Biometrics, 2024, 80(4), ujae105
https://doi.org/10.1093/biomtc/ujae105
Biometric Methodology

OXFORD

ROMI: a randomized two-stage basket trial design to optimize doses for multiple indications

Shuqi Wang [1], Peter F. Thall[1], Kentaro Takeda [2], Ying Yuan [1,*]

[1]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, United States, [2]Astellas Pharma Global Development Inc., Northbrook, IL 60062, United States

[*]Corresponding author: Ying Yuan, Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, United States (yyuan@mdanderson.org).

ABSTRACT

Optimizing doses for multiple indications is challenging. The pooled approach of finding a single optimal biological dose (OBD) for all indications ignores that dose-response or dose-toxicity curves may differ between indications, resulting in varying OBDs. Conversely, indication-specific dose optimization often requires a large sample size. To address this challenge, we propose a Randomized two-stage basket trial design that Optimizes doses in Multiple Indications (ROMI). In stage 1, for each indication, response and toxicity are evaluated for a high dose, which may be a previously obtained maximum tolerated dose, with a rule that stops accrual to indications where the high dose is unsafe or ineffective. Indications not terminated proceed to stage 2, where patients are randomized between the high dose and a specified lower dose. A latent-cluster Bayesian hierarchical model is employed to borrow information between indications, while considering the potential heterogeneity of OBD across indications. Indication-specific utilities are used to quantify response-toxicity trade-offs. At the end of stage 2, for each indication with at least one acceptable dose, the dose with highest posterior mean utility is selected as optimal. Two versions of ROMI are presented, one using only stage 2 data for dose optimization and the other optimizing doses using data from both stages. Simulations show that both versions have desirable operating characteristics compared to designs that either ignore indications or optimize dose independently for each indication.

KEYWORDS: Bayesian hierarchical model; dose optimization; multiple indications; Project Optimus; randomization; utility.

1 INTRODUCTION

Conventional phase I oncology dose-finding designs originally were motivated by trials of cytotoxic agents, where the probabilities of toxicity, $\pi_T(d)$, and response, $\pi_R(d)$, increase with dose, d. This may not hold for targeted molecules or immunotherapies, where $\pi_T(d)$ and $\pi_R(d)$ may take a variety of different shapes. For example, if the delivered dose is saturated in the patient, the $\pi_R(d)$ curve initially increases with d and then flattens to a plateau. In such settings, a phase I maximum tolerated dose (MTD) is undesirable because lower doses achieve similar $\pi_R(d)$ but reduce $\pi_T(d)$ (Sachs et al., 2016). Thus, conventional phase I designs are unsuitable for most targeted agents (Shah et al., 2021; Thall et al., 2023b).

To address these issues, the U.S. Food and Drug Administration (FDA) launched Project Optimus (U.S. Food and Drug Administration, 2022), and released guidance (U.S. Food and Drug Administration, 2024) to shift the dose-finding goal from identifying an MTD to determining an optimal biological dose (OBD) that maximizes a risk-benefit tradeoff. Following the FDA's recommendation to randomize patients among doses, several dose optimization designs, including randomization recently have been proposed. Guo and Yuan (2023) presented a design (DROID) combining the dose-ranging framework of non-oncology trials with oncology dose-finding designs. Yang et al. (2024) developed a multiple-dose randomized trial (MERIT) design that optimizes dose based on toxicity, and provided an algorithm to determine sample size. Thall et al. (2023a) proposed a generalized phase I-II design that uses phase I-II criteria to identify a set of candidate doses based on response and toxicity, randomizes patients among the candidates, and selects the best dose based on long-term treatment success. Zang et al. (2024) extended that approach to a generalized phase I-II-III design, integrating it with a Phase III trial to further enhance the design's efficiency. See Yuan et al. (2024) for a review.

Identifying optimal doses for multiple indications is more difficult because one must account for the possibility that the indications may have different dose-outcome curves, and thus different OBDs. The FDA's guidance indicates that "Different dosages may be needed in different disease settings or oncologic diseases based on potential differences in tumor biology, patient population, treatment setting, and concurrent therapies, among other factors" (U.S. Food and Drug Administration, 2024). While a straightforward approach is to optimize dose independently for each indication, this may lead to a very large sample size.

This paper was motivated by an early phase trial at MD Anderson Cancer Center to identify OBDs of an anti-CD137 agonist in combination with pembrolizumab and nab-paclitaxel for treating metastatic solid tumors. Because the agonist induces re-

Received: June 11, 2024; Revised: August 28, 2024; Accepted: September 10, 2024
© The Author(s) 2024. Published by Oxford University Press on behalf of The International Biometric Society. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

Downloaded from https://academic.oup.com/biometrics/article/80/4/ujae105/7808593 by The University of Texas MD Anderson Cancer Center user on 03 March 2025

sponses in CD8+ T-cells, it was expected to complement and enhance the efficacy of the immune checkpoint blockade pembrolizumab. Doses of pembrolizumab and nab-paclitaxel were fixed at 200 mg and 220 mg/m^2, respectively. The MTD of the CD137 agonist was established in an all-comer dose escalation trial with several indications. The investigator was interested in conducting a dose optimization trial by randomizing patients between the MTD and a lower dose. Four indications were studied: esophageal and gastric cancer, head and neck cancer, Her2-negative breast cancer, and ovarian cancer. Since the treatment might be ineffective in some indications, one aim was to minimize the sample sizes of indications with poor results.

To efficiently identify an OBD for each indication in this setting, the two-stage basket trial design, ROMI, described in this paper was developed. Denote indications by $I_1, \cdots, I_K$ and index stages by $s = 1, 2$. We consider settings where an MTD of a new agent has been provided, possibly based on an earlier phase I trial in one I_k or all-comers. The goal is to identify an OBD for each I_k based on binary toxicity and response. Stage 1 of ROMI focuses on screening a high dose, d_H, which is the MTD that has been provided, in each I_k. Accrual to an I_k is terminated if it is found that $\pi_R(d_H)$ is unacceptably low or $\pi_T(d_H)$ is unacceptably high, compared to fixed limits specified for I_k. In stage 2, the goal is to select an OBD for each I_k, with patients randomized between d_H and a prespecified lower dose, d_L, while doing safety and futility monitoring for each dose in each I_k. To select OBDs, a ROMI design requires elicited numerical utilities of the four possible (toxicity, response) outcome pairs to compute a decision criterion. A Bayesian hierarchical model is assumed that allows the I_k's to have different OBDs, and borrows information between the I_k's. For each I_k, the OBD is the acceptable dose with maximum posterior mean utility. We present two versions of the ROMI design. The first version uses only the randomized stage 2 data to select OBDs. The second version uses the data from both stages, based on an extended hierarchical model accounting for possible bias due to drift of d_H effects between stages 1 and 2.

In Section 2, we present the first version of the ROMI design, including the hierarchical model, descriptions of each stage, an illustrative example, and guidelines for determining sample size. Section 3 presents the second version of the ROMI design, including a model elaboration to account for possible drift of d_H effects between stages. Section 4 reports simulations that evaluate the operating characteristics of the ROMI designs and compare them to designs that choose one dose for all I_k's or conduct separate trials within the I_k's. We close with a discussion in Section 5.

2 NOTATION AND DESIGN ELEMENTS

While a ROMI design can accommodate more than two doses, for simplicity and to control overall sample size, we will restrict attention to the case of two doses, $\{d_L, d_H\}$. A ROMI design with more than two doses is described in Web Appendix A. We consider settings where dose evaluation is based on binary toxicity, Y_T, and binary response, Y_R. In stage 1, all patients are treated with d_H, and I_k's for which d_H is unsafe or ineffective are screened out. I_k's passing stage 1 screening go to stage 2, where patients are randomized between d_H and d_L, and each dose is screened in each I_k. At the end of stage 2, for each I_k with at least one acceptable dose, the OBD is defined as the dose maximizing posterior mean utility.

The remainder of this section will describe the first version of ROMI, where only stage 2 data are used to choose OBDs. The second version, which uses both the stage 1 and stage 2 data to choose OBDs, is presented in Section 3.

2.1 Stage 1 dose screening

Denote the maximum stage s sample size for dose d_ℓ in I_k by $N_{\ell,k,s}$. Because only d_H is evaluated in stage 1, $N_{L,k,1} = 0$ for all k. For I_k, when the maximum sample size $N_{H,k,1}$ of d_H in stage 1 is reached, the acceptability of d_H is evaluated using two screening rules, constructed using the approach of Thall and Russell (1998) and Zhou et al. (2017), which is used by numerous designs. Let $X_{T,H,k,1}$ denote the number of toxicities and $X_{R,H,k,1}$ the number of responses among the $N_{H,k,1}$ patients with indication I_k in stage 1. Denote the stage 1 count data by $\mathcal{D}_1 = \{(N_{H,k,1}, X_{T,H,k,1}, X_{R,H,k,1}), k = 1, \ldots, K\}$, and the marginal outcome probabilities $\pi_{j,\ell,k} = \Pr(Y_j = 1 \mid d_\ell, I_k)$ for $j = R, T$, $\ell = H, L$, and $k = 1, \cdots, K$. For each I_k, $\overline{\pi}_{T,k}$ denotes a fixed maximum acceptable toxicity probability, and $\underline{\pi}_{R,k}$ a fixed minimum response probability, elicited from the clinical investigators. The values of $\overline{\pi}_{T,k}$ may be the same or similar across indications, but values of $\underline{\pi}_{R,k}$ may vary substantially with k due to qualitatively different definitions of response and therapeutic expectations across the I_k's. Accrual to I_k is terminated at the end of stage 1 if d_H is found likely to be excessively toxic, using the posterior safety criterion

$$\Pr(\pi_{T,H,k} > \overline{\pi}_{T,k} \mid \mathcal{D}_1) > c_{T,k,1}, \tag{1}$$

or if it is found likely to be inefficacious, using the posterior futility criterion

$$\Pr(\pi_{R,H,k} < \underline{\pi}_{R,k} \mid \mathcal{D}_1) > c_{R,k,1}. \tag{2}$$

The cutoffs $c_{T,k,1}$ and $c_{R,k,1}$ are fixed at values such as 0.90 or 0.95, calibrated by preliminary simulations to obtain good operating characteristics, including a high probability of stopping accrual to indications where d_H is too toxic, with $\pi^{true}_{T,H,k} > \overline{\pi}_{T,k}$, or inefficacious, with $\pi^{true}_{R,H,k} < \underline{\pi}_{R,k}$.

To evaluate posterior probabilities in the stage 1 monitoring rules (1) and (2), we assume beta-binomial models, with noninformative priors $\pi_{j,H,k} \sim Beta(0.1, 0.1)$, and likelihoods

$$X_{j,H,k,1} \mid \pi_{j,H,k} \sim Binom(N_{H,k,1}, \pi_{j,H,k}), \quad j = R, T.$$

By conjugacy, the posteriors are

$$\pi_{j,H,k} \mid \mathcal{D}_1 \sim Beta(0.1 + X_{j,H,k,1}, 0.1 + N_{H,k,1} - X_{j,H,k,1}).$$

The monitoring rules also may be applied before the end of stage 1, for example, after evaluating $N_{H,k,1}/2$ patients in I_k, and at $N_{H,k,1}$.Each I_k with acceptable response and toxicity rates for d_H at the end of stage 1 is moved to stage 2, otherwise no dose is chosen for I_k.

2.2 Stage 2 dose optimization

In stage 2, patients are randomized between d_H and d_L. The aim is to identify an OBD for each I_k, based on indication-specific utilities $U_k(y_T, y_R)$ for $y_T, y_R \in \{0, 1\}$ and $k = 1, \cdots, K$. For

Downloaded from https://academic.oup.com/biometrics/article/80/4/ujae105/7808593 by The University of Texas MD Anderson Cancer Center user on 03 March 2025

TABLE 1 Example of indication-specific utilities for two binary outcomes.

	Indication 1	
	$Y_R = 1$	$Y_R = 0$
$Y_T = 0$	$U_1(0, 1) = 100$	$U_1(0, 0) = 40$
$Y_T = 1$	$U_1(1, 1) = 60$	$U_1(1, 0) = 0$
	Indication 2	
	$Y_R = 1$	$Y_R = 0$
$Y_T = 0$	$U_2(0, 1) = 100$	$U_2(0, 0) = 20$
$Y_T = 1$	$U_2(1, 1) = 80$	$U_2(1, 0) = 0$
	Indication 3	
	$Y_R = 1$	$Y_R = 0$
$Y_T = 0$	$U_3(0, 1) = 100$	$U_3(0, 0) = 60$
$Y_T = 1$	$U_3(1, 1) = 40$	$U_3(1, 0) = 0$
	Indication 4	
	$Y_R = 1$	$Y_R = 0$
$Y_T = 0$	$U_4(0, 1) = 100$	$U_4(0, 0) = 30$
$Y_T = 1$	$U_4(1, 1) = 70$	$U_4(1, 0) = 0$

each I_k, one may establish $U_k(y_T, y_R)$ by setting $U_k(0, 1) = 100$ for the best outcome (no toxicity, response), $U_k(1, 0) = 0$ for the worst outcome (toxicity, no response), and eliciting $U_k(0, 0)$ and $U_k(1, 1)$ from the physicians. Table 1 gives a numerical example of utilities for four indications. Utility-based phase I-II designs are given by Thall and Nguyen (2012), Guo and Yuan (2017), and Zhou et al. (2019), among many others.

To do utility-based dose optimization for each I_k based on the randomized stage 2 data, denote the joint elementary outcome probabilities for dose d_ℓ in I_k by

$$p_{\ell,k}(y_T, y_R) = \Pr(Y_T = y_T, Y_R = y_R \mid d_\ell, I_k), \quad \text{for } y_T, y_R \in \{0, 1\}. \tag{3}$$

The mean utility of d_ℓ in I_k is the probability weighted average

$$\overline{U}_{\ell,k} = \sum_{y_T=0}^{1} \sum_{y_R=0}^{1} U_k(y_T, y_R)\, p_{\ell,k}(y_T, y_R). \tag{4}$$

Following the utility-based BOIN12 design (Lin et al., 2020), we take a quasi-binominal likelihood approach by defining standardized mean utilities $Q_{\ell,k} = \overline{U}_{\ell,k}/100$, called "quasi-probabilities" because they take values between 0 and 1. For each d_ℓ and I_k, let $X_{\ell,k}(y_T, y_R)$ denote the number of patients in stage 2 who experience the joint outcome (y_T, y_R), and denote the vector of counts for the four elementary outcomes by

$$\mathbf{X}_{\ell,k} = (\, X_{\ell,k}(0, 1), X_{\ell,k}(0, 0), X_{\ell,k}(1, 1), X_{\ell,k}(1, 0)\,), \tag{5}$$

with corresponding joint probability vector $\mathbf{p}_{\ell,k}$. Thus, $\mathbf{X}_{\ell,k} \sim$ *Multinomial*$(N_{\ell,k,2}, \mathbf{p}_{\ell,k})$ for each d_ℓ and I_k. Given the stage 2 data, we define normed utility-weighted average counts

$$Z_{\ell,k} = \frac{1}{100} \sum_{y_T=0}^{1} \sum_{y_R=0}^{1} U_k(y_T, y_R) X_{\ell,k}(y_T, y_R).$$

Each $Z_{\ell,k}$ has domain $(0, N_{\ell,k,2})$, and may take non-integer values. It may be interpreted as the number of "quasi-events" among the $N_{\ell,k,2}$ patients with indication I_k treated with d_ℓ in stage 2. Given the quasi-probability $Q_{\ell,k}$, we denote the distribution of $Z_{\ell,k}$ induced by the multinomial distribution of $\mathbf{X}_{\ell,k}$ by $Z_{\ell,k} \sim \text{Quasi} - \text{Binom}(N_{\ell,k,2}, Q_{\ell,k})$.

To use the stage 2 data to select OBDs, we proceed as follows. We accommodate heterogeneity among indications and facilitate borrowing information between indications by introducing a vector of *latent cluster variables* $\boldsymbol{\zeta} = (\zeta_1, \cdots, \zeta_K)$ (Chu and Yuan, 2018a; Chen and Lee, 2019; Takeda et al., 2022), where $\zeta_k = I[Q_{H,k} \le Q_{L,k}]$, the indicator that d_L has higher mean utility than d_H in I_k. Let $N(\mu, \sigma^2)$ denote a normal distribution with mean μ and variance σ^2, and $IG(a, b)$ an inverse gamma distribution with parameters a and b. Recall that $Z_{\ell,k}$ is the number of quasi-events and $Q_{\ell,k}$ is the quasi-probability for I_k and d_ℓ in stage 2. Denote $\theta_k = \text{logit}(Q_{L,k}) - \text{logit}(Q_{H,k})$, the d_L-versus-d_H effect in I_k, where $\text{logit}(q) = \log\{q/(1-q)\}$ for $q \in [0, 1]$. Thus, θ_k is a function of $\overline{U}_{L,k}$, $\overline{U}_{H,k}$, and the probability vectors $\{p_{\ell,k}\}$. For the stage 2 data, we assume the Bayesian hierarchical model

$$\begin{aligned} &Z_{\ell,k} \mid Q_{\ell,k} \sim \mathit{Quasi{-}Binom}(N_{\ell,k,2}, Q_{\ell,k}), \\ &\quad \text{for } \ell = L, H,\ k = 1, \cdots, K, \\ &\theta_k \mid \zeta_k = g \sim \mathit{iid}\, N(\mu_g, \tau^2), \quad \text{for } g = 0, 1, \\ &\quad \text{and each } k = 1, \cdots, K. \end{aligned} \tag{6}$$

For priors, we assume

$$\mu_g \sim N(\widetilde{\mu}_g, \widetilde{\tau}_g^2), \text{ for } g = 0, 1, \text{ and } \tau^2 \sim IG(a, b),$$
$$Q_{H,k} \sim \mathit{Beta}(c, d), \ \zeta_k \sim \mathit{Bernoulli}(q), \text{ and } q \sim \mathit{Beta}(e, f),$$

with $\widetilde{\mu}_g$, $\widetilde{\tau}_g^2$, a, b, c, d, e, f as fixed hyperparameters. Since $\mathbf{p}_{L,k}$ and $\mathbf{p}_{H,k}$ contribute to the stage 2 likelihood only through the quasi-probabilities $Q_{L,k}$ and $Q_{H,k}$, one only needs to specify priors on these to complete the model. Since normal priors are specified on θ_k for each k, the model is completed by specifying priors on the $Q_{H,k}$'s.

Hyperparameters may be established by applying the approach of Thall and Nguyen (2012) and Guo and Yuan (2017). To do this, expected response and toxicity probabilities are elicited from the clinicians for each combination (d_ℓ, I_k). These provide a basis for calculating a range of utility differences between d_L and d_H on the logit scale, that is, for the θ_k's. One may set $\widetilde{\mu}_0$ to the mean in the subset where $\theta_k < 0$, and set $\widetilde{\mu}_1$ to the mean in the subset where $\theta_k \ge 0$. Once $\widetilde{\mu}_0$ and $\widetilde{\mu}_1$ are established, one may assume a coefficient of variation of 2, which sets $\widetilde{\tau}_g = 2\mu_g$ (Guo and Yuan, 2017). The shrinkage parameter τ^2 can be assigned an inverse gamma prior, such as, IG(0.0001, 0.0001). Gelman (2006) and Chu and Yuan (2018b) noted that the IG(ϵ, ϵ) with $\epsilon \to 0$ does not represent a non-informative prior, but instead imposes strong shrinkage when the number of elements in the hierarchy (indications in our context) is small (eg, ≤ 6) unless the heterogeneity between indications is extremely large. Under our model, this potential problem is mitigated by using $\boldsymbol{\zeta}$ to partition the indications into $\mathcal{I}^0$ and $\mathcal{I}^1$. Since indications in each of these subsets are likely to be homogeneous, the strong shrinkage effect of the prior often enhances the model's performance. As a sensitivity analysis, we consider a Half-Cauchy distribution prior for τ^2 in Section 4.3.

For each I_k where d_H passes the stage 1 screening, in stage 2 patients are randomized between d_H and d_L. If R interim screening analyses are carried out for I_k in stage 2, let $n_{\ell,k,2,r}$ denote the interim sample size for the k^{th} indication at the r^{th} stage 2 look.

Downloaded from https://academic.oup.com/biometrics/article/80/4/ujae105/7808593 by The University of Texas MD Anderson Cancer Center user on 03 March 2025

Let $\mathcal{D}_{2,r}$ denote the data at r^{th} interim look, and $\mathcal{D}_2$ the final data from stage 2. At the r^{th} interim analysis, (Y_T, Y_R) are evaluated for all patients treated at each dose, and a dose is terminated if it is excessively toxic per criteria (1) or ineffective per criteria (2). To reduce bias, futility monitoring relies solely on the stage 2 data. In contrast, safety monitoring pools the stage 1 and stage 2 data, assuming toxicity probabilities will not change between stages.

At the end of stage 2, for each I_k, when the maximum stage 2 sample sizes $N_{L,k,2}$ and $N_{H,k,2}$ are reached for the two doses, a final analysis is conducted to determine the OBD. The toxicity monitoring rule (1) is applied for each dose based on $\mathcal{D}_1 \cup \mathcal{D}_2$, and futility monitoring is done based on the stage 2 data using the rule $\Pr(\pi_{R,\ell,k} < \underline{\pi}_{R,k} | \mathcal{D}_2) > c_{R,k,2}$. For I_k, the OBD is the dose that passes both the toxicity and response requirements and maximizes the posterior mean standardized utility, estimated under the Bayesian hierarchical model. The dose optimization criterion in I_k is denoted by

$$\begin{aligned} \text{OBD}_k &= \underset{\ell=L,H}{\text{argmax}}\ \widehat{Q}_{\ell,k} \\ &= \underset{\ell=L,H}{\text{argmax}}\ E\{Q_{\ell,k} \mid \mathcal{D}_2\}. \end{aligned} \tag{7}$$

2.3 Graphical illustration of trial conduct

Figure 1 presents a schematic of trial conduct using the ROMI design to determine the OBD, if it exists, between two doses d_L and d_H for each of four disease subtypes (indications). In stage 1, all patients are treated with d_H, and toxicity and response are monitored for each I_k. Due to an unacceptably low response rate with d_H, I_1 is dropped, while I_2, I_3, and I_4 are moved forward to stage 2, where patients are randomized between d_H and d_L. A final analysis is conducted to evaluate each dose's safety, response rate, and mean utility. For I_2, both doses have acceptable toxicity and response rates, with d_L selected as the OBD based on posterior mean utility. For I_3 and I_4, d_H is selected as the OBD due to its higher posterior mean utility. Thus, the ROMI design does not identify an OBD for I_1, identifies d_L as the OBD for I_2, and identifies d_H as the OBD for I_3 and I_4.

2.4 Sample size determination

The sample size for each I_k in stage 1 of a ROMI design may be determined to control the false negative decision probability of the futility stopping rule (2). To do this, suppose that, for each I_k, a desirably high response probability $\underline{\pi}_{R,k} + \delta_{R,k}$ can be specified, say for $\delta_{R,k} = 0.15$, 0.20, or 0.25. The cut-off $c_{R,k,1}$ and sample size $N_{H,k,1}$ may be calibrated together by simulation so that, for true response probability $\pi_{R,k}^{true} = \underline{\pi}_{R,k} + \delta_{R,k}$, the false negative early stopping probability is no larger than a specified small value, such as 0.10 or 0.05. In practice, one may fix $c_{R,k,1}$ at a large value, such as 0.90 or 0.95, and do a monotone search for the smallest $N_{H,k,1}$ that ensures the specified false negative early stopping probability.

To determine the sample size for each indication in stage 2, one can first apply the MERIT design (Yang et al., 2024), which gives a structured approach for calculating sample size in randomized phase II dose optimization studies. To do this, for each I_k, one may begin by specifying the lower limit $\underline{\pi}_{R,k}$, a desirably high response probability $\underline{\pi}_{R,k} + \delta_{R,k}$ with $\delta_{R,k} = 0.15$, 0.20, or 0.25 as above, an upper toxicity probability limit $\overline{\pi}_{T,k}$, and a desirably low toxicity probability $\overline{\pi}_{T,k} - \delta_{T,k}$. One then specifies a maximum level, such as 0.10 or 0.15, for the probability of incorrectly accepting an undesirable dose (type I error rate), and a minimum level, such as 0.60, 0.70, or 0.80, for the probability of correctly choosing an acceptable dose (power). The MERIT sample size $N_{\ell,k,2}^M$ for dose d_ℓ and indication I_k may be determined by a numerical search, to find the smallest value that controls the type I error while achieving the desired power. Since MERIT assumes equal randomization, for a ROMI design, one may restrict the randomization by requiring $N_{H,k,2}^M = N_{L,k,2}^M$. Software for calculating sample size using the MERIT design is available at Trial Design (2024).

The MERIT design method may be used to determine the sample size for each indication independently. Compared to a randomized trial assuming homogeneity, however, the ROMI design allows information borrowing between indications, which may reduce the planned overall sample size while still preserving a given level of accuracy in selecting the OBDs at the end of the trial. To exploit this, the stage 2 sample sizes $\{N_{L,k,2}^M\}$ and $\{N_{H,k,2}^M\}$ obtained from the MERIT design may be adjusted by simulating the ROMI design to achieve the desired level of reliability in the final dose selections. For example, with $K = 2$ indications and initial stage 2 sample sizes $(N_{\ell,1,2}^M, N_{\ell,2,2}^M) = (30, 25)$, simulations of the trial using the ROMI design can be conducted with specified stage 1 sample sizes $\{N_{H,k,1}\}$, determined as described above, and several combinations of stage 2 sample sizes, for example, $(N_{\ell,1,2}, N_{\ell,2,2}) = (30, 25), (25, 25), (20, 25), (30, 20), (25, 20), (20, 20)$, to assess operating characteristics. The sample size chosen for stage 2 is based on the tradeoff between the accuracy of the final OBD selection for each I_k and total trial sample size $N = \sum_{k=1}^{K}(N_{H,k,1} + N_{H,k,2} + N_{L,k,2})$. If desired, the $\{N_{H,k,1}\}$ values may be adjusted and the trial simulations repeated.

3 USING DATA FROM BOTH STAGES

Combining data on d_H from both stage 1 and stage 2 may improve the estimate of d_H-versus-d_L effects for the OBD selection in each indication. This is straightforward when it is reasonable to assume that the data from stages 1 and 2 are exchangeable: simply pool the data from both stages when calculating $\mathbf{X}_{\ell,k}$ in (5). However, since there is no randomization in stage 1, and patients are randomized to d_H or d_L in stage 2, there might be drift in the effect of d_H on the outcomes between stages, possibly due to temporal changes in patient characteristics or unknown factors. In this case, simply pooling the data results in bias.

To include stage 1 data on d_H and account for potential temporal drift, we extend the Bayesian hierarchical model, referred to as version 2 of ROMI. The joint distributions $p_{H,k}(y_R, y_T)$, defined earlier, are elaborated to be stage-specific distributions $p_{H,k,s}(y_R, y_T)$ for $s = 1$ and 2 and all I_k. This produces stage-specific mean utilities $\overline{U}_{H,k,1}$ and $\overline{U}_{H,k,2}$, quasi-probabilities $Q_{H,k,1}$ and $Q_{H,k,2}$, and between-dose effects $\theta_{k,s} = \text{logit}(Q_{L,k,s}) - \text{logit}(Q_{H,k,s})$. Since no patients are treated with d_L in stage 1, however, for the stage 2 selection only $\theta_{k,2}$ is relevant for each I_k. We account for the stage by letting $Z_{\ell,k,s}$ denote

Downloaded from https://academic.oup.com/biometrics/article/80/4/ujae105/7808593 by The University of Texas MD Anderson Cancer Center user on 03 March 2025

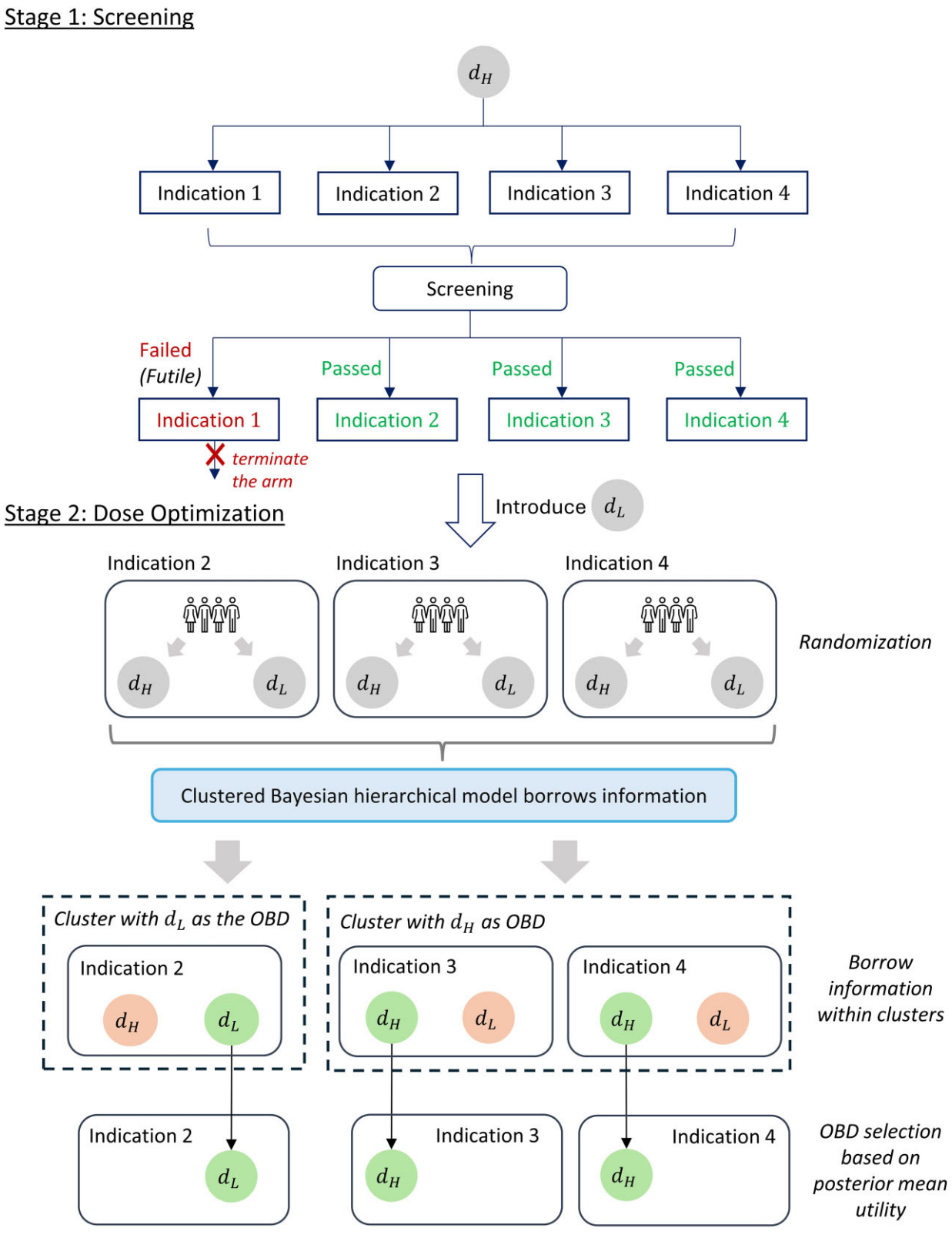

FIGURE 1 A ROMI design example with four indications and two doses, d_H and d_L. OBDs are indicated by green circles.

Downloaded from https://academic.oup.com/biometrics/article/80/4/ujae105/7808593 by The University of Texas MD Anderson Cancer Center user on 03 March 2025

the number of quasi-events and $Q_{\ell,k,s}$ the standardized utility for I_k at dose d_ℓ in stage $s = 1$ or 2. Because d_L is not evaluated in stage 1, each $Z_{L,k,1} = 0$. Thus, for d_L, only $Z_{L,1,2}, \cdots, Z_{L,K,2}$ are defined and used in the stage 2 decisions.

To model stage 1 data on d_H and stage 2 data on $\{d_L, d_H\}$, we assume an extended Bayesian hierarchical model that accounts for the use of stage 1 quasi-event values $Z_{H,k,1}$ with the stage 2 values $Z_{L,k,2}$ and $Z_{H,k,2}$. For each I_k, denoting the drift parameter $\beta_k = \text{logit}(Q_{H,k,1}) - \text{logit}(Q_{H,k,2})$, we assume

$$
\begin{aligned}
&Z_{H,k,1} \mid Q_{H,k,1} \sim Quasi\text{-}Binom(N_{H,k,1}, Q_{H,k,1}), \quad \textbf{(Stage1)}\\
&Z_{\ell,k,2} \mid Q_{\ell,k,2} \sim Quasi\text{-}Binom(N_{\ell,k,2}, Q_{\ell,k,2}), \quad \textbf{(Stage2)}\\
&\quad \text{for } \ell = L, H,\\
&\theta_{k,2} = \text{logit}(Q_{L,k,2}) - \text{logit}(Q_{H,k,2}),\\
&\theta_{k,2} \mid \zeta_k = g \sim iid\, N(\mu_g, \tau^2), \quad \text{for } g = 0, 1, \qquad (8)
\end{aligned}
$$

with priors

$$
\begin{aligned}
&\beta_k \sim \omega N(0, \sigma^2_{spike}) + (1-\omega)N(0, \sigma^2_{slab}),\\
&\mu_g \sim N(\widetilde{\mu}_g, \widetilde{\tau}^2_g), \quad \text{for } g = 0, 1, \quad \tau^2 \sim IG(a, b),\\
&\quad \text{and } \omega \sim U[0, 1],\\
&Q_{H,k,2} \sim Beta(c, d), \quad \zeta_k \sim Bernoulli(q),\\
&\quad \text{and } q \sim Beta(e, f).
\end{aligned}
$$

The variance σ^2_{spike} should be set to a small value, such as 0.01, to concentrate the prior spike's mass near 0, while σ^2_{slab} should be much larger than σ^2_{spike} to allow a broader range of non-zero values for β_k. Following Gelman et al. (2008) and Guo and Yuan (2017), we regularize the prior so that the typical variation of an input variable is unlikely to cause a dramatic change in the response variable. For example, $\beta_k = 1$ corresponds to between-stage drift in $Q_{\ell,k}$ from 0.30 to 0.54. Based on the utility of I_1 in Table 1, a change of 0.24 in $Q_{\ell,k}$ corresponds to large shifts of 0.6 in $\pi_{T,\ell,k}$ or of 0.4 in $\pi_{R,\ell,k}$. Since it is very unlikely that between-stage drift would induce such large changes in the $\pi_{j,\ell,k}$'s, we set $\sigma^2_{slab} = 0.5^2$ to ensure that a change in β_k from one standard deviation (sd) below to one sd above the mean is unlikely to cause a change of $Q_{\ell,k}$ exceeding 0.24.

Decision rules for version 2 of ROMI are as in Section 2.2. The only difference is that the posterior mean of the standardized utility is estimated under the extended model (8), using data from both stages and accounting for possible drift of d_H effects between stages.

4 SIMULATION STUDIES

4.1 Simulation settings

This section reports simulations to evaluate operating characteristics of the ROMI designs, and designs that either ignore the I_k's or conduct separate trials within I_k's. We consider settings with $K = 4$, using dose acceptability limits $\overline{\pi}_{T,k} = 0.40$ and $\underline{\pi}_{R,k} = 0.25$ for all k. For each I_k, the maximum stage 1 sample size is 14, and the maximum stage 2 sample size per dose is 20, with one interim analysis performed when the sample size for each dose reaches 10. We constructed scenarios by varying the number of effective I_k's and the OBD for each I_k. The utility table used for all I_k's corresponds to that given for I_1 in Table 1. To characterize association between Y_R and Y_T, for each dose $\ell = L, H$ and I_k, given marginal probabilities $\pi_{T,\ell,k}$ and $\pi_{R,\ell,k}$, we solved for the joint probabilities $\{p_{\ell,k}(y_T, y_R)\}$ so that

$$\phi = \frac{p_{\ell,k}(0,0)p_{\ell,k}(1,1) - p_{\ell,k}(1,0)p_{\ell,k}(0,1)}{\{\pi_{R,\ell,k}(1-\pi_{R,\ell,k})\pi_{T,\ell,k}(1-\pi_{T,\ell,k})\}^{1/2}} = .25.$$

We set $\widetilde{\mu}_0 = -0.05$, $\widetilde{\mu}_1 = 0.05$, $\widetilde{\tau}_0 = \widetilde{\tau}_1 = c = d = e = f = 0.1$, $\tau^2 \sim IG(0.0001, 0.0001)$, $\sigma^2_{spike} = 0.01$, and $\sigma^2_{slab} = 0.5^2$.

We denote the first version of ROMI design, which uses only stage 2 data for dose optimization, by ROMI-v1, and the second version, which uses data from both stages to optimize dose, by ROMI-v2. To assess the impact of clustering I_k's showing similar dose-outcome probabilities, we define the ROMI-v1-NC design to have the same structure as ROMI-v1 but using the Bayesian hierarchical model without clustering. The first comparator is the Pool design, which ignores I_k's and determines the same OBD for all I_k's based on the utility under a beta-binominal model, $Z_\ell \sim Binom(\sum_k n_{\ell,k}, Q_\ell)$ with a conjugate prior $Q_\ell \sim Beta(0.1, 0.1)$. The second comparator is the Independent design, a two-dose randomized design done independently for each I_k, with the utility of each arm modeled using a beta-binominal model, $Z_{\ell,k} \sim Binom(n_{\ell,k}, Q_{\ell,k})$ with a conjugate prior $Q_{\ell,k} \sim Beta(0.1, 0.1)$.

For a fair comparison, the total maximum sample size for all designs was set to $N = 216$. In the Independent design, patients within each $\{I_1, I_2, I_3, I_4\}$ were randomized between the two doses, with a maximum of 27 patients per dose. For each I_k, one interim analysis was conducted after 14 patients. For the Pool design, one interim analysis was conducted when 108 patients were evaluated. The same interim stopping rules were used for all designs, with cutoffs set to $c_{T,k,1} = c_{R,k,1} = c_{R,k,2} = 0.95$. A total of 2000 simulations were conducted for each combination of design and scenario.

4.2 Simulation results

Table 2 summarizes simulation results of the Pool, Independent, ROMI-v1-NC, ROMI-v1, and ROMI-v2 designs across 11 scenarios, assuming no drift in the effect of d_H between stages. In scenario 1, where no doses are effective for any I_k, the Pool design correctly stops all trials with no OBD selected for any I_k 100% of the time. For each I_k, the stopping percentage with no dose selected is 100 − (% select d_H + % select d_L). The stopping percentage is about 94% for the Independent design and 98% for designs using the ROMI framework, including ROMI-v1-NC, ROMI-v1, and ROMI-v2. Compared to the Pool and Independent designs, the ROMI designs provide substantial sample size savings, with about 42 fewer subjects than the Pool design and 56 fewer than the Independent design. This large sample size reduction for the ROMI designs in scenario 1, where neither dose is effective, is due to the interim screening rule for d_H applied by the ROMI designs after stage 1.

In scenarios 2 and 3, only I_1 responds to treatment. In scenario 2, d_H is the true OBD for I_1. ROMI-v2 and Independent design have the highest OBD correct selection percentages (CSPs),

Downloaded from https://academic.oup.com/biometrics/article/80/4/ujae105/7808593 by The University of Texas MD Anderson Cancer Center user on 03 March 2025

TABLE 2 Simulation results for the Pool, Independent, ROMI-v1-NC, ROMI-v1, and ROMI-v2 designs.

		Probability (%) of selecting the dose as OBD									
		I_1		I_2		I_3		I_4			
Design		d_H	d_L	d_H	d_L	d_H	d_L	d_H	d_L	**CSP**	**N**
Scenario 1											
	$\pi_{T,\ell,k}^{true}$	0.40	0.30	0.40	0.30	0.40	0.30	0.40	0.30		
	$\pi_{R,\ell,k}^{true}$	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05		
	$\overline{U}_{\ell,k}^{true}$	27	31	27	31	27	31	27	31		
Pool		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	NA	113
Independent		2.1	3.3	2.5	4.0	2.5	2.8	2.9	3.3	NA	127
ROMI-v1-NC		1.3	0.9	0.5	0.6	0.6	0.7	1.3	0.7	NA	71
ROMI-v1		1.3	0.9	0.5	0.6	0.6	0.7	1.3	0.7	NA	71
ROMI-v2		1.3	0.9	0.5	0.7	0.7	0.8	1.3	0.8	NA	71
Scenario 2											
	$\pi_{T,\ell,k}^{true}$	**0.2**	0.15	0.40	0.30	0.40	0.30	0.40	0.30		
	$\pi_{R,\ell,k}^{true}$	**0.4**	0.3	0.05	0.05	0.05	0.05	0.05	0.05		
	$\overline{U}_{\ell,k}^{true}$	**56**	52	27	31	27	31	27	31		
Pool		**5.5**	0.8	5.5	0.8	5.5	0.8	5.5	0.8	5.5	128
Independent		**69.8**	30.2	2.6	3.1	2.8	2.8	3.0	3.3	69.8	149
ROMI-v1-NC		**64.5**	35.0	0.7	1.0	0.9	0.8	1.0	0.7	64.5	107
ROMI-v1		**65.0**	34.4	0.7	1.0	0.9	0.8	1.0	0.7	65.0	107
ROMI-v2		**69.7**	29.8	0.7	1.1	0.9	0.8	1.0	0.8	69.7	107
Scenario 3											
	$\pi_{T,\ell,k}^{true}$	0.25	**0.15**	0.40	0.30	0.40	0.30	0.40	0.30		
	$\pi_{R,\ell,k}^{true}$	0.4	**0.40**	0.05	0.05	0.05	0.05	0.05	0.05		
	$\overline{U}_{\ell,k}^{true}$	54	**58**	27	31	27	31	27	31		
Pool		4.0	**4.3**	4.0	4.3	4.0	4.3	4.0	4.3	4.3	135
Independent		32.1	**68.0**	2.4	3.3	2.7	3.3	2.6	3.1	68.0	149
ROMI-v1-NC		30.6	**68.2**	0.9	0.5	0.9	0.6	1.1	1.0	68.2	107
ROMI-v1		30.6	**68.3**	0.9	0.5	0.9	0.6	1.1	1.0	68.3	107
ROMI-v2		30.0	**68.9**	0.9	0.6	0.9	0.7	1.2	1.0	68.9	107
Scenario 4											
	$\pi_{T,\ell,k}^{true}$	**0.2**	0.15	0.40	0.30	0.40	0.30	**0.20**	0.15		
	$\pi_{R,\ell,k}^{true}$	**0.4**	0.3	0.05	0.05	0.05	0.05	**0.40**	0.30		
	$\overline{U}_{\ell,k}^{true}$	**56**	52	27	31	27	31	**56**	52		
Pool		**63.0**	24.4	63.0	24.4	63.0	24.4	**63.0**	24.4	63.0	185
Independent		**68.7**	31.3	2.4	3.2	2.7	3.3	**69.3**	30.7	69.0	170
ROMI-v1-NC		**72.6**	27.0	0.7	0.8	1.0	1.1	**70.6**	28.7	71.6	143
ROMI-v1		**70.3**	29.2	0.7	0.8	1.0	1.1	**66.5**	32.7	68.4	143
ROMI-v2		**72.6**	26.9	0.8	0.8	1.0	1.1	**68.8**	30.5	70.7	143
Scenario 5											
	$\pi_{T,\ell,k}^{true}$	0.25	**0.15**	0.40	0.30	0.40	0.30	0.25	**0.15**		
	$\pi_{R,\ell,k}^{true}$	0.4	**0.4**	0.05	0.05	0.05	0.05	0.40	**0.40**		
	$\overline{U}_{\ell,k}^{true}$	54	**58**	27	31	27	31	54	**58**		
Pool		24.4	**71.7**	24.4	71.7	24.4	71.7	24.4	**71.7**	71.7	202
Independent		32.1	**67.9**	2.4	2.9	2.7	3.3	30.3	**69.8**	68.8	171
ROMI-v1-NC		27.2	**71.7**	0.5	1.2	1.1	0.9	27.4	**71.6**	71.7	143
ROMI-v1		31.4	**67.4**	0.5	1.2	1.1	0.9	29.9	**69.2**	68.3	143
ROMI-v2		31.0	**68.0**	0.5	1.2	1.1	0.9	28.6	**70.5**	69.2	143
Scenario 6											
	$\pi_{T,\ell,k}^{true}$	0.40	0.30	**0.20**	0.15	**0.20**	0.15	**0.20**	0.15		
	$\pi_{R,\ell,k}^{true}$	0.05	0.05	**0.40**	0.30	**0.40**	0.30	**0.40**	0.30		
	$\overline{U}_{\ell,k}^{true}$	27	31	**56**	52	**56**	52	**56**	52		
Pool		71.4	28.6	**71.4**	28.6	**71.4**	28.6	**71.4**	28.6	71.4	210
Independent		3.3	3.2	**70.1**	29.9	**67.9**	32.1	**68.2**	31.9	68.7	192
ROMI-v1-NC		1.2	1.0	**74.0**	25.4	**73.9**	25.2	**73.2**	26.0	73.7	178
ROMI-v1		1.2	1.0	**69.6**	29.8	**68.0**	31.0	**67.8**	31.3	68.5	178
ROMI-v2		1.2	1.0	**72.0**	27.4	**70.8**	28.3	**71.1**	28.1	71.3	178

Downloaded from https://academic.oup.com/biometrics/article/80/4/ujae105/7808593 by The University of Texas MD Anderson Cancer Center user on 03 March 2025

TABLE 2 Continued

Design		Probability (%) of selecting the dose as OBD								CSP	N
		I_1		I_2		I_3		I_4			
		d_H	d_L	d_H	d_L	d_H	d_L	d_H	d_L		
Scenario 7											
	$\pi_{T,\ell,k}^{true}$	0.40	0.30	0.25	**0.15**	0.25	**0.15**	0.25	**0.15**		
	$\pi_{R,\ell,k}^{true}$	0.05	0.05	0.40	**0.40**	0.40	**0.40**	0.40	**0.40**		
	$\overline{U}_{\ell,k}^{true}$	27	31	54	**58**	54	**58**	54	**58**		
Pool		14.8	85.3	14.8	**85.3**	14.8	**85.3**	14.8	**85.3**	85.3	216
Independent		2.7	3.4	31.8	**68.2**	32.0	**68.0**	31.2	**68.9**	68.4	194
ROMI-v1-NC		1.3	1.0	25.4	**73.9**	24.9	**74.0**	23.9	**75.2**	74.4	179
ROMI-v1		1.3	1.0	31.2	**68.0**	30.2	**68.6**	28.0	**71.2**	69.3	179
ROMI-v2		1.3	0.9	28.6	**70.7**	29.0	**69.9**	26.0	**73.2**	71.2	179
Scenario 8											
	$\pi_{T,\ell,k}^{true}$	0.40	0.30	**0.20**	0.15	0.25	**0.15**	0.25	**0.15**		
	$\pi_{R,\ell,k}^{true}$	0.05	0.05	**0.40**	0.30	0.4	**0.40**	0.40	**0.40**		
	$\overline{U}_{\ell,k}^{true}$	27	31	**56**	52	54	**58**	54	**58**		
Pool		30.4	69.7	**30.4**	69.7	30.4	**69.7**	30.4	**69.7**	56.6	215
Independent		2.7	3.6	**68.5**	31.5	31.3	**68.7**	32.2	**67.8**	68.3	193
ROMI-v1-NC		1.4	1.0	**47.7**	51.6	36.2	**62.7**	35.8	**63.3**	57.9	179
ROMI-v1		1.4	1.0	**61.4**	38.0	33.7	**65.2**	33.0	**66.2**	64.3	179
ROMI-v2		1.4	1.0	**62.5**	36.9	32.3	**66.7**	31.1	**68.1**	65.7	179
Scenario 9											
	$\pi_{T,\ell,k}^{true}$	**0.2**	0.15	**0.2**	0.15	**0.2**	0.15	**0.2**	0.15		
	$\pi_{R,\ell,k}^{true}$	**0.4**	0.3	**0.4**	0.3	**0.4**	0.3	**0.4**	0.3		
	$\overline{U}_{\ell,k}^{true}$	**56**	52	**56**	52	**56**	52	**56**	52		
Pool		**81.8**	18.2	**81.8**	18.2	**81.8**	18.2	**81.8**	18.2	81.8	216
Independent		**69.9**	30.1	**69.4**	30.6	**68.6**	31.5	**67.3**	32.7	68.8	214
ROMI-v1-NC		**77.8**	21.6	**78.3**	21.1	**77.7**	21.4	**77.4**	21.8	77.8	214
ROMI-v1		**71.7**	27.8	**70.9**	28.6	**70.8**	28.4	**71.2**	28.0	71.2	214
ROMI-v2		**75.1**	24.4	**74.4**	25.1	**74.4**	24.7	**74.0**	25.2	74.5	214
Scenario 10											
	$\pi_{T,\ell,k}^{true}$	0.25	**0.15**	0.25	**0.15**	0.25	**0.15**	0.25	**0.15**		
	$\pi_{R,\ell,k}^{true}$	0.40	**0.40**	0.40	**0.40**	0.40	**0.40**	0.4	**0.40**		
	$\overline{U}_{\ell,k}^{true}$	54	**58**	54	**58**	54	**58**	54	**58**		
Pool		16.9	**83.1**	16.9	**83.1**	16.9	**83.1**	16.9	**83.1**	83.1	216
Independent		33.0	**67.0**	31.6	**68.4**	33.2	**66.8**	31.3	**68.7**	67.7	216
ROMI-v1-NC		21.3	**77.6**	21.3	**77.8**	22.0	**76.8**	22.2	**76.9**	77.3	214
ROMI-v1		27.0	**71.9**	27.0	**72.2**	29.0	**69.8**	27.0	**72.2**	71.5	214
ROMI-v2		26.0	**72.9**	24.0	**75.2**	26.5	**72.4**	24.3	**74.9**	73.8	214
Scenario 11											
	$\pi_{T,\ell,k}^{true}$	**0.20**	0.15	**0.20**	0.15	0.25	**0.15**	0.25	**0.15**		
	$\pi_{R,\ell,k}^{true}$	**0.40**	0.30	**0.40**	0.30	0.40	**0.40**	0.40	**0.40**		
	$\overline{U}_{\ell,k}^{true}$	**56**	52	**56**	52	54	**58**	54	**58**		
Pool		**50.9**	49.1	**50.9**	49.1	50.9	**49.1**	50.9	**49.1**	50.0	216
Independent		**69.2**	30.8	**68.9**	31.1	30.2	**69.8**	31.6	**68.5**	69.1	215
ROMI-v1-NC		**54.9**	44.4	**54.3**	45.1	43.8	**55.1**	42.4	**56.7**	55.3	214
ROMI-v1		**63.4**	36.0	**62.1**	37.4	36.0	**62.9**	36.0	**63.1**	62.9	214
ROMI-v2		**64.3**	35.1	**64.2**	35.2	35.6	**63.4**	35.2	**64.0**	64.0	214

Abbreviations: CSP: correct selection percentage; N: average total sample size.
Values for the true OBD of each indication are given in boldface. Doses are indexed by $\ell = L, H$ and indications by $k = 1, 2, 3, 4$.

69.7% and 69.8%, respectively. The ROMI-v1 and ROMI-v1-NC designs have CSPs 4.7% and 5.2% lower than ROMI-v2. The Pool design, which ignores indications, stopped 93.7% of trials with a CSP of just 5.5%. Compared to the Independent design, the ROMI designs save 42 subjects on average. A similar sample size saving is seen in scenario 3, where the true OBD for I_1 is d_L. In this case, the Pool design has a very low CSP of 4.3%, while the Independent and ROMI designs have similar CSPs of around 68%.

In scenarios 4 and 5, two indications respond to the treatment. In scenario 4, where the true OBD is d_H for I_1 and I_4, the ROMI designs outperform the Pool and Independent designs in both CSP and sample size saving. ROMI-v2 has a CSP of 70.7%, comparable to the highest CSP of 71.6% achieved by ROMI-v1-NC.

Downloaded from https://academic.oup.com/biometrics/article/80/4/ujae105/7808593 by The University of Texas MD Anderson Cancer Center user on 03 March 2025

TABLE 3 Sensitivity analysis of ROMI designs with efficacy drift of d_H effects between stage 1 and stage 2.

	Probability (%) of selecting the dose as OBD								
	I_1		I_2		I_3		I_4		
Design	d_H	d_L	d_H	d_L	d_H	d_L	d_H	d_L	**CSP**
Positive Drift									
Scenario 9									
ROMI-v1	**72.0**	27.0	**71.8**	27.0	**71.3**	27.2	**72.1**	26.6	71.8
ROMI-v2	**72.1**	26.9	**71.5**	27.4	**72.1**	26.4	**71.7**	27.0	71.8
Scenario 10									
ROMI-v1	28.3	**70.3**	27.2	**71.7**	27.4	**70.7**	27.0	**71.3**	71.0
ROMI-v2	23.6	**75.0**	22.2	**76.8**	22.8	**75.2**	21.9	**76.4**	75.8
Scenario 11									
ROMI-v1	**62.8**	36.0	**61.6**	37.2	35.4	**62.7**	36.6	**61.7**	62.2
ROMI-v2	**60.6**	38.3	**59.9**	38.9	31.7	**66.5**	31.5	**66.8**	63.4
Negative Drift									
Scenario 9									
ROMI-v1	**71.8**	27.8	**71.4**	28.1	**71.7**	27.8	**71.9**	27.5	71.7
ROMI-v2	**77.3**	22.4	**77.4**	22.1	**78.3**	21.2	**77.4**	22.0	77.6
Scenario 10									
ROMI-v1	27.4	**71.9**	26.4	**72.8**	28.4	**71.0**	25.6	**73.9**	72.4
ROMI-v2	28.2	**71.1**	25.1	**74.2**	27.4	**72.0**	25.4	**74.1**	72.8
Scenario 11									
ROMI-v1	**64.9**	34.6	**62.2**	37.4	36.1	**63.2**	36.8	**62.6**	63.2
ROMI-v2	**68.5**	31.1	**67.0**	32.6	39.5	**59.9**	38.9	**60.6**	64.0

Abbreviations: CSP: correct selection percentage.
Values for the true OBD of each indication are given in boldface. Doses are indexed by $\ell = L, H$ and indications by $k = 1, 2, 3, 4$.

It is about 2% higher than ROMI-v1 and Independent, and 7.7% higher than Pool. The ROMI designs save an average of 27 subjects compared to the Independent design and 42 subjects compared to the Pool design. In scenario 5, where d_L is the true OBD, ROMI designs save 28 subjects compared to the Independent design and up to 59 compared to the Pool design. The CSP of ROMI-v2 is 69.2%, and ROMI-v1 is 68.3%, similar to Independent but about 2.5% lower than ROMI-v1-NC and Pool, both with a CSP of 71.7%. Since the Pool design ignores indications and selects the same OBD for all I_k, it has high false-positive rates of selecting ineffective doses for non-responsive indications. In scenario 4, Pool selects an ineffective dose for I_2 and I_3 87.4% of the time, rising to 96.1% in scenario 5. In contrast, the probability of selecting an ineffective dose for these indications is 5.6% with Independent and 1.8% with ROMI designs.

Across scenarios 6, 7, and 8, where I_1 is insensitive to treatment, the ROMI designs save an average of 14 subjects compared to the Independent design and 35 compared to the Pool design. In scenario 6, d_H is the true OBD for I_2, I_3, and I_4, while in scenario 7, d_L is the OBD. Under heterogeneous scenarios where some indications are non-responsive and responsive I_k's share the same OBD, ROMI-v1-NC, and ROMI-v2 show larger CSPs compared to the Independent design, with increases of 5% and 2.5% in scenario 6, and 6% and 2.8% in scenario 7. These improvements show the benefit of borrowing information across indications. ROMI-v1 has a CSP similar to the independent design. The Pool design is effective in selecting the OBD for responsive indications but fails to terminate ineffective doses for non-responsive I_1, with a 100% chance of choosing an ineffective dose. ROMI-v1-NC has the highest CSP due to the strong shrinkage but underperforms in scenarios where the OBD varies across responsive indications, such as scenario 8. In scenario 8, d_H is the true OBD for I_2, while d_L is the true OBD for I_3 and I_4. The CSPs of ROMI-v1 and ROMI-v2 are 4% and 2.6% lower than the Independent design but outperform ROMI-v1-NC, with CSP improvements of 6.4% and 7.8%, respectively. This shows the benefit of clustering indications under the Bayesian hierarchical model in ROMI-v1 and ROMI-v2. The Pool design has the lowest CSP, about 56.6%, and the highest probability of selecting an ineffective dose for I_1.

The advantage of information borrowing increases with the number of responsive indications, shown by scenarios 9, 10, and 11, where all indications respond to treatment, resulting in comparable sample sizes across all designs. In homogeneous scenarios where OBDs are consistent across indications, the Pool and ROMI designs have higher CSPs than the Independent design. For example, in scenario 9, ROMI-v1 shows a 2.4% increase in CSP, ROMI-v2 a 5.7% increase, and ROMI-v1-NC a 9% increase, compared to Independent. The Pool design has the highest CSP of 81.8%, essentially because its homogeneity assumption happens to be correct in this scenario. In the heterogeneous[2] scenarios, the OBD varies across responsive indications. For example, in scenario 11, d_H is the true OBD for I_1 and I_2, and d_L is the true OBD for I_3 and I_4. ROMI-v1 and ROMI-v2 have CSPs about 5% lower than the Independent design, but outperform ROMI-v1-NC by 9%. The Pool design shows the poorest performance, correctly selecting the OBD with only a 50% CSP.

Downloaded from https://academic.oup.com/biometrics/article/80/4/ujae105/7808593 by The University of Texas MD Anderson Cancer Center user on 03 March 2025

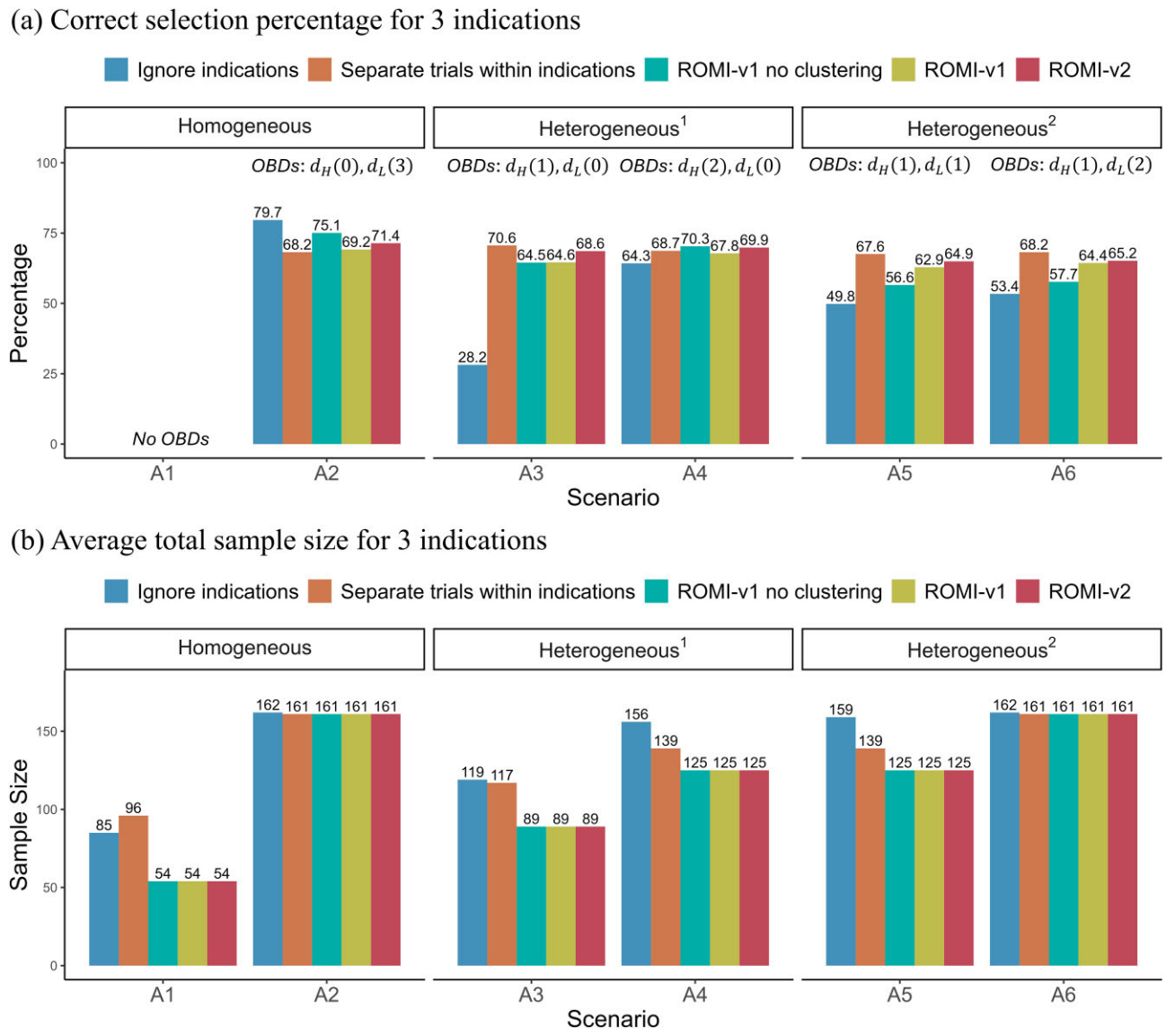

FIGURE 2 For three indications, (a) correct selection percentages and (b) average total sample sizes for the Pool design that ignores indications, Independent design that conducts separate trials within indications, ROMI-v1 with no indication clustering, ROMI-v1 with clustering, and ROMI-v2 with clustering. In homogeneous scenarios, OBDs are identical across indications. Heterogeneous[1] scenarios include some non-responsive indications and identical OBDs among responsive indications. In Heterogeneous[2] scenarios, OBDs vary among responsive indications.

4.3 Sensitivity analyses

We examined the performance of ROMI-v1 and ROMI-v2 in the presence of $\pi_{R,H,k}$ drift for d_H between stages, exploring the impacts of both positive and negative drifts. Table 3 gives simulation results where $\pi_{R,H,k}$ increased by 0.025 from stage 1 to stage 2 in the upper portion of the table, and decreased by 0.025 in the lower portion. This increment corresponds to 25% of the maximum $\pi_{R,H,k} - \pi_{R,L,k}$ difference of .10 in our simulation settings. In each of scenarios 9–11, all I_k's are responsive to both d_H and d_L. Compared to ROMI-v1, ROMI-v2 demonstrates similar or better accuracy in selecting OBD across all scenarios. Thus, ROMI-v2 does a good job of handling drift in response rates between stages.

We also evaluated the performance of the ROMI designs for a trial with either $K = 3$ or $K = 6$ indications, illustrated in Figures 2 and 3. The ROMI designs reduce sample size compared to the Pool and Independent designs when some I_k's are non-responsive to treatment. As expected, the Pool design has the highest CSP when all indications have the same dose-outcome curves but performs very poorly when the dose-outcome curves vary across indications. ROMI-v2 shows similar or superior OBD selection compared to ROMI-v1. For trials with $K = 3$ indications, ROMI-v2 is comparable to the Independent design and outperforms ROMI-v1-NC when OBDs vary across responsive indications in accurately selecting OBDs. The performance of ROMI-v1 and ROMI-v2 improves as the number of

Downloaded from https://academic.oup.com/biometrics/article/80/4/ujae105/7808593 by The University of Texas MD Anderson Cancer Center user on 03 March 2025

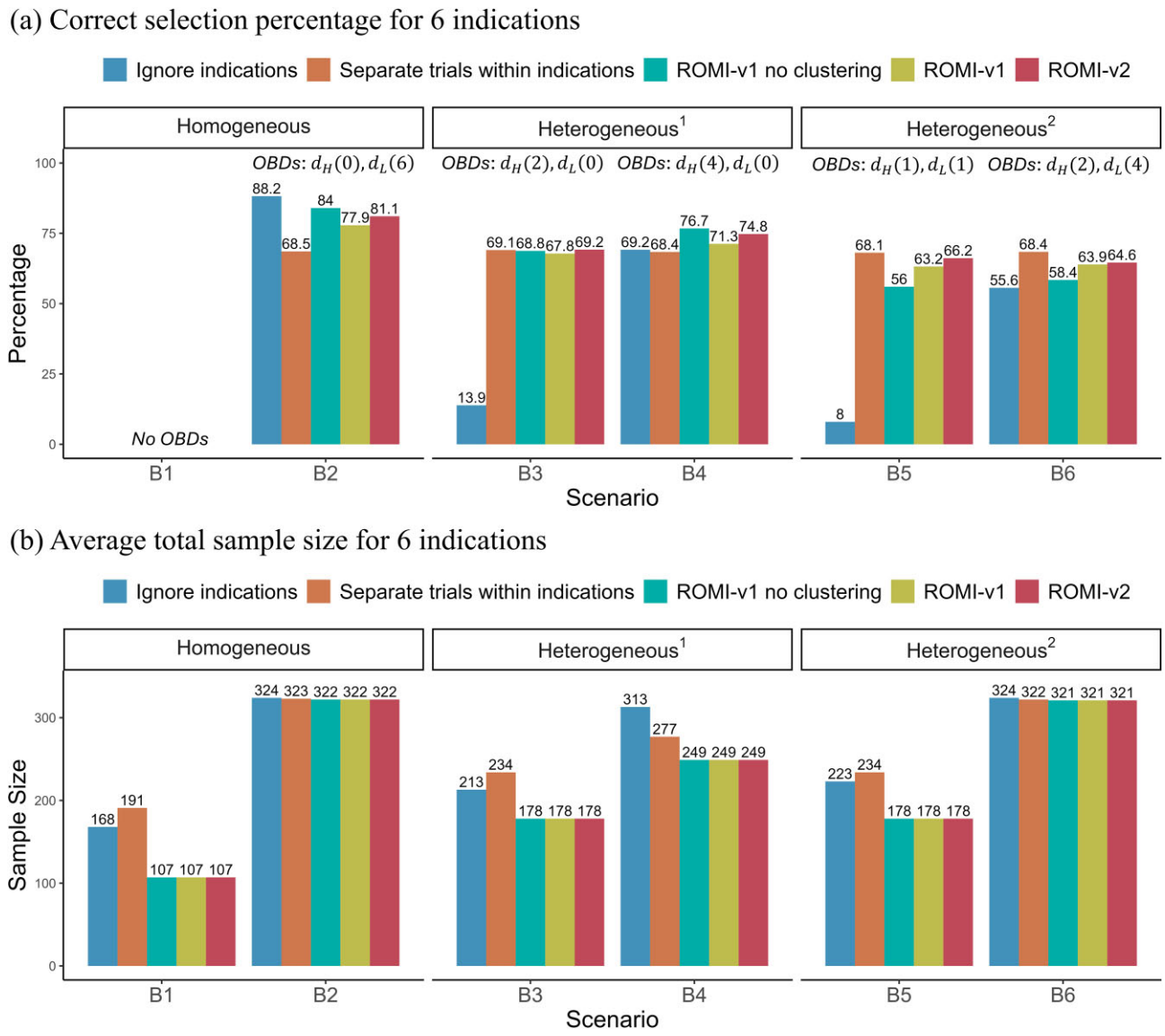

FIGURE 3 For six indications, (a) correct selection percentages and (b) average total sample sizes for the Pool design that ignores indications, Independent design that conducts separate trials within indications, ROMI-v1 with no indication clustering, ROMI-v1 with clustering, and ROMI-v2 with clustering. In homogeneous scenarios, OBDs are identical across indications. Heterogeneous[1] scenarios include some non-responsive indications and identical OBDs among responsive indications. In Heterogeneous[2] scenarios, OBDs vary among responsive indications.

indications increases. In scenario B2, where $K = 6$ indications are responsive, the CSP values of the ROMI-v1 and ROMI-v2 designs are 9.4% and 12.6% higher than the Independent design, respectively. Detailed results are provided in Web Appendix B.

As a final sensitivity analysis, we evaluated the ROMI designs, assuming that the shrinkage parameter follows a Half-Cauchy distribution. Simulation results are given in Web Appendix C. While this provides greater robustness, it reduces information borrowing.

5 DISCUSSION

ROMI effectively identifies and discontinues indications not responsive to treatment, substantially reducing sample size compared to designs that ignore indications or optimize dose independently for each indication. When dose-outcome curves differ between indications, ROMI accurately identifies indication-specific OBDs. The version of ROMI that uses information from both stages shows similar or higher accuracy in OBD selection compared to the version that ignores stage 1 data on d_H. Compared to conducting separate trials within indications, the second version of ROMI has greater accuracy in identifying the OBD if it is the same across indications. When the OBDs vary across indications, the accuracy of the ROMI design is slightly lower than the Independent design, but it still outperforms the design with ROMI structure but does not cluster similar indications. For a larger number of indications, the performance of the ROMI design improves.

As a future study, it may be worthwhile to develop a Bayesian hierarchical model accounting for count variables $\mathbf{X}_{\ell,k}$. Stage 1

Downloaded from https://academic.oup.com/biometrics/article/80/4/ujae105/7808593 by The University of Texas MD Anderson Cancer Center user on 03 March 2025

screening of ROMI is based on the assumption that d_H cannot be less effective than d_L. If this is invalid, stage 1 can be removed, with randomization for all indications throughout. In addition to efficacy and toxicity, endpoints such as pharmacokinetics or quality of life, may be included in the final OBD selection.

ACKNOWLEDGMENTS

The authors thank an Associate Editor and a referee for their valuable comments that substantially improved the article.

SUPPLEMENTARY MATERIALS

Supplementary material is available at *Biometrics* online.

Web Appendices referenced in Sections 2 and 4.3, along with the R code files for the simulation studies, are available with this paper at the Biometrics website on Oxford Academic.

FUNDING

Peter Thall's research was partially funded by NIH/NCI grants 1R01CA261978 and 5P30CA016672. Ying Yuan's research was partially funded by NIH/NCI grants P50CA281701, P50CA127001, P50CA221707, and Bettyann Asche Murray Distinguished Professorship. Shuqi Wang's research was partially funded by U01DK108328 and Bettyann Asche Murray Distinguished Professorship.

CONFLICT OF INTEREST

None declared.

DATA AVAILABILITY

Data sharing is not applicable in this paper as all data in this paper are computer simulated.

REFERENCES

Chen, N. and Lee, J. J. (2019). Bayesian hierarchical classification and information sharing for clinical trials with subgroups and binary outcomes. *Biometrical Journal*, 61, 1219–1231.

Chu, Y. and Yuan, Y. (2018a). BLAST: Bayesian latent subgroup design for basket trials accounting for patient heterogeneity. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 7, 723–740.

Chu, Y. and Yuan, Y. (2018b). A Bayesian basket trial design using a calibrated Bayesian hierarchical model. *Clinical Trials*, 15, 149–158.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, 1, 515–534.

Gelman, A., Jakulin, A., Pittau, M. G. and Su, Y. S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2, 1360–1383.

Guo, B. and Yuan, Y. (2017). Bayesian phase I/II biomarker-based dose finding for precision medicine with molecularly targeted agents. *Journal of the American Statistical Association*, 11, 508–520.

Guo, B. and Yuan, Y. (2023). DROID: dose–ranging approach to optimizing dose in oncology drug developmen. *Biometrics*, 79, 2907–2919.

Lin, R., Zhou, Y., Yan, F., Li, D. and Yuan, Y. (2020). BOIN12: Bayesian optimal interval phase I/II trial design for utility-based dose finding in immunotherapy and targeted therapies. *JCO Precision Oncology*, 4, 1393–1402.

Sachs, J., Mayawala, K., Gadamsetty, S., Kang, S. and Alwis, D. (2016). Optimal dosing for targeted therapies in oncology: drug development cases leading by example optimal dosing for targeted therapies in oncology. *Clinical Cancer Research*, 22, 1318–1324.

Shah, M., Rahman, A., Theoret, M. R. and Pazdur, R. (2021). The drug-dosing conundrum in oncology-when less is more. *The New England Journal of Medicine*, 385, 1445–1447.

Takeda, K., Liu, S. and Rong, A. (2022). Constrained hierarchical Bayesian model for latent subgroups in basket trials with two classifiers. *Statistics in Medicine*, 41, 298–309.

Thall, P. F. and Nguyen, H. Q. (2012). Adaptive randomization to improve utility-based dose-finding with bivariate ordinal outcomes. *Journal of Biopharmaceutical Statistics*, 22, 785–801.

Thall, P. F. and Russell, K. T. (1998). A strategy for dose finding and safety monitoring based on efficacy and adverse outcomes in phase I/II clinical trials. *Biometrics*, 54, 251–264.

Thall, P. F., Zang, Y. and Yuan, Y. (2023a). Generalized phase I-II designs to increase long term therapeutic success rate. *Pharmaceutical Statistics*, 22, 692–706.

Thall, P. F., Zang, Y., Chapple, A., Yuan, Y., Lin, R., Marin, D et al. (2023b). Novel clinical trial designs with dose optimization to improve long term outcomes. *Clinical Cancer Research*, 29, 4549—4554.

U.S. Food and Drug Administration. (2022). *Project Optimus: Reforming the dose Optimization and dose Selection Paradigm in Oncology*, https://www.fda.gov/about-fda/oncology-center-excellence/project-optimus. [Accessed September 15, 2024].

U.S. Food and Drug Administration. (2024). *Optimizing the Dosage of Human Prescription Drugs and Biological Products for the Treatment of Oncologic Diseases*, https://www.fda.gov/regulatory-information/search-fda-guidance-documents/optimizing-dosage-human-prescription-drugs-and-biological-products-treatment-oncologic-diseases. [Accessed September 15, 2024].

Yang, P., Li, D., Lin, R., Huang, B. and Yuan, Y. (2024). Design and sample size determination for multiple-dose randomized phase II trials for dose optimization. *Statistics in Medicine*, 43, 2972–2986.

Trial Design. https://www.trialdesign.org. [Accessed 5 September 2024].

Yuan, Y., Zhou, H. and Liu, S. (2024). Statistical and practical considerations in planning and conduct of dose-optimization trials. *Clinical Trials*, 21, 273–286.

Zang, Y, Thall, P. F. and Yuan, Y. (2024). A generalized phase 1-2-3 design integrating dose optimization with confirmatory treatment comparison. *Biometrics*, 80, ujad022.

Zhou, Y., Lee, J. J. and Yuan, Y. (2019). A utility–based Bayesian optimal interval (U–BOIN) phase I/II design to identify the optimal biological dose for targeted and immune therapies. *Statistics in Medicine*, 38,S5299–S5316.

Zhou, H., Lee, J. J. and Yuan, Y. (2017). BOP2: Bayesian optimal design for phase II clinical trials with simple and complex endpoints. *Statistics in Medicine*, 36, 3302–3314.

Downloaded from https://academic.oup.com/biometrics/article/80/4/ujae105/7808593 by The University of Texas MD Anderson Cancer Center user on 03 March 2025

Received: June 11, 2024; Revised: August 28, 2024; Accepted: September 10, 2024
© The Author(s) 2024. Published by Oxford University Press on behalf of The International Biometric Society. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com