A Simulation Study Evaluating Phase I Clinical Trial Designs for Combinational Agents

Shu Wang\textsuperscript{1,2} and Ji-Hyun Lee\textsuperscript{1,2}

\textsuperscript{1}University of Florida Health Cancer Center
\textsuperscript{2}Department of Biostatistics, University of Florida

Nowadays, more and more clinical trials choose combinational agents as the intervention to achieve better therapeutic responses. However, dose-finding for combinational agents is much more complicated than single agent as the full order of combination dose toxicity is unknown. Therefore, regular phase I designs are not able to identify the maximum tolerated dose (MTD) of combinational agents. Motivated by such needs, plenty of novel phase I clinical trial designs for combinational agents were proposed. With so many available designs, research that compare their performances, explore parameters’ impacts, and provide recommendations is very limited. Therefore, we conducted a simulation study to evaluate multiple phase I designs that proposed to identify single MTD for combinational agents under various scenarios. We also explored influences of different design parameters. In the end, we summarized the pros and cons of each design, and provided a general guideline in design selection.

KEY WORDS: clinical trial, combinational agents, dose-finding, phase I

1 Introduction

Phase I clinical trials using combinational agents are becoming more and more popular. The goal of phase I clinical trials is to identify maximum tolerated dose (MTD) that leads to a pre-specified target toxicity probability. However, dose-finding for combinational agents could be challenging as we only know the partial order of dose toxicity. To fill in this gap, lots of phase I designs for combinational agents have been proposed.
Overall there are 3 categories of designs: algorithm-based, model-based, and model-assisted. Algorithm-based designs do not involve any parametric relationship between dose combinations and their toxicity probabilities. Ivanova and Wang (Ivanova and Wang, 2004) proposed an up-and-down design and used isotonic regression to estimate the MTD. Later Ivanova and Kim (Ivanova and Kim, 2009) updated previous up-and-down design using T-statistics. Lee and Fan (Lee and Fan, 2012) proposed a two dimensional search algorithm to identify MTD. Mander and Sweeting (Mander and Sweeting, 2015) used a product of beta probabilities to identify MTD contour. Algorithm-based designs usually lack of statistically theoretical foundation, and their escalation/de-escalation rules are ad-hoc. Therefore, their performances are not guaranteed.

Model-based designs assume a parametric dose-toxicity relationship. To account for the uncertainty in the beginning of the trial, a start-up phase is usually used before switching to model-based part. The main differences among model-based methods are the choice of dose-toxicity relationship and the scheme of start-up phase. Thall et al. (Thall et al., 2003) proposed to identify MTD contour with a six-parameter logistic regression. Wang and Ivanova (Wang and Ivanova, 2005) proposed a three parameter model to link doses and toxicity probabilities. Yin and Yuan (Yin and Yuan, 2009b) proposed to use latent contingency table method. In addition, Yin and Yuan (Yin and Yuan, 2009a) proposed to use Copula to model toxicity probabilities through marginal toxicity profile of individual agents. Conaway et al. (Conaway et al., 2004) proposed to identify all possible partial orders of dose combination toxicities so that the two dimensional dose-finding can be solved by continuous reassessment method (CRM) (O'Quigley et al., 1990). Built on this method, Wages et al. (Wages et al., 2011a,b) proposed to use a subset of possible partial orders, which is more feasible especially when the total number of doses combinations is large. To avoid pre-specifying fixed partial orders, Lin and Yin (Lin et al., 2016) proposed to dynamically update the ordering. Riviere et al. (Riviere et al., 2014) developed a method based on Bayesian logistic regression. Braun and Jia (Braun and Jia, 2013) used a proportional odds logistic regression fitting models within each “row” of the dose combination matrix, and later join them together. Braun and Wang (Braun and Wang, 2010) proposed a hierarchical model through linking the effective doses with hyper parameter of dose toxicity probabilities. Tighiouart et al. (Tighiouart et al., 2017) extended escalation with overdose control (EWOC) method (Babb et al., 1998) to two-dimensional setting to identify MTD curve. However, there are several limitations of model-based designs: (1) model-based de-
signs are relatively complicated and require constant model updating by statisticians, which posit barriers to clinicians to understand and implement in practice, (2) most designs do not provide a general recommendation for design parameters (e.g., escalation/de-escalation probability cutoff), simulations are needed for parameter calibration. Later we found that it is almost impossible to identify a universal parameter that could help the design achieve optimal performances under different scenarios. (3) Some designs require prior knowledge about agents (e.g., guesses of dose combinations’ toxicities), however, even clinicians cannot guarantee correct inputs.

Model-assisted designs have the advantages of algorithm-based and model-based designs. On one hand, model-assisted designs are able to pre-tabulate escalation and de-escalation rules before trial conduct like algorithm-based designs; on the other hand, model-assisted designs have statistical foundation and performance guarantee like model-based designs. Due to the easier implementation and good performances, model-assisted designs are most idea for practical use. Lin and Yin (Lin and Yin, 2017) extended the BOIN design (Liu and Yuan, 2015; Yuan et al., 2016), Pan et al. (Pan et al., 2020) extended the Keyboard design (Yan et al., 2017) to handle two dimensional dose-findings.

There are designs that incorporate special features while conducting two dimensional dose-finding. Liu and Ning (Liu and Ning, 2013) proposed a design that is able to handle trials with delayed toxicities. Diniz et al. (Diniz et al., 2018) built a Bayesian design for combinational doses upon escalation with overdose control (EWOC) method (Babb et al., 1998) accounting for patient heterogeneity through taking baseline covariates into consideration.

With so many novel methods available, we have relatively little knowledge about which methods are superior under which scenarios. Riviere et al. (Riviere et al., 2015) compared 6 phase I designs for combinational agents that are either algorithm-based or model-based. This paper claimed that all designs were optimized to improve the percentage of correct MTD selection before comparison. However, such claim itself is questionable as it is impossible to achieve optimization using a universal set of parameters under diverse scenarios. Based on the sensitivity analyses this paper conducted, different sets of parameters actually could achieve optimization under different scenarios. Therefore, the parameter selection in the paper seems to be subjective, which makes the design comparison less meaningful. Another limitation of this paper is that it did not discuss in detail about the influences of different model settings, although sensitivity analyses were presented. Hirakawa et
al. (Hirakawa et al., 2015) compared performances of 5 model-based designs for combinational agents. But this paper didn’t explore much about parameter effect beyond cohort size. Harrington et al. (Harrington et al., 2013) just reviewed some algorithm-based and model-based combinational agents designs, and discussed their advantages and limitations without simulation studies. None of above papers included model-assisted designs as they were published in earlier days. Moreover, the feasibility of parameter tuning and influences of design parameters while using model-based designs were not well investigated. To provide a more recent view of phase I clinical trial designs for combinational agents, we conducted a simulation study to evaluate the performances of various designs. Our study is different with previous review papers in several aspects: (1) we included two novel model-assisted designs in the study, (2) for designs containing tuning parameters that are without authors’ recommendations, we tried different sets of parameters to investigate their influences instead of using single subjectively selected set of design parameter, (3) we conducted a separate set of simulations in which early stopping is allowed when 12 patients have been assigned to the same dose combination, and (4) besides summary of each design’s characteristics, we also discussed about potential reasons that led to their performances.

Specifically, we focus on designs that (1) utilize toxicity information only, (2) identify one MTD instead of MTD contour or curve, (3) assume monotonicity within each individual agent, and (4) programming codes/softwares are available. In this way, we selected 9 designs: Dose finding in discrete dose space (Wang and Ivanova, 2005), Bayesian dose finding by copula regression (Yin and Yuan, 2009a), Continual reassessment method for partial ordering (Wages et al., 2011a,b), Hierarchical Bayesian Design (Braun and Wang, 2010), Logistic model-Based Bayesian dose finding design (Riviere et al., 2014), Generalized Continual Reassessment Method (Braun and Jia, 2013), Bootstrap Aggregating Continual Reassessment Method (Lin et al., 2016), combinational Bayesian optimal interval design (Lin and Yin, 2017), and combinational Keyboard design (Pan et al., 2020).

The rest of the paper is organized as follows: in Section 2 we reviewed 9 designs that will be included in our evaluation; Section 3 presents our simulation studies; Section 3.4 is the results; in Section 4 we discussed our findings.
2 Review of Designs

2.1 Notations

Here we define some common notations used in these methods. Assume we have two agents $A$ and $B$, with $J$ and $K$ doses respectively. Define $\pi_{jk}$ to be the true toxicity probability of the dose combination $(j,k)$, $j = 1, 2, \ldots, J, k = 1, 2, \ldots, K$; define $p_j$ to be the true toxicity probability of agent $A$ when used as a monotherapy, $j = 1, 2, \ldots, J$, and $q_k$ to be the true toxicity probability of agent $B$ when used as a monotherapy, $k = 1, 2, \ldots, K$. Define $\phi$ to be the pre-specified target toxicity probability. Define $N$ to be maximum sample size in the trial. Define $n_{jk}$ to be number of subjects that received dose $(j,k)$ and $y_{jk}$ to be number of DLTs observed among those $n_{jk}$ patients.

2.2 Dose finding in discrete dose space (I2D)

This method is a Bayesian design that extends CRM to accommodate dose-finding in two dimensions. It chose the model below for dose-toxicity relationship:

$$\pi_{jk}(\theta) = 1 - (a_j)^\alpha (1 - b_k)^\beta + \gamma \log(1-a_j),$$  \hspace{1cm} (1)

where $\theta = (\alpha, \beta, \gamma)$ and restrict $\alpha > 0, \beta > 0, \gamma < 0$ to satisfy the assumption of toxicity monotonicity; $0 \leq a_1 < \cdots < a_J$ and $0 \leq b_1 < \cdots < b_K$ are constants instead of actual doses of agents. If no interaction between two agents exists in Equation (1) the model becomes

$$\pi_{jk}(\theta) = 1 - (a_j)^\alpha (1 - b_k)^\beta.$$  \hspace{1cm} (2)

In trial conduct, it involves a start-up phase in the beginning, starting with dose combination $(1, 1)$. Then we keep agent $B$ at the lowest dose and escalate agent $A$ if no DLT is observed. If still no DLT is observed when agent $A$ reaches its maximum dose, we escalate agent $B$ to its second lowest dose combining with agent $A$’s $(J - 2)^{th}$ dose. Then if no DLT is observed, we jump to combination where agent $B$ is at its third lowest dose and agent $A$’s $(J - 4)^{th}$ dose, namely the one used in previous combination minus 2. When agent $B$ reaches its maximum dose, if agent $A$ is at its $m^{th}$ dose, we evaluate all combinations from $(m, K), (m+1, K), \ldots, (J, K)$. The start-up phase
ends if at least 1 DLT is observed at any time. After the start-up phase ends, the working model Equation 1 or Equation 2 will be used to obtain toxicity estimates of all combinations. Due to safety concerns, the working model starts at the combination dose where agent B is at its lowest dose and agent A is at the dose that makes the combination’s estimated toxicity probability closest to $\phi$. In terms of dose escalation and de-escalation, if the current combination is $(j, k)$, I2D only considers doses $(i - 1, k), (i + 1, k), (i, k), (i, k + 1), (i, k - 1), (i + 1, k - 1), (i - 1, k + 1)$ prohibiting diagonal moves.

2.3 Bayesian dose finding by copula regression (Copula)

This paper utilizes copula to model the dose-toxicity relationship as copula is able to link the joint distribution and marginal distributions via a dependence parameter. Specifically, the paper proposed the dose-toxicity relationship motivated by Clayton copula:

$$
\pi_{jk} = 1 - \{(1 - p_j^\alpha)^{-\gamma} + (1 - q_k^\beta)^{-\gamma} - 1\}^{-1/\gamma},
$$

where $\alpha$ and $\beta$ are power parameters as in CRM to accommodate the uncertainty, $\gamma > 0$ represents the interaction between two agents. If only one agent is involved, this approach reduces to regular CRM.

In trial conduct, this method also involves a start-up phase beginning with the lowest dose combination (1, 1). Then it goes vertically (1, 2), (1, 3), ..., (1, $K$) until the first toxicity is observed, next it goes horizontally (2, 1), (3, 1), ..., ($J$, 1) until the first toxicity is observed. As long as one toxicity is observed in both directions, formal design starts. This design involves two parameters $c_e$ and $c_d$ that represent the fixed probability cut-offs for dose escalation and de-escalation, respectively, and $c_e + c_d > 1$. Detailed algorithm is laid out as below.

a) If at current dose $(j, k)$, $P(\pi_{jk} < \phi) > c_e$, we will escalate to the dose that belong to $\{(i + 1, j), (i, j + 1), (i + 1, j - 1), (i - 1, j + 1)\}$ and with the toxicity probability closest to $\phi$ and higher than that of $(j, k)$. When current dose is $(J, K)$, then stay at the same combination.

b) If at current dose $(j, k)$, $P(\pi_{jk} > \phi) > c_d$, we will de-escalate to the dose that belong to $\{(i - 1, j), (i, j - 1), (i + 1, j - 1), (i - 1, j + 1)\}$ and with the toxicity probability closest to $\phi$
and lower than that of \((j,k)\). When current dose is \((1,1)\), the trial is terminated.

c Otherwise, stay at the same dose combination.

After \(N\) subjects are exhausted, the MTD is selected to be the one whose estimated probability of toxicity is closest to \(\phi\).

### 2.4 Continual reassessment method for partial ordering (POCRM)

It is well known that the main difficulty of two dimensional dose-finding is that we only know partial order of the dose combinations’ toxicities. To solve this problem, POCRM proposes to pre-specify a subset of possible orderings, then utilize the CRM on each order. In this way, two dimensional dose-finding is reduced to a one dimensional problem. Define \(T\) to be the total number of dose combinations, \(T = J \times K\); \(\pi_t\) to be the true toxicity probability of dose combination \(t\), \(t = 1,2,\ldots,T\); \(d_n\) to be the dose assigned to subject \(n\), \(R(d_n)\) to be the true toxicity probability of \(d_n\), \(y_n\) to be a binary indicator of whether subject \(n\) has toxicity or not, \(n = 1,2,\ldots,N\), \(\Omega_n\) to be data collected after having \(n\) subjects and \(\Omega_n = \{d_1,y_1,\ldots,d_n,y_n\}\). Assume we have \(M\) possible partial ordering in total. Similar to the CRM, \(R(d_n)\) is modeled as below

\[
\pi_n = E(Y_n|x_n) = \psi_m(d_n,a) \tag{4}
\]

After having \(n\) patients, the likelihood under partial order \(m\) is

\[
L_m(a|\Omega_n) = \prod_{i=1}^{n} \psi_{m}^{y_i}(d_i,a) \{1 - \psi_{m}(d_i,a)\}^{(1-y_i)} \tag{5}
\]

We could obtain \(a_m^*\) through maximizing Equation 5. Then obtain the posterior probability of partial order \(m\):

\[
p(m|\Omega_n) = \frac{p(m)L_m(a_m^*|\Omega_n)}{\sum_{l=1}^{M} p(l)L_l(a_l^*|\Omega_n)} \tag{6}
\]

But we do not simply select the partial ordering with the largest posterior probability, instead, we use weighted randomization with \(p(m|\Omega_n)\) being the weight. After selecting the working partial ordering \(m\), we estimate \(\pi_t\) for all \(t \in \{1,2,\ldots,T\}\) through Equation 4 and assign the dose combination that minimizes \(|\hat{\pi}_t - \phi|\) to subject \(n + 1\).
In terms of selecting a subset of partial orders, the paper recommends 6 orders: across rows, up columns, up diagonals, down diagonals, alternative down-up diagonals, and alternative up-down diagonals.

When POCRM is first proposed, it is a single stage method (Wages et al., 2011a). Later it is extended to include a start-up phase (Wages et al., 2011b). The start-up phase partitions the dose combination matrix to different “zones” and starts the first cohort with zone 1, which is the lowest dose combination. If no DLT is observed, then we assign next cohort to doses in zone 2. If there are multiple combinations in zone 2, randomly select one of them and continue to assign next cohort to other combinations in the same zone if no DLT is observed. Moving to the next zone is only allowed when all the dose combinations have been explored in lower zones. The start-up phase ends when 1 DLT is observed. POCRM also allows user to specify their own scheme in the start-up phase as they see appropriate.

2.5 Hierarchical Bayesian Design (Hierarchy)

This method is a single stage design without a start-up phase. It employs a hierarchical model:

$$\pi_{jk} \sim \text{Beta}(\alpha_{jk}, \beta_{jk}),$$

$$\log(\alpha_{jk}) = \theta_0 + \theta_1 a_j + \theta_2 b_k, \log(\beta_{jk}) = \phi_0 + \phi_1 a_j + \phi_2 b_k,$$

where $\theta = \{\theta_0, \theta_1, \theta_2\}$ follows a multivariate normal distribution with mean $\mu = \{\mu_0, \mu_1, \mu_2\}$ and variance covariance matrix $\sigma^2 I$ where $I$ is a $3 \times 3$ identity matrix; $\phi = \{\phi_0, \phi_1, \phi_2\}$ follows a multivariate normal distribution with mean $\omega = \{\omega_0, \omega_1, \omega_2\}$ and the same variance covariance matrix; $a_j$ and $b_k$ are “effective doses” instead of actual clinical values. This method omits the interaction effects between two agents.

The paper also provides recommendations about selecting priors and methods to calculate “effective doses”. They used the fact that $\frac{K \pi_{11}}{K(1-\pi_{11})} = \frac{\exp(\mu_0)}{\exp(\omega_0)}$ to obtain the solutions for $\mu_0$ and $\omega_0$:

$$\mu_0 = \log(K \pi_{11}), \omega_0 = \log(K(1 - \pi_{11})).$$

They suggested setting $\mu_1 = \mu_2 = \omega_1 = \omega_2 = 2\sqrt{\sigma^2}$ and selecting $\sigma^2 \in [5, 10]$. They set $a_1 = b_1 = 0,$
then

\[ a_j = (\mu_1 + \omega_1)^{-1} \log(O\bar{R}_j), b_k = (\mu_2 + \omega_2)^{-1} \log(O\bar{R}_k) \]

where

\[ O\bar{R}_j = \exp\left\{ \frac{\pi_j}{\pi_1}(1 - \pi_j) \right\}, O\bar{R}_k = \exp\left\{ \frac{\pi_k}{\pi_1}(1 - \pi_k) \right\}. \]

Therefore, we need inputs of \( \pi_{j1} \) and \( \pi_{1k} \), \( j = 1, 2, \ldots, J, k = 1, 2, \ldots, K \) from the clinicians.

Below is how a real trial will be conducted:

a) Compute a 95% CI for overall toxicity rate based on observed number of DLTs.

b) If the lower bound of this CI is more than \( \phi \), terminate the trial.

c) If the lower bound of this CI is no more than \( \phi \), use all previous information to obtain posterior mean of \( \pi_{jk} \), \( j = 1, 2, \ldots, J, k = 1, 2, \ldots, K \).

d) Select a dose that belongs to adjacent doses of the current dose combination including diagonal ones and is closest to \( \phi \).

e) Continue until all \( N \) subjects are exhausted.

### 2.6 Logistic model-Based Bayesian dose finding design (DFCOMB)

This method uses logistic regression to link doses and toxicity:

\[ \text{logit}(\pi_{jk}) = \beta_0 + \beta_1 a_j + \beta_2 b_k + \beta_3 a_j b_k, \quad (9) \]

where \( a_j \) and \( b_k \) are “effective doses” instead of actual clinical values, \( \beta_1 > 0, \beta_2 > 0, \beta_1 + \beta_3 b_k > 0, \beta_2 + \beta_3 a_j > 0 \) to ensure monotonicity. They defined “effective doses” as \( a_j = \log\left(\frac{p_j}{1-p_j}\right), b_k = \log\left(\frac{q_k}{1-q_k}\right) \) and recommended a vague normal prior \( N(0, 1) \) for \( \beta_0 \) and \( \beta_3 \), an informative prior \( \exp\{1\} \) for \( \beta_1 \) and \( \beta_2 \).

DFCOMB also has a start-up phase in the beginning, starting from dose \((1, 1)\). If no toxicity is observed, escalate the dose along the diagonal until at least one agent reaches maximum dose. If still no toxicity is observed when one agent reaches its maximum dose, we increase the dose of the other agent until both agents reach maximum doses. The start-up phase ends once the first toxicity is observed and the model-based design starts.
In the model-based design part, the escalation and de-escalation rule is the same as design Copula (Yin and Yuan 2009a). However, DFCOMB utilizes a different method to identify MTD after the trial is completed. The dose combination that has the largest posterior probability \( P(\pi_{jk} \in [\phi - \delta, \phi + \delta]) \) and is used to treat at least one cohort will be selected as the MTD. Parameter \( \delta \) is the length around the target toxicity probability.

2.7 A Generalized Continual Reassessment Method (gCRM)

This method is another generalization of the CRM. Similar to Hierarchical Bayesian design, gCRM does not have a start-up phase as well. It uses proportional odds logistic regression to model the dose-toxicity relationship:

\[
\text{logit}(\pi_{jk}) = \alpha_k + \beta a_j, \tag{10}
\]

where \( \alpha_k \) is agent \( B \) specific intercept, \( k = 1, 2, \ldots, K \); \( \beta \) is a common coefficient across models; \( a_j \) is “effective dose” of agent \( A, j = 1, 2, \ldots, J \). For example, if agent \( B \) has 3 doses, then we will need 3 models: \( \text{logit}(\pi_{j1}) = \alpha_1 + \beta a_j, \text{logit}(\pi_{j2}) = \alpha_2 + \beta a_j, \) and \( \text{logit}(\pi_{j3}) = \alpha_3 + \beta a_j \). Later these “sub” models will be aggregated together through a joint prior distribution that forces correlation among \( (\alpha_1, \alpha_2, \ldots, \alpha_k) \). As we can observe, gCRM assumes no interaction between two agents as well.

In terms of parameters \( \alpha_k \) and \( \beta \), this paper assumes that \( \beta \) follows a Gamma distribution with mean \( \mu_\beta \) and variance \( \sigma_\beta^2 \), \( \alpha_1 \sim N(\mu_\alpha, \sigma_\alpha^2) \), and defines \( \Delta_k = \alpha_k - \alpha_{k-1} \sim N(\delta_k, 2\sigma_\alpha^2) \) for \( k = 2, 3, \ldots, K \) so the joint distribution of \( \alpha = (\alpha_1, \ldots, \alpha_K)^T \) is multivariate normal. If we assume that \( \text{logit}(\pi_{j1}) = E(\alpha_1) + E(\beta) a_j, \) we can obtain \( a_j \approx [\text{logit}(\pi_{j1}) - \mu_\alpha]/\mu_\beta. \) We can approximately obtain \( \delta_k = \text{logit}(\pi_{1k}) - \text{logit}(\pi_{1,k-1}). \) The authors recommended setting \( \mu_\alpha = -8, \mu_\beta = 1, \sigma_\alpha^2 = \sigma_\beta^2 = 1. \) Therefore, with the inputs of \( \pi_{j1} \) and \( \pi_{1k} \) for \( j = 1, 2, \ldots, J, k = 1, 2, \ldots, K \) from clinicians, we could calculate all parameters.

Below are trial conduct details:

a) Treat the first patient at dose \( (1, 1) \).

b) For patient \( 2, 3, \ldots, N \), compute \( \hat{\pi}_{jk} \) from \( \text{logit}(\hat{\pi}_{jk}) = \hat{\alpha}_k + \hat{\beta} a_j \) where \( \hat{\alpha}_k \) and \( \hat{\beta} \) are posterior means.
c From adjacent dose combinations (including diagonal moves), escalate or de-escalate to the
dose whose $\hat{\pi}_{jk}$ is closest to $\phi$.

d Determine if the stopping rule of $P(\pi_{11} > \phi) > 0.95$ is reached or not. If so, terminate the trial.

e Continue until all $N$ subjects are exhausted. Use step (2) to compute the final estimate of MTD.

2.8 Bootstrap Aggregating Continual Reassessment Method (bCRM)

Bootstrap aggregating CRM is similar to POCRM as they both end up with using one-dimensional
CRM to identify the MTD. However, bCRM keeps updating the toxicity ordering of dose combina-
tions rather than pre-specifying fixed orderings.

In bCRM, it assigns a beta prior to $\pi_{jk}$ and obtain its posterior mean $\bar{\pi}_{jk}$. Then bCRM
applies two-dimensional pool-adjacent-violators algorithm (PAVA) on $\bar{\pi}_{jk}$ to obtain
$\tilde{\pi}_{jk}$ to ensure that these estimates satisfy partial ordering. To avoid ties among $\tilde{\pi}_{jk}$, a term $r_{jk}\epsilon$
is added, where $r_{jk}$ is the rank of dose $(j, k)$ and $\epsilon$ is a small positive number. The resulted
estimates are denoted as $\tilde{\pi}_{jk}$ and then we can obtain a new ordering $O$. As denoted by the authors,
such orderings could vary dramatically due to data sparsity. Therefore, they bootstrapped $B$
samples of the data to obtain corresponding orderings $O_b$, and toxicity probability estimates $\hat{\pi}^b_{jk}$,
$b \in 1, 2, \ldots, B$. The final estimate of $\pi_{jk}$ is

$$\hat{\pi}_{jk}^{\text{Bagging}} = \sum_{b=1}^{B} P(O_b|D)\hat{\pi}^b_{jk},$$  \hspace{1cm} (11)

The trial conduct procedure and start-up phase are similar to DFCOMB. After the trial is com-
pleted, we select the combination that has been administered to patients and has the largest
posterior probability of falling into the $\varepsilon$-neighbourhood of $\phi$, where $\varepsilon$ is a small positive number.

2.9 Combinational Bayesian optimal interval design (cBOIN)

Combinational BOIN is a model-assisted design that is generalized from single agent BOIN design
(Liu and Yuan, 2015; Yuan et al., 2016). Boim mainly involves two important parameters $\Delta_L$
and $\Delta_U$ which are pre-specified lower and upper cut-offs. At current dose $j$, the escalation and de-escalation rules are below:

- if $\hat{p}_j \in (\phi - \Delta_L, \phi + \Delta_U)$, then next cohort stays at current dose;
- if $\hat{p}_j \leq \phi - \Delta_L$, then next cohort escalates to dose $j + 1$;
- if $\hat{p}_j \geq \phi + \Delta_U$, then next cohort de-escalates to dose $j - 1$;

where $\hat{p}_j$ is the estimated toxicity probability of dose $j$ in single agent dose-finding and it is simply proportion of patients experiencing toxicities among those who receive dose $j$.

In two dimensional dose-finding, $\hat{p}_{jk}$ is calculated the same way: $\hat{p}_{jk} = y_{jk}/n_{jk}$. As there are multiple options in escalation and de-escalation, admissible dose escalation set is defined to be $A_E = \{(j + 1, k), (j, k + 1)\}$, admissible de-escalation set is define to be $A_D = \{(j - 1, k), (j, k - 1)\}$.

An important task of cBOIN is to determine $\Delta_L$ and $\Delta_U$. Through minimizing the probability of incorrect movement given data at current dose,

$\Delta_L = \phi - \frac{\log\left\{\frac{1-\phi_2}{1-\phi_1}\right\}}{\log\left\{\frac{\phi(1-\phi_1)}{\phi_1(1-\phi)}\right\}}$, $\Delta_U = \frac{\log\left\{\frac{1-\phi}{1-\phi_2}\right\}}{\log\left\{\frac{\phi_2(1-\phi)}{\phi(1-\phi_2)}\right\}} - \phi$.

The authors suggested using $\phi_1 = 0.6\phi$ and $\phi_2 = 1.4\phi$.

The trial conduct is laid out as below:

a Treat the first cohort at dose (1, 1).

b Suppose that current dose is combination $(j, k)$. If $\hat{p}_{jk} \leq \phi - \Delta_L$, escalate to dose combination that belongs to $A_E$ and has the largest $P[p_{jk'} \in (\phi - \Delta_L, \phi + \Delta_U) | y_{jk'}]$.

c If $\hat{p}_{jk} \geq \phi + \Delta_U$, de-escalate to dose combination that belongs to $A_D$ and has the largest $P[p_{jk'} \in (\phi - \Delta_L, \phi + \Delta_U) | y_{jk'}]$.

d Otherwise if $\phi - \Delta_L < \hat{p}_{jk} < \phi + \Delta_U$, stay at current dose.

e Continue until all $N$ subjects are exhausted.

After the trial is completed, isotonic regression will be used to identify the MTD.
2.10 Combinational Keyboard design (cKeyboard)

Similar to combinational BOIN, combinational Keyboard is a model-assisted design as well. Combinational Keyboard design starts with specifying a toxicity interval $\mathcal{J}_t = (\phi - \varepsilon_1, \phi + \varepsilon_2)$, where $\varepsilon_1$ and $\varepsilon_2$ are tolerable deviations from $\phi$. This interval $\mathcal{J}_t$ is called target key. Then a series of equally-width keys are identified along both sides of the target key. In the setting of single agent design, the escalation and de-escalation rules are also straightforward. At current dose $j$,

- if the strongest key based on the posterior distribution of current dose $j$ $\mathcal{J}_{\text{max}} \prec \mathcal{J}_t$, then next cohort escalates to dose $j + 1$;
- if $\mathcal{J}_{\text{max}} \equiv \mathcal{J}_t$, then next cohort stays at current dose;
- if $\mathcal{J}_{\text{max}} \succ \mathcal{J}_t$, then next cohort de-escalates to dose $j - 1$;

In terms of two dimensional dose-finding, the paper defines 5 different strategies of admissible escalate and de-escalate sets. After simulations, strategy 1 when admissible escalation and de-escalation sets are the same with combinational BOIN is recommended. The trial conduct is shown below:

a. Treat the first cohort at dose $(1, 1)$.

b. Suppose that current dose is combination $(j, k)$. If $\mathcal{J}_{\text{max}} \prec \mathcal{J}_t$, escalate to dose combination that belongs to $\mathcal{A}_E$ and has the largest $P[p_{jk} \in \mathcal{J}_t | (n_{jk}, y_{jk})]$.

c. If $\mathcal{J}_{\text{max}} \succ \mathcal{J}_t$, de-escalate to dose combination that belongs to $\mathcal{A}_D$ and has the largest $P[p_{jk} \in \mathcal{J}_t | (n_{jk}, y_{jk})]$.

d. Otherwise if $\mathcal{J}_{\text{max}} \equiv \mathcal{J}_t$, stay at current dose.

e. Continue until all $N$ subjects are exhausted.

After the trial is completed, isotonic regression will be used to identify the MTD.

3 Simulation Studies

In the simulation studies, our goal is to identify single MTD of two combined agents. Agent $A$ has 5 dose levels and agent $B$ has 3. The target toxicity probability is 0.3. Simulation settings are
borrowed from previous study and shown in Figure 1. Maximum sample size is 60. All designs started with the lowest dose combination. Cohort size was set to be 3 for all designs that use cohorts as unit, unless specified. 2000 simulation runs were generated for each scenario.

3.1 Simulation scenarios

All 10 scenarios are displayed in Figure 1. Each scenario is presented as a heatmap where darker blue indicates higher toxicity. Numbers on those figures are true toxicity rates of those dose combinations. Target toxicity rate 0.3 was marked as yellow. From these heatmaps we can observe that: scenario 1 contains multiple MTD combinations that are in the middle of matrix and diagonally connected; scenarios 2 and 4 represent over-toxic situations and scenario 4 is more extreme; scenarios 3 and 5 represent over-conservative situations and scenario 5 is more extreme; scenarios 6 and 7 contain multiple MTD combinations but those MTD combinations are more scattered; scenarios 8, 9, and 10 contain single MTD combination at different locations.

3.2 Design specifications

Table 1 lists design parameters that we explored in simulation. Column "Main setting" contains parameter specifications used to generate Table 2 and Table 3. Column "Alternative setting" contains another set of specifications we used. By comparing results using main setting and alternative setting, we are able to explore the influences of these design parameters. Below are other design-specific settings whose recommendations are available.

For I2D, we set cohort size of start-up phase to be 1 based on suggestions from other simulation studies when target toxicity rate is 0.3 (Ivanova et al., 2003; Wang and Ivanova, 2005), and interaction to be 0 so that it is consistent with the paper’s focus. For Copula, the program fixes escalation and de-escalation probability boundaries as 0.8 and 0.45, respectively. For Hierarchy, we set $\sigma^2$ to be 10 based on the suggestion in the paper. To explore the effect of prior guess of dose combinations’ toxicity probabilities, we tried correct and incorrect prior guesses of toxicity probabilities. For POCRM, We used 6 possible partial ordering based on suggestions (Wages and Varhegyi, 2013; Wages and Conaway, 2013): across rows, across columns, up diagonals, down diagonals; up-down diagonals; down-up diagonals. As we do not have information about which partial ordering is more likely, the prior probabilities of all 6 partial ordering were set to be equal.
The skeleton required by the program was obtained using getprior function from algorithm of Lee and Cheung (Lee and Cheung, 2009) as suggested (Wages and Varhegyi, 2013). For the start-up phase, we used the “zoning” method as suggested (Wages et al., 2011a,b). For cBOIN, the interval boundaries were set to be 0.18 and 0.42 as suggested. For DFCOMB, we set the target toxicity boundaries as 0.18 and 0.42 to be consistent with cBOIN. For gCRM, all parameters used values suggested in the paper. To access the influence of prior guess of combinational doses’ toxicity probability, correct and incorrect toxicity probabilities were used. For bCRM, the skeleton setting is the same with that of POCRM.

For designs POCR, DFCOMB, cBOIN, and cKeyboard, we explored their performances when the early stopping rule of having 12 patients enrolled in the same dose combination is allowed.

3.3 Evaluation metrics

Four evaluation metrics are used: (1) correct MTD selection $S_C$, defined as proportion of simulation runs that correctly identified the MTD among all 2000 simulations; (2) over-toxic MTD selection $S_{OT}$, defined as proportion of simulation runs that identified over-toxic doses as MTD among all 2000 simulations; (3) correct patient assignment $A_C$, defined as the average proportion of patients that were assigned to the MTD during the trial across all 2000 simulations; (4) over-toxic patients assignment $A_{OT}$, defined as the average proportion of patients that were assigned to over-toxic doses during the trial across all 2000 simulations. Metrics $S_C$ and $S_{OT}$ will be used to evaluate the performance of designs in terms of MTD selection. The larger the $S_C$ is, the more accurate the design is in selecting the correct MTD. The larger the $S_{OT}$ is, the more aggressive the design is in selecting MTD. While metrics $A_C$ and $A_{OT}$ will be used to evaluate the characteristics of designs during trial conduct. The larger the $A_{OT}$ is, the more aggressive the design is in dose escalation during the trial. Ideally, a design should show relatively large $S_C$ and $A_C$ but small $S_{OT}$ and $A_{OT}$.

3.4 Results

Table 2 displays $S_C$ and $S_{OT}$; Table 3 displays $A_C$ and $A_{OT}$. Within each scenario, we marked designs whose performances were among the worst as red, and among the best as green. In addition, Table 4 shows performances of some designs when the early stopping rule is allowed. Results from “sensitivity” simulations when using alternative input parameters in Table 1 are shown in
Supplementary Materials.

I2D shows unstable performances in MTD identification across different simulation scenarios. Under extreme conditions like scenario 4 and 5, its $S_C$ is among the best, but under some other scenarios like scenarios 1, 3, 6, 8, and 9, its $S_C$ is among the worst. Partial reason of its unstable performance is that, I2D always starts its model-based part with agent B’s lowest dose no matter what happens in the start-up phase. In this way, the starting dose combination of model-based part could be very far away from the location of true MTDs, which makes I2D harder to identify them. Overall I2D is not an aggressive design as its $S_{OT}$ and $A_{OT}$ are relatively small under most scenarios. With alternative toxicity profile of individual agents ($p_j$ and $q_k$), most scenarios had little changes. Even for those few scenarios that were improved or impaired, such performance fluctuation does not change the overall rank of I2D’s performances among all designs.

Copula performed poorly in MTD identification as its $S_C$ and $S_{OT}$ are among the worst in several scenarios. Especially in scenarios 1, 3, 6, 8, and 10, its $S_{OT}$ is much larger than the second largest ones. This indicates that Copula is quite aggressive as it is more likely to select higher dose combinations as the MTD. Such characteristic is quite dangerous in dose-finding studies. Similar to I2D, Copula’s performances with alternative toxicity profile of individual agents ($p_j$ and $q_k$) do not change much; and the only improved scenario stays at similar rank among all designs. One potential reason of the poor performances is the limitation of not allowing users to change parameters like escalation and de-escalation probability cutoffs in the exe file that runs Copula. Therefore, we can argue that the default values (0.8 and 0.45) are not optimal for some scenarios. But on the other side, as we do not know what the true dose toxicity matrix looks like in real life, obtaining a uniform parameter set to achieve best performances under all scenarios through simulation calibration is not feasible.

Overall, design Hierarchy is quite aggressive in trial conduct as it has the worst $A_{OT}$ under several scenarios. We can observe that using correct $\pi_{j1}$ and $\pi_{1k}$ performed much better than using incorrect ones regarding $S_C$ in most scenarios. However, such improvement is not observed in $A_{OT}$ indicating that even if the clinicians provide accurate estimates of $\pi_{j1}$ and $\pi_{1k}$, Hierarchy is still overly aggressive in trial conduct. A possible reason is that it allows simultaneous dose escalation of both agents during trial conduct while most other designs do not. Another aspect we should notice is that with such high proportion of patients being assigned to over-toxic doses, Hierarchy
did not outperform other designs in terms of $S_C$. Some features of Hierarchy like omitting the interaction effect between agents and no start-up phase may be part of the reason.

POCRM has satisfactory characteristics across all scenarios except low $S_C$ in scenario 5 and slightly high $S_{OT}$ in scenario 4. However, using the alternative skeleton setting in Table 1 improved $S_C$ from 0.54 to 0.73 in scenario 5 and slightly reduced $S_{OT}$ from 0.22 to 0.20 in scenario 4 while keeping other metrics in other scenarios almost the same. Such results indicate that it is likely to calibrate input parameters of POCRM so that it could achieve better performances under different scenarios. Moreover, we observe that POCRM performed well under scenarios (e.g., scenario 9) when the underlying true toxicity orderings of dose combinations are not among any of the 6 partial orderings we used. Such results “validate” the idea of POCRM: in practice we just need to provide orderings close to the correct one instead of exact correct one.

DFCOMB performed poorly in terms of identifying the correct MTD under several scenarios. But overall DFCOMB is not an aggressive design as its $S_{OT}$ and $A_{OT}$ are relatively small under most scenarios. With alternative escalation and de-escalation probability cutoffs, the operating characteristics in most scenarios had little change. However, a few scenarios were impact: $S_C$ in scenario 1 was improved from 0.54 to 0.69, but $S_C$ in scenario 10 was dramatically reduced from 0.47 to 0.14. This indicates that the optimal design parameters are scenario-dependent. Therefore, it is not feasible for us to calibrate the parameters through simulations in real life clinical trails.

Design gCRM overall performs stably well in most scenarios regarding all evaluation metrics. Comparing results using correct $\pi_j$ and $\pi_k$ with in correct ones, We can observe that using correct inputs has slightly better operating characteristics under some scenarios, and similar performances under the other scenarios. Interestingly, although gCRM also allows simultaneous dose escalation of both agents, it did not show much aggressiveness as its $S_{OT}$ and $A_{OT}$ are not among the largest under most scenarios. One possible reason is that gCRM does not use cohorts as patient unit during trial conduct, it deals with every single patient instead. Therefore, every time when an over-toxic dose combination is going to be assigned, only one patient instead of a cohort of several patients will receive it. From this perspective, gCRM could be viewed as more “flexible” in the dose-finding process and such flexibility may dilute the aggressiveness.

Design bCRM is another one whose performances are unstable across different scenarios. Its $S_C$ in scenario 5 and $S_{OT}$ in scenario 4 are among the worst. Metrics in other scenarios are OK.
Similar to DFCOMB, the operating characteristics had little change with alternative escalation and de-escalation probability cutoffs. However, the alternative skeleton setting does have influences as it improved the $S_C$ in scenario 5 and $S_{OT}$ in scenario 4, but diminished $S_C$ in scenario 10 and $S_{OT}$ in scenario 9. Therefore, the issue of unstable performances remains for bCRM even using alternative skeleton setting. Similar to DFCOMB, the optimal design parameters in bCRM are scenario-dependent as well.

Designs cBOIN and cKeyboard always performed well across scenarios. They may not be the top performers, but their operating characteristics are never among the worst. This is especially important as in real life clinical trials, we have no idea which scenario could be the truth. Therefore, cBOIN and cKeyboard are able to guarantee satisfactory performances in practice. The stability may be due to the nature of model-assisted designs.

Table 4 displays results of designs that allow for early stopping if more than a pre-specified number of patients have been assigned to the same dose combination. We set the pre-specified number to be 12. Designs POCRM, DFCOMB, cBOIN, and cKeyboard are included as their programs have the ability of implementing such stopping rule. It is evident that POCRM has the worst $S_{OT}$ and $A_{OT}$ in scenarios 2 and 4, and DFCOMB has the worst $S_C$ and $S_C$ in several scenarios. On the contrary, cBOIN’s performances are always stable across different scenarios and are never the worst compared with others. Design cKeyboard performed slightly worse than cBOIN, but is overall stable as well. Therefore, cBOIN can be seen as the optimal among these four designs. Next, after comparing these results with Table 2 and Table 3 where such early stopping rule is not allowed, we found that every design’s $S_C$ was diminished. This is expected as smaller exhausted sample sizes in Table 4 naturally led to less accurate MTD identification. Among these four designs, performance drop of DFCOMB is the largest. But interestingly, $S_{OT}$ and $A_{OT}$ of DFCOMB were improved using smaller sample sizes whereas $S_{OT}$ and $A_{OT}$ became worse for all other designs.

4 Discussion

Nowadays, with more and more combinational phase I studies and available statistical designs, we actually found that those novel statistical designs were seldomly used. There may be several barriers that limit their use. One of them is that there is no practical guidance to the investigators to choose
one design with so many available. Another one is that model-based designs are quite complicated in implementation as they usually require lots of prior guesses, parameter calibration, and constant involvement of statisticians to update toxicity probabilities estimation. Motivated by these existing barriers, we conducted a simulation study aiming to provide practical recommendations to the investigators in designing phase I clinical trials through comparing performances of multiple designs. In addition, we also investigated the impact of different parameters in running model-based designs.

From our simulation results, it is clear that POCRM, gCRM, cBOIN, and cKeyboard are overall better than the others regarding evaluation metrics. Among these four designs, we recommend using cBOIN and cKeyboard in practice. The reasons are multi-folds. First, as model-assisted designs, cBOIN and cKeyboard need neither parameter calibration nor prior guesses of combination dose toxicity probabilities. Second, cBOIN and cKeyboard are easier to implement in practice as they are able to provide a dose escalation/de-escalation table like 3+3 design. Whereas statisticians need to update next dose to use once previous unit of patients’ toxicity data is available. Such complication may barrier their use in reality. Finally, cBOIN and cKeyboard have better performances than POCRM when early stopping rule of same dose collecting 12 patients is allowed. This indicates that cBOIN and cKeyboard are more reliable with limited sample sizes.

We explored the impact of several parameters that the authors did not provide clear recommendations: dose escalation/de-escalation probability cutoffs, skeleton settings, and monotherapy toxicity probabilities of individual agent. After re-running designs using alternative parameters, we found that dose escalation/de-escalation probability cutoffs actually had little impact on design operating characteristics in all scenarios. Monotherapy toxicity probabilities of individual agent usually have influences on study designs, but different scenarios require different monotherapy toxicity probabilities to achieve a design’s optimal performances. If monotherapy toxicity probabilities are available for both agents in practice, then this is not a concern. However, if information is not available from at least one of the agents, we will not be able to identify a universal set of monotherapy toxicity probabilities through simulation calibration. This becomes a disadvantage for designs I2D, Copula, and DFCOMB that require such inputs. The calibration of skeleton settings in bCRM faces the same issue. However, we were able to identify an optimal skeleton setting that work for all scenarios for POCRM.

In summary, we recommend broader usage of cBOIN and cKeyboard in phase I clinical trials
that aim to identify MTD of combinational agents. Designs cBOIN and cKeyboard have stable and satisfactory operating characteristics in diverse scenarios and more importantly, the implementation convenience could increase the investigators’ willingness of using them in designing future clinical trials.
References

Babb, J., Rogatko, A., and Zacks, S. (1998). Cancer phase i clinical trials: efficient dose escalation with overdose control. *Statistics in medicine*, 17(10):1103–1120.

Braun, T. M. and Jia, N. (2013). A generalized continual reassessment method for two-agent phase i trials. *Statistics in Biopharmaceutical Research*, 5(2):105–115.

Braun, T. M. and Wang, S. (2010). A hierarchical bayesian design for phase i trials of novel combinations of cancer therapeutic agents. *Biometrics*, 66(3):805–812.

Bril, G., Dykstra, R., Pillers, C., and Robertson, T. (1984). Algorithm as 206: isotonic regression in two independent variables. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 33(3):352–357.

Conaway, M. R., Dunbar, S., and Peddada, S. D. (2004). Designs for single-or multiple-agent phase i trials. *Biometrics*, 60(3):661–669.

Diniz, M. A., Kim, S., and Tighiouart, M. (2018). A bayesian adaptive design in cancer phase i trials using dose combinations in the presence of a baseline covariate. *Journal of probability and statistics*, 2018.

Harrington, J. A., Wheeler, G. M., Sweeting, M. J., Mander, A. P., and Jodrell, D. I. (2013). Adaptive designs for dual-agent phase i dose-escalation studies. *Nature Reviews Clinical Oncology*, 10(5):277.

Hirakawa, A., Wages, N. A., Sato, H., and Matsui, S. (2015). A comparative study of adaptive dose-finding designs for phase i oncology trials of combination therapies. *Statistics in medicine*, 34(24):3194–3213.

Ivanova, A. and Kim, S. H. (2009). Dose finding for continuous and ordinal outcomes with a monotone objective function: a unified approach. *Biometrics*, 65(1):307–315.

Ivanova, A., Montazer-Haghighi, A., Mohanty, S. G., and D. Durham, S. (2003). Improved up-and-down designs for phase i trials. *Statistics in medicine*, 22(1):69–82.
Ivanova, A. and Wang, K. (2004). A non-parametric approach to the design and analysis of two-dimensional dose-finding trials. *Statistics in Medicine*, 23(12):1861–1870.

Lee, B. L. and Fan, S. K. (2012). A two-dimensional search algorithm for dose-finding trials of two agents. *Journal of biopharmaceutical statistics*, 22(4):802–818.

Lee, S. M. and Cheung, Y. K. (2009). Model calibration in the continual reassessment method. *Clinical Trials*, 6(3):227–238.

Lin, R. and Yin, G. (2017). Bayesian optimal interval design for dose finding in drug-combination trials. *Statistical methods in medical research*, 26(5):2155–2167.

Lin, R., Yin, G., et al. (2016). Bootstrap aggregating continual reassessment method for dose finding in drug-combination trials. *The Annals of Applied Statistics*, 10(4):2349–2376.

Liu, S. and Ning, J. (2013). A bayesian dose-finding design for drug combination trials with delayed toxicities. *Bayesian analysis*, 8(3):703.

Liu, S. and Yuan, Y. (2015). Bayesian optimal interval designs for phase i clinical trials. *Journal of the Royal Statistical Society: Series C: Applied Statistics*, pages 507–523.

Mander, A. P. and Sweeting, M. J. (2015). A product of independent beta probabilities dose escalation design for dual-agent phase i trials. *Statistics in medicine*, 34(8):1261–1276.

O’Quigley, J., Pepe, M., and Fisher, L. (1990). Continual reassessment method: a practical design for phase 1 clinical trials in cancer. *Biometrics*, pages 33–48.

Pan, H., Lin, R., Zhou, Y., and Yuan, Y. (2020). Keyboard design for phase i drug-combination trials. *Contemporary Clinical Trials*, page 105972.

Riviere, M.-K., Dubois, F., and Zohar, S. (2015). Competing designs for drug combination in phase i dose-finding clinical trials. *Statistics in medicine*, 34(1):1–12.

Riviere, M.-K., Yuan, Y., Dubois, F., and Zohar, S. (2014). A bayesian dose-finding design for drug combination clinical trials based on the logistic model. *Pharmaceutical statistics*, 13(4):247–257.

Thall, P. F., Millikan, R. E., Mueller, P., and Lee, S.-J. (2003). Dose-finding with two agents in phase i oncology trials. *Biometrics*, 59(3):487–496.
Tighiouart, M., Li, Q., and Rogatko, A. (2017). A bayesian adaptive design for estimating the
maximum tolerated dose curve using drug combinations in cancer phase i clinical trials. *Statistics
in medicine*, 36(2):280–290.

Wages, N. A. and Conaway, M. R. (2013). Specifications of a continual reassessment method design
for phase i trials of combined drugs. *Pharmaceutical statistics*, 12(4):217–224.

Wages, N. A., Conaway, M. R., and O’Quigley, J. (2011a). Continual reassessment method for
partial ordering. *Biometrics*, 67(4):1555–1563.

Wages, N. A., Conaway, M. R., and O’Quigley, J. (2011b). Dose-finding design for multi-drug
combinations. *Clinical Trials*, 8(4):380–389.

Wages, N. A. and Varhegyi, N. (2013). pocrm: an r-package for phase i trials of combinations of
agents. *Computer methods and programs in biomedicine*, 112(1):211–218.

Wang, K. and Ivanova, A. (2005). Two-dimensional dose finding in discrete dose space. *Biometrics*,
61(1):217–222.

Yan, F., Mandrekar, S. J., and Yuan, Y. (2017). Keyboard: a novel bayesian toxicity probability
interval design for phase i clinical trials. *Clinical Cancer Research*, 23(15):3994–4003.

Yin, G. and Yuan, Y. (2009a). Bayesian dose finding in oncology for drug combinations by copula
regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 58(2):211–224.

Yin, G. and Yuan, Y. (2009b). A latent contingency table approach to dose finding for combinations
of two agents. *Biometrics*, 65(3):866–875.

Yuan, Y., Hess, K. R., Hilsenbeck, S. G., and Gilbert, M. R. (2016). Bayesian optimal interval
design: a simple and well-performing design for phase i oncology trials. *Clinical Cancer Research*,
22(17):4291–4301.
Figure 1: Simulation Settings
| Design     | Parameter | Main Setting                | Alternative Setting             |
|------------|-----------|-----------------------------|---------------------------------|
| I2D Copula | $p_j$ and $q_k$ | $p_j$: 0.1, 0.2, 0.25, 0.3, 0.35  
$q_k$: 0.1, 0.3, 0.35 | $p_j$: 0.05, 0.1, 0.2, 0.25, 0.3  
$q_k$: 0.1, 0.2, 0.25 |
| DFCOMB     |           |                             |                                 |
| POCRM      | Skeleton setting | half width: 0.05  
MTD position: 11 | half width: 0.03  
MTD position: 13 |
| DFCOMB bCRM| escalation/de-escalation probability cutoffs | 0.85 and 0.45 | 0.75 and 0.35 |
| Hierarchy gCRM | $\pi_{1k}$ and $\pi_{j1}$ | truth of each scenario | incorrect guess |

Table 1: Explored parameter settings
## Simulation Scenario

| Design     | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   | 10  |
|------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| **Selection of correct MTD ($S_C$)** |     |     |     |     |     |     |     |     |     |     |
| I2D        | 0.43| 0.71| 0.41| 0.90| 0.86| 0.34| 0.59| 0.11| 0.15| 0.43|
| Copula     | 0.53| 0.60| 0.52| 0.10| 0.90| 0.44| 0.65| 0.15| 0.32| 0.26|
| Hierarchy.1| 0.61| 0.64| 0.62| 0.10| 0.81| 0.45| 0.45| 0.30| 0.47| 0.52|
| POCRM      | 0.75| 0.71| 0.69| 0.78| 0.54| 0.59| 0.56| 0.59| 0.52| 0.58|
| DFCOMB     | 0.54| 0.76| 0.66| 0.65| 0.54| 0.33| 0.69| 0.48| 0.15| 0.47|
| gCRM.1     | 0.69| 0.65| 0.71| 0.42| 0.81| 0.59| 0.67| 0.34| 0.47| 0.55|
| cBOIN      | 0.70| 0.69| 0.70| 0.62| 0.72| 0.58| 0.74| 0.38| 0.40| 0.45|
| cKeyboard  | 0.67| 0.70| 0.70| 0.60| 0.72| 0.56| 0.71| 0.38| 0.40| 0.45|
| bCRM       | 0.72| 0.75| 0.66| 0.76| 0.52| 0.51| 0.63| 0.51| 0.37| 0.50|

| Selection of over-toxic MTD ($S_{OT}$) |     |     |     |     |     |     |     |     |     |     |
|----------------------------------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| I2D                                    | 0.18| 0.18| 0.18| 0.10| 0   | 0.29| 0.07| 0.26| 0.39| 0.32|
| Copula                                 | 0.32| 0.19| 0.36| 0.09| 0   | 0.43| 0.22| 0.58| 0.55| 0.63|
| Hierarchy.1                            | 0.22| 0.18| 0.21| 0.16| 0   | 0.30| 0.31| 0.30| 0.17| 0.20|
| POCRM                                  | 0.12| 0.24| 0.08| 0.22| 0   | 0.11| 0.19| 0.17| 0.18| 0.26|
| DFCOMB                                 | 0.18| 0.08| 0.17| 0.06| 0   | 0.27| 0.09| 0.27| 0.57| 0.32|
| gCRM.1                                 | 0.15| 0.13| 0.13| 0.10| 0   | 0.18| 0.10| 0.31| 0.11| 0.33|
| cBOIN                                  | 0.16| 0.21| 0.15| 0.17| 0   | 0.19| 0.13| 0.21| 0.13| 0.31|
| cKeyboard                              | 0.17| 0.21| 0.14| 0.17| 0   | 0.20| 0.14| 0.21| 0.12| 0.31|
| bCRM                                   | 0.08| 0.22| 0.05| 0.24| 0   | 0.11| 0.15| 0.20| 0.20| 0.38|

Table 2: Performance of MTD Selection of Designs across Scenarios
| Simulation Scenario | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   | 10  |
|---------------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| I2D                 | 0.38| 0.57| 0.45| 0.71| 0.58| 0.23| 0.5 | 0.07| 0.05| 0.33|
| Copula              | 0.33| 0.57| 0.31| 0.72| 0.46| 0.29| 0.41| 0.12| 0.31| 0.09|
| Hierarchy.1         | 0.42| 0.48| 0.43| 0.59| 0.68| 0.31| 0.35| 0.21| 0.3  | 0.35|
| POCRWM             | 0.53| 0.53| 0.47| 0.64| 0.33| 0.37| 0.43| 0.35| 0.3  | 0.36|
| DFCOMB              | 0.23| 0.45| 0.37| 0.91| 0.39| 0.16| 0.33| 0.23| 0.13| 0.16|
| gCRM.1              | 0.46| 0.45| 0.49| 0.75| 0.69| 0.36| 0.46| 0.27| 0.29 | 0.36|
| cBOIN               | 0.43| 0.49| 0.4  | 0.72| 0.43| 0.34| 0.46| 0.21| 0.26 | 0.2 |
| cKeyboard           | 0.42| 0.49| 0.4  | 0.72| 0.43| 0.33| 0.44| 0.21| 0.25 | 0.2 |
| bCRM                | 0.41| 0.51| 0.36| 0.67| 0.24| 0.24| 0.32| 0.23| 0.15 | 0.22|

| Patient receiving over-toxic doses during trials ($A_{OT}$) | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   | 10  |
|------------------------------------------------------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| I2D                                                         | 0.29| 0.27| 0.19| 0.29| 0   | 0.27| 0.16| 0.37| 0.35| 0.31|
| Copula                                                      | 0.17| 0.13| 0.19| 0.28| 0   | 0.18| 0.13| 0.31| 0.23| 0.41|
| Hierarchy.1                                                | 0.34| 0.31| 0.31| 0.28| 0   | 0.36| 0.35| 0.4  | 0.29| 0.38|
| POCRWM                                                     | 0.16| 0.32| 0.11| 0.36| 0   | 0.15| 0.24| 0.24| 0.21| 0.38|
| DFCOMB                                                     | 0.11| 0.05| 0.14| 0.09| 0   | 0.14| 0.08| 0.23| 0.25| 0.23|
| gCRM.1                                                     | 0.27| 0.27| 0.24| 0.25| 0   | 0.29| 0.21| 0.34| 0.23| 0.42|
| cBOIN                                                      | 0.20| 0.27| 0.18| 0.28| 0   | 0.22| 0.20| 0.27| 0.21| 0.39|
| cKeyboard                                                  | 0.20| 0.27| 0.17| 0.28| 0   | 0.22| 0.21| 0.27| 0.20 | 0.38|
| bCRM                                                       | 0.13| 0.28| 0.06| 0.33| 0   | 0.12| 0.19| 0.25| 0.19 | 0.35|

Table 3: Performance of Patient Assignment of Designs across Scenarios
## Simulation Scenario

| Design   | 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  | 10 |
|----------|----|----|----|----|----|----|----|----|----|----|
| Selection of correct MTD ($S_C$) | | | | | | | | | | |
| POCR | 0.61 | 0.50 | 0.58 | 0.70 | 0.43 | 0.45 | 0.49 | 0.44 | 0.39 | 0.40 |
| DFCOMB | 0.27 | 0.21 | 0.50 | 0.79 | 0.46 | 0.24 | 0.41 | 0.27 | 0.14 | 0.19 |
| cBOIN 0.56 | 0.56 | 0.57 | 0.63 | 0.63 | 0.47 | 0.62 | 0.32 | 0.36 | 0.35 |
| cKeyboard | 0.50 | 0.51 | 0.54 | 0.61 | 0.60 | 0.41 | 0.54 | 0.27 | 0.33 | 0.22 |
| Selection of over-toxic MTD ($S_{OT}$) | | | | | | | | | | |
| POCR | 0.17 | 0.34 | 0.09 | 0.31 | 0.00 | 0.15 | 0.22 | 0.23 | 0.21 | 0.38 |
| DFCOMB | 0.10 | 0.02 | 0.14 | 0.02 | 0.00 | 0.13 | 0.03 | 0.20 | 0.25 | 0.12 |
| cBOIN | 0.20 | 0.25 | 0.17 | 0.24 | 0.00 | 0.22 | 0.18 | 0.26 | 0.18 | 0.40 |
| cKeyboard | 0.21 | 0.24 | 0.18 | 0.25 | 0.00 | 0.24 | 0.20 | 0.29 | 0.24 | 0.46 |
| Patients receiving correct MTD during trials ($A_C$) | | | | | | | | | | |
| POCR | 0.41 | 0.39 | 0.34 | 0.56 | 0.19 | 0.26 | 0.36 | 0.22 | 0.2 | 0.23 |
| DFCOMB | 0.15 | 0.24 | 0.29 | 0.93 | 0.26 | 0.12 | 0.22 | 0.14 | 0.12 | 0.07 |
| cBOIN | 0.31 | 0.40 | 0.28 | 0.69 | 0.26 | 0.24 | 0.34 | 0.15 | 0.19 | 0.14 |
| cKeyboard | 0.29 | 0.39 | 0.27 | 0.69 | 0.25 | 0.22 | 0.31 | 0.14 | 0.18 | 0.10 |
| Patients receiving over-toxic doses during trials ($A_{OT}$) | | | | | | | | | | |
| POCR | 0.16 | 0.42 | 0.09 | 0.44 | 0.00 | 0.13 | 0.22 | 0.26 | 0.19 | 0.43 |
| DFCOMB | 0.09 | 0.03 | 0.13 | 0.07 | 0.00 | 0.11 | 0.06 | 0.23 | 0.27 | 0.22 |
| cBOIN | 0.18 | 0.26 | 0.15 | 0.31 | 0.00 | 0.20 | 0.20 | 0.26 | 0.21 | 0.40 |
| cKeyboard | 0.17 | 0.24 | 0.14 | 0.31 | 0.00 | 0.19 | 0.20 | 0.25 | 0.21 | 0.43 |

Table 4: Performances of Designs If Early Stop (Having 12 Patients Receiving A Dose) Is Allowed