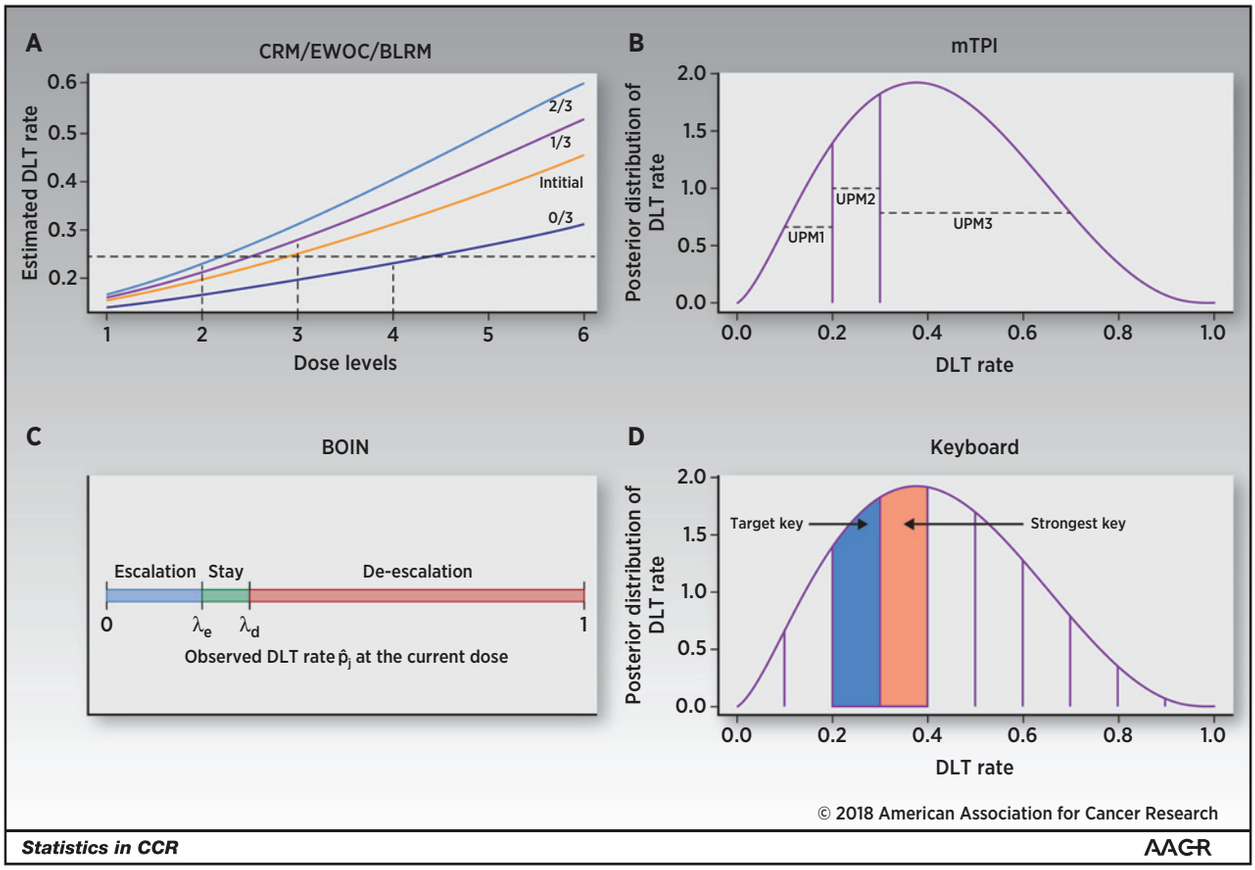

Figure 1.

Decision of dose escalation and de-escalation under the CRM/EWOC/BLRM, mTPI, BOIN, and keyboard designs. **A,** CRM/EWOC/BLRM uses the estimated dose–toxicity curve that is continuously updated on the basis of accumulative data; the curve labeled "Initial" is the initial estimate of the dose–toxicity curve before the first cohort is treated; and curve labels "0/3," "1/3," and "2/3" represent the updated estimate of the dose–toxicity curve when 0/3 and 1/3 and 2/3 patients had DLTs, respectively. **B,** mTPI calculates and compares the UPMs of the underdosing, proper dosing, and overdosing intervals. **C,** BOIN compares the observed DLT rate at the current dose with the prespecified dose-escalation boundary λ_e and de-escalation boundary λ_d. **D,** The keyboard design forms a series of equal-width keys and bases the decision on the position of the strongest key with respect to the target key.

overdosing intervals, respectively. The mTPI design determines dose escalation/de-escalation as follows:

- If UPM1 is the maximum of the three UPMs, escalate to the next higher dose;
- If UPM2 is the maximum of the three UPMs, stay at the current dose;
- If UPM3 is the maximum of the three UPMs, de-escalate to the next lower dose.

Because the UPM can be calculated given the number of treated patients and the observed number of toxicities, dose escalation and de-escalation can be determined before the onset of the trial, which makes the mTPI design user-friendly for practitioners. The trial continues in this manner until the prespecified sample size is exhausted. At that point, the MTD is selected as the dose for which the isotonic estimate (15) of the DLT probability is closest to ϕ. In the mTPI design, a dose exclusion/early stopping rule is included: If the observed data suggest that the posterior probability that the DLT rate of the current dose is greater than the target ϕ exceeds 0.95, that is, $\Pr(p_{cur} > \phi|data) > 0.95$, the current dose and higher doses are excluded from the trial; if the lowest dose is excluded, the trial is terminated.

Keyboard design. Yan and colleagues proposed the keyboard design to improve the performance of the mTPI design, noting that the latter has a high risk of overdosing patients due to the use of the UPM to guide dose escalation (9). Unlike the mTPI, which specifies three unequal-width dosing intervals, the keyboard design constructs a series of equal-width dosing intervals, referred to as keys, to guide dose escalation and de-escalation (see Fig. 1D). The keyboard design starts by eliciting the proper dosing interval (referred to as the target key) from clinicians, and then forms a series of equal-width keys on both sides of the target key. For example, in Fig. 1D, after specifying the target key as (0.2, 0.3), we then form two keys with the width of 0.1 on the left side of the target key, including (0, 0.1) and (0.1, 0.2), and seven keys with the width of 0.1 on the right side of the target key, including (0.3, 0.4), . . ., (0.9, 1). The keyboard design makes the decision of dose escalation and de-escalation based on the location of the "strongest" key, defined as the key that has the largest area under the posterior distribution curve of p_{cur} (see Fig. 1D), which can be

Downloaded from http://aacrjournals.org/clincancerres/article-pdf/24/18/4357/2047511/4357.pdf by guest on 27 August 2022

easily obtained using the beta-binomial model, as with the mTPI. Statistically, the strongest key represents the interval in which p_{cur} is most likely located. For example, if the strongest key is (0.3, 0.4), this means that the DLT probability of the current dose is most likely in (0.3, 0.4). Therefore, if the strongest key is on the left side of (i.e., smaller than) the target key, it means that the current dose is underdosing patients, and if the strongest key is on the right side of (i.e., greater than) the target key, it means that the current dose is overdosing patients. This intuitive interpretation of the strongest key naturally leads to the following keyboard dose escalation and de-escalation rule:

- If the strongest key is on the left side of the target key, escalate to the next higher dose;
- If the strongest key is the target key, stay at the current dose level;
- If the strongest key is on the right side of the target key, de-escalate to the next lower dose.

The trial continues until the prespecified sample size is exhausted and the MTD is selected as the dose for which the isotonic estimate (15) of the DLT probability is closest to ϕ. The keyboard imposes the same dose exclusion/early stopping rule as follows: If $\Pr(p_{cur} > \phi|data) > 0.95$ and at least 3 patients have been treated, the current and higher doses are eliminated from the trial. The trial is terminated if the lowest dose is eliminated.

BOIN design. Compared with the mTPI and keyboard designs, which require calculating the posterior distribution of p_{cur}, the BOIN design is more straightforward and transparent. Let $\hat{p}_{cur}$ denote the observed DLT rate at the current dose, defined as $\hat{p}_{cur} =$ (the number of patients who experienced DLTs at the current dose)/(the number of patients treated at the current dose). The BOIN design makes the decision of dose escalation and de-escalation simply by comparing the observed DLT rate $\hat{p}_{cur}$ with a pair of fixed, predetermined dose escalation and de-escalation boundaries (denoted by λ_e and λ_d, respectively) as follows (see Fig. 1C):

- If $\hat{p}_{cur} \leq \lambda_e$, escalate to the next higher dose;
- If $\hat{p}_{cur} \geq \lambda_d$, de-escalate to the next lower dose;
- Otherwise, stay at the current dose.

Table 1 provides the values of λ_e and λ_d for some commonly used target DLT probabilities. The trial continues in this manner until the prespecified sample size is exhausted. At that point, the MTD is selected on the basis of the isotonic estimates of the DLT probabilities as described before. The BOIN design imposes a dose elimination/early stopping rule as follows: If $\Pr(p_{cur} > \phi|data) > 0.95$ and at least 3 patients have been treated, the current and higher doses are eliminated from the trial. The trial is terminated if the lowest dose is eliminated.

The respective dose escalation and de-escalation boundaries λ_e and λ_d are derived from a pair of prespecified toxicity probability thresholds ϕ_1 and ϕ_2, where ϕ_1 is the highest DLT probability that is deemed to be underdosing such that dose escalation is needed; ϕ_2 is the lowest DLT probability that is deemed to be overdosing such that dose de-escalation is needed. Liu and Yuan (13) provided general guidance to specify ϕ_1 and ϕ_2 with recommended default values of $\phi_1 = 0.6\phi$ and $\phi_2 = 1.4\phi$, and the formula to calculate λ_e and λ_d based on ϕ_1 and ϕ_2. As ϕ_1 and ϕ_2 represent the unacceptable DLT probabilities such that dose escalation and de-escalation are required, they are different from the proper dosing interval (δ_1, δ_2) specified in the BLRM, mTPI, and keyboard designs. In general, $\phi_1 < \delta_1$ and $\phi_2 > \delta_2$.

Table 1. Dose escalation and de-escalation boundaries of the BOIN design

	Target DLT probability (ϕ)					
Boundary	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

As the observed DLT rate $\hat{p}_{cur}$ is the most intuitive and also optimal (i.e., the maximum likelihood) estimate of the true DLT probability at the current dose, using $\hat{p}_{cur}$ to determine dose escalation and de-escalation makes the BOIN particularly transparent and assessable for nonstatisticians. Specifically, because the BOIN design guarantees de-escalating the dose when $\hat{p}_{cur}$ is higher than the de-escalation boundary λ_d, clinicians and regulatory agents can easily assess the safety of the trial. 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 $\lambda_d = 0.298$ (i.e., the default de-escalation boundary). 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. For example, suppose that 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 design automatically guarantees de-escalating the dose if the observed DLT rate $\hat{p}_{cur} > \lambda_d = 0.25$. Such flexibility and transparency are an important practical advantage of the BOIN design.

Monte Carlo experiment

Simulation settings. We conducted a Monte Carlo experiment to compare the performance of the CRM, EWOC, BLRM, mTPI, BOIN, and keyboard designs with respect to the 3 + 3 design. We considered three target DLT probabilities $\phi = 0.20$, 0.25, and 0.30, with six dose levels and a maximum sample size of 36. The starting dose level is 1. We considered five model-based designs: CRM (forbids dose skipping), CRM-DS (allows dose skipping), BLRM (with the overdose control rule), BLRM-NOC (with no overdose control rule), and EWOC. The detailed configurations for these designs are provided in Supplementary Appendix SA. We set the proper dosing interval $(\delta_1, \delta_2) = (\phi - 0.05, \phi + 0.05)$ for the mTPI, keyboard, BLRM, and BLRM-NOC designs, and $\phi_1 = 0.6\phi$ and $\phi_2 = 1.4\phi$ for BOIN, as recommended by these designs. The 3 + 3 design often completes (e.g., when 2/3 or 2/6 had DLTs) before reaching its maximum sample size. For comparability, after the 3 + 3 design selects the MTD, an expansion cohort is treated at the MTD to reach the total sample size of 36. To avoid cherry-picking and inadvertent selection biases, we randomly generated 1,000 true dose–toxicity scenarios (or curves) using the pseudouniform algorithm (16) for comparing the designs. Under each scenario, we conducted 2,000 simulated trials. Supplementary Fig. S1 (in Supplementary Appendix SB) shows 25 randomly selected scenarios that display various shapes of the dose–toxicity curve, and Supplementary Appendix SC shows eight representative scenarios with the corresponding simulation results. We considered cohort sizes of 3 and 1 for all designs except the 3 + 3 design. As the results are generally similar, below we focus on the cohort size of 3 with the target DLT

Downloaded from http://aacrjournals.org/clincancerres/article-pdf/24/18/4357/2047511/4357.pdf by guest on 27 August 2022

probability of 0.25. The results for cohort size 1 (Supplementary Appendix SD) and target DLT probabilities of 0.20 (Supplementary Appendix SE) and 0.30 (Supplementary Appendix SF) are provided in the Supplementary Appendix.

Performance metrics.

- Accuracy
  1. The percentage of correct selection (PCS), which is defined as the percentage of simulated trials in which the target dose is correctly selected as the MTD. When all the dose levels are above the MTD (i.e., the DLT probability of the lowest dose > $\phi + 0.1$), PCS is defined as the percentage of early termination of trials.
  2. The average percentage of patients who are assigned to the MTD across the simulated trials. When all the dose levels are above the MTD (i.e., the DLT probability of the lowest dose > $\phi + 0.1$), we use the average percentage of patients not enrolled into the trial for this metric.
- Safety
  3. The percentage of simulated trials in which a toxic dose with the true DLT probability $\geq$33% is selected as the MTD.
  4. The average percentage of patients assigned to the toxic doses with true DLT probability $\geq$ 33%.
- Reliability
  5. The risk of overdosing, defined as the percentage of simulated trials with more than x% of patients treated at doses above the MTD. In our simulation study, we set $x\% = 50\%$, that is, measuring the likelihood of a design assigning more than half of the patients to doses above the MTD.
  6. The risk of poor allocation, defined as the percentage of simulated trials in which fewer than 6 patients are treated at the MTD.
  7. The risk of irrational dose assignment, defined as the percentage of times that the design fails to de-escalate the dose when 2/3 or $\geq$3/6 patients had DLTs at a dose.

Metrics 5 to 7 measure the likelihood of a design demonstrating extreme problematic behaviors (e.g., treating 50% or more patients at toxic doses, or fewer than 6 patients at the MTD), that is, the reliability of the design. Although these metrics are of great practical importance, they are largely overlooked in the existing literature. Note that these reliability metrics are not covered by other metrics. For example, the percentage of patients overdosed (i.e., metric 4) does not cover the risk of overdosing (i.e., metric 5). Two designs can have a similar percentage of patients overdosed but rather different risks of overdosing 50% of the patients (see "Results"). Statistically, metric 4 measures the mean of overdosing, whereas metric 5 measures the tail probability of overdosing. To compare the relative performance of the designs, we used the 3 + 3 design as a benchmark and report the difference between each of the designs and the 3 + 3 design for each metric. For example, the PCS for the CRM is reported as (the PCS of the CRM) – (the PCS of the 3 + 3 design).

Results

Accuracy

Figure 2A and B shows distributions of the PCS and the average percentage of patients treated at the MTD, respectively, for the investigational designs relative to the 3 + 3 design across 1,000 scenarios. As each dose–toxicity scenario generates a value of the performance metric (e.g., PCS), we obtained a total of 1,000 values for each of the metrics across the 1,000 scenarios. The boxplot reflects the distribution of the metric across the 1,000 scenarios. In terms of the accuracy of correctly selecting the MTD, the CRM, mTPI, BOIN, and keyboard designs are comparable, although notable difference is observed when target DLT probability is 0.2 (see Supplementary Appendix SE), and substantially outperform the 3 + 3 design. The BLRM and EWOC perform the worst, with the average PCS similar to that of the 3 + 3 design. The EWOC also has the largest variation in the PCS. The poor accuracy of the BLRM can be addressed by removing the overdose control rule; the BLRM-NOC has the highest average PCS. However, by doing so, the resulting BLRM-NOC becomes overly aggressive and treats a large percentage of patients above the MTD (as shown later). The CRM-DS, which allows dose skipping, has a slightly higher PCS than the CRM but at the cost of increasing the risk of overdosing patients (shown later). The results for the number of patients treated at the MTD are similar to those for the PCS. The CRM, mTPI, BOIN, and keyboard designs are generally comparable and substantially outperform the 3 + 3 design. The mTPI and CRM designs allocate slightly more patients to the MTD than the BOIN and keyboard designs, but the latter two designs are less variable, as shown by the shorter boxes in the boxplot (Fig. 2B). The BLRM and EWOC perform the worst, and BLRM-NOC and CRM-DS perform well, with the highest average percentage of patients treated at the MTD. The EWOC is the most variable method in terms of treating patients at the MTD. To illustrate the performance of the designs under certain specific dose–toxicity curves, Supplementary Appendix SC shows the results under eight representative scenarios. The results are generally similar to Fig. 2.

Safety

As shown in Fig. 2C, the CRM, mTPI, BOIN, and keyboard designs are comparable in terms of the percentage of selecting a toxic dose (with DLT probability $\geq$33%) as the MTD, but CRM and mTPI are slightly more variable than the BOIN and keyboard designs. BLRM-NOC not only has the highest chance of selecting a toxic dose as the MTD but also is the most variable. The BLRM and EWOC designs are the most conservative and least likely to select a toxic dose as the MTD. In terms of the percentage of patients treated at a toxic dose with DLT probability $\geq$33%, BLRM-NOC and CRM-DS stand out as the most aggressive designs, see Fig. 2D. These two designs treat substantially more patients at toxic doses than the other designs and exhibit the largest variation. On average, the CRM, mTPI, BOIN, and keyboard designs are comparable, but BOIN and keyboard show smaller variations.

Reliability

In terms of the risk of overdosing 50% or more of the patients (Fig. 3A), the BLRM, BOIN, and keyboard designs perform the best, and BLRM-NOC performs the worst, with significantly higher (i.e., about 10% higher on average) risk. The performances of the CRM and mTPI designs are similar and rank in between the performances of these other designs. The EWOC has similar averaged risk of overdosing patients as the BOIN and keyboard designs but is much more variable. We note that the CRM, mTPI, BOIN, and keyboard, on average, overdose similar percentages of patients (Fig. 2D) but have different risks of overdosing 50% or more of the patients (Fig. 3A). This indicates that the risk of

Downloaded from http://aacrjournals.org/clincancerres/article-pdf/24/18/4357/2047511/4357.pdf by guest on 27 August 2022

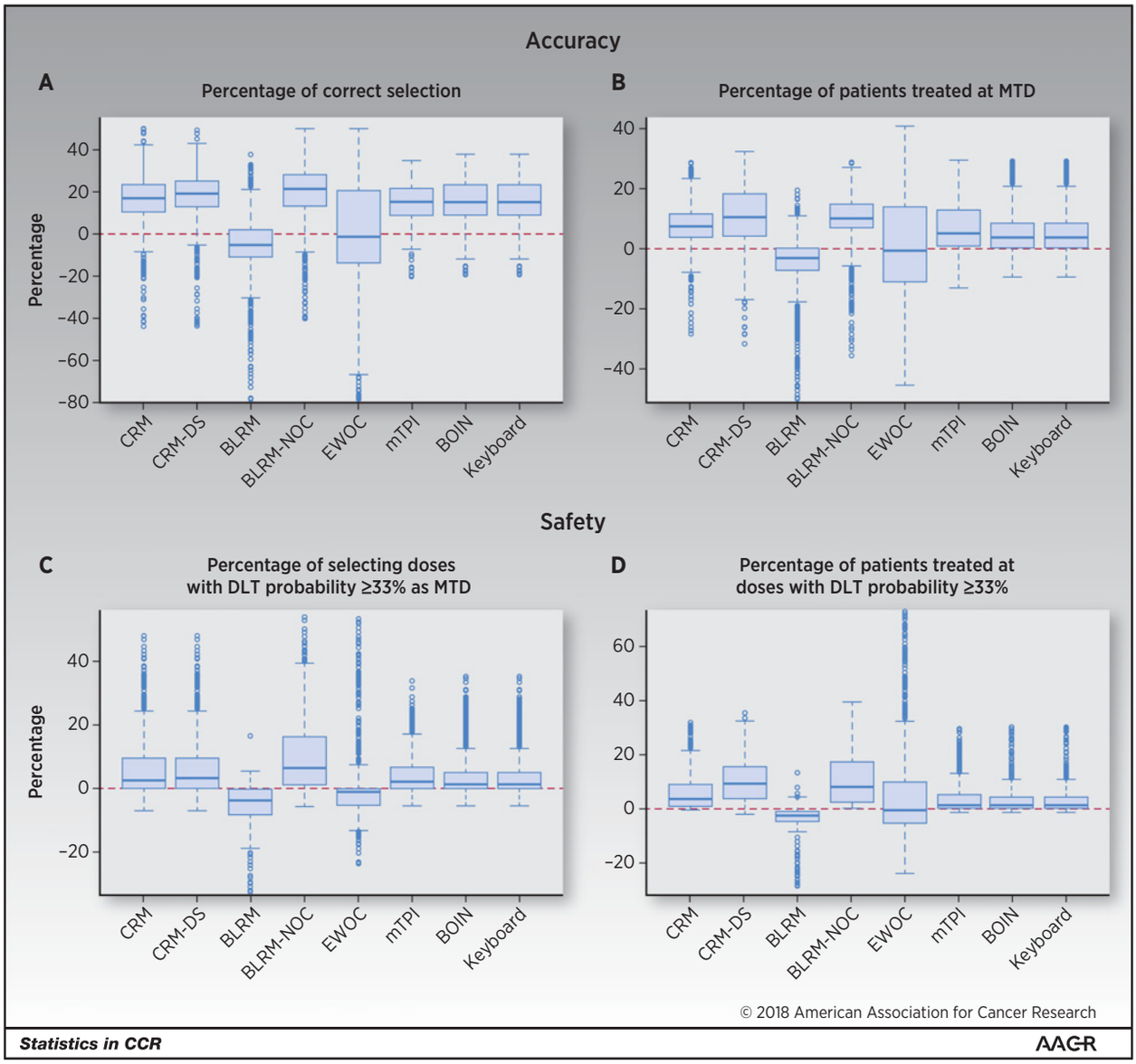

Figure 2.

Accuracy and safety of the eight designs with respect to the 3 + 3 design, including PCS of the MTD (**A**), percentage of patients treated at the MTD (**B**), percentage of selecting doses with DLT probability ≥33% as the MTD (**C**), and percentage of patients treated at doses with DLT probability ≥33% (**D**). For **A** and **B**, a larger value indicates better performance; positive value means that the design outperforms the 3 + 3 design. For **C** and **D**, a smaller value indicates better performance; negative value means that the design outperforms the 3 + 3 design.

overdosing (50% or more patients) and the average percentage of patients overdosed indeed measure different aspects of a design, and it is thus important to consider both metrics when evaluating a design. Compared with the CRM, CRM-DS had about 5% higher risk of overdosing 50% or more of the patients on average due to its aggressive dose skipping. In terms of the risk of poor allocation (i.e., treating fewer than 6 patients at the MTD, see Fig. 3B), BLRM and EWOC perform the worst, with a significantly higher risk than the other designs. The CRM, CRM-DS, BLRM-NOC, BOIN, and keyboard designs have comparable risks of poor allocation, and the keyboard design (thus mTPI-2 as well) improves the mTPI design.

In terms of the risk of irrational dose assignment (Fig. 3C), the model-assisted designs outperform the model-based designs. The model-based designs (i.e., CRM, BLRM, and EWOC) have 8% to 55% chance of failing to de-escalate the dose when 2/3 or ≥3/6 patients had DLTs, whereas such irrational dose assignments never occur in the mTPI, BOIN, and keyboard designs. To the best of our knowledge, this result is new and no literature has studied such in-trial behavior of designs. Our result discloses a disturbing, yet unsurprising, behavior of model-based designs. The model-based designs rely on the assumed model to make the decision of dose assignment. When the model is misspecified, the estimates can be biased, and thus irrational dose assignment

Downloaded from http://aacrjournals.org/clincancerres/article-pdf/24/18/4357/2047511/4357.pdf by guest on 27 August 2022

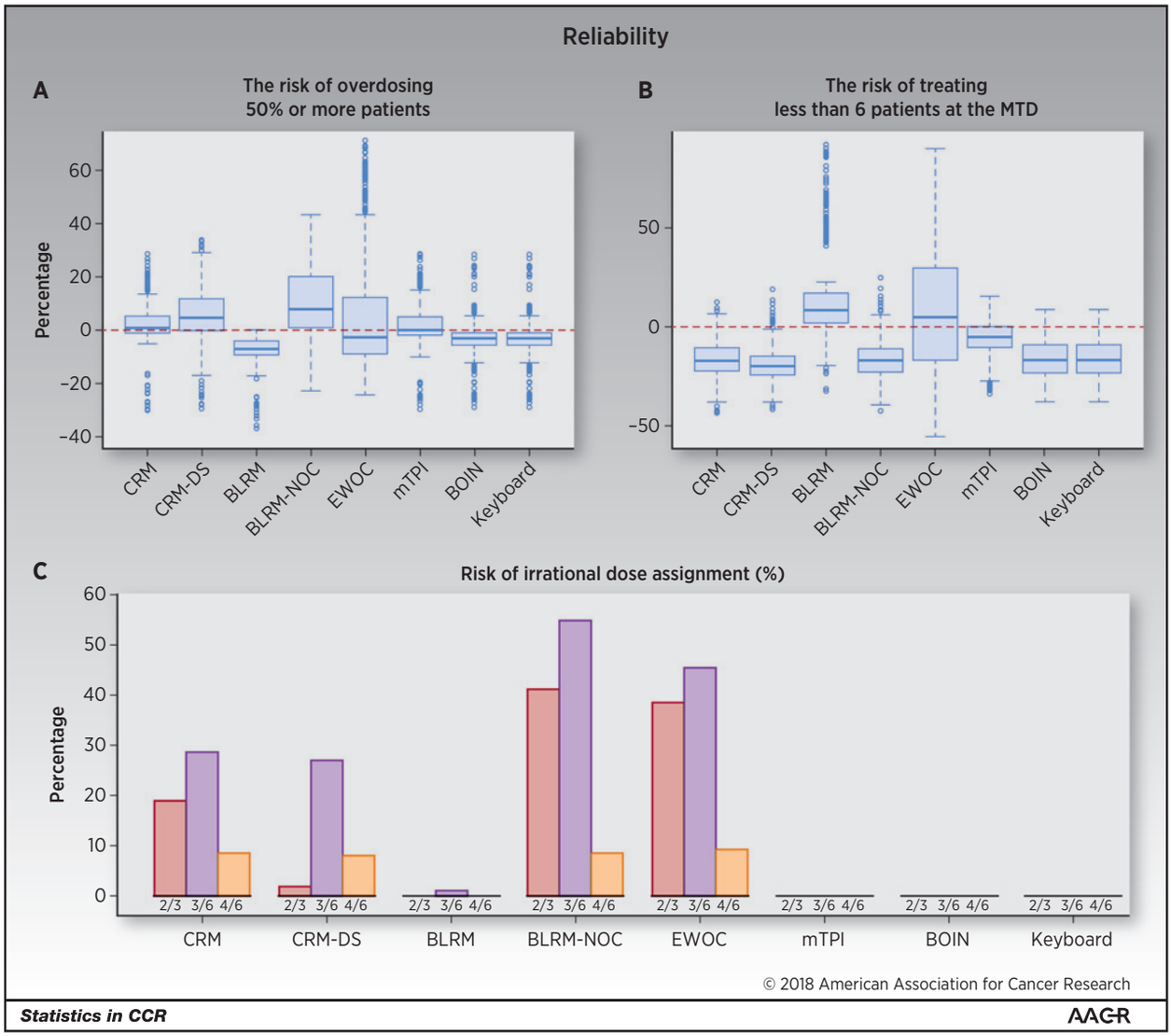

Figure 3.

Reliability of the eight designs with respect to the 3 + 3 design, including risk of overdosing 50% or more patients (**A**), risk of treating <6 patients at the MTD (**B**), and risk of irrational dose assignments (**C**). A smaller value indicates better performance; negative value means that the design outperforms the 3 + 3 design.

arises. The model-assisted designs are free of that issue because they do not impose any model assumption on the dose–toxicity curve. For example, by its dose escalation/de-escalation rule, the BOIN guarantees de-escalating the dose if the observed DLT rate at the current dose is higher than 29.8% given the target DLT rate of 25%.

Discussion

We evaluated the operating characteristics of some novel Bayesian phase I trial designs in terms of accuracy, overdose control, and reliability. Compared with the 3 + 3 design, most of these novel designs yield better accuracy for identifying the MTD and allocate more patients to the MTD. Overall, CRM performs well in most metrics. Allowing dose skipping slightly improves the accuracy of identifying the MTD and the allocation of patients to the MTD but at the cost of substantially increasing the number of overdosed patients and decreasing the design reliability (i.e., a higher risk of overdosing a large percentage of patients). Thus, dose skipping in CRM is generally not recommended, and we suggest restricting dose escalation and de-escalation to one dose level at a time. The performance of BLRM is mixed. BLRM (with the overdose control rule) is excessively conservative and has poor accuracy to identify the MTD and allocate patients to the MTD. Removing the overdose control rule (i.e., BLRM-NOC) improves the accuracy to identify the MTD and allocate patients to the MTD but at the cost of substantially reduced safety (i.e., treating a high percentage of patients above the MTD) and reliability (i.e., high risk of overdosing a large percentage of patients). The overdose control rule commonly used in BLRM seems to be too conservative, and a more appropriate overdose control rule may be needed to make BLRM work

Downloaded from http://aacrjournals.org/clincancerres/article-pdf/24/18/4357/2047511/4357.pdf by guest on 27 August 2022

appropriately. The EWOC appears overly conservative; it is safe but has poor accuracy to identify the MTD. The EWOC has similar average performance as the BLRM but has larger variation. The BOIN and keyboard model-assisted designs yield good performance that is generally comparable with that of CRM in terms of accuracy and safety while often providing smaller variation and better reliability. The mTPI performs well in identifying the MTD and allocating patients to the MTD when the target DLT probability is 0.25, but it has lower reliability with a higher risk of overdosing a large percentage of patients and poor allocation of patients to the MTD. The mTPI has a relatively low accuracy to identify the MTD when the target DLT probability is 0.2 (see Supplementary Appendix SE). Given that BOIN and keyboard are more transparent and easy to implement, they provide attractive approaches to designing phase I clinical trials. The BOIN and keyboard designs have virtually the same performance in every metric. As BOIN uses the observed DLT rate to determine dose escalation and de-escalation, it is more transparent and assessable for nonstatisticians, and it is easier to calibrate to fit the design goal. In addition, as noted by a referee, the BOIN has both Bayesian and frequentist interpretations. Its decision rule is equivalent to using the likelihood ratio test to determine dose escalation/de-escalation (13), making it appealing to wider audiences. In contrast, the mTPI/mTPI2 and keyboard designs have only a Bayesian interpretation and require specification of the prior and calculation of the posterior distribution.

In our Monte Carlo experiment, we used the default design parameters recommended by the designs that are tailored to the "noninformative" case where limited prior knowledge is available on the toxicity profile of the investigational drug. This is appropriate for evaluating and comparing the general performance of the designs across a variety of toxicity profiles, and for first-in-human drug trials. For the "me-too" or same-family drugs with a better known toxicity profile, the design parameters should be calibrated on the basis of the available prior information to fit the trial under consideration. For example, if the prior information suggests that the investigational drug is relatively safe, we can choose the design parameters that encourage more aggressive dose escalation to find the MTD quickly.

The designs reviewed here focus on single-agent trials and require that before enrolling the next cohort of new patients, patients who were enrolled into the trial have completed their DLT assessment. This requirement is troublesome when toxicity is late onset or the accrual is fast. Extension of these novel designs has been developed to address the late-onset toxicity, for example, the TITE-CRM (6) and data augmentation CRM (17), and to handle drug combination trials (18–20).

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Disclaimer

This article reflects the views of the authors and should not be construed to represent the FDA's views or policies.

Authors' Contributions

Conception and design: H. Zhou, Y. Yuan, L. Nie
Development of methodology: H. Zhou, Y. Yuan, L. Nie
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): Y. Yuan
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): H. Zhou, Y. Yuan
Writing, review, and/or revision of the manuscript: H. Zhou, Y. Yuan, L. Nie
Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): H. Zhou
Study supervision: L. Nie

Acknowledgments

Y. Yuan was supported in part by the NIH under award number P50CA098258. The authors thank the three reviewers for their very helpful comments and suggestions.

Received January 15, 2018; revised February 24, 2018; accepted April 12, 2018; published first April 16, 2018.

References

1. Storer BE. An evaluation of phase I clinical trials designs in the continuous dose-response setting. Stat Med 2001;20:2399–408.
2. Le Tourneau C, Lee JJ, Siu LL. Dose escalation methods in phase I cancer clinical trials. J Natl Cancer Inst 2009;101:708–20.
3. O'Quigley J, Pepe M, Fisher L. Continual reassessment method: a practical design for phase I clinical trials in cancer. Biometrics 1990;46:33–48.
4. Babb J, Rogatko A, Zacks S. Cancer phase I clinical trials: efficient dose escalation with overdose control. Stat Med 1998;17:1103–20.
5. Neuenschwander B, Branson M, Gsponer T. Critical aspects of the Bayesian approach to phase I cancer trials. Stat Med 2008;27:2420–39.
6. Cheung YK, Chappell R. Sequential designs for phase I clinical trials with late-onset toxicities. Biometrics 2000;56:1177–82.
7. Yin G, Yuan Y. Bayesian model averaging continual reassessment method in phase I clinical trials. J Am Stat Assoc 2009;104:954–68.
8. Iasonos A, Wilton AS, Riedel ER, Seshan VE, Spriggs DR. A comprehensive comparison of the continual reassessment method to the standard 3 + 3 dose escalation scheme in phase I dose-finding studies. Clin Trials 2008;5:465–77.
9. 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.
10. Zhou H, Murray TA, Pan H, Yuan Y. Comparative review of novel model-assisted designs for phase I clinical trials. Stat Med 2018;37:2208–22.
11. Ji Y, Liu P, Li Y, Bekele N. A modified toxicity probability interval method for dose-finding trials. Clin Trials 2010;7:653–63.
12. 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.
13. Liu S, Yuan Y. Bayesian optimal interval designs for phase I clinical trials. J R Stat Soc Ser C Appl Stat 2015;64:507–23.
14. 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;22:4291–301.
15. Barlow RE, Bartholomew DJ, Bremner J, Brunk HD. Statistical inference under order restrictions: the theory and application of isotonic regression. New York: Wiley; 1972.
16. Clertant M, O'Quigley J. Semiparametric dose finding methods. J R Stat Soc Ser B Stat Methodol 2017;79:1487–508.
17. Liu S, Yin G, Yuan Y. Bayesian data augmentation dose finding with continual reassessment method and delayed toxicity. Ann Appl Stat 2013;4:2138–56.
18. Yin G, Yuan Y. Bayesian dose-finding in oncology for drug combinations by copula regression. J Royal Stat Soc 2009;58:211–24.
19. Wage NA, Conaway MR, O'Quigley J. Continual reassessment method for partial ordering. Biometrics 2011;67:1555–63.
20. Lin R, Yin G. Bayesian optimal interval design for dose finding in drug-combination trials. Stat Methods Med Res 2017;26:2155–67.

Downloaded from http://aacrjournals.org/clincancerres/article-pdf/24/18/4357/2047511/4357.pdf by guest on 27 August 2022