Contemporary Clinical Trials 71 (2018) 162–172

Contents lists available at ScienceDirect

Contemporary Clinical Trials

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

Efficacy/toxicity dose-finding using hierarchical modeling for multiple populations

Kristen M. Cunanan*, Joseph S. Koopmeiners

Memorial Sloan Kettering Cancer Center, Department of Epidemiology and Biostatistics, 485 Lexington Avenue 2nd Floor, New York, NY 10017, United States

ARTICLE INFO

Keywords:
Phase I-II
Dose-finding
Multiple populations
Continual reassessment method

ABSTRACT

Traditionally, Phase I oncology trials evaluate the safety profile of a novel agent and identify a maximum tolerable dose based on toxicity alone. With the development of biologically targeted agents, investigators believe the efficacy of a novel agent may plateau or diminish before reaching the maximum tolerable dose while toxicity continues to increase. This motivates dose-finding based on the simultaneous evaluation of toxicity and efficacy. Previously, we investigated hierarchical modeling in the context of Phase I dose-escalation studies for multiple populations and found borrowing strength across populations improved operating characteristics. In this article, we discuss three hierarchical extensions to commonly used probability models for efficacy and toxicity in Phase I-II trials and adapt our previously proposed dose-finding algorithm for multiple populations to this setting. First, we consider both parametric and non-parametric bivariate models for binary outcomes and, in addition, we consider an under-parameterized model that combines toxicity and efficacy into a single trinary outcome. Our simulation results indicate hierarchical modeling increases the probability of correctly identifying the optimal dose and increases the average number of patients treated at the optimal dose, with the under-parameterized hierarchical model displaying desirable and robust operating characteristics.

1. Introduction

Phase I oncology trials are primarily dose-escalation studies to evaluate the safety of a novel treatment and identify the maximum tolerable dose (MTD), defined as the highest dose with probability of dose limiting toxicity (DLT) less than some pre-specified threshold. Typically, efficacy is not examined until Phase II. Historically, clinicians believed the probabilities of toxicity and efficacy increase monotonically with dose and, subsequently, the highest dose with acceptable toxicity was thought to have the best chance to succeed in future trials. However, for contemporary biologically targeted agents, investigators often believe a drug's potential efficacy may level off or diminish before reaching the MTD, while potential toxicity increases with dosage. This motivates dose-finding based on the simultaneous evaluation of toxicity and efficacy. Furthermore, given the limited sample sizes in Phase I oncology trials, incorporating efficacy into dose-finding may improve identifying the optimal dose used in subsequent trials. Gooley et al. [8] were among the first to propose a dose-finding design based on simultaneous evaluation of toxicity and efficacy. Their results suggest that the additional dose-efficacy curve adds complexity (i.e., model parameters) to the dose-finding algorithm which is a cost that should be considered when designing a Phase I-II trial. Consequently, Thall and Russell [20] proposed a design combining toxicity and efficacy into one variable, reducing the parameter space. Alternatively, Braun [3] extends the continual reassessment method to account for two competing outcomes, while Thall and Cook [17] take a similar approach but also define a trade-off contour to guide dose-finding. A number of extensions to this basic approach have been discussed over the last decade [9, 12–15, 18, 19, 21–23]. Researchers are often interested in evaluating a novel treatment in a number of patient populations, which may have different background standards-of-care. For example, researchers at University of Minnesota College of Veterinary Medicine Animal Cancer Care and Research Program are interested in completing a Phase I-II trial of a novel targeted toxin. The trial will enroll dogs in two cohorts: a cohort focused on hemangiosarcomoa, for which the drug has previously shown promising results [2], and a cohort for other solid tumors. In this case, the hemangiosarcoma cohort will not utilize information found in the solid tumor cohort, resulting in a potential loss of efficiency, while the solid tumor cohort will collapse across multiple tumor types with potentially heterogeneous dose-response relationships. An alternate approach would be to use hierarchical modeling (HM) to allow each population to have separate dose-response relationships, while borrowing strength across populations to gain efficiency. Previously, we investigated HM in the context of Phase I dose-

* Corresponding author.
E-mail address: kristenmay206@gmail.com (K.M. Cunanan).

https://doi.org/10.1016/j.cct.2018.06.012
Received 11 December 2017; Received in revised form 5 May 2018; Accepted 13 June 2018
Available online 21 June 2018

escalation studies. We proposed extensions to commonly used dose-toxicity models and proposed dose-finding guidelines that protect patient safety, while allowing the design to fully realize the potential of HM [5]. Our simulation results indicate incorporating HM into Phase I dose-finding increases the probability of correctly identifying the MTD and the average number of patients treated at the MTD, with little impact on the rate of DLTs. In this article, we propose a Bayesian adaptive Phase I-II dose-escalation design that uses HM to estimate population-specific biologically optimal doses (BODs), while sharing both dose-toxicity and -efficacy information across populations.

2. Models

In this section, we present hierarchical extensions of three joint probability models for efficacy and toxicity that have been proposed for use in Phase I-II dose-finding trials. In each case, we define a two-level Bayesian hierarchical model where the first level specifies the population-level parameters and the second level facilitates borrowing across populations. Existing joint probability models for Phase I-II clinical trials can be broadly classified into two groups: bivariate outcome models, where separate dose-response models are specified for efficacy and toxicity and the correlation between efficacy and toxicity is incorporated into the model using a copula model or some other approach [3, 17, 21], and trinomial models, where efficacy and toxicity are combined into a trinomial outcome and a dose-response relationship is specified for the trinomial outcome [20, 23]. We begin by discussing hierarchical extensions of two bivariate binary outcome models and then discuss a hierarchical extension of the trinomial model proposed by Zhang et al. [23].

2.1. Bivariate binary outcomes

We use the following notation throughout Section 2.1. First, let T_{ikj} be a binary indicator for the presence or absence of DLT in subject i treated at dose j in population k, which takes the value 1 with probability $\pi_{T,\,kj}$, and let E_{ikj} be a binary indicator for the probability of tumor response in subject i treated at dose j in population k, which takes the value 1 with probability $\pi_{E,\,kj}$. We will consider two approaches for specifying a bivariate outcome model. First, we consider a parametric approach, where parametric dose-response models are specified for efficacy and toxicity. Next, we consider a non-parametric model that imposes a monotonicity constraint on the dose-toxicity model but avoids a formal parametric model.

2.1.1. Parametric model

For our parametric model, we extend a simple one-parameter power model for toxicity and a more flexible, quadratic logistic regression model for efficacy. Our hierarchical model for toxicity is specified as:

$$pr(T_{ikj} = 1 \mid population = k, dose = j) = \pi_{T,kj} = p_j^{exp(\alpha_k)} \tag{1}$$

$$\alpha_k \mid \mu_\alpha, \sigma_\alpha^2 \sim N(\mu_\alpha, \sigma_\alpha^2)$$

$$\mu_\alpha \sim Normal(0, 2^2) \quad \text{and} \quad \sigma_\alpha \sim Uniform(0.39, 3)$$

for dose level $j = 1, \ldots, D$ and population $k = 1, \ldots, K$. The vector $(p_1, \ldots, p_D)$ is referred to as the skeleton and its components are monotonically increasing and take values between 0 and 1. For our simulation results presented in Section 4, we set the power model skeleton equal to $(0.05, 0.15, 0.25, 0.35, 0.45)$. Our hierarchical model for efficacy is specified as:

$$pr(E_{ikj} = 1 \mid population = k, dose = j) = \pi_{E,kj} = \beta_{0k} + \beta_{1k}(dose - 1) + \beta_{2k}(dose - 1)^2 \tag{2}$$

$$\beta_{lk} \mid \mu_{\beta l}, \sigma_{\beta l}^2 \sim Normal(\mu_{\beta l}, \sigma_{\beta l}^2)$$

$$\mu_{\beta l} \sim Normal(m_l, s_l^2) \quad \text{and} \quad \sigma_{\beta l} \sim Uniform(0.39, 3),$$

for $l = 0, 1, 2$, dose level $j = 1, \ldots, D$ and population $k = 1, \ldots, K$. We originally fixed the intercept equal to -3 to reduce the number of unknown parameters, as suggested by Goodman et al. [7]. This reflects a 5% probability of tumor response at dose level 1, but we found that this model did not provide enough flexibility when the true optimal dose resides in the higher dose levels. The unknown m_0, m_1, and m_2 are the shared mean hyper-parameters for the intercept, linear and quadratic terms and are set equal to -2, 0.1, and 0, respectively, with shared variance hyper-parameters set to ${s_0}^2 = 4$, ${s_1}^2 = 9$, and ${s_2}^2 = 4$. This corresponds to a conservative, monotonic prior efficacy-skeleton of 0.12, 0.13, 0.14, 0.15, 0.17 for dose levels 1, 2, 3, 4, 5, respectively. The ${\sigma_{\beta l}}^2$ are our hierarchical variance parameters that control the amount of borrowing across populations, with smaller values indicating more borrowing. We specify a uniform prior distribution on the standard deviation, rather than the log standard deviation, as in [5], since this prior is well-received for other hierarchical applications and we are interested in exploring its use further in our dose-finding setting. In our previous investigation, a uniform prior on the standard deviation with a lower bound of 0 produced poor convergence and identifiability, given the small sample sizes early in a trial. The lower bound of our uniform prior was set to 0.39, based on our simulation results, which suggested that a lower bound < 0.39 results in over-borrowing and poor trial operating characteristics in settings where the true optimal dose varies by population. The toxicity and efficacy outcomes in Phase I-II clinical trials are thought to be correlated and a number of approaches have been proposed for jointly modeling efficacy and toxicity in Phase I-II clinical trials [3, 17, 21]. Recently, Iasonos et al. [11] provided an extensive evaluation of the effect of dimensionality on trial operating characteristics in early phase dose-finding studies. They found that more parsimonious models typically result in improved operating characteristic, even when some aspects of the data generating process are misspecified. A number of other authors have come to similar conclusions with respect to estimating the correlation between efficacy and toxicity in Phase I-II clinical trials [4, 10, 15, 21]. Therefore, we will proceed assuming independence between the toxicity and efficacy outcome for our parametric model.

2.1.2. Non-parametric model

The second model we consider is a hierarchical extension of the non-parametric model proposed by Yin et al. [21]. They specify a dose-response relationship for toxicity and efficacy through the following transformations. For population $k = 1, \ldots, K$, the dose-response model for toxicity is specified as,

$$\phi_{k1} = logit(\pi_{T,k1}), \qquad \phi_{kj} = log\left(\frac{\pi_{T,kj}}{1 - \pi_{T,kj}} - \frac{\pi_{T,k(j-1)}}{1 - \pi_{T,k(j-1)}}\right)$$

for $j = 2, \ldots, D$, and for efficacy, let

$$\psi_{k1} = logit(\pi_{E,k1}), \qquad \psi_{kj} = log\left(\frac{\pi_{E,kj}}{1 - \pi_{E,kj}}\right) - log\left(\frac{\pi_{E,k(j-1)}}{1 - \pi_{E,k(j-1)}}\right)$$

for $j = 2, \ldots, D$. The primary difference between the two parameterizations is that the model for toxicity enforces a monotonicity constraint on the dose-response relationship for toxicity, whereas the model for efficacy does not. Yin et al. [21] originally specified a bivariate normal prior for the efficacy and toxicity parameters to allow a priori correlation between the model parameters but found that setting the off-diagonal covariance elements to zero did not impact their results. We will specify independent normal priors for ϕ_{kj} and ψ_{kj} and facilitate borrowing strength across populations by specifying a

hierarchical model on ϕ_{kj} and ψ_{kj} as follows:

$$\phi_{kj}|\mu_{\phi j}, \sigma^2_{\phi j} \sim Normal(\mu_{\phi j}, \sigma^2_{\phi j}) \quad (3)$$

$$\mu_{\phi j} \sim Normal(0, 50) \quad \text{and} \quad \sigma_{\phi j} \sim Uniform(0.39, 3)$$

and

$$\psi_{kj}|\mu_{\psi j}, \sigma^2_{\psi j} \sim Normal(\mu_{\psi j}, \sigma^2_{\psi j}) \quad (4)$$

$$\mu_{\psi j} \sim Normal(0, 50) \quad \text{and} \quad \sigma_{\psi j} \sim Uniform(0.39, 3)$$

for dose level $j = 1, \ldots, D$ and population $k = 1, \ldots, K$. Yin et al. [21] specify a $Normal(0, 100)$ prior on ϕ_j and ψ_j. Our design, which uses HM to share information across populations, will have a smaller sample size for each population than would typically be used in an independent Phase I-II design. To accommodate the smaller sample size, we reduce the prior variance to 50. Similar to the parametric binary bivariate model, ${\sigma_{\phi j}}^2$ and ${\sigma_{\psi j}}^2$ are our hierarchical variance parameters, controlling the amount of sharing across populations, and were specified using simulation studies, as described at the end of this section. The dose-response models specified above provide the marginal probabilities of toxicity and efficacy. Yin et al. [21] induce correlation between the toxicity and efficacy outcomes using the global cross-ratio model proposed by Dale [6]. Define $\pi_{xy,\,kj} = Pr(T_{kj} = x, E_{kj} = y|\text{population} = k, \text{dose} = j)$ with $x \in \{0, 1\}$ and $y \in \{0, 1\}$. Under the global cross-ratio model, the toxicity-efficacy odds ratio (OR) is defined as follows:

$$\theta_{kj} = \frac{\pi_{00,kj}\pi_{11,kj}}{\pi_{01,kj}\pi_{10,kj}}$$

where θ_{kj} quantifies the association between the two outcomes for population k at dose level j. Yin et al. [21] specify a $logNormal(0, 10)$ prior distribution for each θ_j and assume all θ_j's are independent to ease computation. To reduce our parameter space, we define $\theta_{kj} \sim logNormal(0, 5)$, rather than define a hierarchical structure to share information across populations when estimating the odds ratio. Recall that we will share information across populations for estimating the probability of toxicity and efficacy through the hierarchical models specified in Eqs. (3) and (4) and feel that specifying an additional hierarchy for the odds ratio is unnecessary. Finally, we reduce the prior variance for the odds ratio to accommodate a smaller sample size for each population compared to an independent design for each population. After accounting for the correlation induced by the global cross-ratio model, the joint toxicity and efficacy outcomes for dose j and population k follow a multinomial distribution with response probabilities $(\pi_{11,\,kj}, \pi_{10,\,kj}, \pi_{01,\,kj}, \pi_{00,\,kj})$ and a sample size of n_{kj} patients, where the response probabilities are defined as follows [6]:

$$\pi_{11,kj} = \begin{cases} \left(a_{kj} - \sqrt{a_{kj}^2 + b_{kj}}\right)/\{2(\theta_{kj} - 1)\}, & \text{for } \theta_{kj} \neq 1 \\ \pi_{T,kj}\pi_{E,kj}, & \text{for } \theta_{kj} = 1 \end{cases}$$

$$\pi_{10,kj} = \pi_{T,kj} - \pi_{11,kj}$$
$$\pi_{01,kj} = \pi_{E,kj} - \pi_{11,kj}$$
$$\pi_{00,kj} = 1 - \pi_{T,kj} - \pi_{E,kj} + \pi_{11,kj},$$

with $a_{kj} = 1 + (\pi_{T,\,kj} + \pi_{E,\,kj})(\theta_{kj} - 1)$ and $b_{kj} = (-4)\theta_{kj}(\theta_{kj} - 1)\pi_{T,\,kj}\pi_{E,\,kj}$.

2.2. Trinary outcome

The last model we consider is a hierarchical extension of the triCRM proposed by Zhang et al. [23]. Rather than separately modeling bivariate binary outcomes, they collapse the four possible outcomes into a single variable with three outcomes: no efficacy or toxicity, efficacy without toxicity and toxicity with or without efficacy. An advantage to this approach is that the model is simple relative to the other models, since we do not have to model separate dose-response models for the two outcomes, but a disadvantage is that the marginal probability of efficacy is no longer identifiable. However, our primary interest is identifying doses with sufficient efficacy and acceptable toxicity, mitigating the impact of this disadvantage. Denote the probabilities of the three possible outcomes (no efficacy or toxicity, efficacy without toxicity, toxicity with or without efficacy) as $\omega_0, \omega_1, \omega_2$, respectively, which by definition sum to 1. We can define a hierarchical extension of the continuation-ratio model proposed by Zhang et al. [23] as follows:

$$log\left(\frac{\omega_{1,kj}}{\omega_{0,kj}} \mid population = k, dose = j\right) = \zeta_{1k} + \zeta_{2k} + \eta_{1k}(dose) \quad (5)$$

$$logit(\omega_{2,kj} \mid population = k, dose = j) = \zeta_{1k} + \eta_{2k}(dose)$$

for dose levels $j = 1, \ldots, D$ and population $k = 1, \ldots, K$, with hierarchical priors defined as follows:

$$\zeta_{tk}|\mu_{\zeta t}, \sigma^2_{\zeta t} \sim Normal(\mu_{\zeta t}, \sigma^2_{\zeta t})$$
$$\eta_{tk}|\mu_{\eta t}, \sigma^2_{\eta t} \sim Normal(\mu_{\eta t}, \sigma^2_{\eta t})$$
$$\mu_{\zeta t} \sim Normal(u_t, c_t^2) \quad \text{and} \quad \sigma_{\zeta t} \sim Uniform(0.39, 3)$$
$$\mu_{\eta t} \sim Normal(v_t, b_t^2) \quad \text{and} \quad \sigma_{\eta t} \sim Uniform(0.39, 3)$$

for $t = 1, 2$ with $u_1 = -3$, $u_2 = 2$, $v_1 = 0.5$, $v_2 = 1$, and $c_1 = c_2 = b_1 = b_2 = 4$. The second level mean specifications correspond to a prior toxicity skeleton, i.e., ω_2, of 0.12, 0.27, 0.50, 0.73, 0.88 for dose levels 1, 2, 3, 4, 5, respectively; and a prior skeleton for efficacy with no toxicity, i.e., ω_1, of 0.33, 0.37, 0.31, 0.20, 0.10 for dose levels 1, 2, 3, 4, 5, respectively, which results in a prior skeleton for no response, i.e., ω_0, of 0.55, 0.37, 0.19, 0.07, 0.02 for dose levels 1, 2, 3, 4, 5, respectively. As in the binary bivariate models, ${\sigma_{\zeta t}}^2$ and ${\sigma_{\eta t}}^2$ are our hierarchical variance parameters, controlling the amount of sharing across populations, and were specified using simulation studies, as described in Section 2.3. There is one major difference between our hierarchical model and the original model proposed by Zhang et al. [23]. For simplicity and computational ease, Zhang et al. [23] define a $Uniform(-10, 5)$ prior for the common intercept ζ_{1k}, a $Uniform(0, 10)$ prior on the second intercept ζ_{2k}, and $Uniform(0, 10)$ priors on the slope parameters η_{1k} and η_{2k}. These priors impart the following restrictions on the model: (i) the probability of no response, ω_0, decreases monotonically with dose, (ii) the probability of toxicity with or without efficacy, ω_2, increases monotonically with dose, and (iii) the probability of efficacy without toxicity, ω_1, may or may not be monotone with dose. In contrast, our hierarchical model has no such restrictions. We originally considered hierarchical prior specifications that maintained these restrictions but found them to be difficult to implement computationally. Furthermore, our simulation results suggest that our model performs well without these restrictions and, hence, we proceed with the hierarchical model presented above.

2.3. Hyperparameter specification

The hyperparameters for the models discussed above were determined by simulation, as follows. The second-level mean hyperparameters were determined by separately varying each parameter in the corresponding independence model and selecting the combination with the most robust performance, as evaluated by simulation. We specify a uniform prior for the second-level standard deviation. The lower bound for the uniform prior distribution was set greater than zero due to the small sample sizes found in Phase I clinical trials, especially early in the trial, in which case the model cannot rule out a population variance of zero, resulting in an invalid distribution for our first level probability model. We explored different lower and upper bounds for the uniform prior and found that decreasing the lower bound resulted in poor BOD selection probabilities for heterogeneous scenarios but favorable BOD selection probabilities for homogeneous scenarios; and for a large enough lower bound, the design performs similar to independent designs. On the contrary, increasing the upper bound has the reverse effect but is

less dramatic than increasing the lower bound. We note that simulations were similar for a larger lower bound for the non-parametric bivariate binary model, however, we chose the smaller value to be consistent with the other models. After fixing the mean hyperparameters, we selected the hyper-parameters for the standard deviation for each model by progressively increasing the lower bound for our uniform prior and selecting the value with the most robust operating characteristics, as evaluated by simulation.

3. Dose-finding algorithm

In this section, we discuss dose-finding when using hierarchical modeling to share information across populations in Phase I-II clinical trials. We expect enrollment to be staggered and randomly distributed across populations. In our previous investigation, we discussed three dose-finding guidelines that define when to allow dose-escalation within a population taking into account the number of patients observed in other populations. These guidelines are incorporated for patient safety but our simulation results suggest that our guidelines result in improved operating characteristics based on a number of metrics, compared to unrestricted dose-finding. In this investigation, we consider only a single dose-finding guideline based on our previous results [5]. We identify a set of acceptable doses for each population, assuming admissibility criteria and minimum performance levels as elicited from clinicians. For each population, we determine the optimal dose from the set of acceptable doses by maximizing each population's posterior mean probability of efficacy without toxicity, $\pi_{01,\ kj}$, following the work of Yin et al. [21]. For the parametric bivariate model, the two binary outcomes are assumed independent and the probability of efficacy with no toxicity for each dose is simply the product of the marginal probability of efficacy and the marginal probability of no toxicity. For the non-parametric bivariate model, the optimal dose is determined by π_{01}, the multinomial probability for efficacy with no toxicity. For the two bivariate binary outcome models described in Section 2.1, a dose is acceptable if the posterior probability of a DLT being less than the clinician-specified target toxicity level and the posterior probability of an efficacious response exceeding the clinician-specified minimum threshold for efficacy both exceed pre-specified minimum thresholds, i.e.,

$$Pr\left(\pi_{T_k} < \overline{\pi}_T \mid Data, Dose\right) > \gamma_T \quad \text{and} \quad Pr\left(\pi_{E_k} > \underline{\pi}_E \mid Data, Dose\right) > \gamma_E \tag{6}$$

where $\overline{\pi}_T$ is the maximum acceptable probability of DLT, $\underline{\pi}_E$ is the minimum acceptable probability of efficacy, and γ_T and γ_E are the minimum pre-specified thresholds for toxicity and efficacy, respectively. These are admissibility criteria for toxicity and efficacy proposed by Thall and Cook [17]. The thresholds γ_T and γ_E are typically chosen between 0.05 and 0.20 and can be thought of as tuning parameters to achieve desired trial operating characteristics [1]. We cannot use the acceptability criteria described above for the trinary model because, although the marginal probability of toxicity can be estimated, the marginal probability of efficacy is not identifiable. Instead, we use two decision functions proposed by Zhang et al. [23] to determine the set of acceptable doses and, among those found to be acceptable, the optimal dose. We denote $\widehat{\omega}_{0,kj}$, $\widehat{\omega}_{1,kj}$, $\widehat{\omega}_{2,kj}$ to be the posterior mean probabilities of no toxicity or efficacy, efficacy without toxicity, and toxicity with or without efficacy, respectively. The first decision rule determines the set of acceptably safe doses using:

$$\delta_{1,kj} = I\left(\widehat{\omega}_{2,kj} < \overline{\pi}_T\right) \tag{7}$$

Given that a dose is acceptable, i.e., $\delta_{1,\ kj} = 1$, Yin et al. [21] determine the optimal dose from the set of acceptable doses by maximizing the toxicity-adjusted treatment success rate,

$$\delta_{2,kj} = \widehat{\omega}_{1,kj} - \lambda\widehat{\omega}_{2,kj}, \tag{8}$$

where $0 \le \lambda \le 1$ is a weight for the posterior mean probability of toxicity, $\widehat{\omega}_{2,kj}$. If we set $\lambda = 0$, the decision rule to determine the optimal dose is the dose maximizing the posterior mean probability of efficacy with no toxicity. We can also consider using $\delta_{1,\ kj}$ for the bivariate binary outcome models presented in Section 2.1. However, these models are over-specified and using $\delta_{1,\ kj}$ with these models results in more trials stopping early due to poor estimation. Finally, we also modify Zhang et al. [23]'s decision function, $\delta_{1,\ kj}$, to require a minimum performance level for efficacy,

$$\delta^*_{1,kj} = I\left(\widehat{\omega}_{2,kj} < \overline{\pi}_T\right) \times I\left(\frac{\widehat{\omega}_{1,kj}}{\widehat{\omega}_{0,kj} + \widehat{\omega}_{1,kj}} > \underline{\pi}_{E|T^c}\right) \tag{9}$$

That is, we require the posterior mean probability of efficacy conditional on no toxicity to be greater than some minimum pre-specified threshold in addition to the toxicity decision criteria found in (7). In a standard Phase I-II clinical trial, the initial cohort of (typically) three patients is treated at the lowest dose level and subsequent cohorts are treated at the current estimate of the optimal dose based on the outcomes for all previous subjects under the restriction that no untried dose-level may be skipped when escalating. Extending this approach to hierarchical modeling with multiple populations is not straightforward. One approach would be to escalate in cohorts of three patients regardless of the population, but this would be too aggressive and potentially result in a patient being treated at a dose-level before other patients from the same population are treated at a lower dose-level. Alternately, escalation could occur using cohorts of three patients within a population but this would not take full advantage of sharing information across populations using hierarchical modeling. Instead, we will use the "$m/(m+1)$" dose finding guideline (DFG) proposed in [5]. The "$m/(m+1)$" DFG provides a run-in period for each population and indicates when a population is able to escalate to an untried dose, but the ultimate decision to escalate is based on the current estimate of the optimal dose. Formally, the "$m/(m+1)$" DFG allows escalation to dose-level $j+1$ for population $k = 1, \ldots, K$ if:

- **m** patients in population k,
- Or **m** + **1** patients overall (and at least 1 patient in population k),

have been treated at dose level j, for $j = 1, \ldots, D-1$. This DFG will encourage escalation, when appropriate, but was shown in our previous investigation to effectively limit the number of DLTs and patients treated at overly toxic dose-levels. Yin et al. [21] propose that the trial should escalate to the next untried dose level if there is high posterior probability that the probability of DLT for the highest tried dose is less than the target toxicity level, i.e.,

$$Pr\left(\pi_{T_k} < \overline{\pi}_T \mid Data, Dose_{max}\right) > p \tag{10}$$

where $p \ge \gamma_T$. With the complex models used in Phase I-II clinical trials, it can be difficult to estimate the dose-response curves when there are multiple untried dose-levels. Criterion (10) encourages escalation when there are untried dose levels that appear to be sufficiently safe. We will implement the above escalation rule once a population is allowed to escalate, as determined by the "$m/(m+1)$" DFG, to assure adequate exploration of all dose levels with an acceptable probability of toxicity. In summary, our proposed Phase I-II design will proceed as follows:

1. Treat the first patient in each population at the lowest dose level.
2. When a new patient is enrolled, determine if their population is allowed to escalate (or de-escalate) as determined by the "$m/(m+1)$" DFG and if so, update the posterior distribution for all model parameters using all available data. Otherwise, treat the next patient at the current dose-level.
3. If escalation is allowed, determine the set of acceptable dose-levels for the current population using Criteria (6) for the bivariate models or Criteria (9) for the trinomial model. The trial terminates for futility if all dose-levels are unacceptable and Criterion (10) is not

satisfied.

4. Otherwise, treat the next patient at the dose level that maximizes either $\pi_{01,\ kj}$, for the bivariate models, or $\delta_{2,\ kj}$, for the trinomial model, under the restriction that untried dose-levels cannot be skipped when escalating. If Criterion (10) is satisfied, the next patient will be treated at the lowest untried dose-level.
5. Repeat steps 2–4 until the maximum overall sample size is reached. Within each population, the acceptable dose that maximizes either $\pi_{01,\ kj}$ or $\delta_{2,\ kj}$ (depending on the model) at study completion is considered the optimal dose.

4. Simulation study

We conducted a simulation study to evaluate the operating characteristics of a Phase I-II clinical that incorporates hierarchical modeling using the models discussed in Section 2. Design performance was summarized by (i) the probability of correctly identifying each population's biologically optimal dose (BOD), (ii) the percent of patients experiencing a DLT at the BOD and across all dose levels, and (iii) the average number of patients treated at each population's BOD. Simulations were completed assuming the following design parameters. The target toxicity level was set to $\overline{\pi}_T = 0.3$ and the minimum efficacy level was set to $\underline{\pi}_E = 0.3$. Gatekeepers defining acceptable doses were set to $\gamma_T = 0.25$, $\gamma_E = 0.1$ for Criteria (6) and $\underline{\pi}_{E|T^c} = 0.1$ for Criteria (9). We set $p = .5$ for Criterion (10), which encourages escalation to untried doses if all tried doses have been shown to be safe. We assume $K = 5$ populations with uniform enrollment across populations and $D = 5$ dose levels for investigation. We assume a maximum sample size of 100 total patients, with a minimum of 3 patients per population. This corresponds to an average of 20 patients in each population, which is smaller than the sample sizes typically needed to implement a Phase I-II clinical tria. Gibbs and slice sampling were completed in JAGS via R using *rjags* [16]. Posterior inference was completed using 5000 MCMC samples following a period of 1000 iterations for burn-in. 1000 simulated trials were completed for each scenario. For comparison, we also evaluate the operating characteristics assuming independent trials were completed for each population. Simulations were completed with models analogous to the three models discussed in Section 2. In each case, we fit the model specified in Section 2 without the second level of the hierarchy. A maximum of 20 subjects per population was used for the independent designs.

4.1. Scenarios

Data were simulated from one of eight scenarios. Each scenario included five populations. The dose-toxicity and dose-efficacy curves for each population were generated from one of the thirteen cases discussed by Yin et al. [21]. The dose-toxicity and -efficacy curves for each case are presented in Fig. 1. We define the BOD to be the dose that maximizes π_{01} (black dot) or $\delta_2(\lambda = 0)$, with δ_1^* defining acceptable doses (square), depending on the model. Other decision rules for selecting the optimal dose include minimizing the toxicity-efficacy odds ratio (open circle) and maximizing $\delta_2(\lambda = 0)$ with Zhang et al. [23]'s original decision function δ_1 to define acceptable doses (x). We note that in some cases the optimal dose varies by selection criteria. For example, in Case 7, the dose-toxicity and dose-efficacy curves increase at the same rate for the first three dose levels and consequently, minimizing the toxicity-efficacy odds ratio selects dose level 1 while all other decision criteria select dose level 3. In Scenario 1, the same combination of dose-toxicity and -efficacy curves were used for all five populations (Case 1), where the lowest dose is the optimal dose. In Scenario 2, population one uses the curves from Case 3, where toxicity and efficacy increase at the same rate (odds ratio = 1) and there is no optimal dose, and the other four populations use the curves from Case 13, where dose 4 is optimal. In Scenario 3, the optimal dose differs by population but is clustered around the intermediate dose-levels. Scenario 4 is comprised of Cases 6, 7, 8, 10 and 13, which have optimal doses of dose level 4, 3, 3, 2, and 4, respectively. In Scenario 4, the populations are more heterogeneous and the optimal dose varies substantially across populations. Specifically, Scenario 4 is comprised of Cases 9, 3, 1, 8 and 11, which have no optimal dose, no optimal dose, optimal dose equal to dose 1, optimal dose equal to dose 3, and optimal dose equal to dose 3, respectively. Scenario 5 is the same as Scenario 4, however, there is an additional population with no optimal dose (Population 3). Here, the populations assume dose-response curves from Cases 9, 3, 3, 8, and 11. Similarly, Scenario 6 is the same as Scenario 5 but an additional futile population is added (Population 4). This scenario is comprised of Cases 9, 9, 3, 3, and 8. In the last two scenarios (Scenarios 7 and 8), the populations are very heterogeneous with the true BOD spanning across the entire dose set. Scenario 7 is comprised of Cases 1, 1, 7, 12, and 12 with true optimal dose levels of dose 1, dose 1, dose 3, dose 5, and dose 5, respectively. Scenario 8 is the same as Scenario 7, however, Population 2 now has an optimal dose level of dose 3. In Table 1, the true case numbers and optimal dose levels are clearly displayed for all scenarios.

4.2. Results

Fig. 2 shows results for Scenarios 1 and 2. Fig. 2 presents the probability of correctly identifying the optimal dose (top plots; in the upper right plot corner: average probability of correctly identifying true BOD across all k), the percent of patients experiencing a DLT at the BOD and across all dose levels (middle plots), and the average number of patients treated at the optimal dose (bottom plots). In Scenario 1, all populations have the same dose-toxicity and -efficacy curves, resulting in the same optimal dose for all populations (dose level 1). We see that hierarchical modeling results in an increase in the probability of correctly identifying the optimal dose with the parametric bivariate model (displayed as “PLR HM”) showing the best performance, around 10% improvement in the trinomial model (displayed as “Tri HM”), while the non-parametric bivariate model showed relatively little improvement (displayed as “OR HM”). We see similar trends for the average percent correct selection across all populations. A possible concern related to our design is that our more aggressive dose-finding algorithms might increase DLTs, but our results suggest that the rate of DLTs is similar across models both with and without HM. In addition to increasing the probability of correctly identifying the optimal dose, hierarchical modeling also increased the average number of patients treated at the optimal dose, although in this case, the hierarchical trinomial model displayed better performance than the PLR HM. These results highlight the benefits of hierarchical modeling: when the populations are homogenous and it is appropriate to share across populations, HM increases the probability of correctly identifying the optimal dose and the number of patients treated at the optimal dose with limited impact on the number of DLTs observed in the trial.

In Scenario 2, the last four populations are homogeneous (Case 13, where dose 4 is the optimal dose), while the first population has no dose level with an acceptable efficacy/toxicity trade-off (Case 3). This is a challenging scenario because the last four populations will encourage borrowing across populations but this could result in incorrectly borrowing strength from the first population, where no dose is acceptable. For the last four populations, HM greatly out-performs the independent models in correctly identifying the optimal dose and in treating more patients on average at the optimal dose, with limited impact on the probability of DLTs. This behavior is to be expected because the larger pooled sample size results in a more precise estimate of the BOD and allows individual populations to escalate more quickly than the independent designs. Comparing across hierarchical models, we see that the OR HM model has the best performance and largest improvement over its corresponding independent design, while the PLR HM and Tri HM have similar performance. For the first population, where no dose has an acceptable efficacy/toxicity trade-off, we see that HM decreases

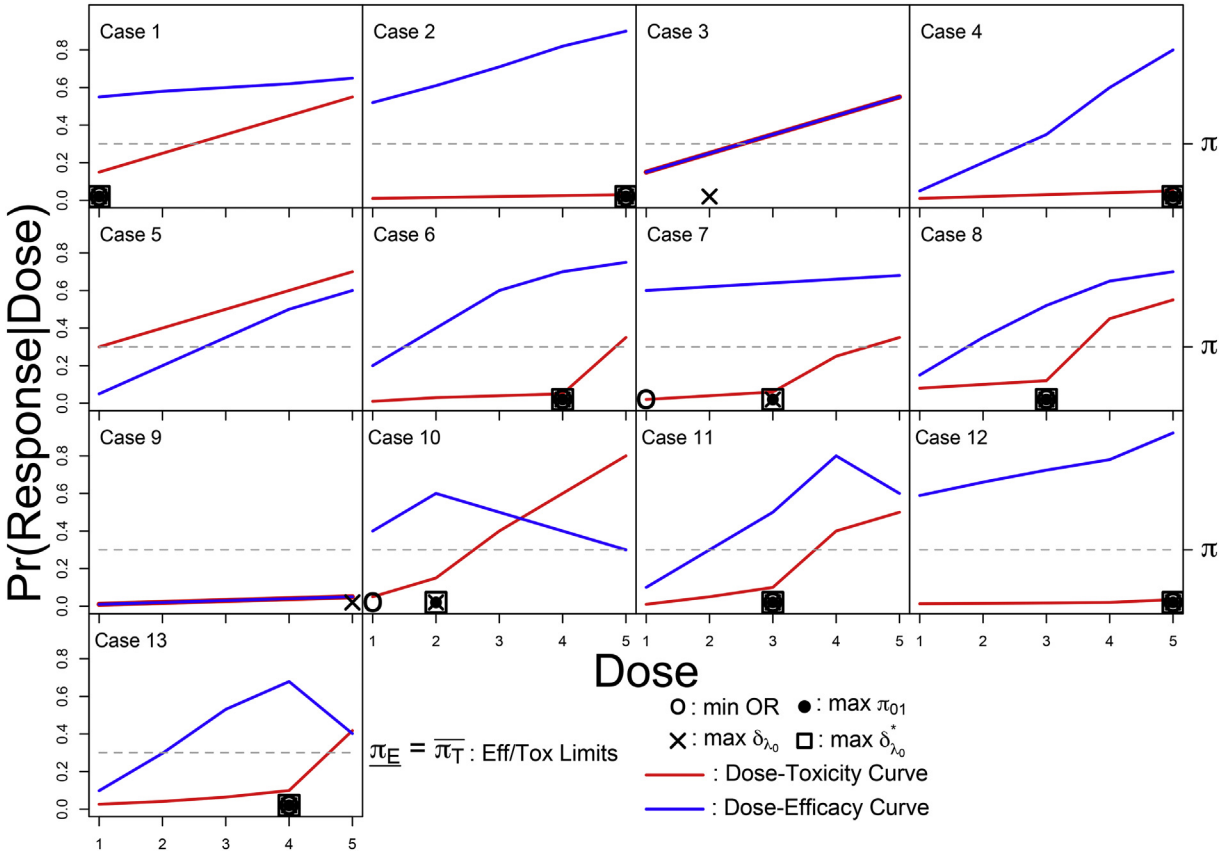

Fig. 1. Thirteen combinations of dose-toxicity and dose-efficacy curves from Yin et al. [21]. The optimal dose based on different decision criteria are displayed above the pre-specified dose levels on the x-axis, and denoted as: (open circle) minimize toxicity-efficacy odds ratio, (black dot) maximize joint posterior probability of no toxicity with efficacy, (x) maximize Zhang et al. [23] decision rule with $\lambda = 0$, and (square) maximize Zhang et al. [23] decision rule with $\lambda = 0$ and additional acceptability criterion for posterior probability of efficacy conditional on no toxicity. The dose-toxicity and dose-efficacy curves are represented with red and blue lines, respectively. The gray line displays the upper toxicity and lower efficacy limits for our posterior probabilities. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1
For 8 scenarios, the true biological optimal dose (BOD) level is in bold and below: the case index (see Fig. 1 for dose-response curves) for each population $k = 1, \ldots, 5$.

Scenario		Population				
		1	2	3	4	5
1	BOD	1	1	1	1	1
	Case	1	1	1	1	1
2	BOD	0	4	4	4	4
	Case	3	13	13	13	13
3	BOD	4	3	3	2	4
	Case	6	7	8	10	13
4	BOD	0	0	1	3	3
	Case	9	3	1	8	11
5	BOD	0	0	0	3	3
	Case	9	3	3	8	11
6	BOD	0	0	0	0	3
	Case	9	9	3	3	8
7	BOD	1	1	3	5	5
	Case	1	1	7	12	12
8	BOD	1	3	3	5	5
	Case	1	7	7	12	12

the probability of drawing the correct conclusion, with the two bivariate models exhibiting worse performance than the Tri HM model. This is to be expected due to the more stringent acceptability criterion used by the trinomial model. Finally, we note that while the HM designs decreased the probability of correctly concluding that no dose level is acceptable, HM had little impact on the number of DLTs observed with little increase over the independent designs (with the exception of OR HM). Results for Scenarios 3 and 4 can be found in Fig. 3. In Scenario 3, there is modest heterogeneity in the optimal dose with the optimal dose varying from dose level 2 to dose level 4. We believe that this scenario best represents what we might expect to see in practice, for two reasons. First, investigators are advised to select their dose range such that the optimal dose is likely to be an intermediate dose based on pre-clinical data. Second, this scenario represents the case where the optimal dose is similar across populations but there is some variability due to different patient characteristics for each population. In this case, we see that the hierarchical extensions are more likely to correctly identify the optimal dose and treat more patients, on average, at the optimal dose in all populations. Among the hierarchical designs, the trinomial model performs the best across all populations. We see similar trends for the average percent correct selection across all populations. Again, HM has only a modest impact on the DLT rate compared to the independent designs, with the Tri HM again having the lowest DLT rate from among the three hierarchical designs and the PLR HM design having the highest. A possible solution to decreasing the DLT rate for the PLR HM design is to increase γ_T when using the parametric bivariate model but the impact of increasing γ_T on the other operating characteristics is a potential concern. Scenario 4 is another difficult case, where all doses have an unacceptable efficacy/toxicity trade-off in the first two populations (Cases 3 and 9, respectively), dose level 1 is the optimal dose in the third population (Case 1) and dose level 3 is the optimal dose for the last two populations (Cases 8 and 11, respectively). The results for Scenario 4 are consistent with our previous results. The three

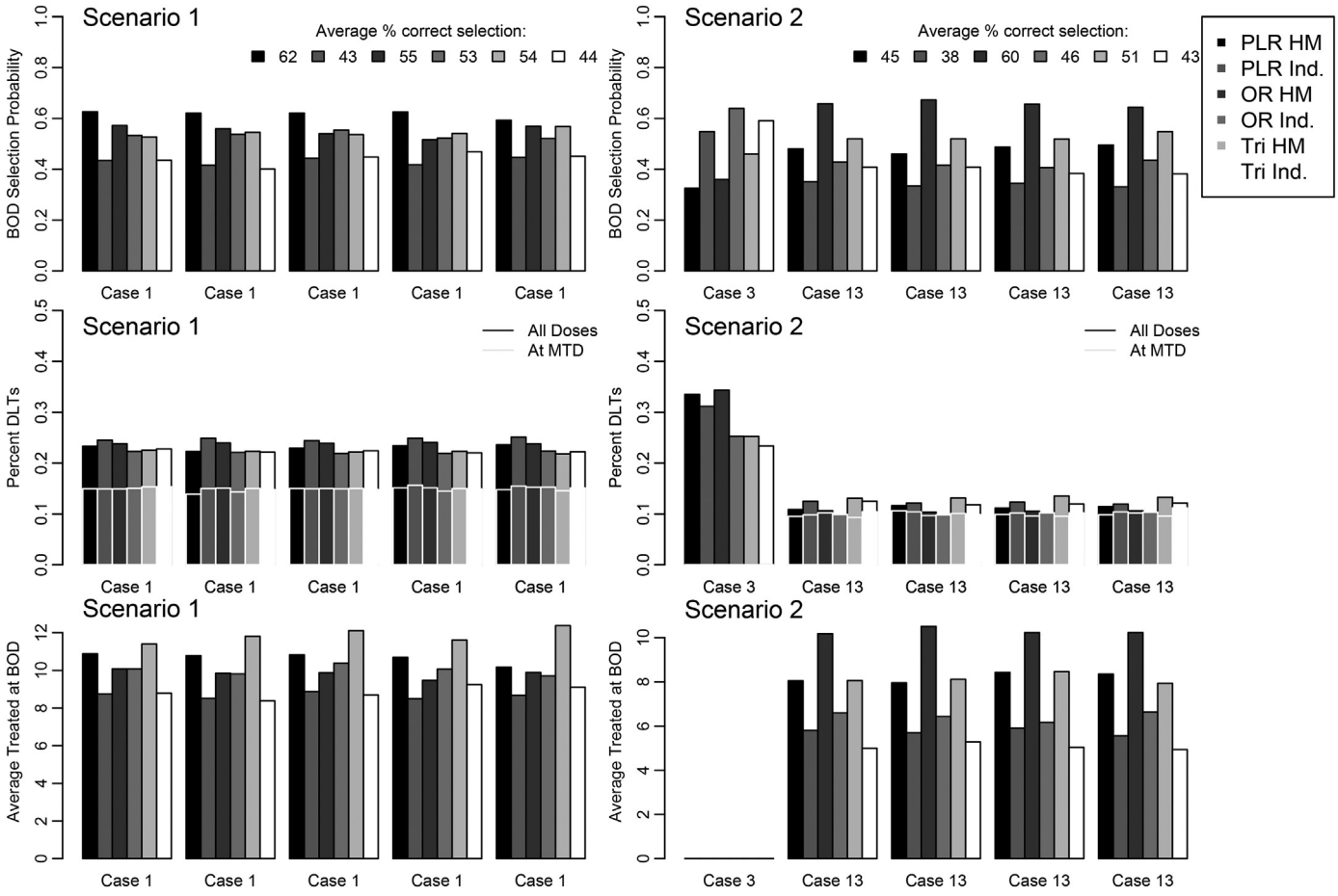

Fig. 2. (Left column) Trial operating characteristics from 1000 simulated trials for Scenario 1 by population: (top) selection probability for the population-specific biologically optimal dose (BOD) and in the upper right corner: average probability of correctly identifying true BOD across all *k*; (middle) percentage of dose-limited toxicities across all doses (black outline) and at BOD (gray outline); (bottom) average number of patients treated at the population specific BOD, for the three hierarchical and independence designs. The first two bars (black and dark gray; labeled "PLR HM" and "PLR Ind.", respectively) present results using the parametric bivariate binary models. The next two bars (darker gray and gray; labeled "OR HM" and "OR Ind.", respectively) present results using the non-parametric bivariate binary models. The last two bars (light gray and white; labeled "Tri HM" and "Tri Ind.", respectively) present results for the parametric trinary model. For each scenario, the population's case-index used for data generation is presented on the x-axis. (Right column) Simulation results for Scenario 2.

hierarchical models are more likely to correctly identify the optimal dose and treat more patients, on average, at the optimal dose when an optimal dose exists (last three populations) but the HM approaches are also more likely to incorrectly conclude that an optimal dose exists when no dose level has an acceptable efficacy/toxicity trade-off, although we again note that the impact on the total number of DLTs is minimal for PLR HM and Tri HM, with the PLR HM having the highest DLT rate and Tri HM having the lowest DLT rate from among the three HM designs. Finally, comparing across hierarchical models, we see that the Tri HM design has the best performance of the three hierarchical models, treating more patients at the optimal dose and fewer patients above the optimal dose than the other two models. Fig. 4 displays results for Scenarios 5 and 6. In these scenarios, a majority of the populations have no optimal dose. Results for Scenario 5 are similar to the results for Scenario 4 in that HM shows an improvement over the independent models when an optimal dose exists, however, the difference between the HM and independent models is less dramatic due to the additional population with no optimal dose (Population 3). In addition, we note that the ability to identify the correct optimal dose (dose level 3) for Populations 4 and 5 did not diminish. Interestingly, OR HM performed the best in this scenario for the average probability of correctly selecting the optimal dose across all populations. This is due to superior performance in the last two populations, which was previously seen in Scenario 4 but was not as prominent due to Population 3 having a BOD of dose level 1. In Scenario 6, four of the populations do not have a BOD and consequently, HM performs similar to the independence designs but our ability to identify the BOD (dose level 3) in the last population was diminished relative to Scenario 5. In both scenarios, we again see that the PLR HM displays the largest DLT rate across all dose levels. Lastly, Fig. 5 displays results for Scenarios 7 and 8. These are very heterogeneous settings, wherein all populations have a BOD but the true dose level spans across all possible doses. In Scenario 7, the first two populations have dose level 1 and both PLR HM and Tri HM outperform the independence designs. Similar to Scenario 1, PLR HM displays slightly higher selection probability but Tri HM displays slightly lower DLT rate across all doses and slightly higher average number of patients treated at BOD. For Population 3 (Case 7), we see that all HM models perform worse than the independent models with the bivariate models showing worse performance than the trivariate model. This is due to the increased heterogeneity across populations. The last two populations have a true BOD of dose level 5 and we see the Tri HM has an approximately 20% improvement in the BOD selection probability and 4 more patients treated at the BOD, on average, compared to the Tri Ind. and OR HM designs (and a 40% BOD selection probability improvement and 6 additional patients treated at the BOD, on average, compared to the PLR HM design). Scenario 8 has the same degree of heterogeneity, however, Population 2 has a higher true BOD level than in Scenario 7. Here, we see very similar results to Scenario 7,

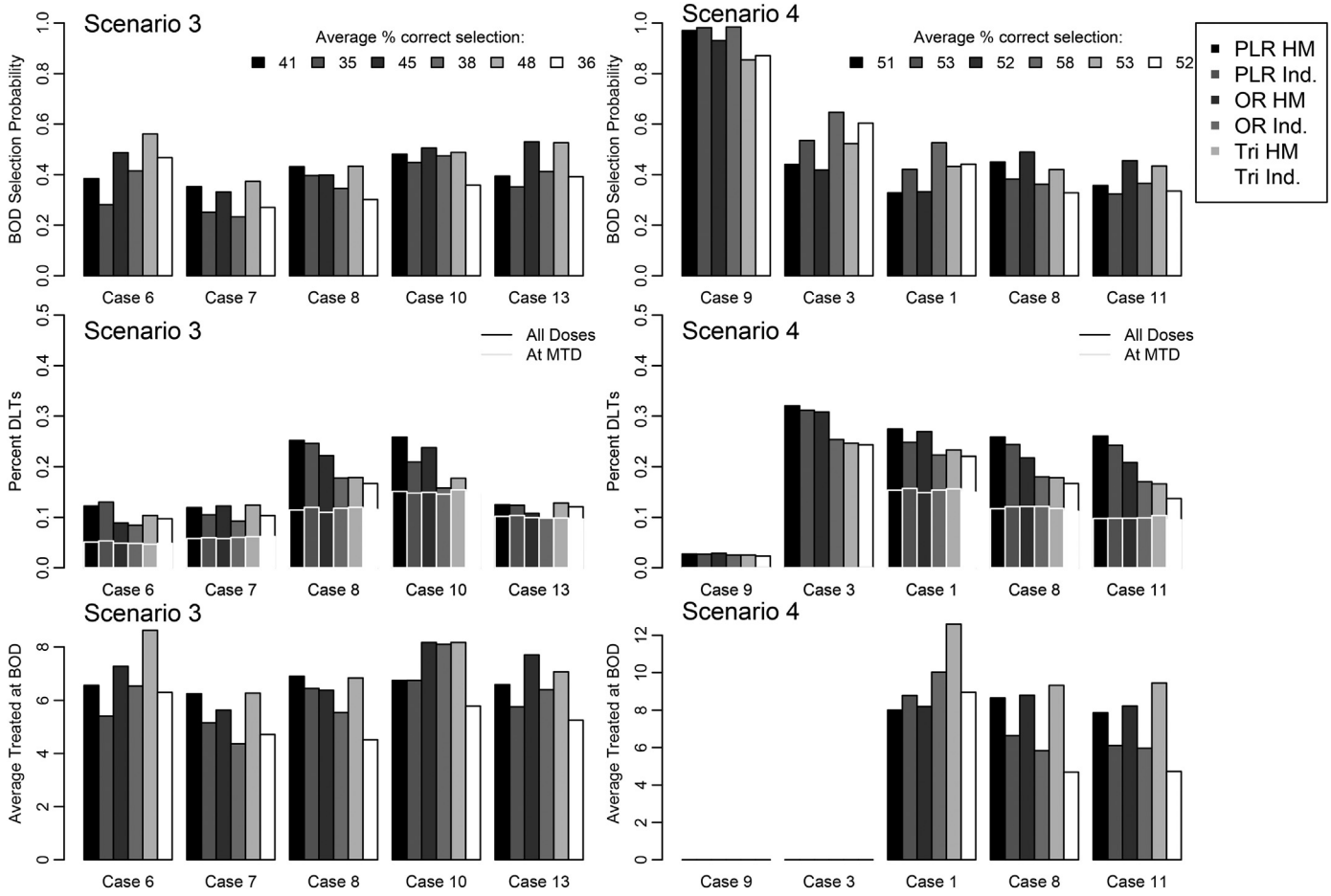

Fig. 3. (Left column) Trial operating characteristics from 1000 simulated trials for Scenario 3 by population: (top) selection probability for the population-specific biologically optimal dose (BOD) and in the upper right corner: average probability of correctly identifying true BOD across all *k*; (middle) percentage of dose-limited toxicities across all doses (black outline) and at BOD (gray outline); (bottom) average number of patients treated at the population specific BOD, for the three hierarchical and independence designs. The first two bars (black and dark gray; labeled "PLR HM" and "PLR Ind.", respectively) present results using the parametric bivariate binary models. The next two bars (darker gray and gray; labeled "OR HM" and "OR Ind.", respectively) present results using the non-parametric bivariate binary models. The last two bars (light gray and white; labeled "Tri HM" and "Tri Ind.", respectively) present results for the parametric trinary model. For each scenario, the population's case-index used for data generation is presented on the x-axis. (Right column) Simulation results for Scenario 4.

except that improved performance is observed for the Tri HM in Case 7 due to the additional common population. Based on these results, we see that the trinay models are more likely to identify the optimal dose than the bivariate models when the highest dose level is optimal, with superior performance observed for the HM relative to the independent models.

In summary, our simulation results highlight the benefits of HM in Phase I-II dose-finding trials. Incorporating HM resulted in an increased probability of correctly identifying the optimal dose and treating more patients at the optimal dose than the independent designs, with only a modest increase in the probability of DLT. Comparing the three hierarchical models discussed in Section 2, the trinomial hierarchical model resulted in the best trade-off between correctly sharing information across populations when the populations were homogeneous and oversharing when the populations were heterogeneous. In addition, the trinomial model treated more patients at the optimal dose than the other two models, in most cases. In contrast, our simulation results suggest that implementing HM in the bivariate binary models and the parametric model, in particular, is not ideal. Both models were less likely to correctly declare all doses futile or unsafe and resulted in more DLTs when populations were heterogeneous compared to the trinomial model. Furthermore, they did not provide substantially more benefits than the trinomial model when populations were homogeneous.

4.3. Smaller K

Initially, we evaluated the HM models presented in Section 2 with five populations, but we also completed additional simulations assuming three populations to evaluate the robustness of our conclusions to a smaller number of populations. In this simulation study, we assume the same design parameters as previously given except with $K = 3$, so that on average 33 patients are treated in each population (max 33 patients/population for independent designs). We evaluate four different scenarios: (1) all populations have BOD equal to dose 3, (2) Population 1 has no BOD and Populations 2–3 have BOD equal to dose 3, (3) Populations 1–2 have no BOD and Population 3 has BOD equal to dose 3, (4) Populations 1–3 have BOD equal to doses 1, 3, 5, respectively. The selection probability for each population-specific BOD from 1000 simulated trials is presented in Fig. 6. We see that our results with 3 populations are consistent with the results observed with 5 populations. That is, when an optimal dose exists (Scenarios 1 and 4), the Tri HM performs better than the independence design and the other HM approaches. When no optimal dose exists (Scenarios 2–3), the PLR HM performs the best, however, in these scenarios the Tri HM still performs better than the Tri Ind. design. Due to the small number of populations, the non-parametric approach performs the worst across all scenarios, except for OR Ind. in Scenario 4 for Case 1 (BOD equal to dose 1). These results suggest parametric HM can be used in a Phase I-II trial with as

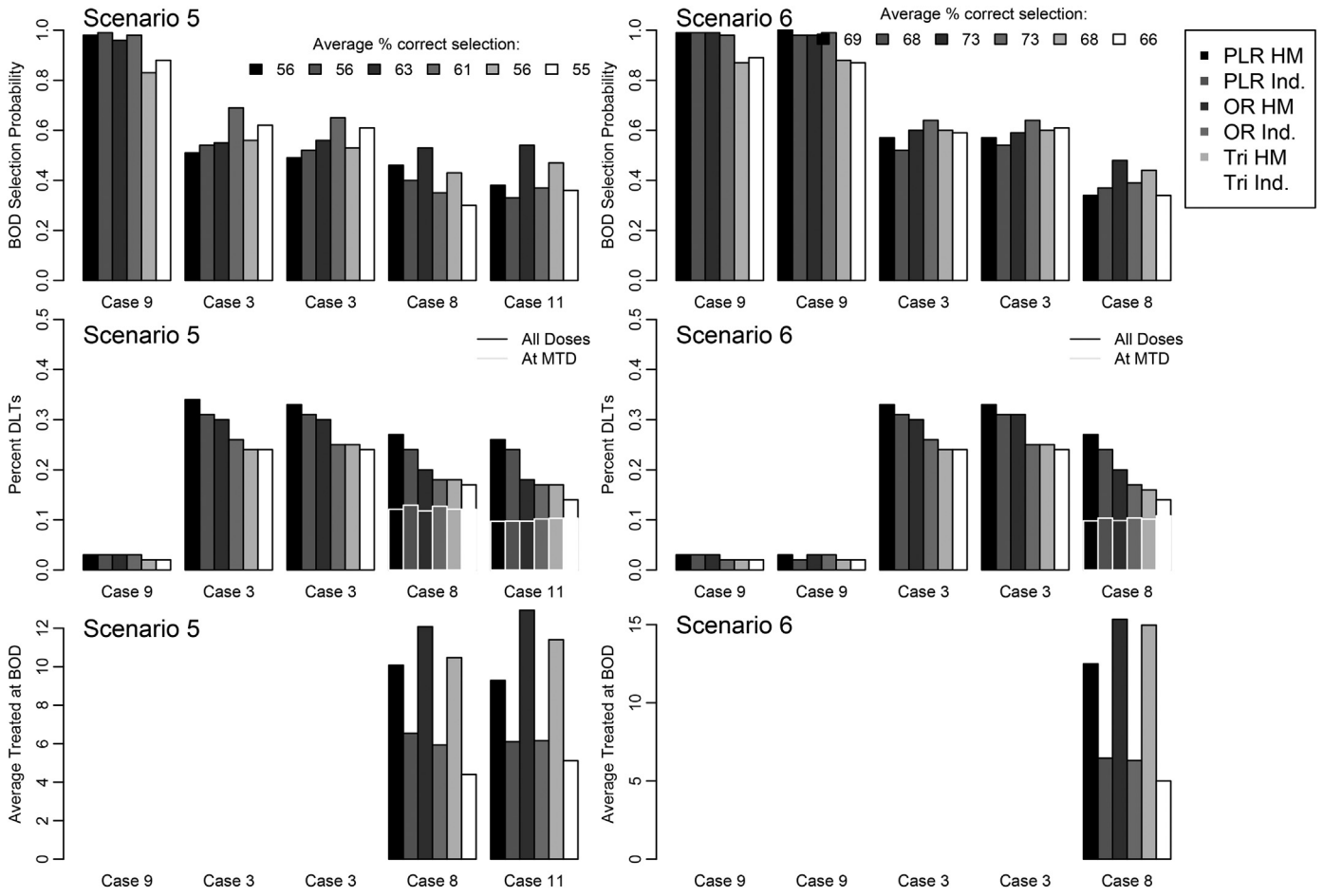

Fig. 4. (Left column) Trial operating characteristics from 1000 simulated trials for Scenario 5 by population: (top) selection probability for the population-specific biologically optimal dose (BOD) and in the upper right corner: average probability of correctly identifying true BOD across all *k*; (middle) percentage of dose-limited toxicities across all doses (black outline) and at BOD (gray outline); (bottom) average number of patients treated at the population specific BOD, for the three hierarchical and independence designs. The first two bars (black and dark gray; labeled "PLR HM" and "PLR Ind.", respectively) present results using the parametric bivariate binary models. The next two bars (darker gray and gray; labeled "OR HM" and "OR Ind.", respectively) present results using the non-parametric bivariate binary models. The last two bars (light gray and white; labeled "Tri HM" and "Tri Ind.", respectively) present results for the parametric trinary model. For each scenario, the population's case-index used for data generation is presented on the x-axis. (Right column) Simulation results for Scenario 6.

small as three populations.

5. Discussion

In this article we discussed HM for sharing information across populations in Phase I-II clinical trials. First, we presented two bivariate models, one parametric and one non-parametric, for modeling the dose-response relationship for efficacy and toxicity. Dose-finding using these models implemented the acceptability criteria defined by Thall and Cook [17] to identify acceptable dose-levels with the optimal dose defined as the one maximizing the probability of efficacy with no toxicity [21]. Next, we presented a hierarchical extension of the trinary outcome model proposed by Zhang et al. [23], which combined the two binary outcomes into a single, trinary outcome. Reducing the two binary outcomes to a single trinary outcome precluded the direct application of acceptability criteria defined by Thall and Cook [17] and instead we adapted the decision functions proposed by Zhang et al. [23] for identifying the optimal dose with the trinary model. Our simulation results suggest that the two hierarchical bivariate outcome models outperformed the trinary model when the populations are homogeneous and, in particular, the non-parametric bivariate model performed very well when the populations are homogeneous and the optimal dose is one of the higher dose levels. On the other hand, the two bivariate outcome models did not perform as well when the populations were heterogeneous, and performed particularly poorly when no dose was acceptable. In contrast the trinary model emerged as simpler and, as a result, exhibited more consistent performance than the other two models. Furthermore, in our simulations for scenarios where a true optimal dose exists, the trinary hierarchical model consistently outperformed the trinary independent model; and in almost all scenarios, the trinary hierarchical model treated more patients at the true optimal dose compared to the trinary independent model. In settings where the true optimal dose varied across populations, we found the bivariate hierarchical models performed worse than the bivariate independent models for populations with an optimal dose at a lower dose level. In these settings, the non-parametric hierarchical model treated less patients on average at the true optimal dose compared to the non-parametric independent model; whereas in some cases, the parametric hierarchical model still treated more patients on average at the true optimal dose compared to the parametric independent model. The trinary hierarchical model displayed more desirable properties as compared to the bivariate hierarchical models due to the smaller number of parameters needing to be modeled and estimated. A possible concern with the trinary model is performance when all tested dose levels are safe but efficacy plateaus, since the marginal probability of efficacy is non-identifiable. However, in this case, we believe if the marginal probability of efficacy is constant over the plateau, then the probably of efficacy without toxicity will decrease and the trinary

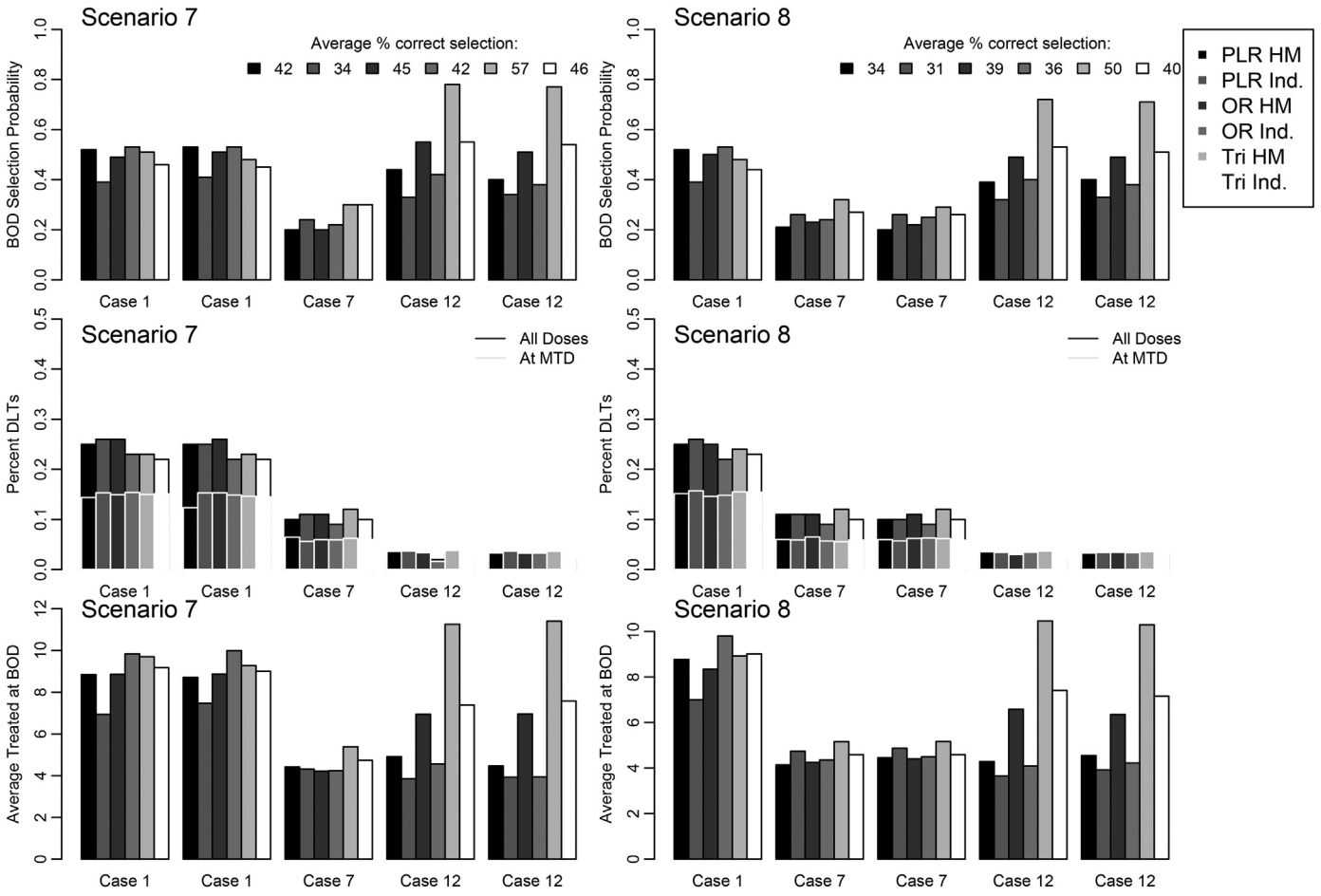

Fig. 5. (Left column) Trial operating characteristics from 1000 simulated trials for Scenario 7 by population: (top) selection probability for the population-specific biologically optimal dose (BOD) and in the upper right corner: average probability of correctly identifying true BOD across all *k*; (middle) percentage of dose-limited toxicities across all doses (black outline) and at BOD (gray outline); (bottom) average number of patients treated at the population specific BOD, for the three hierarchical and independence designs. The first two bars (black and dark gray; labeled "PLR HM" and "PLR Ind.", respectively) present results using the parametric bivariate binary models. The next two bars (darker gray and gray; labeled "OR HM" and "OR Ind.", respectively) present results using the non-parametric bivariate binary models. The last two bars (light gray and white; labeled "Tri HM" and "Tri Ind.", respectively) present results for the parametric trinary model. For each scenario, the population's case-index used for data generation is presented on the x-axis. (Right column) Simulation results for Scenario 8.

model will still perform well. We note in some settings, such as infrequent or low grade toxicity endpoints, it may make sense to collapse and model the toxicity outcome as a single pooled population (rather than using a hierarchical model), thus reducing the dimension of the parameter space for the bivariate models. In this case, distributions other than normal for hyper-parameters may perform better and be easier to specify, such as a uniform distribution. With this adaptation, the bivariate pooled models may exhibit better performance than the bivariate hierarchical models, especially in homogeneous cases, but we expect these improvements to be marginal. Furthermore, this approach may lead to an increased DLT rate in scenarios wherein only one population has an unacceptably high dose-toxicity curve. The methods presented in Section 2.1 implement the design parameters from Yin et al. [21]. Our results suggest that additional tuning is required for the parametric bivariate model, especially for the toxicity acceptability criterion, γ_T. As with any Bayesian analysis, the results presented in Section 4.2 depend on the prior distribution and prior dose-response skeleton. While the trinomial model's prior skeleton corresponds to a higher optimal dose level a priori, this method exhibits the best performance with respect to preventing escalation to overly toxic dose-levels. In the future, we would like to investigate more prior distributions for the pooling variance parameter of the hierarchical trinomial model. Furthermore, while we have assumed common prior skeletons for efficacy and toxicity across populations, this assumption is not required and could be relaxed. That said, it is not clear if hierarchical modeling would be appropriate if researchers believe, a priori, that the dose-toxicity and dose-efficacy curves are different across populations. This issue also requires further investigation. While our primary interest was in the performance of the three hierarchical models, the comparison of the three independent designs is also of interest because, to the best of our knowledge, these three models have never been compared using the same scenarios. While the dose-finding algorithms are different, it is interesting to note how the simpler yet conservative method performs against the more complex and flexible methods. Comparing both dose-finding designs under the same scenarios motivated the alteration to Zhang et al. [23]'s decision function defining acceptable doses. We incorporate an additional criterion for the probability of efficacy conditional on no toxicity. It would be interesting to further investigate altering the minimum response rate for efficacy conditional on no toxicity and the acceptability threshold, i.e., a smaller quantile of the posterior for the probability of efficacy conditional on no toxicity. An important practical issue to consider when implementing our design is variability in accrual rates across populations. In our simulations, we allowed accrual to be random, which will result in variable accrual patterns for each realized study, but it is also possible that the expected accrual rates could differ by populations (if one population is a rare disease, for example). We considered this possibility when developing our dose-finding guidelines, and expect our guidelines to perform well

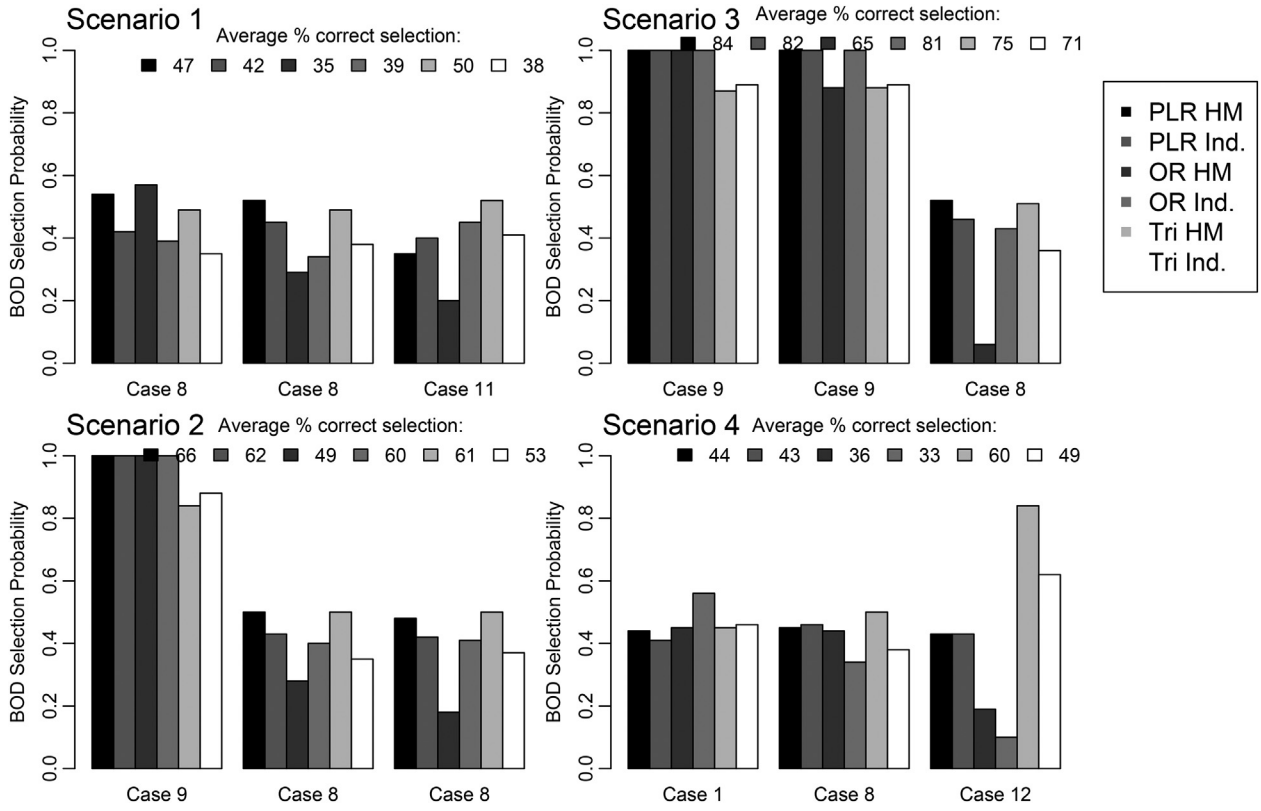

Fig. 6. For three populations: Selection probability for the population-specific biologically optimal dose (BOD) from 1000 simulated trials for Scenarios 1–4 by population; and in the upper right corner: average probability of correctly identifying true BOD across all *k*; The first two bars (black and dark gray; labeled "PLR HM" and "PLR Ind.", respectively) present results using the parametric bivariate binary models. The next two bars (darker gray and gray; labeled "OR HM" and "OR Ind.", respectively) present results using the non-parametric bivariate binary models. The last two bars (light gray and white; labeled "Tri HM" and "Tri Ind.", respectively) present results for the parametric trinary model. For each scenario, the population's case-index used for data generation is presented on the x-axis.

when accrual rates vary by population. If other populations have already explored higher doses, a slow accruing population could potentially escalate faster than in an independence design as long as there is strong evidence that higher doses are safe in other populations and the current dose level appears safe in the slow accruing population. On the other hand, for a fast accruing population dose-escalation is not delayed due to lack of patients in other populations and can mirror a traditional dose-finding study (with possible information sharing with other fast accruing populations). Nevertheless, slow accrual will likely result in a lower total sample size for a population and we expect that slow enrollment would limit our ability to correctly identify the population-specific optimal dose, especially when the true optimal dose is an extreme dose level (i.e., no optimal dose or highest dose level considered).

References

[1] S.M. Berry, B.P. Carlin, J.J. Lee, P. Muller, Bayesian Adaptive Methods for Clinical Trials, vol. 38, CRC Press, 2010.

[2] A. Borgatti, J.S. Koopmeiners, A.L. Sarver, A.L. Winter, K. Stuebner, D. Todhunter, A.E. Rizzardi, J.C. Henriksen, S. Schmechel, C.L. Forster, et al., Safe and effective sarcoma therapy through bispecific targeting of EGFR and uPAR, Mol. Cancer Ther. 16 (2017) 956–965.

[3] T. Braun, The bivariate continual reassessment method: extending the CRM to phase I trials of two competing outcomes, Control. Clin. Trials 23 (2002) 240–256.

[4] K. Cunanan, J.S. Koopmeiners, Evaluating the performance of copula models in phase I–II clinical trials under model misspecification, BMC Med. Res. Methodol. 14 (2014) 51.

[5] K.M. Cunanan, J.S. Koopmeiners, Hierarchical models for sharing information across populations in phase I dose-escalation studies, Stat. Methods Med. Res. (2017) (0962280217703812).

[6] J.R. Dale, Global cross-ratio models for bivariate, discrete, ordered responses, Biometrics (1986) 909–917.

[7] S.N. Goodman, M.L. Zahurak, S. Piantadosi, Some practical improvements in the continual reassessment method for phase I studies, Stat. Med. 14 (1995) 1149–1161.

[8] T.A. Gooley, P.J. Martin, L.D. Fisher, M. Pettinger, Simulation as a design tool for phase I/II clinical trials: an example from bone marrow transplantation, Control. Clin. Trials 15 (1994) 450–462.

[9] N. Houede, P.F. Thall, H. Nguyen, X. Paoletti, A. Kramar, Utility-based optimization of combination therapy using ordinal toxicity and efficacy in phase I/II trials, Biometrics 66 (2010) 532–540.

[10] A. Iasonos, J. O'Quigley, Dose expansion cohorts in phase I trials, Stat. Biopharm. Res. 8 (2016) 161–170.

[11] A. Iasonos, N.A. Wages, M.R. Conaway, K. Cheung, Y. Yuan, J. O'Quigley, Dimension of model parameter space and operating characteristics in adaptive dose-finding studies, Stat. Med. 35 (2016) 3760–3775.

[12] A. Ivanova, A new dose-finding design for bivariate outcomes, Biometrics 59 (2003) 1001–1007.

[13] J.S. Koopmeiners, J. Modiano, A Bayesian adaptive phase I-II clinical trial for evaluating efficacy and toxicity with delayed outcomes, Clin. Trials 11 (2014) 38–48.

[14] B. Nebiyou Bekele, Y. Shen, A Bayesian approach to jointly modeling toxicity and biomarker expression in a phase I/II dose-finding trial, Biometrics 61 (2005) 343–354.

[15] J. O'Quigley, M.D. Hughes, T. Fenton, Dose-finding designs for HIV studies, Biometrics (2001) 1018–1029.

[16] M. Plummer, rjags: Bayesian Graphical Models Using MCMC, (2011) (R package version 3-10).

[17] P.F. Thall, J.D. Cook, Dose-finding based on efficacy-toxicity trade-offs, Biometrics 60 (2004) 684–693.

[18] P.F. Thall, H.Q. Nguyen, T.M. Braun, et al., Using joint utilities of the times to response and toxicity to adaptively optimize schedule–dose regimes, Biometrics 69 (2013) 673–682.

[19] P.F. Thall, H.Q. Nguyen, E.H. Estey, Patient-specific dose finding based on bivariate outcomes and covariates, Biometrics 64 (2008) 1126–1136.

[20] P.F. Thall, K.E. Russell, A strategy for dose-finding and safety monitoring based on efficacy and adverse outcomes in phase I/II clinical trials, Biometrics (1998) 251–264.

[21] G. Yin, Y. Li, Y. Ji, Bayesian dose-finding in phase I/II clinical trials using toxicity and efficacy odds ratios, Biometrics 62 (2006) 777–787.

[22] Y. Yuan, G. Yin, Sequential continual reassessment method for two-dimensional dose finding, Stat. Med. 27 (2008) 5664–5678.

[23] W. Zhang, D.J. Sargent, S. Mandrekar, An adaptive dose-finding design incorporating both toxicity and efficacy, Stat. Med. 25 (2006) 2365–2383.