Biostatistics (2022) **23**, 3, *pp.* 721–737
doi: 10.1093/biostatistics/kxaa054
Advance Access publication on January 4, 2021

A benchmark for dose-finding studies with unknown ordering

PAVEL MOZGUNOV*

Department of Mathematics and Statistics, Lancaster University, Lancaster, UK

XAVIER PAOLETTI

Université Versailles St Quentin & INSERM U900 STAMPM, Institut Curie, Paris, France

THOMAS JAKI

Department of Mathematics and Statistics, Lancaster University, Lancaster, UK and MRC Biostatistics Unit, University of Cambridge, Cambridge, UK

p.mozgunov@lancaster.ac.uk

Summary

An important tool to evaluate the performance of a dose-finding design is the nonparametric optimal benchmark that provides an upper bound on the performance of a design under a given scenario. A fundamental assumption of the benchmark is that the investigator can arrange doses in a monotonically increasing toxicity order. While the benchmark can be still applied to combination studies in which not all dose combinations can be ordered, it does not account for the uncertainty in the ordering. In this article, we propose a generalization of the benchmark that accounts for this uncertainty and, as a result, provides a sharper upper bound on the performance. The benchmark assesses how probable the occurrence of each ordering is, given the complete information about each patient. The proposed approach can be applied to trials with an arbitrary number of endpoints with discrete or continuous distributions. We illustrate the utility of the benchmark using recently proposed dose-finding designs for Phase I combination trials with a binary toxicity endpoint and Phase I/II combination trials with binary toxicity and continuous efficacy endpoints.

Keywords: Benchmark; Combination trial; Dose finding; Partial ordering; Power likelihood

1. Introduction

There has been growing interest in combination dose-finding trials of several agents administered simultaneously. Whilst coadministration can induce improved activity, designing such trials is more challenging compared to single-agent ones. Many single-agent dose-finding designs are based on the assumption that toxicity increases monotonically with the dose. However, in a combination study, there are combinations that cannot be ordered with respect to increasing toxicity. As a result, many novel model-based (see reviews

*To whom correspondence should be addressed.

by Riviere *and others*, 2015; Hirakawa *and others*, 2015, and references therein) and curve-free methods (e.g. Mozgunov and Jaki, 2019, 2020) were proposed to relax this assumption. Similarly to single-agent designs, the performance of these methods is conventionally assessed by simulation studies. These studies use combination-toxicity relationships, scenarios, which are chosen by the researchers themselves. This adds subjectivity to the assessment as the performance depends on the chosen scenario. The problem of selecting the scenarios is of relevance in dose-finding trials generally. To reduce the subjectivity, O'Quigley *and others* (2002) proposed an evaluation tool, the *nonparametric optimal benchmark*, that provides a scenario-specific evaluation of the performance in terms of the proportion of correct selections (PCS) of single-agent designs. When no strong prior information is used, the benchmark provides the highest PCS *a design* can achieve under the given simulation scenario. Occasionally, dose-finding methods can result in PCS that exceeds the PCS provided by the benchmark under certain scenarios. This is known as super-efficiency (Paoletti *and others*, 2004) and might be an indication of the design favoring particular doses (either due to the prior information or design specification) which the benchmark can reveal.

The benchmark was proposed under the assumption of monotonically increasing toxicity which typically holds in single-agent trials. Whilst the original benchmark can be also applied to dose-finding studies with unknown orderings (Mozgunov *and others*, 2020), the obtained upper bound for the PCS is expected to be less sharp compared to the setting when the monotonicity assumption holds.

To illustrate this, consider a hypothetical setting of a dual-agent combination study without early stopping. Assume that there are three increasing doses of agent A denoted by a_1, a_2, a_3 and three increasing doses of agent B denoted by b_1, b_2, b_3. Denote the combination of two doses a_k and b_l by d_{kl} where the first index refers to the kth dose of agent A, and the second index refers to the lth dose of agent B. There are nine drug combinations in the trial. Assume that the toxicity of combinations increases within each agent. This corresponds to at least one of the subscripts in d_{kl} increasing. However, some of the combinations cannot be ordered, for example, it is unknown whether d_{12} is more or less toxic than d_{31} as the dose of A is increased while the dose of B is decreased. Due to this uncertainty, there are 42 *complete* orderings of these combinations (see Supplementary materials available at *Biostatistics* online) that satisfy the monotonicity assumption within each agent. We will call the orderings satisfying this assumption the *feasible* orderings. The term "complete" refers to the feasible orderings of *all* nine combinations with respect to increased toxicity. The term "partial ordering" will refer to an ordered subset of combinations that could be arranged in increasing toxicity order (Wages *and others*, 2011a).

Consider a binary toxicity endpoint—occurrence of dose-limiting toxicity (DLT). The toxicity of d_{kl} is characterized by toxicity probability $p_{kl}, k, l = 1, 2, 3$. Suppose that the objective is to find the combination with the toxicity probability closest to 30% and assume a sample size of $n = 36$ patients. We would like to evaluate a design under the two scenarios given in Table 1.

The distance between the probabilities closest to the target is nearly the same under both scenarios, although the locations of the target combinations are different. Under Scenario 1, the target combinations are d_{12} and d_{21}. Following the monotonicity assumption, one of these combinations must be in the *second* position in any feasible complete ordering. Under Scenario 2, there are more possibilities of the location of the target combinations. Combination d_{13} can be in the *third, fourth, fifth*, *sixth*, or *seventh* positions of the complete orderings, while d_{22} can be in the *fourth, fifth,* and *sixth* positions (Table 6 in Supplementary material available at *Biostatistics* online). Therefore, one can expect that it is more challenging to find the target combinations under Scenario 2. However, the original benchmark (implemented by Wages and Varhegyi, 2017) disregards the uncertainty in the ordering and treats these scenarios similarly providing nearly the same PCS (Table 1).

In this article, we propose an extension of the benchmark for studies with unknown ordering. The novel benchmark accounts for both the uncertainty in the target combination locations within each feasible ordering and distribution of these orderings. We show that, compared to the original benchmark, the

Table 1. *Toxicity probabilities at each combination and corresponding proportions (in %) of each combination selection by the original benchmark based on* 10^4 *replications under two combination-toxicity scenarios with nine drug combinations. The target toxicity level and the selection of the target combinations are in bold.*

Toxicity probability									
		Drug B					Drug B		
Scenario 1		b_1	b_2	b_3	Scenario 2		b_1	b_2	b_3
	a_1	0.15	**0.30**	0.45		a_1	0.05	0.15	**0.30**
Drug A	a_2	**0.30**	0.45	0.55	Drug A	a_2	0.15	**0.30**	0.45
	a_3	0.55	0.60	0.65		a_3	0.45	0.55	0.60
Selection proportions									
		Drug B					Drug B		
Scenario 1		b_1	b_2	b_3	Scenario 2		b_1	b_2	b_3
	a_1	12.0	**36.5**	7.3		a_1	0.0	5.9	**36.2**
Drug A	a_2	**36.3**	7.3	0.2	Drug A	a_2	6.0	**36.7**	7.2
	a_3	0.3	0.0	0.0		a_3	7.4	0.5	0.0

proposal can provide a sharper bound on the performance of dose-finding designs relaxing monotonicity assumption while capturing the whole distribution of selections. In contrast to the recent benchmark proposal for dual-agent combination dose-finding trials by Guo and Liu (2018), the novel approach uses the original concept of complete information by O'Quigley *and others* (2002), which assumes that outcomes of each patient can be observed at all combinations. The benchmark, therefore, uses all available information about each patient, while accounting for the fact that combinations that cannot be ordered carry limited information about each other.

In line with extensions of the original benchmark to categorical and continuous endpoints (Cheung, 2014; Mozgunov *and others*, 2020), the proposal allows for an arbitrary number of endpoints having either discrete or continuous distributions. We demonstrate how the novel benchmark can be applied to a Phase I/II dual-agent combination study evaluating a binary toxicity endpoint and a Phase I/II combination study with binary toxicity and continuous efficacy endpoints.

The rest of the manuscript proceeds as follows. We review the benchmark by O'Quigley *and others* (2002) in Section 2. The construction of the benchmark for partial ordering in the combination setting with a single binary endpoint is given in Section 3 and extended to trials with multiple endpoints in Section 4. Section 5 demonstrates applications of the proposed benchmark before we conclude with a discussion.

2. The benchmark for single-agent studies with binary endpoint

Consider a Phase I clinical trial with a binary toxicity outcome, DLT or no DLT, n patients and M increasing doses of a drug, $c_1, \dots, c_M$. Let $Y_j^{(i)}$ be a Bernoulli random variable taking value $y_j^{(i)} = 0$ if patient i has experienced no DLT at dose c_j and $y_j^{(i)} = 1$ otherwise. This distribution of $Y_j^{(i)}$ is characterized by probability p_j such that $p_j = \mathbb{P}\left(Y_j^{(i)} = 1\right)$ for $j = 1, \dots, m$ and any i. The goal of the trial is to find the maximum tolerated dose (MTD) defined as the dose having the probability of toxicity closest to the target level, γ, typically between 20% and 35%.

The benchmark uses the concept of *complete information*. For a given patient, the complete information consists of the vector of outcomes (DLT or no DLT) at all doses (in contrast to an actual trial, in which patients can only be assigned to one) assuming that $p_1, \dots, p_m$ are known. Formally, the information about the DLT of patient i at each dose is summarized in a single value, $u^{(i)} \in (0,1)$, which is drawn from a uniform distribution, $\mathcal{U}(0,1)$, and is known as a toxicity profile of patient i. The variable $u^{(i)}$ is transformed to response $y_j^{(i)} = 0$ for doses with $p_j < u^{(i)}$ and to $y_j^{(i)} = 1$ otherwise. The procedure is repeated for n patients, which results in the vector of responses for each dose level $\boldsymbol{y}_j = (y_j^{(1)}, \dots, y_j^{(n)})$, $j = 1, \dots, M$. Note that the procedure is not sequential—responses for previous patients are not required to compute the complete information for the next ones. Therefore, there is no assignment criterion used by the benchmark. Let $R(\boldsymbol{y}_j, \gamma)$ be a summary statistic for the dose c_j, upon which the decision about the MTD selection is based. For example, in many Phase I trials with binary outcomes, $R(\boldsymbol{y}_j, \gamma) = \left| \frac{\sum_{i=1}^{n} y_j^{(i)}}{n} - \gamma \right|$ is a conventional choice. Therefore, c_j, for which $R(\boldsymbol{y}_j, \gamma)$ is minimized among all $j = 1, \dots, M$, is declared as the MTD in a single trial. The procedure is repeated for Z simulated trials. For each dose, the proportion of simulated trials that choose this dose as the MTD is computed. This proportion is the benchmark's estimate for the upper bounds of the PCS. Importantly, the benchmark is an evaluation tool and is not obtainable in actual trials. It can however be used at the planning stage to evaluate the performance of a dose-finding design.

3. Benchmark for Phase I combination studies with binary endpoint

3.1. *Setting*

Using the notations above, consider a Phase I dual-agent trial with $a_1, \dots, a_K$ doses of drug A, $b_1, \dots, b_L$ doses of drug B, their combinations d_{kl}, and a binary toxicity outcome, DLT or no DLT. Similarly to the single-agent setting, let $Y_{kl}^{(i)}$ be a Bernoulli random variable taking value $y_{kl}^{(i)} = 0$ if patient i has experienced no DLT at combination d_{kl} and $y_{kl}^{(i)} = 1$ otherwise, $k = 1, \dots, K, l = 1, \dots, L$. The distributions of $Y_{kl}^{(i)}$ are characterized by probabilities $p_{kl} = \mathbb{P}\left(Y_{kl}^{(i)} = 1 | d_{kl}\right)$ that increase with the dose of each compound. The goal is to find the maximum tolerated combination (MTC), the combination corresponding to a risk of toxicity closest to the target value γ. We use the following example throughout this section to demonstrate the novel benchmark construction.

Example Consider the simplest dual-agent trial with 2 doses of drugs A and B with $a_1 < a_2$ and $b_1 < b_2$, and four combinations, $d_{11}, d_{12}, d_{21}, d_{22}$, and suppose the target toxicity is 20%. There are two *complete* orderings satisfying the monotonicity assumption within each agent

$$(a)\ d_{11} \rightarrow d_{21} \rightarrow d_{12} \rightarrow d_{22} \qquad \text{and} \qquad (b)\ d_{11} \rightarrow d_{12} \rightarrow d_{21} \rightarrow d_{22}. \tag{3.1}$$

Then, the *partial* orderings are $d_{11} \rightarrow d_{12} \rightarrow d_{22}$ and $d_{11} \rightarrow d_{21} \rightarrow d_{22}$.

To provide an upper bound for the PCS, the benchmark for unknown ordering proposed in this work answers two questions: (i) "What is the probability of finding the true MTC if the ordering is known?" and (ii) "What is the probability of an ordering being identified as a correct one?." The original benchmark answers the first question only, and hence, provides a less accurate PCS upper bound. The general construction of our proposal is outlined below.

- Assuming the true ordering is known, obtain the patients' responses at each combination;
- Fixing these responses but not using the information about the true ordering, compute the probability that these responses were obtained from a given ordering;

- Under the given ordering, find the combination selected and assign the corresponding probability of this ordering being identified as a correct one to this combination selection.

3.2. *Construction: generating responses*

To address the first question, we start by following the original benchmark. Assume that p_{kl} is known for all k, l. We will refer to these probabilities as the *true scenario*. In the simulation setting, this sequence is known.

As before, assume that the toxicity profile of patient i is summarized in a single value $u^{(i)} \sim \mathcal{U}(0, 1)$ meaning that patient i can tolerate combinations d_{kl} with $p_{kl} < u^{(i)}$ and would experience a DLT if given combinations $d_{k'l'}$ associated with $p_{k'l'} > u^{(i)}$. Then, the patient's response can be written as $y_{kl}^{(i)} = 1$ for $p_{kl} > u^{(i)}$ and as $y_{kl}^{(i)} = 0$, otherwise. Assume that there is a sample of n patients with tolerances $u^{(1)}, \ldots, u^{(n)}$ and denote the number of DLTs for these n patients at each combination by $x_{kl} = \sum_{i=1}^{n} y_{kl}^{(i)}$, $k = 1, \ldots, K, l = 1, \ldots, L$. Estimates of the probabilities of toxicity at d_{kl} can then be found as $\hat{p}_{kl} = \frac{x_{kl}}{n}$. Note that the patient outcomes are generated using the true scenario and, hence, a true ordering.

Example (Continued) Assume that the true probabilities of toxicity p_{kl} for $d_{11}, d_{12}, d_{21}, d_{22}$ are given by $p_{11} = .10, p_{12} = .30, p_{21} = .20, p_{22} = .40$.

$$\begin{bmatrix} 0.10 & 0.30 \\ 0.20 & 0.40 \end{bmatrix}$$

implying that the ordering (a) in Equation (3.1) is correct. Assume that $n = 10$ patients with toxicity profiles $u^{(1)} = 0.59, u^{(2)} = 0.01, u^{(3)} = 0.29, u^{(4)} = 0.28, u^{(5)} = 0.81, u^{(6)} = 0.26, u^{(7)} = 0.72, u^{(8)} = 0.31, u^{(9)} = 0.95, u^{(10)} = 0.11$ were generated. This corresponds to the following numbers of DLTs at each combination $x_{11} = 1, x_{12} = 5, x_{21} = 2, x_{22} = 6$.
We now fix the number of DLTs obtained and find how likely is that they were drawn from each of the feasible orderings.

3.3. *Construction: identifying the probability of each ordering*

Fixing the values of the true toxicity probabilities and the number of DLTs at each combination, consider now S complete feasible orderings for these values. We assume that the values of toxicity probabilities are known but we do not know which probability goes with which combination. Denote the probability of DLT given d_{kl} under ordering s by $q_{kl}^{(s)}$, and let $s^{\star}$ be a correct ordering. Consequently, $q_{kl}^{(s^{\star})} = p_{kl}$ for all k, l. Probabilities $q_{kl}^{(s)}$ are constructed as all possible permutations (with respect to the complete feasible orderings) of the true probabilities p_{kl}.

Example (Continued) There are two feasible orderings in the considered example, $S = 2$. Consequently, $q_{kl}^{(s)}, s = 1, 2$ are

$$q^{(1)} = \begin{bmatrix} 0.10 & \underline{0.30} \\ \underline{0.20} & 0.40 \end{bmatrix}, \quad q^{(2)} = \begin{bmatrix} 0.10 & \underline{0.20} \\ \underline{0.30} & 0.40 \end{bmatrix},$$

where the values corresponding to the uncertainty in the monotonic ordering are underlined.
The second question to be answered by the benchmark can be reformulated as "How likely it is that the sequence of $q_{kl}^{(s)}$ (also referred to as ordering s) is a correct one, given the observed responses x_{kl}?." Using

the data generated for all n patients, the proposed benchmark computes

$$\mathbb{P}\left(P_{kl} = q_{kl}^{(s)} | x_{kl}\right), \ s = 1, \ldots, S \tag{3.2}$$

for all d_{kl}. Note that the probability of toxicity, P_{kl}, is now considered as a random variable itself, which can take a discrete number of values which are defined by the true toxicity probabilities that are feasible at the position (a_k, b_l).

Using Bayes Theorem, the probability (3.2) is proportional to the likelihood of observing X_{kl} given the DLT probability $P_{kl} = q_{kl}^{(s)}$, which equals

$$\mathbb{P}\left(X_{kl} = x_{kl} | P_{kl} = q_{kl}^{(s)}\right) = \text{Bin}\left(x_{kl}, n, q_{kl}^{(s)}\right) = \binom{n}{x_{kl}} q_{kl}^{(s)x_{kl}} \left(1 - q_{kl}^{(s)}\right)^{n-x_{kl}},$$

where Bin $(\cdot)$ is the density function of the binomial random variable. Let t_{kl} be the number of values P_{kl} can take, and let $h_{kl}^{(s)} = \mathbb{P}\left(P_{kl} = q_{kl}^{(s)}\right)$ be a prior probability that the toxicity probability at d_{kl} under ordering s is $q_{kl}^{(s)}$ such that $\sum_{s=1}^{S} h_{kl}^{(s)} = 1$. If all feasible values corresponding to combination d_{kl} are a priori equally likely then $h_{kl}^{(s)} = \frac{1}{t_{kl}}$. Then, the posterior probability that the DLTs at d_{kl} were obtained from the probability $q_{kl}^{(s)}$ given DLTs x_{kl} is proportional to

$$\mathbb{P}\left(P_{kl} = q_{kl}^{(s)} | x_{kl}\right) \propto \binom{n}{x_{kl}} q_{kl}^{(s)x_{kl}} \left(1 - q_{kl}^{(s)}\right)^{n-x_{kl}} \times h_{kl}^{(s)}. \tag{3.3}$$

Using these posterior probabilities for each combination corresponding to some ordering s', we find the probability of this ordering to be identified as a correct one. We allow for different importances of the contributions of various combinations to the posterior probability of the responses to be obtained from ordering s'. Specifically, we assume that it is proportional to

$$\mathbb{P}\left(s = s' | x_{11}, \ldots, x_{KL}\right) \propto \prod_{k,l} \left[\binom{n}{x_{kl}} q_{kl}^{(s')^{x_{kl}}} \left(1 - q_{kl}^{(s')}\right)^{n-x_{kl}} \times h_{kl}^{(s')}\right]^{w_{kl}}, \tag{3.4}$$

where w_{kl} is a weighting parameter corresponding to combination d_{kl}. The RHS in (3.4) is the power likelihood with parameter $w_{kl} > 0$ used in Bayesian analysis to control the learning rate of Bayesian update (Holmes and Walker, 2017). Values $0 < w_{kl} < 1$ give less prominence to the data than the Bayesian model. In the context of the study with uncertainty in monotonic ordering, the weights w_{kl} represent different contributions the combinations provide about the probability of complete ordering. Intuitively, one learns about the combinations within the same partial ordering more than about combinations that cannot be ordered. Then, the probability of ordering s' can be written as

$$\mathbb{P}\left(s = s' | x_{11}, \ldots, x_{KL}\right) = \frac{\prod_{k,l} \left[\text{Bin}\left(x_{kl}, n, q_{kl}^{(s')}\right) \times h_{kl}^{(s')}\right]^{w_{kl}}}{\sum_{s=1}^{S} \prod_{k,l} \left[\text{Bin}\left(x_{kl}, n, q_{kl}^{(s)}\right) \times h_{kl}^{(s)}\right]^{w_{kl}}}. \tag{3.5}$$

Below, we consider the following form of the weight function

$$w_{kl} = \left(1 + \#\{\text{combinations that cannot be ordered wrt } d_{kl}\}\right)^{-1} \tag{3.6}$$

corresponding to a higher weight if one has less uncertainty about the toxicity probability of combination d_{kl} with respect to other combinations. Note that the form of the weight function above is an arbitrary choice and other forms of the weight function that resembles the idea of assigning less weight to combinations carrying less information can be used.

Example (Continued) Under 1, 5, 2, 6 DLTs observed at combinations $d_{11}, d_{12}, d_{21}, d_{22}$, the probabilities of observing these responses x_{kl} under $q^{(1)}$ and $q^{(2)}$ in $n = 10$ patients are

$$\text{Bin}\left(x_{11} = 1, P_{11} = q_{11}^{(1)} = q_{11}^{(2)} = 0.10\right) \approx 0.39,\ \ \text{Bin}\left(x_{22} = 6, P_{22} = q_{22}^{(1)} = q_{22}^{(2)} = 0.40\right) \approx 0.11,$$

$$\text{Bin}\left(x_{12} = 5, P_{12} = q_{12}^{(1)} = 0.30\right) \approx 0.10,\ \text{Bin}\left(x_{12} = 5, P_{12} = q_{12}^{(2)} = 0.20\right) \approx 0.03,$$

$$\text{Bin}\left(x_{21} = 2, P_{21} = q_{21}^{(1)} = 0.20\right) \approx 0.30,\ \text{Bin}\left(x_{21} = 2, P_{21} = q_{21}^{(2)} = 0.30\right) \approx 0.23.$$

The weight values for each combinations (3.6) are equal to $w_{11} = w_{22} = 1$ and $w_{12} = w_{21} = \frac{1}{2}$. The weight w_{11} represents that the responses at d_{11} and d_{22} provides information for all four combinations, while the responses d_{12} and d_{21} do not provide information about each other. Assume that a priori any of the probability values specified in the true scenario at the anti-diagonal elements of the combination-toxicity matrix are equally likely, $h_{12}^{(1)} = h_{12}^{(2)} = h_{21}^{(1)} = h_{21}^{(2)} = \frac{1}{2}$. Then, the probabilities of each ordering can be found as $\mathbb{P}(s = 1|\cdot) = 0.69$ and $\mathbb{P}(s = 2|\cdot) = 0.31$.

Note that the posterior probabilities of the orderings in Equation (3.5) should not be used to select a single correct ordering to base further inference on. Instead, these probabilities will define each ordering's contribution to the selection probabilities obtained by the novel benchmark.

3.4. *Construction: computing the proportion of selections under the benchmark*

Once the probability of each ordering $s = 1, \dots, S$ is found, the benchmark proceeds as follows. Fix the ordering s' and find the estimates of the toxicity probabilities at combination d_{kl}, $\hat{q}_{kl}^{(s')}$ under this ordering using the toxicity profiles $u^{(1)}, \dots, u^{(n)}$ generated before and computed as $\hat{q}_{kl}^{(s')} = \frac{\sum_{i=1}^{n} \mathbb{I}\left(u^{(i)} < q_{kl}^{(s')}\right)}{n}$. Under ordering s', the MTC is selected using

$$R(\hat{q}_{kl}^{(s')}, \gamma) = \left|\hat{q}_{kl}^{(s')} - \gamma\right|. \tag{3.7}$$

The combination which minimizes criterion (3.7), is selected with the probability that the ordering s' is selected, $P(s = s'|\cdot)$. Using the same toxicity profiles, the procedure is repeated for all S orderings. The resulting estimates are the probability of selection of each combination.

Example (Continued) If ordering $s = 1$ is selected, then the estimates of the toxicity probabilities are $\hat{q}_{11}^{(1)} = 0.10, \hat{q}_{12}^{(1)} = 0.50, \hat{q}_{21}^{(1)} = 0.20, \hat{q}_{22}^{(1)} = 0.60$. Targeting the toxicity probability of 20%, the combination d_{21} is selected using criterion (3.7). As ordering $s = 1$ is selected with probability 0.69, then d_{12} is also selected with probability 0.69. Similarly, if the ordering $s = 2$ is selected, then the estimates are $\hat{q}_{11}^{(1)} = 0.10, \hat{q}_{12}^{(1)} = 0.20, \hat{q}_{21}^{(1)} = 0.50, \hat{q}_{22}^{(1)} = 0.60$, and d_{12} is selected with probability 0.31. Therefore, the probability of combinations $d_{11}, d_{12}, d_{21}, d_{22}$ being selected in this simulated trial with the observed DLTs are $(0.00, 0.31, 0.69, 0.00)$, respectively.

Finally, by generating Z simulated trials (each with n new toxicity profiles), the probability of each combination selection can be found in every simulated trial; the mean probability over Z simulations will be the benchmark's estimate of the combinations' selections. These proportions of each combinations' selections are used to obtain the proportion of *correct* selections (PCS) for a given definition of a "correct" combination set by the clinicians in a trial.

A step-by-step guide on how the benchmark for studies with unknown ordering and a binary endpoint can be constructed based on Z simulated trials is given in Algorithm 1.

Algorithm 1 Computing a partial ordering benchmark for a single binary outcome

```
1. Specify $S$ feasible complete orderings and toxicity probabilities $p_{kl}$ for all combinations, $k = 1 \ldots, K$, $l = 1 \ldots, L$.
2. Generate a sequence of patients' profiles $\{u_i\}_{i=1}^{n}$ from $\mathcal{U}(0,1)$, transform $u_i$ to $y_{kl}^{(i)} = 1$ if $p_{kl} > u_i$ and store $x_{kl} = \sum_{i=1}^{n} y_{kl}^{(i)}$, $k = 1 \ldots, K$, $l = 1 \ldots, L$, $\mathbb{X} = [x_{11}, \ldots, x_{KL}]$.
3. Compute the probability of ordering $s'$ being selected, $\mathbb{P}\left(s = s' | \mathbb{X}\right)$, $s' = 1, \ldots, S$.
4. For each ordering $s'$, $s' = 1, \ldots, S$, compute estimates $\hat{q}_{kl}^{(s')}$, the criterion $R(\hat{q}_{kl}^{(s')}, \gamma)$, and find the target combination $d_{k^{\star}l^{\star}}$ under ordering $s'$ and set $Q_{k^{\star}l^{\star}}(z) = \mathbb{P}\left(s = s' | \mathbb{X}\right)$.
5. Repeat steps 2–4 for $z = 1, \ldots, Z$ simulated trials.
6. Use $\hat{Q}_{kl} = \sum_{z=1}^{Z} Q_{kl}(z)/Z$ as the selection proportion of $d_{kl}$, $k = 1 \ldots, K$, $l = 1 \ldots, L$.
```

An application of the proposed benchmark to evaluate a dose-finding design for a Phase I dual-agent combination study is provided in Section 5.1.

4. Benchmark for combination studies with multiple endpoints

We now extend the proposed benchmark to accommodate a growing number of combination studies evaluating more than a single toxicity endpoint. For example, there are several novel designs for Phase I/II combination studies evaluating binary toxicity and binary or continuous efficacy simultaneously (Hirakawa, 2012; Wages *and others*, 2014; Yuan *and others*, 2016). For this, we build on the benchmark for continuous endpoints (Mozgunov *and others*, 2020).

Consider a Phase I/II trial with toxicity outcome $T_{kl}^{(i)}$ and efficacy outcome $E_{kl}^{(i)}$ with Cumulative Density Functions (CDFs) $F_{t,kl}$ and $F_{e,kl}$, respectively, at d_{kl} for patient i. Assume that $F_{t,kl}$ and $F_{e,kl}$ are parametrized by $\theta_{t,kl}$ and $\theta_{e,kl}$, respectively, and $f_{t,kl}(\cdot), f_{e,kl}(\cdot)$ are the corresponding density functions.

For patient i, the toxicity profile is given by $u_t^{(i)} \in (0,1)$ and the efficacy profile is given by $u_e^{(i)} \in (0,1)$. Then, following Mozgunov *and others* (2020), the toxicity and efficacy responses, $t_{kl}^{(i)}$ and $e_{kl}^{(i)}$, patient i would have at combination d_{kl} can be found as $t_{kl}^{(i)} = F_{t,kl}^{-1}\left(u_t^{(i)}\right)$, and $e_{kl}^{(i)} = F_{e,kl}^{-1}\left(u_e^{(i)}\right)$. Repeating the procedure for n patients, one can obtain the vectors $\boldsymbol{t}_{kl} = \left(t_{kl}^{(1)}, \ldots, t_{kl}^{(n)}\right)$, $\boldsymbol{e}_{kl} = \left(e_{kl}^{(1)}, \ldots, e_{kl}^{(n)}\right)$ for each d_{kl}.

Fixing the *values* of the toxicity and efficacy parameters, $\theta_{t,kl}, \theta_{e,kl}$, and the toxicity and efficacy responses $\boldsymbol{t}_{kl}, \boldsymbol{e}_{kl}$, consider now, S_t orderings of the values of $\theta_{t,kl}$, and S_e orderings of the values of $\theta_{e,kl}$. We assume that the values of parameters $\theta_{t,kl}, \theta_{e,kl}$ are known, but similar to the setting above, we do not know which parameters go with which combination. For example, in the setting with binary toxicity and efficacy responses, these parameters are probabilities of toxicity and efficacy, respectively. Denote the toxicity parameter associated with combination d_{kl} under ordering s_t by $\lambda_{t,kl}^{(s_t)}$, the efficacy parameter associated with combination d_{kl} under ordering s_e by $\lambda^{(s_e)}_{e,kl}$, and let $s_t^\star, s_e^\star$ be the true orderings (true scenario). Consequently, $\lambda^{(s_t^\star)}_{t,kl} = \theta_{t,kl}$ and $\lambda^{(s_e^\star)}_{e,kl} = \theta_{e,kl}$ for all k, l. As before, parameters $\lambda^{(s_t)}_{t,kl}, \lambda^{(s_e)}_{e,kl}$ are constructed as all possible permutations of the true parameter values $\theta_{t,kl}, \theta_{e,kl}$ with respect to complete feasible orderings, respectively.

Again, in the setting of the benchmark, one would like to answer the question "What is the probability of identifying correct orderings $s_t^\star$ and $s_e^\star$ among all feasible orderings given the responses $\boldsymbol{t}_{kl}, \boldsymbol{e}_{kl}$, $k = 1, \dots, K$, and $l = 1, \dots, L$?." The probability of ordering $s_t = s_t'$ being identified as a correct one is

$$\mathbb{P}\left(s_t = s_t' | \boldsymbol{t}_{11}, \dots, \boldsymbol{t}_{KL}\right) = \frac{\prod_{k,l}\left[\mathcal{L}\left(\boldsymbol{t}_{kl}, \lambda^{(s_t')}_{t,kl}\right) \times h^{(s_t')}_{t,kl}\right]^{w_{kl}}}{\sum_{s_t=1}^{S_t}\prod_{k,l}\left[\mathcal{L}\left(\boldsymbol{t}_{kl}, \lambda^{(s_t)}_{t,kl}\right) \times h^{(s_t)}_{t,kl}\right]^{w_{kl}}}, \tag{4.8}$$

where $\mathcal{L}$ is the likelihood function $\mathcal{L}\left(\boldsymbol{t}_{kl}, \lambda^{(s_t')}_{t,kl}\right) = \prod_{i=1}^{n} f\left(t^{(i)}_{kl}, \lambda^{(s_t')}_{t,kl}\right)$ and $h^{(s_t')}_{t,kl}$ is the prior probability that $\theta_{t,kl}$ equals $\lambda^{(s_t')}_{t,kl}$ under s'. Similarly, one can find $\mathbb{P}\left(s_e = s_e' | \boldsymbol{e}_{11}, \dots, \boldsymbol{e}_{KL}\right)$. Then, the probability of identifying orderings s_t' and s_e' simultaneously, is

$$\mathbb{P}\left(s_t = s_t', s_e = s_e' | \cdot\right) = \frac{\mathbb{P}\left(s_t = s_t' | \cdot\right) \times \mathbb{P}\left(s_e = s_e' | \cdot\right)}{\sum_{u,v}^{S_t,S_e} \mathbb{P}\left(s_t = u | \cdot\right) \times \mathbb{P}\left(s_e = v | \cdot\right)}.$$

Note that the weights w_{kl} have the same interpretation as above, and the function in the form given in Equation (3.6) is studied further.

Under each combination of orderings s_t and s_e, using previously generated responses $\boldsymbol{t}_{kl}, \boldsymbol{e}_{kl}$, one can find the target combination (TC) that optimizes some decision criterion $R(\cdot)$. Then, this combination is selected with probability $\mathbb{P}(s_t, s_e|\cdot)$. The procedure repeats for Z simulated trials. Algorithm 2 provides step-by-step guidance on how the benchmark for studies with partial ordering and Q endpoints with discrete or continuous distributions can be constructed.

Algorithm 2 Computing a partial ordering benchmark for studies with several endpoints

```
1. Specify CDFs $F_{q,kl}$ for $q = 1, \dots, Q$ endpoints and all combinations $k = 1 \dots, K$, $l = 1 \dots, L$.
   Specify $S_1, \dots, S_Q$ orderings for each endpoints, and criterion $R(\cdot)$.
2. Generate profiles $u^{(i)}_q$ for all patients $i = 1, \dots, n$ and all endpoints $q = 1, \dots, Q$.
3. Apply the quantile transformation $y^{(i)}_{q,kl} = F^{-1}_{q,kl}\left(u^{(i)}_q\right)$ for $i = 1, \dots, n$, $q = 1, \dots, Q$, $k = 1, \dots, K$
   and $l = 1, \dots, L$, and store $\boldsymbol{y}_{q,kl}$.
4. Compute the probability (4.8) of ordering $s_q = 1, \dots, S_q$ being a correct one, $q = 1, \dots, Q$.
5. For each combination of orderings $(s_1', \dots, s_Q')$, compute the values of the criterion $T(\cdot)$, find the
   target combination $d_{k^\star l^\star}$ and set $Q_{k^\star l^\star}(z) = \mathbb{P}\left(s_1 = s_1', \dots, s_Q = s_Q' | \cdot\right)$.
6. Repeat steps 2–5 for $z = 1, \dots, Z$ simulated trials.
6. Use $\hat{Q}_{kl} = \sum_{z=1}^{Z} Q_{kl}(z)/Z$ as the selection proportion of $d_{kl}$ for $k = 1 \dots, K$, $l = 1 \dots, L$.
```

Note that the construction of the benchmark above concerns a general case of an arbitrary (and possibly different) number of orderings of toxicities and efficacies. However, there are cases in which it might be reasonable to assume that the order of toxicities is the same as the order of efficacies, $s_t = s_e = s$. Then, the construction of the probabilities of orderings for a pair of endpoints reduces to the computation of the probability of orderings for a single endpoint but using both toxicity and efficacy data. Specifically, in case of the toxicity and efficacy orderings being the same, probability (4.8) can be found as

$$\mathbb{P}\left(s=s'|\cdot\right)=\frac{\prod_{k,l}\left[\mathcal{L}\left(\boldsymbol{t}_{kl},\lambda_{t,kl}^{(s')}\right)\times\mathcal{L}\left(\boldsymbol{e}_{kl},\lambda_{e,kl}^{(s')}\right)\times h_{kl}^{(s')}\right]^{w_{kl}}}{\sum_{s=1}^{S}\prod_{k,l}\left[\mathcal{L}\left(\boldsymbol{t}_{kl},\lambda_{t,kl}^{(s)}\right)\times\mathcal{L}\left(\boldsymbol{e}_{kl},\lambda_{e,kl}^{(s)}\right)\times h_{kl}^{(s)}\right]^{w_{kl}}},$$

where $h_{t,kl}^{(s_t)}=h_{e,kl}^{(s_e)}=h_{kl}^{(s)}$. Applications of the proposed benchmark to evaluate a Phase I/II dual-agent design for binary toxicity and continuous efficacy when toxicity and efficacy orderings can differ is provided in Section 5.2, and an evaluation in the setting of binary toxicity and efficacy endpoints with coinciding orderings is given in Supplementary material available at *Biostatistics* online.

5. Examples

Below, we provide two examples of how the novel benchmark can be used at the planning stage of a trial to provide a more accurate evaluation of a design to be used in the study. Specifically, we consider a Phase I combination clinical trial with a binary toxicity endpoint, and a Phase I/II clinical trial with binary toxicity and continuous efficacy endpoints.

5.1. *Evaluation of dose-finding designs for combination studies with binary toxicity*

The original benchmark for single-agent trials was found to provide an accurate upper bound for the model-based design, continual reassessment method (O'Quigley *and others*, 2002, CRM). Therefore, it is of interest to evaluate how the extension of CRM relaxing the monotonicity assumption proposed by Wages *and others* (2011a), Partial Ordering CRM (POCRM), performs compared to the novel benchmark for partial ordering. Additionally, we also evaluate the Bayesian I2D design by Wang and Ivanova (2005).

5.1.1. *Setting* Consider a dual-agent combination study with three doses of drug A and five doses of drug B (resulting in fifteen combinations), $n = 60$ patients, and a binary toxicity endpoint. The goal of the trial is to identify the MTC corresponding to the target probability of toxicity $\gamma = 0.30$. We consider ten combination-toxicity scenarios (Table 2) considered by Riviere *and others* (2015) in their review of dose-finding designs for combination studies

On top of the true probabilities of toxicity, one needs to specify the feasible toxicity orderings to apply the proposed benchmark. The total number of orderings satisfying the monotonicity assumption within each agent is $S = 6006$ (see Supplementary material available at *Biostatistics* online for procedures computing the orderings), and we assume that all orderings are equally likely prior to the trial. Finally, in line with the objective function of the dose-finding designs under evaluation, we consider absolute distance decision criterion (3.7) for the dose selection.

We evaluate the maximum likelihood (two-stage) version of the POCRM design proposed by Wages *and others* (2011b) and the I2D design by Wang and Ivanova (2005). The core idea of the POCRM is to run several CRM models under different orderings and allocate patients sequentially based on the most likely ordering. The maximum likelihood POCRM requires a sequence of initial patients' allocations to be used until at least one DLT and one non-DLT have been observed. After this, the combination selection will be governed by the POCRM. The initial escalation phase as proposed by Wages (2015) is considered. Furthermore, the POCRM requires the specification of a set of orderings that will be tried by the design. We consider six orderings as proposed by Wages and Conaway (2013); Wages (2015) that were found to lead to good operational characteristics. The two-stage design is implemented in R-package `pocrm` (Wages and Varhegyi, 2013). We also evaluate the I2D design as specified by Riviere *and others* (2015).

Table 2. *Ten considered combination-toxicity scenarios. The MTC is in bold.*

Drug A	Drug B					Drug B				
	b_1	b_2	b_3	b_4	b_5	b_1	b_2	b_3	b_4	b_5
			Scenario 1					Scenario 2		
a_1	0.05	0.10	0.15	**0.30**	0.45	0.15	**0.30**	0.45	0.50	0.60
a_2	0.10	0.15	**0.30**	0.45	0.55	**0.30**	0.45	0.50	0.60	0.75
a_3	0.15	**0.30**	0.45	0.50	0.60	0.45	0.55	0.60	0.60	0.80
			Scenario 3					Scenario 4		
a_1	0.02	0.07	0.10	0.15	**0.30**	**0.30**	0.45	0.60	0.70	0.80
a_2	0.07	0.10	0.15	**0.30**	0.45	0.45	0.55	0.65	0.75	0.85
a_3	0.10	0.15	**0.30**	0.45	0.55	0.50	0.60	0.70	0.80	0.90
			Scenario 5					Scenario 6		
a_1	0.01	0.02	0.08	0.10	0.11	0.05	0.08	0.10	0.13	0.15
a_2	0.03	0.05	0.10	0.13	0.15	0.09	0.12	0.15	**0.30**	0.45
a_3	0.07	0.09	0.12	0.15	**0.30**	0.15	**0.30**	0.45	0.50	0.60
			Scenario 7					Scenario 8		
a_1	0.07	0.10	0.12	0.15	**0.30**	0.02	0.10	0.15	0.50	0.60
a_2	0.15	**0.30**	0.45	0.52	0.60	0.05	0.12	**0.30**	0.55	0.70
a_3	**0.30**	0.50	0.60	0.65	0.75	0.08	0.15	0.45	0.60	0.80
			Scenario 9					Scenario 10		
a_1	0.005	0.01	0.02	0.04	0.07	0.05	0.10	0.15	**0.30**	0.45
a_2	0.02	0.05	0.08	0.12	0.15	0.45	0.50	0.60	0.65	0.70
a_3	0.15	**0.30**	0.45	0.55	0.65	0.70	0.75	0.80	0.85	0.90

Table 3. *Comparison of POCRM and I2D against the benchmark for partial ordering, the original benchmark, and the GL benchmark.*

Scenario	1	2	3	4	5	6	7	8	9	10
POCRM	72.8	69.2	69.7	81.0	69.6	59.4	50.0	54.6	51.8	54.1
I2D	68.0	73.7	66.9	89.7	83.7	37.2	41.9	50.4	5.1	13.0
Benchmark	84.1	84.0	84.1	91.1	92.3	84.3	84.2	83.1	83.2	83.2
PO-Benchmark	73.8	78.2	75.9	91.1	92.3	65.5	66.3	57.7	56.0	54.4
GL	73.3	75.0	75.1	84.6	94.6	77.7	89.6	83.8	82.2	76.3

We also include the benchmark proposed by Guo and Liu (2018) for trials with a single binary endpoint. We refer to this benchmark as "GL." It is based on the critical information introduced by the authors that is argued to offer a middle ground between the complete information and data available in actual trial. The GL as specified in the original work is used in the evaluation.

5.1.2. *Numerical results* Table 3 shows the PCS for I2D, POCRM, the original benchmark (Benchmark), the novel benchmark for partial ordering (PO-Benchmark), and the benchmark by Guo and Liu (2018) (GL). The results of I2D are extracted from Table 2 in the original review, and the results of POCRM are extracted from Table 1 in the comment by Wages (2015).

Comparing the proposed PO-Benchmark and GL approach under scenarios in which the target combination is located on a diagonal (Scenarios 1–3), they provide similar PCS. Under Scenario 4, PO-Benchmark results in 7% higher PCS than the GL approach and performs similar to the original benchmark as there is little uncertainty in the monotonicity associated with the target combination. Under Scenario 5, a similar behavior for the PO-Benchmark is found but the GL approach now corresponds to higher PCS than both the PO-Benchmark and original benchmark that employs the monotonicity assumption.

Under Scenarios 6 and 7, while PO-Benchmark implies that it is more challenging to locate the target combination than, for example, in Scenarios 1–3, the PCS of 65–66% against 73–78%, the GL approach suggests otherwise: PCS of 77–89% against 73–75%. This is counter-intuitive due to fewer target combinations (Scenarios 6) and a more complex interaction mechanism of the compounds (Scenario 7). Under Scenario 7, the GL approach again results in higher PCS than the original benchmark.

Finally, differences between the PO-Benchmark and GL approach can be seen under Scenarios 8–10 with a single target combination. While the PO-Benchmark suggests that these are the most challenging scenarios to find the target combination, the GL suggests that it is, in fact, easier than, for example, Scenario 1 with three target combinations located on the same diagonal. Once more the GL approach, under Scenarios 8 and 9, results in slightly higher or nearly the same PCS as the original benchmark. Consequently, the GL approach does not provide as sharp an upper bound under a number of scenarios, and the PO-Benchmark might provide a more accurate guidance on how challenging each scenarios is under the uncertainty in the ordering.

Under all considered scenarios, the two dose-finding designs result in lower PCS compared to both the original benchmark and the benchmark for partial ordering. Importantly, the original benchmark considered all scenarios with the MTC being not the first or last combination (Scenarios 4 and 5, respectively) as equally difficult with nearly 84% PCS. However, this does not reflect the true challenges that these scenarios impose as they have a different number of the MTCs located at different places on the combination grid. The benchmark for partial ordering recognizes these differences and provides a sharper upper bound for the PCS. Specifically, under Scenario 1, the POCRM and I2D result in 72.8% and 68.0% PCS, respectively. This corresponds to the ratios (with respect to the PO-Benchmark) of 72.8/73.8 = 98.6% and 68.0/73.8 = 92.1%, respectively. At the same time, under Scenario 6, both POCRM and I2D result in a much lower PCS 59.4% and 37.2%. Looking at these values alone (or using the original benchmark) can result in the conclusion that these designs perform poorer in this case compared to Scenario 1. However, the ratio of PCS with respect to the PO-Benchmark is 59.4/65.5 = 90.7% for POCRM and 37.2/65.5 = 56.8% for I2D. Therefore, POCRM still corresponds to a relatively accurate performance, while the I2D design does have potential problems under these scenarios but not as severe as one might conclude by considering the PCS alone.

Regarding the overall performance, POCRM corresponds to a ratio of PCS (compared to PO-Benchmark) of at least 88% under 8 out of 10 scenarios. Under the other two scenarios, Scenario 5 and Scenario 7, the ratio is around 75% which is still relatively high. While further calibration of the model parameters can result in less diverse values of ratios, this is an indication that the POCRM design under the proposed specification is properly calibrated and results in accurate selections under many different scenarios. The I2D design results in the ratio above 87% in 6 out of 10 scenarios. For scenarios 6–7 and 9–10, the I2D design corresponds to ratios of 56.8%, 63.8%, 8.9%, and 23.9%, respectively. This implies that further tuning of the I2D design is required before the design can be applied to an actual clinical trial.

Overall, the novel benchmark has provided noticeable added value over the original benchmark. It leads to the conclusion that the POCRM design results in a good performance in many different scenarios while I2D requires further attention. We refer the reader to Supplementary material available at *Biostatistics* online for another example of the POCRM evaluation with three doses of each drug.

Table 4. *True values of* (p_{kl}, μ_{kl}) *for each combination of two agents, A and B. The TC is in bold.*

		b_1	b_2	b_3	b_4
Scenario 1	a_1	$(0.01, 0.5)$	$(0.10, 0.0)$	$(0.40, -1.5)$	$(0.50, -2.5)$
	a_2	$(0.05, -1.5)$	$\mathbf{(0.15,–2.0)}$	$(0.45, -3.5)$	$(0.55, -4.5)$
Scenario 2	a_1	$(0.01, 0.0)$	$(0.05, -0.5)$	$\mathbf{(0.15,–3.5)}$	$(0.45, -5.5)$
	a_2	$(0.45, -1.0)$	$(0.50, -1.5)$	$(0.60, -4.5)$	$(0.90, -6.5)$
Scenario 3	a_1	$(0.01, 0.0)$	$\mathbf{(0.15,–2.0)}$	$(0.40, -2.0)$	$(0.50, -2.0)$
	a_2	$(0.05, 0.0)$	$(0.20, -2.0)$	$(0.45, -2.0)$	$(0.55, -2.0)$

5.2. *Evaluation of Phase I/II Design for Binary Toxicity and Continuous Efficacy*

Below, we evaluate the Phase I/II design for combination trials with binary toxicity and continuous efficacy endpoints proposed by Hirakawa (2012). We refer the reader to Supplementary material available at *Biostatistics* online for the evaluation of Phase I/II design for binary endpoints.

Hirakawa (2012) considered Phase I/II cervical carcinoma trial, in which the squamous cell carcinoma antigen (SCCA) was used as a marker of effect on a continuous scale. Among others, a combination setting with two compounds (A and B) was considered. There were two doses of drug A and four doses of drug B. The efficacy outcome was "change in log-transformed SCCA levels from baseline and end of treatment." Consequently, the lower values of the efficacy outcomes correspond to better performance. It was assumed that the efficacy endpoints has a normal distribution $\mathcal{N}(\mu_{kl}, 1)$ at combination d_{kl}. The toxicity was evaluated as a binary endpoint characterized by the probability p_{kl} at combination d_{kl}. The goal of the combination trial was to find the TC defined as the safe and efficacious combination having the highest efficacy. The upper toxicity bound is $\phi = 0.3$, and the upper efficacy bound is $\psi = 0$ corresponding to no changes in SCCA levels. To find the target combination, Hirakawa (2012) proposed a model-based approach with a four-parameter combination-toxicity model and an Emax-type seven-parameter combination-efficacy model. The combination selection was based on a Mahalanobis-type distance representing the trade-off between toxicity and efficacy and computed using the posterior distribution of the parameters. We will adopt the notation "Emax" for this design.

The proposed benchmark requires all feasible orderings to be specified. Assuming that the toxicity and efficacy increases with the dose, there are 14 feasible orderings (see Supplementary material available at *Biostatistics* online). Then, the benchmark as in Algorithm 2 with weight function (3.6) and with the binomial likelihood for the toxicity endpoints and the normal likelihood for the efficacy endpoint, assuming that all the orderings are equally likely a priori, can be applied. The following decision criterion is used by Hirakawa (2012)

$$R(\boldsymbol{y}_{1,kl}, \boldsymbol{y}_{2,kl}) = \frac{\sum_{i=1}^{n} y_{2,kl}^{(i)}}{n} \times \mathbb{I}\left(\int_0^{+\infty} g_{2,kl}(v|\boldsymbol{y}_{2,kl})\mathrm{d}v < \eta_2\right) \times \mathbb{I}\left(\int_{0.3}^{1} g_{1,kl}(v|\boldsymbol{y}_{1,kl})\mathrm{d}v < \eta_1\right). \tag{5.9}$$

Three scenarios considered in the original work are given in Table 4, and proportions of each combination selections by the Emax design and respective benchmarks are given in Table 5.

Under Scenario 1, the original benchmark selects d_{22} in almost all trials due to the known ordering of toxicities and efficacies. At the same time, the benchmark for partial ordering selects the target combination in 88% of trials with d_{21} having the second largest proportion of selections. It also selects d_{12} and d_{13} with small probabilities. This is in fact in line with the proportion of selections by the Emax design. The ratio

Table 5. *Comparison of the Emax design and the respective benchmark for partial ordering and the original benchmark. The selections of the TC is in bold.*

Design		b_1	b_2	b_3	b_4
Scenario 1					
Benchmark	a_1	0.0	0.0	0.0	0.0
	a_2	0.0	**99.9**	0.1	0.0
PO-Benchmark	a_1	0.0	2.2	1.6	0.0
	a_2	8.1	**88.1**	0.0	0.0
Emax	a_1	0.1	5.1	3.4	0.0
	a_2	14.8	**70.1**	4.7	0.0
Scenario 2					
Benchmark	a_1	0.0	0.0	**99.9**	0.1
	a_2	0.0	0.0	0.0	0.0
PO-Benchmark	a_1	0.0	0.0	**99.9**	0.1
	a_2	0.0	0.0	0.0	0.0
Emax	a_1	4.3	11.5	**78.6**	3.1
	a_2	0.0	0.1	0.3	0.0
Scenario 3					
Benchmark	a_1	0.0	**50.1**	1.2	0.0
	a_2	0.0	47.7	0.0	0.0
PO-Benchmark	a_1	0.0	**46.8**	2.3	0.0
	a_2	4.2	47.3	0.0	0.0
Emax	a_1	0.9	**44.9**	2.8	0.0
	a_2	4.3	45.3	1.8	0.0

of PCS with respect to the PO-Benchmark is nearly 80% against approximately 70% using the original benchmark. Under Scenario 2, both benchmarks lead to the same evaluation of the design resulting in the conclusion that the unknown ordering does not cause any additional obstacles for a design to select the target combination. The ratio of PCS is again nearly 80%. Under Scenario 3, the original benchmark recommends d_{12} and d_{22} in almost 99% of trials, and never selects d_{21} and d_{31} as the complete ordering is known. The PO-Benchmark, however, shows that the unknown ordering makes a correct selection more challenging, and selects corresponding suboptimal combinations in 6.5% of trials. This, again, is in line with the Emax design which selects the TC in 44.9% of trials (against 46.8% for the PO-Benchmark—the ratio of PCS is 95%) and combinations d_{21} and d_{13} in 7.1% of trials.

Overall, the evaluation of the Emax design using the novel benchmark provides the conclusion that the design has high accuracy in all three considered scenarios with the ratio of PCS being above 80%. At the same time, the original benchmark would reveal some problems with the design under Scenario 1, while the performance is as good as under Scenario 2.

6. Discussion

A novel benchmark for dose-finding studies with unknown ordering is proposed. The novel benchmark is a generalization of the original proposal by O'Quigley *and others* (2002) for the setting with unknown ordering. The distinguishing feature of the proposal is that it assesses the complexity of scenarios taking into account not only the uncertainty about the parameters but also the uncertainty about the ordering of these parameters. The proposed benchmark computes the proportions of each combination selection for a given scenario (that might have several combinations with either the same or close toxicity probabilities). It is found that the novel benchmark can provide a more accurate evaluation of dose-finding designs for combination studies than the analysis compared to the original benchmark. The novel approach is easy to implement and does not require any additional information other than that which is available in a simulation study. Finally, the proposed benchmark is computationally feasible even under a large number of orderings as obtaining the benchmark under each ordering has low computational costs.

The proposed benchmark does not select a correct ordering, but in line with the main objective of many Phase I trials, selects the MTC. Moreover, the probability of the ordering being identified as a correct one in itself is not necessarily a useful measure of a good procedure for the MTC selection objective as there may exist multiple orderings that are identical up to the point of the MTC. Either one of these orderings can result in recommending a correct MTC. Consequently, the probability of each ordering is used to compute the probability of the selection under this ordering rather than to select the single ordering and make the inference solely based on it.

Similarly to the original benchmark, the partial ordering benchmark is an evaluation tool that can be used to comprehensively assess the performance of a design that might be considered for a trial. Being a theoretical tool, the benchmark should be used at the planning stage of the trial. Importantly, for the fair and meaningful comparison, the benchmark should use the same criterion for the combination selection as the design under evaluation. The benchmark can also stimulate discussions about the sample size (Cheung, 2013). If in some scenarios, one observes a low PCS under the benchmark, this might indicate that the change in the sample size/number of doses should be explored. At the same time, low PCS should not be interpreted outside of the context as the benchmark accounts for the difficulty of the scenarios. The clinical plausibility of each scenario should be accounted for when interpreting the benchmark' results. An investigation of the link between the sample sizes and the benchmark performance is subject to future research.

Exploring the behavior of the designs under various assumptions on the correlation between these endpoints and interaction between the compounds might be of interest at the planning stage. The benchmark includes the correlation in its assessment through the algorithm to generate the complete information using the prespecified value of the correlation coefficient. Similarly, the interaction is accounted for implicitly via the simulation scenarios themselves by specifying the toxicity probabilities. In this sense, the proposed benchmark is universal as allows for the assessment of each of these aspects.

While our examples of the benchmark concerned the setting where each of the orderings is equally likely a priori, the benchmark construction allows for prior information about each ordering to be incorporated. As the number of complete orderings can be large, we propose to include this information through the prior information of each combination location in the complete ordering. For example, eliciting the information about the second combination in the complete ordering can be phrased as "What is the probability that the second-lowest dose is d_{12}?."

The original benchmark provides an upper bound for the proportion of correct selections as it employs the complete information about each patient. However, it is known that a particular method can provide a higher PCS than the original benchmark under a given scenario if the prior information used is strong enough (Paoletti *and others*, 2004). The same applies to the benchmark for the partial ordering. Additionally, the proposed benchmark depends on the choice of weight function, w_{kl}. Whilst we have found that the proposed weight function results in an accurate upper bound for a dose-finding method's performance in many scenarios, it is possible that the PCS of the evaluated method is greater than the benchmark due to the choice of weight function. Nevertheless, the benchmark still provides a basis for standardization of the PCS that cannot be achieved if analyzing PCS alone—if the ratio of PCSs (compared to the proposed benchmark) is noticeably higher under one scenario than under others, it implies that the design as specified favors the selection of the target combinations under this scenarios.

Finally, it is important to mention that while the proposed benchmark is a useful tool for assessing the performance of any given dose-finding method for combination studies, similar to the benchmark for single-agent studies, it does not capture all aspects of the evaluation. For instance, it does not provide information on the distribution of dose allocation or the average number of DLTs. Developments in these directions are of great value for a more comprehensive assessment of dose-finding designs.

7. Software

Software in the form of R code is available on GitHub (https://github.com/dose-finding/combo-benchmark).

Supplementary material

Supplementary material is available at is http://biostatistics.oxfordjournals.org.

Acknowledgments

The authors would like to thank Dr Helen Barnett for reading an earlier version of the manuscript, and Beibei Guo for providing the code implementing the GL approach.

Conflict of Interest: None declared.

Funding

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie (633567). The Institut National du Cancer (French NCI) grant SHS-2015 Optidose immuno project to X.P.; This report is independent research supported by the National Institute for Health Research (NIHR Advanced Fellowship, Dr Pavel Mozgunov, NIHR300576; and Prof. Jaki's Senior Research Fellowship, NIHR-SRF-2015-08-001). The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health and Social Care (DHCS). UK Medical Research Council (MC_UU_0002/14 to T.J.).

References

Cheung, Y. K. (2013). Sample size formulae for the Bayesian continual reassessment method. *Clinical Trials* **10**, 852–861.

Cheung, Y. K. (2014). Simple benchmark for complex dose finding studies. *Biometrics* **70**, 389–397.

Guo, B. and Liu, S. (2018). Optimal benchmark for evaluating drug-combination dose-finding clinical trials. *Statistics in Biosciences* **10**, 184–201.

Hirakawa, A. (2012). An adaptive dose-finding approach for correlated bivariate binary and continuous outcomes in phase I oncology trials. *Statistics in Medicine* **31**, 516–532.

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**, 3194–3213.

Holmes, C. C. and Walker, S. G. (2017). Assigning a value to a power likelihood in a general Bayesian model. *Biometrika* **104**, 497–503.

Mozgunov, P. and Jaki, T. (2019). An information theoretic phase I–II design for molecularly targeted agents that does not require an assumption of monotonicity. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **68**, 347–367.

MOZGUNOV, P. AND JAKI, T. (2020). An information theoretic approach for selecting arms in clinical trials. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **82**, 1223–1247.

MOZGUNOV, P., JAKI, T. AND PAOLETTI, X. (2020). A benchmark for dose finding studies with continuous outcomes. *Biostatistics* **21**, 189–201.

O'QUIGLEY, J., PAOLETTI, X. AND MACCARIO, J. (2002). Non-parametric optimal design in dose finding studies. *Biostatistics* **3**, 51–56.

PAOLETTI, X., O'QUIGLEY, J. AND MACCARIO, J. (2004). Design efficiency in dose finding studies. *Computational Statistics & Data Analysis* **45**, 197–214.

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**, 1–12.

WAGES, N. A., CONAWAY, M. R. AND O'QUIGLEY, J. (2011a). Continual reassessment method for partial ordering. *Biometrics* **67**, 1555–1563.

WAGES, N. A., O'QUIGLEY, J. AND CONAWAY, M. R. (2014). Phase I design for completely or partially ordered treatment schedules. *Statistics in Medicine* **33**, 569–579.

WAGES, N. A. (2015). Comments on competing designs for drug combination in phase I dose-finding clinical trials by MK. Riviere, F. Dubois, S. Zohar. *Statistics in Medicine* **34**, 18.

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**, 217–224.

WAGES, N. A, CONAWAY, M. R. AND O'QUIGLEY, J. (2011b). Dose-finding design for multi-drug combinations. *Clinical Trials* **8**, 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**, 211–218.

WAGES, N. A. AND VARHEGYI, N. (2017). A web application for evaluating phase I methods using a non-parametric optimal benchmark. *Clinical Trials* **14**, 553–557.

WANG, K. AND IVANOVA, A. (2005). Two-dimensional dose finding in discrete dose space. *Biometrics* **61**, 217–222.

YUAN, Y., NGUYEN, H. Q. AND THALL, P. F. (2016). *Bayesian designs for phase I–II clinical trials*. New York: Chapman and Hall/CRC.

[*Received April 16, 2020; revised September 25, 2020; accepted for publication November 9, 2020*]