CFO: Calibration-free odds design for phase I/II clinical trials

Huaqing Jin ⓘ and Guosheng Yin ⓘ

Abstract

Recent revolution in oncology treatment has witnessed emergence and fast development of the targeted therapy and immunotherapy. In contrast to traditional cytotoxic agents, these types of treatment tend to be more tolerable and thus efficacy is of more concern. As a result, seamless phase I/II trials have gained enormous popularity, which aim to identify the optimal biological dose (OBD) rather than the maximum tolerated dose (MTD). To enhance the accuracy and robustness for identification of OBD, we develop a calibration-free odds (CFO) design. For toxicity monitoring, the CFO design casts the current dose in competition with its two neighboring doses to obtain an admissible set. For efficacy monitoring, CFO selects the dose that has the largest posterior probability to achieve the highest efficacy under the Bayesian paradigm. In contrast to most of the existing designs, the prominent merit of CFO is that its main dose-finding component is model-free and calibration-free, which can greatly ease the burden on artificial input of design parameters and thus enhance the robustness and objectivity of the design. Extensive simulation studies demonstrate that the CFO design strikes a good balance between efficiency and safety for MTD identification under phase I trials, and yields comparable or sometimes slightly better performance for OBD identification than the competing methods under phase I/II trials.

Keywords

Bayesian method, dose finding, maximum tolerated dose, oncology trial, optimal biological dose

Introduction

In conventional dose finding for oncology treatment, a common assumption is that both efficacy and toxicity of the drug increase monotonically with the dose. Traditional phase I clinical trials mainly focus on toxicity with the goal to determine the maximum tolerated dose (MTD) based on the target dose-limiting toxicity (DLT) rate.[1] However, due to the revolution of the targeted therapy and immunotherapy in cancer treatment,[2] many new agents in clinical oncology violate the monotonic dose–efficacy relationship. For some immunotherapy agents, a higher dose may yield lower efficacy, which leads to an umbrella-shape dose–efficacy curve.[3] An example of a plateau-shape efficacy curve can be observed for the efficacy of PTK/ZK, an orally active inhibitor of vascular endothelial growth factor receptor tyrosine kinases. Its efficacy initially increases with the dose but then remains unchanged after reaching a threshold,[4] which results in a plateau-shape curve. It becomes commonplace to incorporate efficacy evaluation in dose recommendation for oncology clinical trials. By incorporating both efficacy and toxicity data, such dose-finding trials are typically referred to as seamless phase I/II trials, which aim to identify the optimal biological dose (OBD), defined as the dose with the highest efficacy probability while controlling the DLT rate.[5] Due to violation of the monotonic dose–efficacy relationship, the traditional phase I trial designs, such as the "3+3" design,[6] the continual reassessment method (CRM)[7] and non-parametric overdose control (NOC) design,[8] are not applicable any more. Following the trend of phase I/II trials, abundant adaptive designs have been proposed to determine the OBD. Gooley et al. (1994)[9] proposed three two-stage designs for conducting phase I/II trials. Thall and Russell (1998)[10] developed a parametric Bayesian design for phase I/II trials, where a trinary variable was adopted to account for both toxicity and efficacy. As an extension, Thall and Cook (2004)[11] further modified the logistic model and proposed the efficacy–toxicity (EffTox) design which outperformed the original method under a wide range of dose–outcome scenarios. Braun (2002)[12] extended the CRM to

Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong

Corresponding author:
Guosheng Yin, Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong.
Email: gyin@hku.hk

monitor the toxicity and efficacy outcomes simultaneously. Under the Bayesian framework, Yin et al. (2006)[13] proposed a phase I/II trial design using the odds ratio of the efficacy and toxicity as a measure of desirability. Yuan and Yin (2009)[14] developed a Bayesian phase I/II design by jointly modelling the efficacy and toxicity as time-to-event outcomes. Through combining the features from CRM and order restricted inference, Wages and Tait (2015)[15] developed a seamless phase I/II adaptive design. Based on a Bayesian dynamic model, Liu and Johnson (2016)[16] introduced a robust Bayesian design for monitoring efficacy and toxicity outcomes simultaneously. Xu et al. (2016)[17] developed a Bayesian two-stage phase I/II design based on a model adaptation method. By reformulating dose finding as a Bayesian decision-making problem under several simple hypotheses, Lin and Yin (2017)[18] developed a Bayesian interval phase I/II design, named as STEIN (Simple Toxicity and Efficacy INterval design). Riviere et al. (2018)[19] adopted a logistic model with a plateau parameter to investigate drugs with a plateau-shape dose–efficacy relationship in phase I/II trials. Zhou et al. (2019)[20] developed a utility-based Bayesian optimal interval design to determine the OBD in phase I/II trials.

However, all the aforementioned methods either rely upon a parametric model assumption or require tedious specification of design parameters. Either misspecification of the model or inappropriate tuning of design parameters would lead to compromised or even poor trial performance. Our goal is to develop a model-free and calibration-free approach to dose finding, which does not require calibration of any essential design parameters. In general, early stopping rules are not intrinsic part of a design, which serve as external monitoring schemes for safety and futility. The model-free and calibration-free features guarantee our design to be robust and simple for practical use.

Our research is motivated by a collaboration with clinicians on a phase I dose-escalation study of the CD19 chimeric antigen receptor (CAR) induced-T-to-natural-killer (ITNK) cell therapy. The objective of the study was to assess the safety as well as the efficacy of CD19 CAR-ITNK cell therapy in adult patients with relapsed or refractory diffuse large-B-cell lymphoma. Three prespecified doses were considered in the trial, 5×10^5, 7.5×10^5, 10×10^5 CAR-ITNK cells/kg body weight. The DLTs were defined as neurotoxicity or cytokine release syndrome with grade ≥ 3. The efficacy response would be assessed by the 2014 Lugano classification for non-Hodgkin's lymphoma.[21] A related phase I/II trial for the CAR-NK cell therapy in patients with relapsed or refractory CD19-positive cancer had been conducted under a similar protocol,[22] which also investigated three doses: 1×10^5, 1×10^6, 1×10^7 cells/kg. Applying the EffTox design[11] by jointly evaluating the bivariate outcomes, the trial concluded that the MTD was not reached and 73% of patients responded to treatment with no major toxic effect. Given only three dose levels under investigation, it is challenging to apply a model-based design because parametric regression may not fit the data well.

Another motivating example is a phase I/II trial of lenalidomide in combination with high-dose melphalan for patients with relapsed or progressive multiple myeloma.[23] There were four doses of lenalidomide in the dose escalation phase: 25, 50, 75 and 100 mg, while the dose of melphalan was fixed. The goal of the trial was to identify the OBD of lenalidomide in terms of the trade-off between toxicity and efficacy. The DLTs were defined as regimen-related death, graft failure, grade 3 or 4 atrial fibrillation as well as the grade 4 deep venous thrombosis or pulmonary embolism before day 30 after the autologous hematopoietic stem cell transplantation (auto-HCT). The efficacy outcome was defined as being alive in complete response on day 90 after auto-HCT.

The major difficulty in these phase I/II studies is to determine the shapes of dose–efficacy and dose–toxicity curves without adequate prior information. The model-based dose-finding designs may be at risk of violation of parametric assumptions and thus lead to unreliable dose assignment and incorrect OBD identification. Further, most of the existing phase I/II designs require calibration of certain design parameters prior to the implementation. However, due to a lack of preclinical information in the first-in-human study, it is challenging to specify the design parameters suitable for such phase I/II designs. To avoid the potential risk of model misspecification and alleviate the burden of parameter calibration, we propose a calibration-free odds (CFO) design to identify the OBD. The CFO design bypasses all the parametric model assumptions, which is thus model-free or curve-free. Before the dose assignment for each new cohort of patients, CFO casts the current dose level in competition with its two neighboring (left and right) dose levels based on evidence in the form of odds to determine an admissible set. An incoming cohort is then assigned to the dose level that has the largest posterior probability to achieve the highest efficacy rate among the dose levels in the admissible set. The CFO design is calibration-free in the sense that its implementation does not require prespecification of any essential design parameter except for the target DLT rate ϕ and the minimal acceptable efficacy rate ψ which are the external rather than intrinsic part of the design. When only considering the toxicity outcomes, the CFO design can also be applied to a phase I trial focusing on the MTD identification. Extensive simulation studies show that CFO delivers robust performance and the operating characteristics are satisfactory compared with existing phase I and phase I/II trial designs for both MTD and OBD identification tasks.

The rest of the paper is organized as follows. In the next section, we introduce the CFO design for both MTD and OBD identification. We then present the simulation studies to evaluate the operating characteristics of the new method and compare CFO with several phase I and phase I/II designs in *Simulation Studies* section. An application to the phase I/II trial of lenalidomide is provided in *Real Trial Application* section. The paper is concluded with some discussion.

Methodology of the CFO Design

Identification of the MTD

Suppose that a clinical trial is initiated to investigate K dose levels with the monotonically increasing DLT rates, $p_1 < \cdots < p_K$. The corresponding efficacy probabilities of the K doses are denoted by $\{q_k\}_{k=1}^K$, which do not satisfy any monotonic assumption. Let ϕ be the target DLT rate of the trial, and let d_i be the dose level at which the ith cohort of patients is treated.

After enrolling n cohorts of patients, we observe the cumulative data, $D_n = \{(x_k, y_k, m_k)\}_{k=1}^K$, where the triplet (x_k, y_k, m_k) represents the numbers of observed DLTs, efficacy outcomes and patients at dose level k, respectively. Given the nth cohort treated at dose level d_n, the DLT rates of dose levels $(d_n - 1, d_n, d_n + 1)$ are denoted as (p_L, p_C, p_R) based on their left, central, and right positions, and (x_L, x_C, x_R) and (m_L, m_C, m_R) are the corresponding number of DLTs and number of patients, respectively.

We first illustrate the CFO design for a phase I trial, which aims to determine the MTD of the drug and its dose level satisfies

$$k_{\mathrm{MTD}} = \mathrm{argmin}_{k=1,\ldots,K} |p_k - \phi|.$$

Upon observing the cumulative data D_n with the enrolled n cohorts, we need to determine the dose level for the $(n + 1)$th cohort of patients. We define the odds of $p_k > \phi$ as

$$O_k = \frac{\Pr(p_k > \phi | x_k, m_k)}{\Pr(p_k \leq \phi | x_k, m_k)}$$

for $k = L, C, R$ corresponding to left, current/central and right doses. The reciprocal $\bar{O}_k = 1/O_k$ represents the odds of $p_k \leq \phi$. Under the Bayesian paradigm, a noninformative Beta$(\phi, 1 - \phi)$ prior distribution is adopted for each DLT probability p_k.

Intuitively, the odds O_k measures the evidence in favor of $p_k > \phi$. When O_k is large, the corresponding dose level is unlikely to be selected for the $(n + 1)$th cohort due to its over-toxicity. As shown in the left panel of Figure 1, the odds of $p_C > \phi$ is so large that we know the corresponding dose level d_n is overly toxic. Similarly, the odds $\bar{O}_k$ represents the evidence in favor of $p_k \leq \phi$ and a large value of $\bar{O}_k$ indicates that the corresponding dose is overly tolerable.

The key issue for dose finding is to determine how large the value of O_k (or $\bar{O}_k$) is adequate in order to claim the dose is overly toxic (or safe), which triggers the dose movement. Without introducing any design parameter, we make the current dose level compete against its two neighboring dose levels and aggregate the comparison results to select the next dose level.

Specifically, a large value of O_C means the current dose d_n is overly toxic, while a large value of $\bar{O}_L$ indicates that dose level $d_n - 1$ is overly safe (too low). This situation is similar to a combat between two game players, while one tries to push the dose down and the other tries to push the dose up. If $O_C/\bar{O}_L$ is large, it indicates the evidence in O_C is stronger than that in $\bar{O}_L$, as the case shown in the left panel of Figure 1, so we should de-escalate the dose; otherwise as shown by the case in
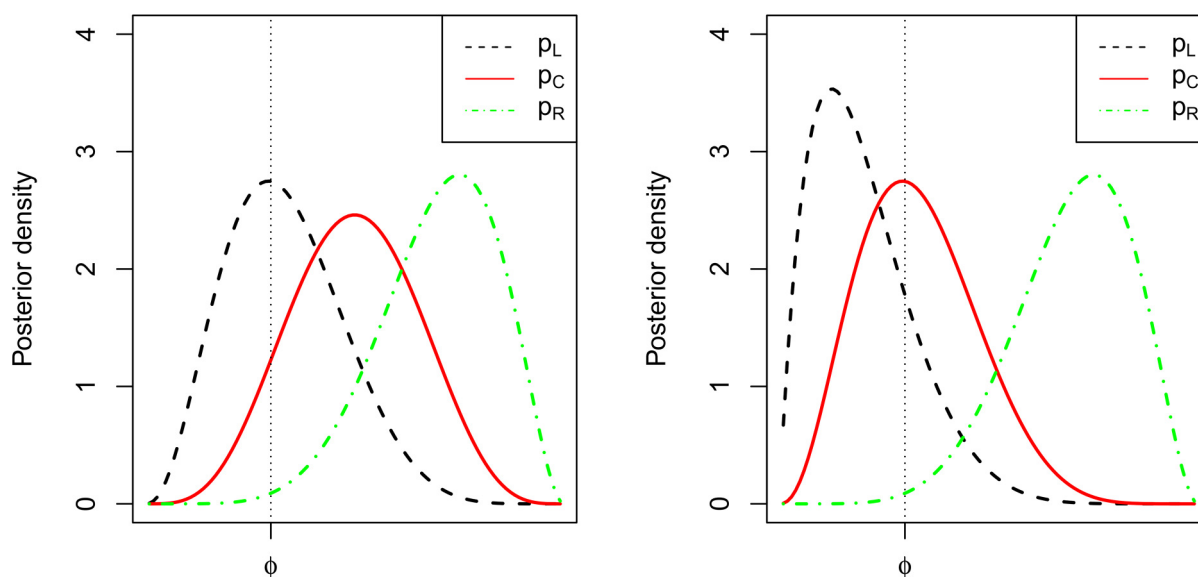

Figure 1. Illustration of the posterior distributions of the DLT probabilities for the left, current and right doses, (p_L, p_C, p_R), with the left panel corresponding to large $O_C/\bar{O}_L$ and the right panel to small $O_C/\bar{O}_L$. The dotted line indicates the target DLT rate ϕ.

the right panel of Figure 1, it suggests that the information supports dose $d_n - 1$ is overly safe, and thus de-escalation is not the appropriate move. Therefore, by comparing the ratio $O_C/\bar{O}_L$ with some threshold value γ_L, we can obtain a vote between de-escalation and staying at the current dose d_n.

In addition, when making O_C compete with $\bar{O}_L$, we further take the monotonic relationship $p_L < p_C$ into consideration. By accounting for such monotonicity, the marginal posterior density functions for p_L and p_C can be derived,

$$f_L(p_L|x_L, x_C) \propto f_\beta(p_L; a_L, b_L) \int_{p_L}^1 f_\beta(p_C; a_C, b_C) \mathrm{d}p_C$$

$$f_C(p_C|x_L, x_C) \propto f_\beta(p_C; a_C, b_C) \int_0^{p_C} f_\beta(p_L; a_L, b_L) \mathrm{d}p_L,$$

where $f_\beta(\,\cdot\,; a_k, b_k)$ is the density function of Beta(a_k, b_k), with $a_k = \phi + x_k$ and $b_k = 1 - \phi + m_k - x_k$ for $k = L, C$, i.e., the posterior distribution of p_k given the data (x_k, m_k) without incorporating the monotonic relationship. The odds O_C and $\bar{O}_L$ can be obtained via numerical integration using the Gaussian quadrature or the Monte Carlo method.

The essential step is to choose a suitable threshold γ_L in a totally data-driven manner. We denote the true values of p_L and p_C as p_{0L} and p_{0C}, respectively. Intuitively, if $p_{0C} = \phi$ and $p_{0L} < \phi$, we should avoid de-escalation, i.e., we prefer to the threshold satisfying $\gamma_L \geq O_C/\bar{O}_L$; while if $p_{0L} = \phi$ and $p_{0C} > \phi$, then de-escalation is more desirable, i.e., we prefer to $\gamma_L < O_C/\bar{O}_L$. Following this principle, we propose to minimize the probability of the incorrect vote to obtain γ_L,

$$\begin{aligned}
V_L(\gamma_L) &= \Pr\left(O_C/\bar{O}_L > \gamma_L | p_{0C} = \phi, p_{0L} < \phi\right) \\
&\quad + \Pr\left(O_C/\bar{O}_L \leq \gamma_L | p_{0L} = \phi, p_{0C} > \phi\right) \\
&= \sum_{i=0}^{m_C} \sum_{j=0}^{m_L} I(O_C/\bar{O}_L > \gamma_L) \Pr(x_C = i | p_{0C} = \phi) \Pr(x_L = j | p_{0L} < \phi) \\
&\quad + \sum_{i=0}^{m_C} \sum_{j=0}^{m_L} I(O_C/\bar{O}_L \leq \gamma_L) \Pr(x_C = i | p_{0C} > \phi) \Pr(x_L = j | p_{0L} = \phi),
\end{aligned}$$

where $I(\cdot)$ is the indicator function.

Given $p_{0C} = \phi$ and $p_{0L} = \phi$, it is obvious that

$$\Pr(x_C = i | p_{0C} = \phi) = \binom{m_C}{i} \phi^i (1 - \phi)^{m_C - i},$$

$$\Pr(x_L = j | p_{0L} = \phi) = \binom{m_L}{j} \phi^j (1 - \phi)^{m_L - j}.$$

We adopt a Uniform$(0, \phi)$ prior distribution for p_{0L} when $p_{0L} < \phi$, and a Uniform$(\phi, 2\phi)$ prior distribution for p_{0C} when $p_{0C} > \phi$. Thus, $\Pr(x_L = j | p_{0L} < \phi)$ and $\Pr(x_C = i | p_{0C} > \phi)$ can be calculated via the Gaussian quadrature,

$$\Pr(x_L = j | p_{0L} < \phi) = \int_0^\phi \frac{1}{\phi} \binom{m_L}{j} p_{0L}^j (1 - p_{0L})^{m_L - j} \mathrm{d}p_{0L},$$

$$\Pr(x_C = i | p_{0C} > \phi) = \int_\phi^{2\phi} \frac{1}{\phi} \binom{m_C}{i} p_{0C}^i (1 - p_{0C})^{m_C - i} \mathrm{d}p_{0C}.$$

With a similar discussion of $\bar{O}_C/O_R$ on the right side of the current dose, we can derive another threshold value γ_R by minimizing

$$\begin{aligned}
&V_R(\gamma_R) \\
&= \Pr\left(\bar{O}_C/O_R > \gamma_R | p_{0C} = \phi, p_{0R} > \phi\right) \\
&\quad + \Pr\left(\bar{O}_C/O_R \leq \gamma_R | p_{0R} = \phi, p_{0C} < \phi\right) \\
&= \sum_{i=0}^{m_C} \sum_{j=0}^{m_R} I(\bar{O}_C/O_R > \gamma_R) \Pr(x_C = i | p_{0C} = \phi) \Pr(x_R = j | p_{0R} > \phi) \\
&\quad + \sum_{i=0}^{m_C} \sum_{j=0}^{m_R} I(\bar{O}_C/O_L \leq \gamma_R) \Pr(x_C = i | p_{0C} < \phi) \Pr(x_R = j | p_{0R} = \phi),
\end{aligned}$$

and attain the vote of staying at the same dose or escalation. The two votes are then aggregated together to determine the dose level for the $(n + 1)$th cohort, based on the decision rule summarized in Table 1.

Given the target ϕ, the two thresholds γ_L and γ_R are functions of (m_L, m_C) and (m_C, m_R) respectively. For ease of implementation, we can calculate the values of (γ_L, γ_R) beforehand as shown in Figure 2, where the values of (γ_L, γ_R) vary from 1 to 30 under $\phi = 0.3$. In general, the value of γ_R is larger than that of γ_L. The value of γ_L typically falls in the range between 0 and 1 which tends to be smaller for $m_C < m_L$, while γ_R mainly falls between 0 and 2.2 and its value tends to be larger for $m_C > m_R$.

Under the dose movement rule in Table 1, the CFO design for MTD identification is described as follows.

(i) Start the trial by treating the first cohort of patients at the lowest dose or a prespecified initial dose.
(ii) After enrolling n cohorts, compute the ratios of odds between the central dose versus the left and the central dose versus the right, $(O_C/\bar{O}_L, \bar{O}_C/O_R)$.
(iii) Select the dose for the next cohort following the rules in Table 1.
(iv) Repeat steps (ii) and (iii) until the maximal sample size is reached or the early stopping criteria are met.

Table 1 includes the case with $O_C/\bar{O}_L > \gamma_L$ and $\bar{O}_C/O_R > \gamma_R$, i.e., the information from two odds ratios is contradictory with each other; the former suggests dose de-escalation while the latter suggests dose escalation. Although such case may happen theoretically, it is rarely encountered in practice. In our simulation studies with random scenarios, there is no occurrence of such event over more than 10000 repetitions.

Identification of the OBD

As an essential part of the outcomes collected in phase I/II trials, efficacy data need to be incorporated in dose finding under the CFO design. Upon the arrival of the $(n + 1)$th cohort of patients, CFO adopts two steps to determine the dose level for the new cohort. An admissible set $\mathcal{A}_n$ is first determined via the dose escalation rules for the MTD in Table 1:

- If the decision is to de-escalate the dose, then $\mathcal{A}_n = \{1, \ldots, d_n - 1\}$;
- If the decision is to stay at the current dose, then $\mathcal{A}_n = \{1, \ldots, d_n\}$.
- If the decision is to escalate the dose, then $\mathcal{A}_n = \{1, \ldots, d_n + 1\}$.

The admissible set is constructed using toxicity data alone and no dose skipping is allowed during dose escalation, while dose skipping is permitted for dose de-escalation due to jointly modelling both toxicity and efficacy data.

Given the current data D_n, we select from the admissible set $\mathcal{A}_n$ the next dose level d_{n+1} which has the maximal posterior probability to yield the highest efficacy,

$$d_{n+1} = \operatorname{argmax}_{k \in \mathcal{A}_n} \Pr\left(q_k = \max_{j \in \mathcal{A}_n} \{q_j\} | D_n\right). \tag{1}$$

We adopt Jeffreys' prior Beta(0.5, 0.5) distribution for each q_k, so that the observed data dominate the posterior estimation. We use the Monte Carlo method to calculate $\Pr(q_k = \max_{j \in \mathcal{A}_n} \{q_j\} | D_n)$ for $k \in \mathcal{A}_n$. Specifically, we first generate 10000 random samples $\{(\tilde{q}_k^{(i)})_{k \in \mathcal{A}_n}\}_{i=1}^{10000}$ from the distribution of $(q_k | D_n)_{k \in \mathcal{A}_n}$, and then calculate the empirical probability of $\tilde{q}_k^{(i)}$ being the largest among $(\tilde{q}_j^{(i)})_{j \in \mathcal{A}_n}$.

Following the above dose movement decisions when accounting for both toxicity and efficacy, the proposed phase I/II dose-finding procedure for the OBD proceeds as follows.

(i) Start the trial by treating the first cohort of patients at the lowest dose or a prespecified initial dose.
(ii) After enrolling n cohorts, determine the admissible set $\mathcal{A}_n$ via the dose escalation rule for the MTD.

Table 1. Dose escalation and de-escalation rules of the CFO design in searching for the MTD.

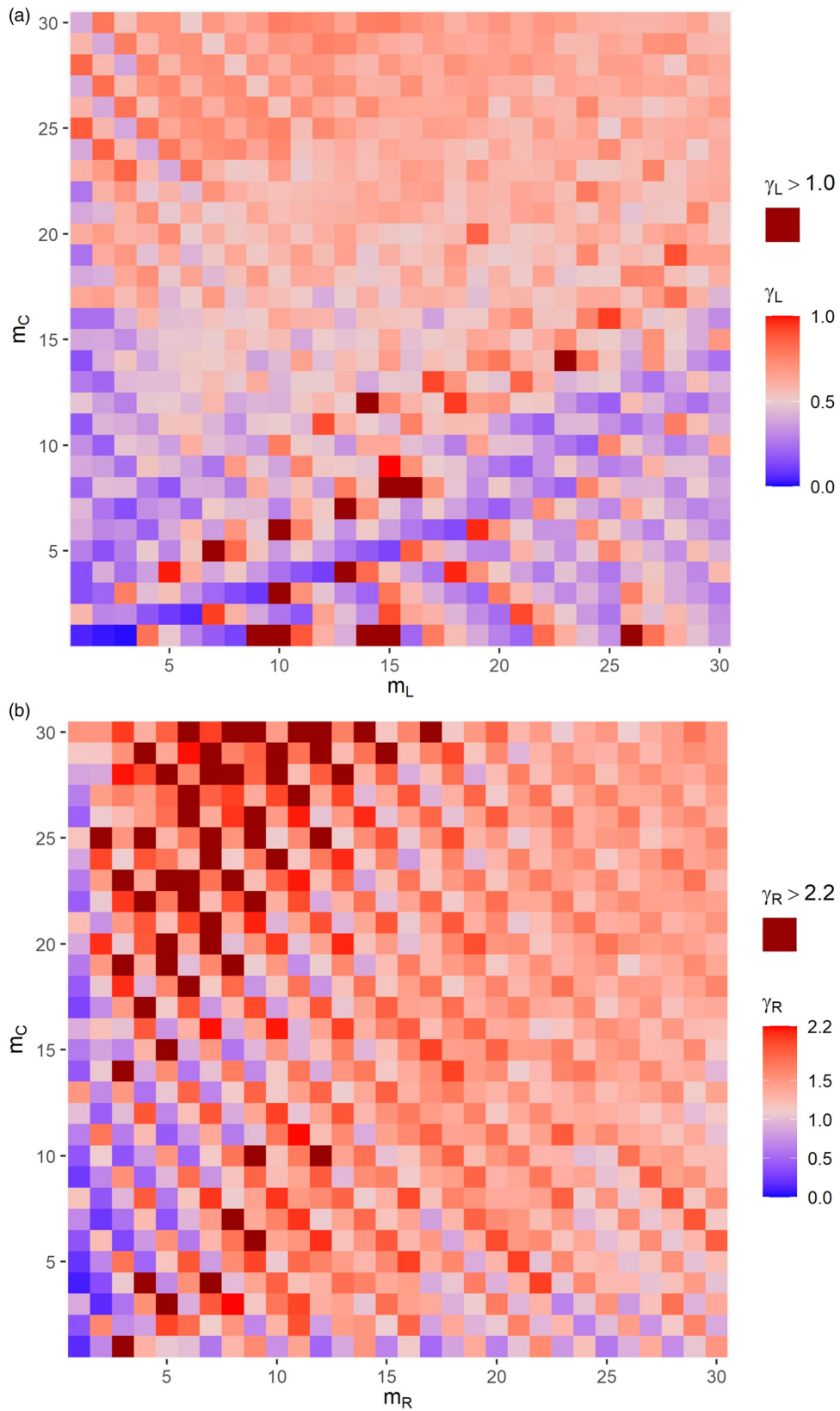|	p_C against p_L	$O_C/\bar{O}_L > \gamma_L$	
p_C against p_R		Yes (De-escalation)	No (Stay)
$\bar{O}_C/O_R > \gamma_R$	Yes (Escalation)	Stay	Escalation
	No (Stay)	De-escalation	Stay

Figure 2. The threshold values of (γ_L, γ_R) when the numbers of patients treated at the left, current, and right doses (m_L, m_C, m_R) vary from 1 to 30 given the target DLT rate $\phi = 0.3$.

(iii) The dose level for the next cohort is determined by (1).

(iv) Repeat steps (ii) and (iii) until the maximal sample size is reached or the early stopping criteria are met.

At the beginning of the trial, there is no information for the neighboring dose levels, while the CFO design can still work normally because we assign non-informative priors to the DLT and efficacy rates of each dose. An example in Appendix C.3 demonstrates how CFO works at the beginning of a trial.

Early Stopping and Final Selection

During the implementation of the CFO design, it is preferable to impose some early stopping criteria to ensure the safety and benefit for the patients. For toxicity monitoring, we eliminate the dose level when there is strong evidence to corroborate its over-toxicity. In particular, we eliminate dose level k and all the dose levels above from the trial if $\Pr(p_k > \phi | x_k, m_k \geq 3) > 0.95$. If the posterior probability of the lowest dose level satisfies $\Pr(p_1 > \phi | x_1, m_1 \geq 3) > 0.95$, then we terminate the entire trial for safety.

For the phase I/II trial design, we further consider the efficacy data to terminate the trial early if none of the admissible dose levels shows adequate efficacious effect. Given the lowest acceptable efficacy rate ψ, the trial would be terminated early for futility if $\Pr(q_k < \psi | y_k, m_k \geq 3) > 0.9$ for all the admissible dose levels.

In our simulation studies and real data application, the two cutoff values for toxicity and efficacy early stopping are set as 0.95 and 0.9 respectively, which yield satisfactory performances. Nevertheless, the cutoff values can be adopted to meet practical needs in real trials. For selecting a suitable cutoff value for toxicity, we can randomly generate a large number of over-toxic scenarios without an MTD as well as typical scenarios with an MTD using the scheme in Section B of Appendix. The CFO design is then applied to these scenarios to choose a cutoff value that strikes a balance for the non-selection rates between both types of scenarios. A similar strategy can be applied to selecting the cutoff value for futility stopping.

After the trial is completed, to guarantee the monotonically increasing trend of the dose–toxicity curve, an isotonic regression[24] is performed on the observed DLT rates to obtain the final estimates $\{\hat{p}_k\}_{k=1}^K$ through the pool-adjacent-violators algorithm. In a phase I trial searching for the MTD, the MTD level k_{MTD} is selected as

$$k_{\mathrm{MTD}} = \mathrm{argmin}_{k=1,\ldots,K} |\hat{p}_k - \phi|.$$

In a phase I/II trial searching for the OBD, the OBD level k_{OBD} is determined as

$$k_{\mathrm{OBD}} = \mathrm{argmax}_{k \leq k_{\mathrm{MTD}}} \Pr\left(q_k = \max_{j \leq k_{\mathrm{MTD}}} \{q_j\} | D\right).$$

where D is the observed data throughout the trial.

Simulation Studies

Toxicity Evaluation Under Random/Fixed Scenarios

As determination of the MTD is an essential part of the CFO design, we first conduct extensive simulation studies in the context of identifying the MTD. We compare CFO with BOIN[25] and CRM[7]. The target DLT rate is $\phi = 0.33$ and there are five dose levels under investigation with the maximum sample size of 30 and a cohort size of 3. For the BOIN method, we adopt the default parameters suggested in the original paper. Following Lin and Yin (2017, 2018),[8,26] the CRM takes the power model formulation, $p_k = a_k^{\exp(\alpha)}$, where the skeleton a_k is chosen by the model calibration method of Lee and Cheung (2009)[27] with a halfwidth of the indifference interval of 0.05 and the initial guess of MTD at dose level $\lceil K/2 \rceil$. The detailed settings of the compared methods are given in Appendix A.1 and we also discuss selection of the half-width of the indifference interval for the CRM in Appendix C.2. To avoid cherry-picking cases, we randomly generate dose–toxicity scenarios following Paoletti et al. (2004).[28] The detailed scheme on generating the phase I scenarios is presented in Appendix B.1. The average probability difference around the target is controlled at 0.05, 0.07, 0.1 and 0.15 respectively, and under each configuration, we replicate 5000 simulations.

Six performance statistics are used to assess the operating characteristics of the three designs. The two main measurements, reflecting the accuracy and efficiency of a design, are the percentage of MTD selection and the percentage of patients treated at the MTD, for which the larger the better. The remaining four measurements quantify the safety aspects of a trial, which include the percentage of trials of selecting overdoses as the MTD, the percentage of patients allocated to overdoses, the risk of high toxicity (defined as the percentage of trials leading to the DLT rates greater than ϕ), and

the percentage of patients experiencing DLT. A design with smaller values of these four safety statistics is considered more ethical and desirable.

The results on the MTD identification are shown in Figure 3. When the average probability difference around the target increases, all the three methods lead to better performances in terms of the six measurements because the MTD is more easily distinguishable from its neighboring doses. In terms of the two main measurements on accuracy and efficiency, the CRM design performs the best, while the CFO method ranks the second. The gap diminishes when the average probability difference around the target increases. When the average probability difference is 0.15, the CFO design yield the highest percentage of the MTD allocation. Regarding the four safety measurements, the CFO design yields the best performance and CRM appears to be the most aggressive, as it yields significantly higher percentages in the four safety metrics.

To better evaluate the characteristics of CFO, BOIN and CRM, we further investigate the operating characteristics of the three designs under six fixed representative dose–toxicity scenarios. The metrics of evaluation are the percentage of MTD selection, the number of patients allocated to each dose level and the percentage of patients experiencing DLT. For consistent comparisons, we adopt the same settings as the random scenarios. We also include the non-parametric optimal design as the benchmark,[29,30] for which the non-selection rule is incorporated for a fair comparison. For each scenario, we replicate 5000 simulations and summarize the results in Table 2. Overall, the two algorithm-based methods, CFO and BOIN, yield more robust performances across the six scenarios. In particular, CFO performs slightly better than BOIN in terms of both the MTD selection and patient allocation in the first five scenarios. The model-based CRM appears to be sensitive to the parametric modeling structure, i.e., the matching between the model skeleton and the truth. For example, in scenario 3 where the truth is close to the CRM model skeleton, the CRM performs better than the other two methods with an increment of around 3% in the MTD selection percentage. However, in scenario 4 where the model skeleton seriously deviates from the truth, the performance of the CRM deteriorates dramatically and there is a gap of around 10% in the percentage of MTD selection between CRM and the other two methods. In addition, the CRM design tends to select an over-toxic dose as the MTD, which is consistent with our observation in the random scenario setting. In the over-toxic scenario (scenario 6), the BOIN design has the best performance.

We also investigate the influential factors which affect the result of the dose-finding trial in terms of the percentage of MTD selection via the analysis of variance (ANOVA) method used by Cangul et al. (2009)[31] in Appendix C.1. The results also indicate that the CFO design strikes a good balance between efficiency and safety in our settings.

Toxicity and Efficacy Evaluation Under Random/Fixed Scenarios

We further compare the CFO design for identification of the OBD with the WT design,[15] STEIN[18] and model adaptation (MADA) design[17] in phase I/II clinical trials. We consider $K = 5$ dose levels with the maximal sample size of 60 and a cohort size of 3. The target DLT rate is $\phi = 0.3$, while the minimal acceptable efficacy rate is set as $\psi = 0.3$. The detailed settings of the MADA, STEIN and WT designs are given in Appendix A.2. Among the three competitors, the WT design is a model-based method, and the STEIN design is a model-free method, while the MADA design is an adaptive method which can switch between beta–binomial and regression models.

To assess the four designs comprehensively, we evaluate them under the randomly generated phase I/II scenarios. We first consider the umbrella-shape and plateau-shape dose–efficacy curves separately, and then we mix the two types of curves together to show the overall performance of the four designs. For the dose–toxicity curve, we still follow the generation method of Paoletti et al. (2004)[28] and control the average probability difference around ϕ at 0.05, 0.07, 0.1 and 0.15 respectively. Under each configuration, we replicate 5000 simulated trials. The detailed scheme on generating the phase I/II scenarios is given in Appendix B.2.

The comparison mainly focuses on two important metrics: the percentages of OBD selection and OBD allocation. The results under the random scenarios are presented in Figure 4. The top row of Figure 4 shows the percentages of the OBD selection and allocation for the umbrella-shape scenarios. Among the four methods, MADA has the overall best performance in the OBD selection percentage, while CFO also shows satisfactory results. The WT design performs the best when the probability difference is 0.15, while its performance deteriorates when the probability difference shrinks. In terms of the OBD allocation, the performance of MADA is much worse than its counterparts, because MADA has two stages and in stage one it only considers toxicity. The STEIN design has the highest OBD allocation percentage among the four methods. The results of the plateau-shape curve are presented in the middle row of Figure 4. The CFO design has the best performance in terms of the OBD selection in general, while MADA is clearly worse than other methods. The WT design shows a similar trend to that under the umbrella-shape curve, i.e., the relative performance deteriorates when the probability difference is diminished. With regard to the OBD allocation, the results are similar to those under the umbrella-shape curves. We then combine results for both types of curves at the bottom of Figure 4. Overall, the CFO design has the highest OBD
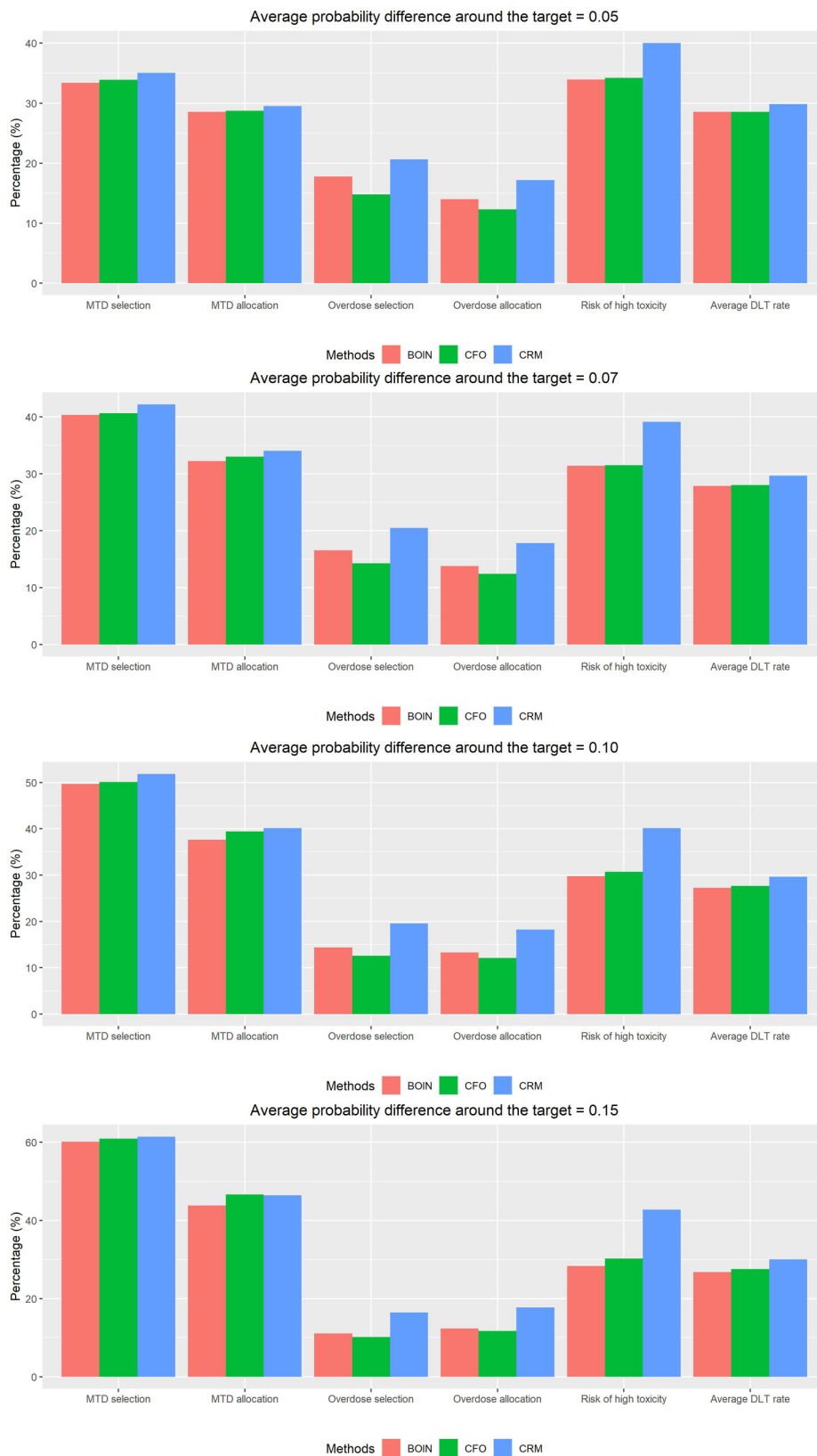
Figure 3. Simulation results for the MTD identification based on 5000 randomly generated dose–toxicity scenarios with the average probability difference of 0.05, 0.07, 0.10 and 0.15 (from top to bottom panels) around the target toxicity probability $\phi = 0.33$.

Table 2. The percentage of MTD selection (the number of patients treated at each dose) under the CFO design in comparison with the BOIN and CRM under six fixed scenarios with the target toxicity probability 0.33 in boldface. None represents the percentage of trials of non-selection. Benchmark indicates the results under the non-parametric optimal design with complete information.

Design	Dose Level					DLT (%)	None (%)
	1	2	3	4	5		
			Scenario 1				
p_k	**0.33**	0.45	0.58	0.70	0.80		
CFO	63.8 (19.6)	20.8 (6.9)	1.4 (1.0)	0.1 (0.1)	0 (0)	37.0	13.9
BOIN	58.7 (18.4)	20.6 (6.5)	1.7 (1.2)	0.1 (0.1)	0 (0)	37.2	18.9
CRM	62.0 (19.3)	21.2 (6.2)	2.2 (1.7)	0 (0.2)	0 (0)	37.7	14.6
Benchmark	74.3 (30)	20.6 (30)	1.2 (30)	0 (30)	0 (30)	57.2	3.9
			Scenario 2				
p_k	0.18	**0.33**	0.52	0.60	0.70		
CFO	25.2 (10.9)	61.2 (14.4)	11.7 (4.1)	1.1 (0.5)	0.1 (0)	30.6	0.7
BOIN	24.5 (11.5)	60.1 (13.2)	12.7 (4.3)	1.0 (0.5)	0 (0)	30.4	1.6
CRM	18.9 (10.6)	60.5 (12.3)	18.5 (5.9)	1.1 (0.9)	0 (0.1)	32.5	0.9
Benchmark	16.9 (30)	72.2 (30)	10.6 (30)	0.3 (30)	0 (30)	46.6	0.0
			Scenario 3				
p_k	0.12	0.20	**0.33**	0.40	0.50		
CFO	3.4 (5.9)	29.7 (9.9)	43.1 (9.5)	18.7 (3.7)	5.1 (1.0)	25.9	0.1
BOIN	3.1 (6.1)	29.1 (10.1)	41.1 (8.7)	20.7 (3.9)	5.7 (1.1)	25.8	0.4
CRM	1.3 (5.6)	18.7 (7.1)	46.0 (9.7)	26.8 (5.4)	6.9(2.1)	28.5	0.3
Benchmark	1.0 (30)	20.1 (30)	48.0 (30)	23.5 (30)	7.4 (30)	31.0	0
			Scenario 4				
p_k	0.01	0.02	0.03	**0.33**	0.50		
CFO	0 (3.1)	0 (3.2)	11.2 (5.1)	70.4 (13.8)	18.5 (4.8)	24.1	0
BOIN	0 (3.1)	0 (3.2)	14.3 (7.3)	67.5 (11.7)	18.2 (4.7)	21.6	0
CRM	0 (3.1)	0 (3.0)	6.2 (4.0)	58.7 (9.6)	35.1 (10.3)	28.5	0
Benchmark	0 (30)	0.0 (30)	0.1 (30)	86.3 (30)	13.5 (30)	17.8	0
			Scenario 5				
p_k	0.00	0.00	0.05	0.10	**0.33**		
CFO	0 (3.0)	0 (3.0)	0.2 (3.7)	17.4 (6.1)	82.4 (14.2)	18.3	0
BOIN	0 (3.0)	0 (3.0)	0.3 (3.7)	17.3 (7.4)	82.4 (12.8)	17.1	0
CRM	0 (3.0)	0 (3.0)	0 (3.0)	6.7 (4.0)	93.3 (16.9)	20.5	0
Benchmark	0 (30)	0 (30)	0.1 (30)	4.2 (30)	95.8 (30)	9.6	0
			Scenario 6				
p_k	0.45	0.55	0.65	0.75	0.85		
CFO	46.5 (19.2)	3.3 (2.5)	0.1 (0.2)	0 (0)	0 (0)	46.2	50.1
BOIN	40.9 (17.0)	3.1 (2.5)	0.1 (0.2)	0 (0)	0 (0)	46.3	55.9
CRM	45.5 (18.8)	2.8 (2.3)	0.1 (0.5)	0 (0)	0 (0)	46.7	51.6
Benchmark	61.8 (30)	1.9 (30)	0.1 (30)	0 (30)	0 (30)	65.0	36.2

selection percentage when the probability difference is not very large. The performance of the WT design varies dramatically, as it performs the best when the probability difference is 0.15, but almost worst when the probability difference is 0.05. The results under the random scenarios demonstrate the robustness of the CFO design. It is model-free and calibration-free, and thus it yields satisfactory performance under different settings.

We further assess the four designs under six fixed scenarios as shown in Figure 5, which include the plateau-shape (scenarios 1 and 2), umbrella-shape (scenarios 3 and 4) and monotone increasing (scenario 5) dose–efficacy relationships as well as the over-toxic (scenario 6) case. We adopt the same settings as the random scenarios and report the percentage of OBD selection and the number of patients allocated to each dose level as well as the percentage of patients experiencing DLT, the percentage of patients showing efficacy outcomes and the non-selection rate (i.e., the percentage of trials that do not select any dose as the OBD). To facilitate the comparison, we also add the non-parametric optimal design[29,32,33] as the benchmark. Under each scenario, we carry out 5000 repetitions and Table 3 summarizes the simulation results.

In scenarios 1 and 2 where the dose–efficacy curves are plateau-shape, the WT design yields the highest percentage of OBD selection while CFO ranks the second. The CFO design has a relatively small percentage of DLT in scenario 1 and the
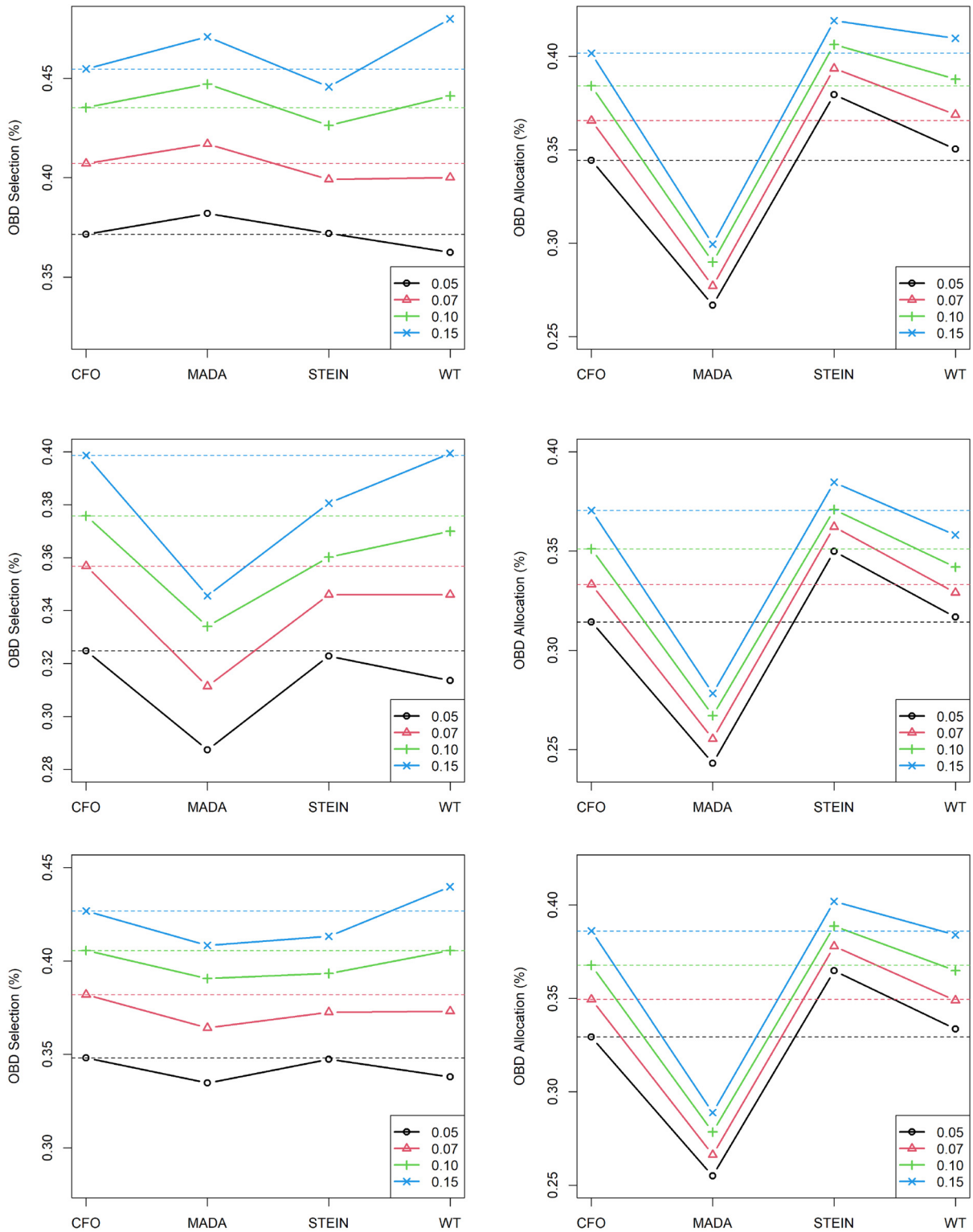
Figure 4. Simulation results for the OBD identification based on 5000 randomly generated phase I/II scenarios with the average probability difference of 0.05, 0.07, 0.10 and 0.15 around the target toxicity probability $\phi = 0.30$ under the umbrella-shape (top), plateau-shape (middle) and mixed (bottom) dose–efficacy curves. The minimal acceptable efficacy rate is $\psi = 0.3$ and the maximal sample size is 60 with a cohort size of 3. The dashed lines indicate the results for the CFO design.

WT design appears to be the safest in scenario 2. The MADA design also leads to satisfactory results for the two plateau-shape scenarios. The STEIN design performs well in scenario 2 but poorly in scenario 1.

Under the umbrella-shape dose–efficacy curves corresponding to scenarios 3 and 4, similarly, the WT performs the best in terms of the OBD selection while CFO yields the second highest percentage of OBD selection. With regard to the safety, the WT design yields the best result in scenario 3 and CFO has the smallest percentage of DLT in scenario 4. The MADA and STEIN designs also demonstrate satisfactory results, but they are consistently worse than the CFO and WT designs. In scenario 5 where the MTD and OBD are identical, MADA has a significantly higher percentage of OBD selection than the other three designs, while CFO still delivers a decent performance in comparison with the WT and STEIN methods. The WT design performs rather poorly under this scenario, which may be due to the model misspecification because it is a model-based method. When all the dose levels are overly toxic as in scenario 6, CFO leads to the highest non-selection rate, while the performances of STEIN and WT are comparable. The MADA design has an extremely low non-selection rate and it selects the first dose level for most of the times, which is due to the fact the MADA design has no early stopping rule for futility. In the first five scenarios, there are large gaps between the four designs and the non-parametric optimal benchmark. Under the over-toxic scenario, the CFO, STEIN and WT designs have comparable results with the benchmark.

Aggregating results under both the random and fixed scenarios, it can be concluded that overall the WT and CFO designs perform the best in phase I/II trials. However, the performance of the WT design depends on the scenarios which may yield rather poor performance under some specific cases due to the potential risk of assuming a model-based structure. Because of its model-free and calibration-free nature, the CFO design leads to a more robust performance in the OBD-identification task in contrast to the other three methods. Although the STEIN design is also a model-free approach, it still requires to specify some design parameters, and thus it is still sensitive to certain dose–response scenarios. The performance of the MADA design varies dramatically as the scenarios change and it yields fairly low percentages of the OBD allocation because it is a two-stage design.

Real Trial Application

As an illustration, we apply the proposed CFO to redesign the aforementioned phase I/II trial of lenalidomide in combination with the high-dose melphalan. The trial enrolled a total of 57 patients with relapsed or progressive multiple myeloma.[23] Patients were sequentially assigned to one of the four prespecified doses of lenalidomide $\{25, 50, 75, 100\}$ mg, while the dose of melphalan was fixed. Based on the observed data in the trial, the estimated DLT and efficacy rates were $\{(p_1, q_1), \ldots, (p_4, q_4)\} = \{(0.02, 0.03), (0.02, 0.02), (0.04, 0.17), (0.04, 0.16)\}$.

We rerun this trial on the basis of the estimated DLT and efficacy rates using the CFO design, for which we set the target DLT rate as $\phi = 0.2$ and the minimal acceptable efficacy rate as $\psi = 0.15$. Patients were treated with a cohort size of 3. As illustrated by the trial conduct in Figure 6, the first cohort was treated at dose level 1, and there was no DLT or efficacy outcome observed. It yielded $\bar{O}_C/O_R = 4.44 > \gamma_R = 0.02$, $\mathcal{A}_1 = \{1, 2\}$ and $\{\Pr(q_k = \max_{j=1,2}\{q_j\})\}_{k=1}^2 = (0.19, 0.81)$. Thus, the next cohort was treated at dose level 2, and again there was no DLT or efficacy outcome. Consequently, the trial escalated to dose level 3, where we observed one efficacy response but no DLT. We obtained $(O_C/\bar{O}_L, \bar{O}_C/O_R) = (0.00, 4.44)$ and $(\gamma_L, \gamma_R) = (0.14, 0.02)$, which led to $\mathcal{A}_3 = \{1, 2, 3, 4\}$ and $\{\Pr(q_k = \max_{j=1,2,3,4}\{q_j\})\}_{k=1}^4 = (0.05, 0.05, 0.34, 0.56)$. As a result, dose level 4 was selected for the next cohort. The two subsequent cohorts were both treated at dose level 4 and there was no DLT outcome while two efficacy responses were observed. We obtained $O_C/\bar{O}_L = 0.00 \leq \gamma_L = 0.196$, $\mathcal{A}_5 = \{1, 2, 3, 4\}$ and $\{\Pr(q_k = \max_{j=1,2,3,4}\{q_j\})\}_{k=1}^4 = (0.07, 0.06, 0.45, 0.42)$. Therefore, the trial de-escalated to dose level 3, where the next four cohorts were all treated. Among those four cohorts, no DLT outcome was observed and five efficacy responses occurred, which yielded a small left-side odds ratio $O_C/\bar{O}_L = 0.00$ and a large right-side odds ratio $\bar{O}_C/O_R = 3.43 \times 10^5$. As a result, the admissible set was $\mathcal{A}_{10} = \{1, 2, 3, 4\}$ with $\{\Pr(q_k = \max_{j=1,2,3,4}\{q_j\})\}_{k=1}^4 = (0.07, 0.07, 0.41, 0.46)$. Again, the next four cohorts were all assigned to dose level 4, where two DLTs and three efficacy outcomes occurred. After 14 cohorts were treated, we had $O_C/\bar{O}_L = 0.00 \leq \gamma_L = 0.26$, $\mathcal{A}_{14} = \{1, 2, 3, 4\}$ and $\{\Pr(q_k = \max_{j=1,2,3,4}\{q_j\})\}_{k=1}^4 = (0.09, 0.09, 0.53, 0.30)$, and the trial de-escalated to dose level 3. Following the same procedure, the remaining five cohorts were treated back and forth either at dose level 3 or 4. Finally, upon the completion of the trial, the observed data were

$$\text{Patient} : \{m_1, m_2, m_3, m_4\} = \{3, 3, 27, 24\},$$

$$\text{DLT} : \{x_1, x_2, x_3, x_4\} = \{0, 0, 1, 2\},$$

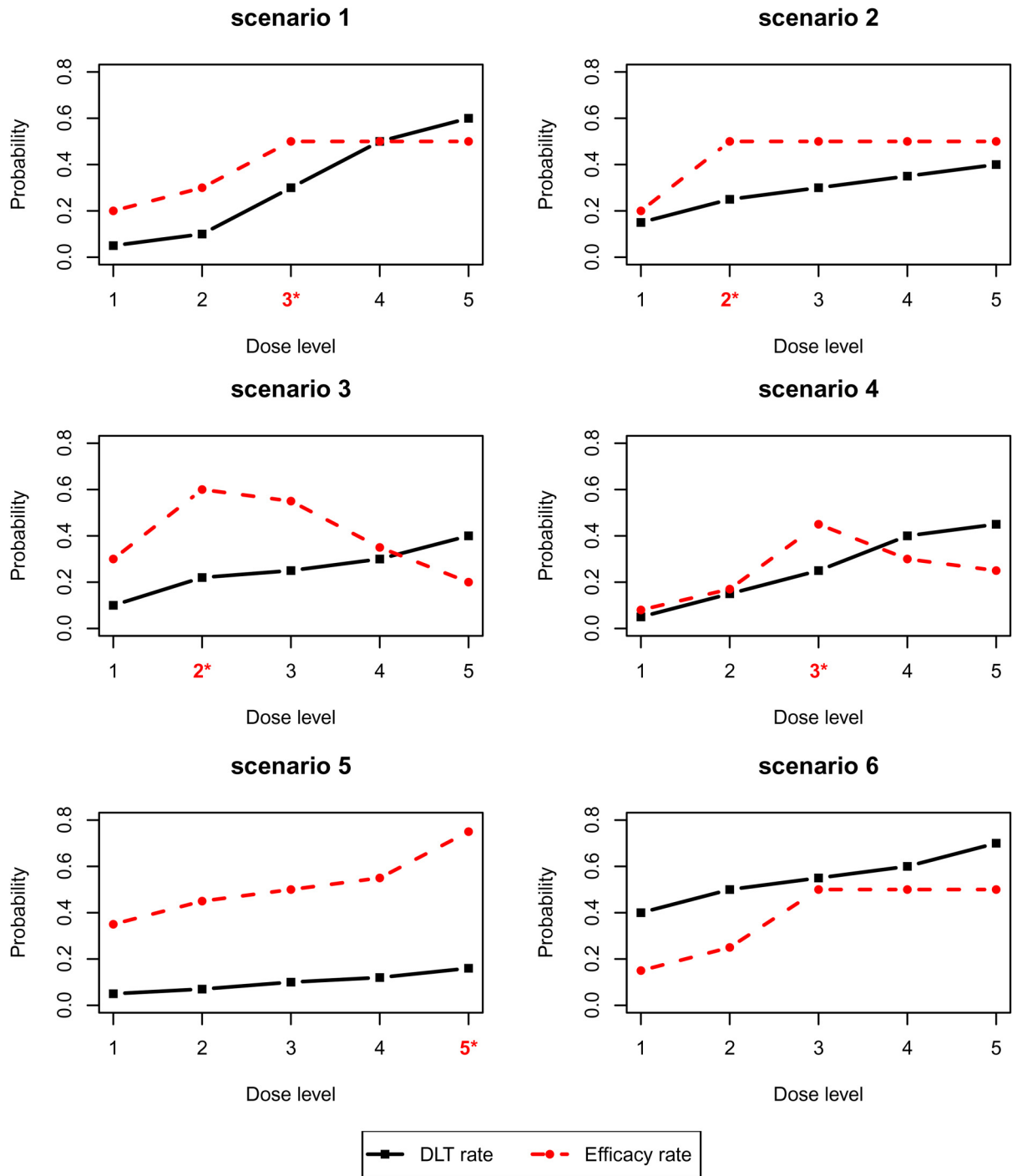$$\text{Efficacy} : \{y_1, y_2, y_3, y_4\} = \{0, 0, 6, 5\},$$

Figure 5. Six simulation scenarios for assessing the CFO design in identification of the optimal biological dose (OBD). The dashed line is the dose–efficacy curve while the solid line is the dose–toxicity curve. The OBD is highlighted by asterisk in the x-axis.

which led to $\{\Pr(q_k = \max_{j=1,2,3,4}\{q_j\})\}_{k=1}^4 = (0.14, 0.15, 0.38, 0.32)$. Thus, we selected dose level 3 (i.e., the dose of 75 mg) as the OBD for this trial, because it yielded the highest efficacy with tolerable toxicity among the four doses.

Discussion

We have proposed a new calibration-free odds design for phase I/II clinical trials to find the OBD for the targeted therapy and immunotherapy treatments. Identification of the MTD is a by-product of the CFO design, if we monitor the toxicity

Table 3. The percentage of OBD selection (the number of patients treated at each dose) under the CFO design in comparison with existing phase I/II dose-finding methods under six fixed scenarios in Figure 5. None represents the percentage of trials of non-selection. Benchmark indicates the results under the non-parametric optimal design with complete information.

Design	Dose Level					DLT/Efficacy (%)	None (%)
	1	2	3	4	5		
			Scenario 1				
(p_k, q_k)	(0.05, 0.20)	(0.10, 0.30)	**(0.30, 0.50)**	(0.50, 0.50)	(0.60, 0.50)		
CFO	13.6 (13.9)	23.0 (16.1)	58.4 (26.6)	3.1 (3.0)	0.1 (0.2)	19.8/37.6	1.8
MADA	2.8 (9.9)	41.7 (21.0)	54.4 (23.5)	1.0 (4.7)	0 (0.9)	20.9/38.1	0.0
STEIN	1.2 (8.4)	53.8 (18.0)	43.7 (29.5)	0.6 (3.8)	0 (0.2)	21.8/39.8	0.8
WT	6.7 (10.9)	24.2 (17.8)	66.1 (27.6)	1.7 (3.4)	0 (0.1)	20.8/38.6	1.4
Benchmark	0 (60)	0.9 (60)	98.4 (60)	0 (60)	0 (60)	31.0/40.0	0.8
			Scenario 2				
(p_k, q_k)	(0.15, 0.20)	**(0.25, 0.50)**	(0.30, 0.50)	(0.35, 0.50)	(0.40, 0.50)		
CFO	9.9 (15.4)	59.4 (31.3)	17.8 (9.1)	3.8 (2.1)	0.4 (0.3)	23.6/42.0	8.6
MADA	24.2 (18.8)	51.6 (23.9)	20.0 (11.6)	3.1 (3.9)	0.5 (1.4)	23.9/40.5	0.6
STEIN	18.8 (12.5)	52.3 (31.0)	17.7 (11.0)	3.6 (2.7)	0.3 (0.6)	24.4/43.5	7.3
WT	11.3 (19.0)	62.9 (28.7)	13.9 (8.3)	1.0 (1.1)	0 (0.1)	22.7/40.0	10.8
Benchmark	0.1 (60)	97.6 (60)	0 (60)	0 (60)	0 (60)	29.0/44.0	2.3
			Scenario 3				
(p_k, q_k)	(0.10, 0.30)	**(0.22, 0.60)**	(0.25, 0.55)	(0.30, 0.35)	(0.40, 0.20)		
CFO	8.6 (11.9)	68.8 (36.4)	20.0 (9.6)	0.8 (1.2)	0.1 (0.3)	20.4/52.4	1.7
MADA	12.4 (15.6)	64.8 (24.1)	21.9 (13.3)	0.7 (4.8)	0.1 (2.2)	21.0/47.6	0.1
STEIN	13.5 (11.0)	67.6 (37.7)	17.0 (9.2)	1.1 (1.4)	0 (0.3)	20.6/52.9	0.8
WT	5.8 (14.1)	74.6 (34.6)	18.9 (10.4)	0.2 (0.8)	0 (0)	19.9/51.9	0.4
Benchmark	0.1 (60)	99.8 (60)	0 (60)	0 (60)	0 (60)	25.4/40.0	0.1
			Scenario 4				
(p_k, q_k)	(0.05, 0.08)	(0.15, 0.17)	**(0.25, 0.45)**	(0.40, 0.30)	(0.45, 0.25)		
CFO	5.2 (10.9)	13.8 (14.3)	66.5 (29.3)	4.6 (3.8)	0.6 (0.6)	20.1/30.2	9.3
MADA	2.6 (9.0)	29.0 (18.2)	64.8 (24.2)	3.5 (6.7)	0.1 (2.0)	21.4/28.7	0.0
STEIN	0.6 (6.8)	22.2 (12.4)	65.7 (32.0)	3.2 (6.0)	0.2 (1.2)	22.5/32.8	8.1
WT	0.7 (8.3)	10.7 (14.4)	72.0 (30.0)	2.6 (4.2)	0 (0.3)	20.8/31.3	14.0
Benchmark	0 (60)	0 (60)	97.1 (60)	0 (60)	0 (60)	26.0/25.0	2.9
			Scenario 5				
(p_k, q_k)	(0.05, 0.35)	(0.07, 0.45)	(0.10, 0.50)	(0.12, 0.55)	**(0.16, 0.75)**		
CFO	7.9 (10.8)	14.6 (13.0)	14.5 (11.7)	16.6 (9.9)	46.3 (14.6)	10.3/53.1	0.1
MADA	1.2 (4.8)	3.2 (6.9)	6.9 (9.5)	12.5 (12.7)	76.3 (26.1)	12.4/60.1	0.0
STEIN	2.5 (8.0)	14.8 (14.4)	25.7 (15.7)	28.6 (12.4)	28.4 (9.5)	10.0/51.8	0.0
WT	16.4 (14.8)	31.3 (19.6)	22.6 (13.7)	13.1 (7.1)	16.7 (4.8)	8.5/47.4	0.0
Benchmark	0.1 (60)	0.4 (60)	2.2 (60)	6.3 (60)	91.0 (60)	10.0/52.0	0
			Scenario 6				
(p_k, q_k)	(0.40, 0.15)	(0.50, 0.25)	(0.55, 0.50)	(0.60, 0.50)	(0.70, 0.50)		
CFO	3.2 (29.7)	0.8 (3.0)	0.2 (0.4)	0 (0)	0 (0)	41.2/16.3	95.9
MADA	53.6 (33.4)	1.0 (6.7)	1.2 (1.5)	3.3 (1.0)	7.4 (0.9)	43.1/19.4	33.5
STEIN	7.0 (21.3)	2.0 (3.8)	0.4 (0.6)	0 (0)	0 (0)	41.9/17.4	90.6
WT	10.1 (29.8)	0.1 (1.7)	0 (0.1)	0 (0)	0 (0)	40.7/15.6	89.8
Benchmark	0.1 (60)	0 (60)	0 (60)	0 (60)	0 (60)	55.0/38.0	99.9

alone. Unlike other methods which monitor the toxicity data by considering either the current dose level only (e.g., the 3 + 3 and BOIN designs) or all dose levels (e.g., the CRM), our method adopts the game competition idea which compares the evidence supporting the current dose with that of its two neighboring doses. Similar to a two-player game, one tries to push the dose up and the other tries to push it down, and once it reaches the equilibrium, the corresponding dose is the MTD. In this way, the CFO method avoids introducing any essential design parameters to calibrate, which guarantees its robustness and ease for implementation in practice. The efficacy monitoring is conducted in a simple and intuitive manner by choosing the dose which is most probable to possess the highest efficacy rate. Thus, the whole procedure of the CFO design is model-free and calibration-free and it helps to bypass the risk of model misspecification and alleviate the effect of parameter
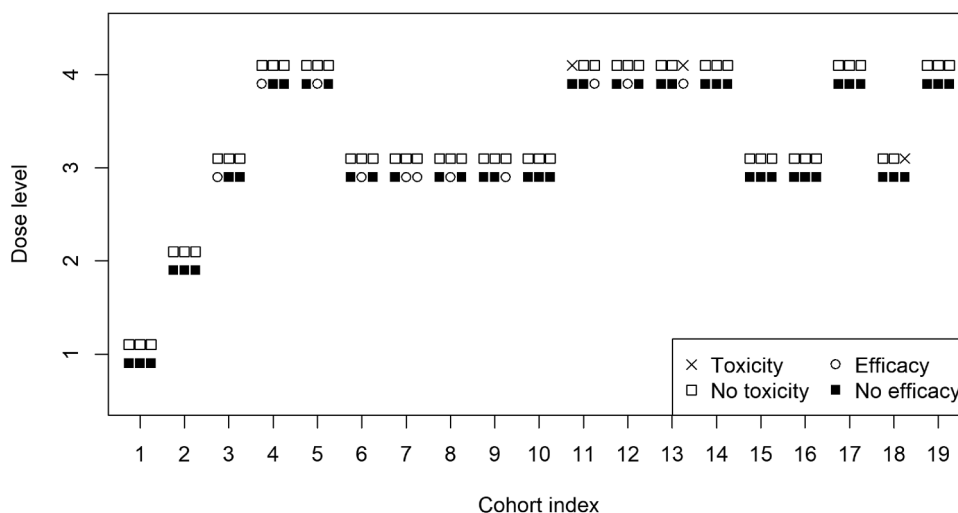
Figure 6. Dose allocations and the corresponding toxicity and efficacy outcomes for the redesigned trial.

calibration. The simulation studies show that the CFO design has robust performance in contrast to other existing methods in both MTD- and OBD-identification tasks. For phase I trials, the CFO design strikes a good balance between efficiency and safety, and for phase I/II trials, it yields similar or sometimes slightly better performance compared with the competing methods as shown by our simulations with random scenarios.

Although minimization of $V_L(\gamma_L)$ and $V_R(\gamma_R)$ seems complicated, the computation of the CFO design is fast due to the small sample size of a phase I/II trial. Using a laptop with Intel i7-10510U CPU, it only takes 0.17 second to implement the CFO design for a phase I trial with sample size 30 and 1.5 seconds for a phase I/II trial with sample size 60. Moreover, as γ_L and γ_R only depend on the numbers of patients treated at relevant dose levels as well as the target DLT rate ϕ, their values can be determined beforehand as shown in Figure 2.

The early stopping rules used in CFO are not internal components of the design, and other rules may be adopted for safety and futility stopping.[15] Our stopping rules follow the work of Yin et al. (2013) and Yin and Yang (2020),[18,25] which deliver robust and good performances with the toxicity and futility cutoff values of 0.95 and 0.9. In practice, other cutoff values can be adopted according to the characteristics of the trial. Before the trial starts, the cutoff values can be selected using simulation studies to achieve overall good trial performance.

In the development of the CFO method, we only consider the case where the efficacy and DLT outcomes are ascertainable quickly after the treatment. However, it is straightforward to extend the CFO design for the late-onset endpoints; for example, we can combine the CFO design with the so-called factional imputation method[35,34] for the late-onset endpoints, which warrants further development.

The R code for reproducing the simulation results is available at https://github.com/JINhuaqing/CFO-simu, and the one-trial implementation of the CFO design is accessible at https://github.com/JINhuaqing/CFO.

Acknowledgements

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

ORCID iDs

Huaqing Jin https://orcid.org/0000-0002-1932-1336
Guosheng Yin https://orcid.org/0000-0003-3276-1392

Supplemental material

Supplementary material for this article is available online.

References

1. Yin G. *Clinical trial design: Bayesian and frequentist adaptive methods*. vol. 876. Hoboken, NJ: John Wiley & Sons, 2012.
2. Paoletti X and Postel-Vinay S. Phase I–II trial designs: how early should efficacy guide the dose recommendation process? *Ann Oncol* 2018; **29**(3): 540–541.
3. Reynolds AR. Potential relevance of bell-shaped and u-shaped dose-responses for the therapeutic targeting of angiogenesis in cancer. *Dose Response* 2010; **8**: 253–284.
4. Morgan B, Thomas AL, Drevs J et al. Dynamic contrast-enhanced magnetic resonance imaging as a biomarker for the pharmacological response of ptk787/zk 222584, an inhibitor of the vascular endothelial growth factor receptor tyrosine kinases, in patients with advanced colorectal cancer and liver metastases: results from two phase I studies. *J Clin Oncol* 2003; **21**: 3955–3964.
5. Hoering A, Mitchell A, LeBlanc M, et al. Early phase trial design for assessing several dose levels for toxicity and efficacy for targeted agents. *Clinical Trials* 2013; **10**: 422–429.
6. Storer BE. Design and analysis of phase I clinical trials. *Biometrics* 1989; **45**: 925–937.
7. O'Quigley J, Pepe M and Fisher L. Continual reassessment method: a practical design for phase 1 clinical trials in cancer. *Biometrics* 1990; **46**: 33–48.
8. Lin R and Yin G. Nonparametric overdose control with late-onset toxicity in phase I clinical trials. *Biostatistics* 2017; **18**: 180–194.
9. Gooley TA, Martin PJ, Fisher LD, et al. Simulation as a design tool for phase I/II clinical trials: an example from bone marrow transplantation. *Control Clin Trials* 1994; **15**: 450–462.
10. Thall PF and Russell KE. A strategy for dose-finding and safety monitoring based on efficacy and adverse outcomes in phase I/II clinical trials. *Biometrics* 1998; **54**: 251–264.
11. Thall PF and Cook JD. Dose-finding based on efficacy–toxicity trade-offs. *Biometrics* 2004; **60**: 684–693.
12. Braun TM. The bivariate continual reassessment method: extending the crm to phase I trials of two competing outcomes. *Control Clin Trials* 2002; **23**: 240–256.
13. Yin G, Li Y and Ji Y. Bayesian dose-finding in phase I/II clinical trials using toxicity and efficacy odds ratios. *Biometrics* 2006; **62**: 777–787.
14. Yuan Y and Yin G. Bayesian dose finding by jointly modelling toxicity and efficacy as time-to-event outcomes. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 2009; **58**: 719–736.
15. Wages NA and Tait C. Seamless phase I/II adaptive design for oncology trials of molecularly targeted agents. *J Biopharm Stat* 2015; **25**: 903–920.
16. Liu S and Johnson VE. A robust Bayesian dose-finding design for phase I/II clinical trials. *Biostatistics* 2016; **17**: 249–263.
17. Xu J, Yin G, Ohlssen D, et al. Bayesian two-stage dose finding for cytostatic agents via model adaptation. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 2016; **65**: 465–482.
18. Lin R and Yin G. STEIN: A simple toxicity and efficacy interval design for seamless phase I/II clinical trials. *Stat Med* 2017; **36**: 4106–4120.
19. Riviere MK, Yuan Y, Jourdan JH, et al. Phase I/II dose-finding design for molecularly targeted agent: plateau determination using adaptive randomization. *Stat Methods Med Res* 2018; **27**: 466–479.
20. Zhou Y, Lee JJ and Yuan Y. A utility-based Bayesian optimal interval (U-BOIN) phase I/II design to identify the optimal biological dose for targeted and immune therapies. *Stat Med* 2019; **38**: S5299–S5316.
21. Cheson BD, Fisher RI, Barrington SF, et al. Recommendations for initial evaluation, staging, and response assessment of hodgkin and non-hodgkin lymphoma: the lugano classification. *J Clin Oncol* 2014; **32**: 3059.
22. Liu E, Marin D, Banerjee P et al. Use of car-transduced natural killer cells in cd19-positive lymphoid tumors. *N Engl J Med* 2020; **382**: 545–553.
23. Shah N, Thall PF, Fox PS et al. Phase I/II trial of lenalidomide and high-dose melphalan with autologous stem cell transplantation for relapsed myeloma. *Leukemia* 2015; **29**: 1945–1948.
24. Bril G, Dykstra R, Pillers C, et al. Algorithm as 206: isotonic regression in two independent variables. *Journal of the Royal Statistical Society Series C (Applied Statistics)* 1984; **33**: 352–357.
25. Liu S and Yuan Y. Bayesian optimal interval designs for phase I clinical trials. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 2015; **64**: 507–523.
26. Lin R and Yin G. Uniformly most powerful Bayesian interval design for phase I dose-finding trials. *Pharm Stat* 2018; **17**: 710–724.
27. Lee SM and Cheung YK. Model calibration in the continual reassessment method. *Clinical Trials* 2009; **6**: 227–238.
28. Paoletti X, O'Quigley J and Maccario J. Design efficiency in dose finding studies. *Comput Stat Data Anal* 2004; **45**: 197–214.
29. O'quigley J, Paoletti X and Maccario J. Non-parametric optimal design in dose finding studies. *Biostatistics* 2002; **3**: 51–56.
30. Wages NA and Varhegyi N. A web application for evaluating phase I methods using a non-parametric optimal benchmark. *Clinical Trials* 2017; **14**: 553–557.
31. Cangul M, Chretien YR, Gutman R, et al. Testing treatment effects in unconfounded studies under model misspecification: Logistic regression, discretization, and their combination. *Stat Med* 2009; **28**: 2531–2551.
32. Cheung YK. Simple benchmark for complex dose finding studies. *Biometrics* 2014; **70**: 389–397.
33. Mozgunov P, Jaki T and Paoletti X. A benchmark for dose finding studies with continuous outcomes. *Biostatistics* 2020; **21**: 189–201.
34. Yin G and Yang Z. Fractional design: An alternative paradigm for late-onset toxicities in oncology dose-finding studies. *Contemporary Clinical Trials Communications* 2020; **19**: 100650.
35. Yin G, Zheng S and Xu J. Fractional dose-finding methods with late-onset toxicity in phase I clinical trials. *J Biopharm Stat* 2013; **23**: 856–870.