Received: 6 April 2016 | Revised: 20 September 2016 | Accepted: 4 November 2016

DOI 10.1002/pst.1797

WILEY

MAIN PAPER

A simulation study of methods for selecting subgroup-specific doses in phase 1 trials

Satoshi Morita[1] | **Peter F. Thall**[2] | **Kentaro Takeda**[3,4]

[1]Department of Biomedical Statistics and Bioinformatics, Kyoto University Graduate School of Medicine, Kyoto, Japan
[2]Department of Biostatistics, The University of Texas, M. D. Anderson Cancer Center Houston, TX, USA
[3]Department of Biostatistics and Epidemiology, Yokohama City University, Yokohama, Japan
[4]Biostatistics Group, Data Science, Global Development, Astellas Pharma INC, Tokyo, Japan

Correspondence
Satoshi Morita, Department of Biomedical Statistics and Bioinformatics, Kyoto University Graduate School of Medicine, 54 Kawahara-cho, Shogoin, Sakyo-ku, Kyoto 606-8507, Japan.
Email: smorita@kuhp.kyoto-u.ac.jp

Funding information
Grant-in-Aid for Scientific Research, Grant/Award Number: C-24500345; NIH/NCI, Grant/Award Number: RO1 CA 83932

Patient heterogeneity may complicate dose-finding in phase 1 clinical trials if the dose-toxicity curves differ between subgroups. Conducting separate trials within subgroups may lead to infeasibly small sample sizes in subgroups having low prevalence. Alternatively, it is not obvious how to conduct a single trial while accounting for heterogeneity. To address this problem, we consider a generalization of the continual reassessment method on the basis of a hierarchical Bayesian dose-toxicity model that borrows strength between subgroups under the assumption that the subgroups are exchangeable. We evaluate a design using this model that includes subgroup-specific dose selection and safety rules. A simulation study is presented that includes comparison of this method to 3 alternative approaches, on the basis of nonhierarchical models, that make different types of assumptions about within-subgroup dose-toxicity curves. The simulations show that the hierarchical model-based method is recommended in settings where the dose-toxicity curves are exchangeable between subgroups. We present practical guidelines for application and provide computer programs for trial simulation and conduct.

KEYWORDS

Bayesian study design, conditionally independent hierarchical model, continual reassessment method, phase 1 clinical trial, subgroup-specific dose-finding

1 | INTRODUCTION

Patient heterogeneity may complicate phase 1 clinical trials in oncology. The goal may be either to determine a single optimal dose, or possibly different optimal doses within subgroups. For example, subgroups may be determined by disease subtypes, biomarkers targeted by the agent being studied, or known prognostic variables. Preclinical and clinical data often suggest that patient subgroups may have different dose-toxicity relationships, but the order in the tolerability of the subgroups is not known. Conventionally, even if such subgroups have been identified, most often a phase 1 trial ignoring subgroups is conducted. If the dose-toxicity curves differ between subgroups, however, a single dose chosen for all subgroups may be either subtherapeutic or excessively toxic in some subgroups. Moreover, ignoring subgroup effects during the dose-finding process may lead to undesirable interim dose assignments and adaptive decisions. An alternative approach is to conduct a separate dose-finding study within each subgroup. This may not be feasible in subgroups for which the prevalence is too low to reliably identify an optimal dose within a reasonable time frame. A more refined approach is to conduct 1 trial with the goal to find optimal doses that may differ between subgroups. Ideally, the underlying model should borrow strength between subgroups so that the data obtained from each subgroup may help inform the decisions in another subgroups.

We apply a hierarchical logistic regression model used by Morita et al[1] to illustrate the computation of a prior equivalent sample size in hierarchical models. The hierarchical model-based method generalizes the continual reassessment method (CRM) proposed by O'Quigley et al[2] by allowing different doses to be chosen within subgroups, while borrowing strength between subgroups. In this paper, we use

the same model as Morita et al[1] and develop in detail its uses for subgroup-specific dose finding. We review the model and dose-finding method, which we call the Hierarchical Bayesian CRM (HB-CRM), compare it to 3 alternative approaches, each on the basis of a nonhierarchical model, give practical guidelines, and provide a computer program for simulation and trial conduct. All 4 methods considered here address the problem of determining an optimal dose, or optimal subgroup-specific doses, on the basis of toxicity in settings where K subgroups have been identified. Each method uses a CRM-type criterion for optimality. For each subgroup, indexed by $k = 1, \cdots, K$, denote the probability of toxicity with dose x by $\pi_k(x, \theta_k)$, where θ_k is the model's parameter vector. The HB-CRM assumes a hierarchical structure for $\theta_1, \cdots, \theta_K$, which implies a priori that $\pi_1(x, \theta_1), \cdots, \pi_K(x, \theta_K)$ are exchangeable for each x and conducts a single trial including all subgroups. Each of the 3 alternative model-based comparators relies on a nonhierarchical model. The first alternative completely ignores subgroups and conducts a single trial using the same logistic dose-toxicity model for all subgroups. That is, it is the CRM on the basis of a logistic dose-toxicity model. The second alternative assumes K different subgroup-specific models and conducts K separate trials. The third alternative conducts 1 trial, assuming a dose-toxicity model with K different parameters θ_k to account for the inter-subgroup variability, and allows different optimal doses to be chosen within subgroups.

To help motivate the problem, it is worthwhile to consider a simple example in which there are $K = 3$ subgroups, with true dose-toxicity probability curves $\pi_1(x)^{\text{true}}$, $\pi_2(x)^{\text{true}}$, $\pi_3(x)^{\text{true}}$, given by Figure 1. If the aim of a phase I trial is to find a dose having mean toxicity probability $\pi^* = .30$ then, as shown by Figure 1, the true optimal doses are different for the 3 subgroups. Any method that finds 1 optimal dose x^{opt} ignores this possibility, and giving the same x^{opt} to all patients has the consequence of underdosing patients in subgroup 1, and overdosing patients in subgroup 3.

Several authors have addressed the problem of accounting for patient heterogeneity in phase 1 trials. O'Quigley et al[3] and O'Quigley and Paoletti[4] proposed a parametric model-based 2-sample CRM to find the optimal dose for each of 2 possibly ordered subpopulations of patients. Ivanova and Wang[5] proposed a nonparametric design with bivariate isotonic regression to address the same problem. Yuan and Chappell[6] compared 3 dose-finding methods, which respectively extended the up-and-down design,[7] the CRM,[2] and the isotonic design,[8] to deal with multiple risk subgroups which can be ordered according to their risk of toxicity. Thall et al[9] proposed a phase 1 to 2 design including covariates accounting for patient subpopulations on the basis of a trade-off between efficacy and toxicity. Liu et al[10] proposed an extended CRM with multiple skeletons of toxicity probabilities to deal with dose-finding in different ethnic populations. All of these approaches assume that the probability of toxicity is monotonically ordered for subgroups, so that in particular the subgroups are not exchangeable and do not represent qualitatively different subgroups for which no ordering may be assumed. In particular, O'Quigley et al[3] reported that the 2-sample CRM in this setting was comparable to conducting 2 independent trials for each of the 2 subgroups separately using the 1-sample CRM. This comparison is similar to that between our second and third alternatives. The main objective of our simulation study is to determine advantages and disadvantages of the hierarchical model-based method, in comparison with these alternative approaches, in a setting with more than 2 nonordered subgroups. We also examine the effects of the subgroup proportions, or prevalences, on how each of the methods behaves.

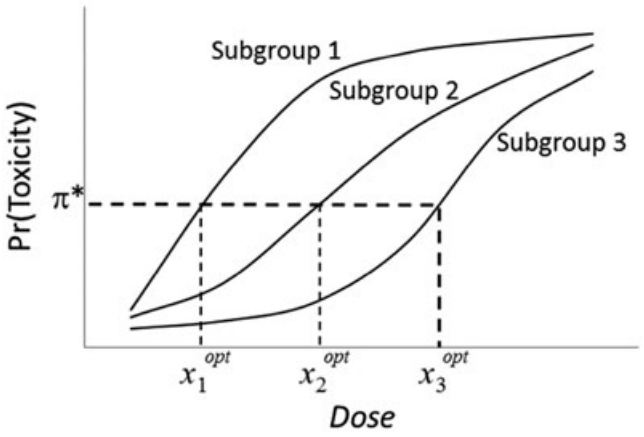

FIGURE 1 An example of 3 patient subgroups with different true dose-toxicity curves (y-axis, toxicity probability; x-axis, dose). Given the fixed target toxicity probability, $\pi^* = 0.30$, the 3 subgroups have different true optimal doses

In Section 2, we present probability models and prior specification for dose-finding methods to account for patient heterogeneity. We evaluate the operating characteristics of HB-CRM and each of the 3 alternative methods by simulation in Section 3. Section 4 gives guidelines for constructing study designs. We close with a brief discussion in Section 5.

2 | DOSE-FINDING METHODS

2.1 | Preliminaries

Denote the population proportions (prevalences) of the K subgroups by $\xi = (\xi_1, \ldots, \xi_K)$, that is, a patient belongs to subgroup k with probability ξ_k. In the trial, each patient in each subgroup receives 1 of J doses, denoted by $d_1 < \ldots < d_J$. We formulate the models using standardized doses $x_j = \log(d_j) - J^{-1} \sum_{l=1}^{J} \log(d_l)$. For the i^{th} patient in subgroup k, denote the assigned dose by $x_{[k,i]}$, the indicator $Y_{k,i} = 1$ if the patient suffers toxicity, 0 if not, and the toxicity probability

$$\pi_k(x_{[k,i]}, \theta_k) = \Pr(Y_{k,i} = 1 | x_{[k,i]}, \theta_k), \quad k = 1, \ldots, K.$$

At any given point during the trial, let $\mathcal{D}_n$ denote the data for the first n patients and $\mathbf{n} = (n_1, n_2, \cdots, n_K)$ the current sample sizes within the subgroups, so $n = n_1 + \cdots + n_K$.

Let $\mathbf{N} = (N_1, N_2, \cdots, N_K)$ denote the final subgroup sample sizes at the end of a trial, so the final total sample size is $N = N_1 + \cdots + N_K$. Given a planned maximum total sample size, N_{max}, because of the use of early stopping rules it may be the case that $N < N_{max}$. An important point, which will play a central role in determining the properties of the designs, is that both $\mathbf{n}$ and $\mathbf{N}$ depend on $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_K)$, as well as the particular design being used. Given N, temporarily ignoring the effects of early stopping, the expected final subgroup sample sizes are $\xi_1 N, \ \ldots, \ \xi_K N$. Due to both random variation and the use of adaptive rules, however, each achieved N_k may differ substantially from its mean.

2.2 | Hierarchical Bayesian CRM

In the HB-CRM (Morita, et al.),[1] the parameter vector in subgroup k is $\boldsymbol{\theta}_k = (\alpha_k, \beta)$ and model's linear components are

$$logit\{\pi_k(x_{[k,i]}, \alpha_k, \beta)\} = \alpha_k + \beta x_{[k,i]} \tag{1}$$

for $k = 1, \ldots, K$. For Level 1 priors, it is assumed that $\alpha_1, \ldots, \alpha_K$ are i.i.d. $\mathrm{N}(\tilde{\mu}_\alpha, \tilde{\sigma}_\alpha^2)$ and that β follows a $\mathrm{N}(\tilde{\mu}_\beta, \tilde{\sigma}_\beta^2)$ prior. For Level 2 priors (hyperpriors), it is assumed that $\tilde{\mu}_\alpha$ follows a normal hyperprior and, following the recommendation of Gelman,[11] $\tilde{\sigma}_\alpha$ follows a uniform prior on the interval .01 to U_ϕ, denoted $\tilde{\sigma}_\alpha \ \sim \mathrm{U}(0.01, \ U_\phi)$. In summary, the model assumptions are as follows:

$$\begin{aligned}
\text{Sampling model} \quad Y_{k,i} &\sim \text{Bernoulli}(\pi_k(x_{[k,i]}, \alpha_k, \beta)) \quad \text{indep. for all} \quad k \\
\text{Priors} \qquad\quad \alpha_k &\sim \text{i.i.d.} \quad \mathrm{N}\left(\tilde{\mu}_\alpha, \tilde{\sigma}_\alpha^2\right) \qquad\quad \text{for all} \quad k \\
\beta &\sim \mathrm{N}\left(\tilde{\mu}_\beta, \tilde{\sigma}_\beta^2\right) \\
\text{Hyperpriors} \qquad \tilde{\mu}_\alpha &\sim \mathrm{N}\left(\mu_{\alpha,\phi}, \sigma_{\alpha,\phi}^2\right) \\
\tilde{\sigma}_\alpha &\sim \mathrm{U}(0.01, U_\phi).
\end{aligned} \tag{2}$$

We do not impose the constraint that $\beta > 0$ with probability 1 to ensure that each $\pi_k(x, \alpha_k, \beta)$ increases in x or, alternatively, assume a lognormal prior for β. In practice, appropriate calibration of the hyperparameters $\tilde{\mu}_\beta$ and $\tilde{\sigma}_\beta^2$ in Equation 2 ensures this monotonicity. Specifically, if $\tilde{\mu}_\beta$ is a large enough positive value and $\tilde{\sigma}_\beta^2$ is sufficiently small, then no constraint on β is needed because all computed posterior values of β will be positive.

While a hierarchical prior structure is assumed for $(\alpha_1, \cdots, \alpha_K)$, the dose effect parameter β shared by all K marginal toxicity probabilities has a usual prior without an additional hyperprior on its hyperparameters $(\tilde{\mu}_\beta, \tilde{\sigma}_\beta)$, which are fixed. Collecting terms, the $K + 1$ sampling model parameters that characterize the marginal probabilities of toxicity in the K subgroups are $\boldsymbol{\theta} = (\alpha_1, \cdots, \alpha_K, \beta)$, the 4 hyperparameters that characterize the priors on $\boldsymbol{\theta}$ are $\tilde{\boldsymbol{\theta}} = (\tilde{\mu}_\alpha, \tilde{\sigma}_\alpha, \tilde{\mu}_\beta, \tilde{\sigma}_\beta)$, and the fixed hyperparameters that characterize the hyperpriors on $(\tilde{\mu}_\alpha, \tilde{\sigma}_\alpha^2)$ are $\boldsymbol{\phi} = (\mu_{\alpha,\phi}, \sigma_{\alpha,\phi}, .01, U_\phi)$. Consequently, to complete the Bayesian model one must specify numerical values for a total of 5 parameters, the 2 hyperparameters $(\tilde{\mu}_\beta, \tilde{\sigma}_\beta)$ and for the 3-fixed hyperprior parameters $(\mu_{\alpha,\phi}, \sigma_{\alpha,\phi}, U_\phi)$. In Equation 2 for the hierarchical model, "priors" may be called "level 1 priors" because they are distributions on the parameters $\boldsymbol{\theta}$ of the sampling model, while "hyperpriors" may be called "level 2 priors" because they are distributions on the 2 parameters $(\tilde{\mu}_\alpha, \tilde{\sigma}_\alpha^2)$ of the level 1 priors.

Under this hierarchical model, a priori, the parameter vectors $(\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_K)$ of the marginal toxicity probability models for the K subgroups are exchangeable. This is the property that if the subgroup indices $(1, \ldots, K)$ are replaced by any permutation $(j_1, \ldots, j_K)$ then the joint prior distribution of the permuted vector $(\boldsymbol{\theta}_{j_1}, \cdots, \boldsymbol{\theta}_{j_K})$ is the same as that of $(\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_K)$. Exchangeability is a useful property for probability models of random quantities corresponding to qualitatively different objects, where the order in which the objects are indexed to identify them is arbitrary. A well-known special case is a vector of random quantities that are independent and identically distributed (iid), which trivially must be exchangeable. The joint distribution of an exchangeable random vector is more general, however, because the random quantities need not be independent. The prior exchangeability of $(\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_K)$ implies that, a priori, the toxicity probabilities $\{\pi_1(x, \boldsymbol{\theta}_1), \cdots, \pi_K(x, \boldsymbol{\theta}_K)\}$ are exchangeable for each x.

The hierarchical model given above induces prior association among $\alpha_1, \cdots, \alpha_K$. To see this, denoting the level 1 priors by p_1 and level 2 priors (hyperpriors) by p_2, we obtain the unconditional prior of the sampling model parameters by averaging over the hyperprior $p_2(\tilde{\mu}_\alpha, \tilde{\sigma}_\alpha | \boldsymbol{\phi})$, as

$$\begin{aligned}
p(\beta, \alpha_1, \cdots, \alpha_K \ | \ \tilde{\mu}_\beta, \tilde{\sigma}_\beta, \boldsymbol{\phi}) &= p_1(\beta \ | \ \tilde{\mu}_\beta, \tilde{\sigma}_\beta) \int \prod_{k=1}^{K} p_1(\alpha_k \ | \ \tilde{\mu}_\alpha, \tilde{\sigma}_\alpha) \, p_2(\tilde{\mu}_\alpha, \tilde{\sigma}_\alpha \ | \ \boldsymbol{\phi}) d\tilde{\mu}_\alpha d\tilde{\sigma}_\alpha \\
&= p_1(\beta \ | \ \tilde{\mu}_\beta, \tilde{\sigma}_\beta) \ \ p_{1,2}(\alpha_1, \cdots, \alpha_K \ | \ \boldsymbol{\phi}),
\end{aligned} \tag{3}$$

where $p_{1,2}$ denotes the marginal prior of $(\alpha_1, \cdots, \alpha_K)$ obtained by averaging over the hyperprior of $(\tilde{\mu}_\alpha, \tilde{\sigma}_\alpha)$. The

prior association among $\alpha_1,\cdots,\ \alpha_K$ in $p_{1,2}$ induces prior association among the toxicity probabilities $\pi_1(x,\ \theta_1),\cdots,\ \pi_K(x, \theta_K)$ for any x. In this regard, the dose effect parameter β shared by these probabilities also induces positive association among them. These prior associations shrink the posteriors of the $\pi_k(x, \theta_k)$'s toward each other. Equation 3 also shows why numerical values of $(\tilde{\mu}_\beta, \tilde{\sigma}_\beta, \boldsymbol{\phi})$ must be specified to complete the model.

Medical settings where this hierarchical model is appropriate include trials in which there are K qualitatively different disease subtypes, different solid tumors types, or subgroups defined by biomarkers. The common feature is that one's prior uncertainty about $\boldsymbol{\theta}_1,\ \cdots,\ \boldsymbol{\theta}_K$ would not be changed if the subgroups were reindexed in a different order. In contrast, for example, the hierarchical model is not appropriate if the subgroups correspond to prognostic risk of toxicity from any agent, such as good (low risk, $k = 1$), intermediate ($k = 2$), and poor (high risk, $k = 3$). In this case, $\pi_k(x, \boldsymbol{\theta}_k)$ is stochastically increasing in k for any dose x, and the exchangeability assumption is not valid.

For each decision during trial conduct, the HB-CRM defines the optimal dose $x^{\text{opt}}_{[k]}$ in subgroup k to be that for which the posterior mean of $\pi_k(x_j, \theta_k)$ is closest to a given fixed target, π^*. Formally, given $\mathcal{D}_n$, the dose chosen for subgroup k is

$$x^{\text{opt}}_{[k]} = \underset{j=1,\cdots,J}{\text{argmin}} \qquad |E\{\pi_k(x_j, \alpha_k, \beta)|\mathcal{D}_n\} - \pi^*|. \tag{4}$$

If desired, different target values for the subgroups may be used, although we will not explore that case here.

A safety rule imposed on the method is that, within each subgroup, the HB-CRM may not skip an untried doses when escalating. In addition, to control overdosing, HB-CRM does not escalate within subgroup k if

$$\Pr(\pi_k(x_j, \boldsymbol{\theta}_k) > \pi^{\text{odc}}|\mathcal{D}_n) > \psi^{\text{odc}}$$

where π^{odc} is a fixed upper limit, and ψ^{odc} is a probability cutoff. This rule supersedes the criterion (Equation 4) for dose escalation. The design parameter ψ^{odc} must be calibrated along with the prior parameters to obtain a design with desirable operating characteristics.

The HB-CRM in Equation 2 relies on the prior assumption that the toxicity probabilities of the patient subgroups are exchangeable, because $\alpha_1,\cdots,\ \alpha_K$ are conditionally i.i.d., given $\tilde{\mu}_\alpha$ and $\tilde{\sigma}^2_\alpha$. Since one must average over the hyperpriors of $\tilde{\mu}_\alpha$ and $\tilde{\sigma}^2_\alpha$ to compute posteriors, given the observed data $\mathcal{D}_n$ the intercept parameters $\alpha_1,\cdots,\alpha_K$ are positively correlated. This, and the $\pi_k(x_j, \alpha_k, \beta)$'s share the common slope parameter β, induces positive correlation among all KJ toxicity probabilities. This induces association among the posterior means in Equation 4, which in turn shrinks the chosen doses $x^{\text{opt}}_{[1]},\cdots,x^{\text{opt}}_{[K]}$ toward each other. In this way, conducting a single trial with this hierarchical model provides a basis for borrowing strength across patient subgroups.

The hyperpriors on $\tilde{\mu}_\alpha$ and $\tilde{\sigma}^2_\alpha$ play key roles in how the HB-CRM design behaves. Thus, their fixed parameters, $\mu_{\alpha,\phi}, \sigma^2_{\alpha,\phi}$, and U_ϕ, must be calibrated carefully, along with the fixed level 1 prior parameters $\tilde{\mu}_\beta, \tilde{\sigma}^2_\beta$. As our simulations will show, this may lead to more accurate within-subgroup dose selection compared to what is obtained by either conducting separate trials within subgroups or conducting 1 trial but ignoring subgroups to obtain a design with good operating characteristics.

2.3 | Nonhierarchical models and designs

For the 3 alternative dose-finding methods, we assume nonhierarchical logistic models, with different parameterizations of their intercepts and slopes to account for subgroups. That is, for each of the following models, usual priors are assumed, and there are no hyperpriors. The first alternative method assumes complete patient homogeneity under the model $\pi(x_j, \alpha, \beta) = \Pr(Y_i = 1|x_j, \alpha, \beta)$ with

$$logit\{\pi(x_j, \alpha, \beta)\} = \alpha + \beta x_j \tag{5}$$

for all subgroups, where α and β follow $\text{N}(\tilde{\mu}_\alpha, \tilde{\sigma}^2_\alpha)$ and $\text{N}(\tilde{\mu}_\beta, \tilde{\sigma}^2_\beta)$ priors, respectively. For this model, numerical values of the 4 prior hyperparameters $\tilde{\boldsymbol{\theta}} = (\tilde{\mu}_\alpha, \tilde{\sigma}_\alpha, \tilde{\mu}_\beta, \tilde{\sigma}_\beta)$ must be specified. This method conducts a single trial and treats all newly enrolled patient at the same recommended dose, defined under the usual CRM criterion as the x_j having estimated posterior mean $\text{E}\{\pi(x_j, \boldsymbol{\theta})|\mathcal{D}_n\}$ closest to π^*, where $\boldsymbol{\theta} = (\alpha, \beta)$. This is a usual CRM criterion, computed under a 2-parameter logistic model. For comparability, and to ensure an ethical trial, the CRM as defined here also includes a "do-not-skip" rule and a rule for overdose control, but applied overall rather than within subgroups.

The second alternative method, which we denote by K-CRM-1-trial, uses the CRM in 1 trial accounting for K subgroups. This method accounts for patient heterogeneity by modeling the within-subgroup probability of toxicity using the same logistic form with linear term $\alpha_k + \beta x$ as given in Equation 1, but without hierarchical borrowing of strength among $\alpha_1, \ldots, \alpha_K$ through a common hyperparameter. For this model, the assumed priors are $\alpha_1,\ \ldots,\ \alpha_K \sim$ i.i.d. $\text{N}(\tilde{\mu}_\alpha, \tilde{\sigma}^2_\alpha)$ and $\beta \sim \text{N}(\tilde{\mu}_\beta, \tilde{\sigma}^2_\beta)$. Again, since there are only level 1 priors and no level 2 priors (hyperpriors) in this model, only numerical values of the 4 prior hyperparameters $\tilde{\boldsymbol{\theta}} = (\tilde{\mu}_\alpha, \tilde{\sigma}_\alpha, \tilde{\mu}_\beta, \tilde{\sigma}_\beta)$ must be specified. This model also has prior exchangeablity across patient subgroups, but it borrows strength between the subgroups only through the common slope parameter, β. Like the HB-CRM, the K-CRM-1-trial conducts a single trial and treats each newly enrolled patient in subgroup k at the subgroup-specific dose for which $\text{E}\{\pi_k(x_j, \alpha_k, \beta)|\mathcal{D}_n\}$ is closest to π^*. Thus, the chosen doses for the K subgroups are obtained by averaging the $\pi_k(x_j, \alpha_k, \beta)$'s with respect to the parameters $(\alpha_1,\ \cdots,\ \alpha_K,\ \beta)$. This method applies the same subgroup-specific, do-not-skip, and overdose-control rules as those used in HB-CRM. That is, the K-CRM-1-trial uses

TABLE 1 Summary of study designs of hierarchical versus nonhierarchical model and the linear term of the logistic model for the probability of toxicity as a function of dose x and prognostic subgroup $k = 1, \ldots, K$

	Single trial		K Separate trials
Bayesian model structure	$\alpha + \beta x$	$\alpha_k + \beta x$	$\alpha_k + \beta_k x$
Nonhierarchical	CRM[a]	K-CRM-1-trial[b]	1-CRM-K-trials[d]
Hierarchical	—	HB-CRM[c]	—

[a]Ordinary CRM ignoring subgroups, conduct 1 trial.

[b]K-subgroup CRM in 1 trial, assuming different intercepts $\alpha_1, \cdots, \alpha_K$ without a hierarchical structure, conduct 1 trial.

[c]Hierarchical model-based CRM assuming different intercepts $\alpha_1, \cdots, \alpha_K$ with a hierarchical structure, conduct 1 trial.

[d]Ordinary CRM conducted in each of K separate trials, assuming independent subgroup-specific parameters (α_k, β_k).

In the linear terms, α and β denote the intercept and slope parameters, respectively.

precisely the same decision rules as the HB-CRM, but assumes a different, nonhierarchical model.

The third alternative method, 1-CRM-K-trials, conducts separate trials in the K subgroups using the ordinary 1-sample CRM in each trial, and it does not assume exchangeability across patient subgroups or borrow strength in any way between subgroups. For subgroup k, this method assumes the model $\pi_k(x_j, \alpha_k, \beta_k) = \Pr(Y_i = 1 | x_j, \alpha_k, \beta_k)$ with

$$logit\{\pi_k(x_j, \alpha_k, \beta_k)\} = \alpha_k + \beta_k x_j, \tag{6}$$

with priors $\alpha_1, \cdots, \alpha_K \sim$ i.i.d. $N(\tilde{\mu}_\alpha, \tilde{\sigma}_\alpha^2)$ and $\beta_1, \cdots, \beta_K \sim$ i.i.d. $N(\tilde{\mu}_\beta, \tilde{\sigma}_\beta^2)$. For this third model, numerical values of the 4 prior hyperparameters $\tilde{\theta} = (\tilde{\mu}_\alpha, \tilde{\sigma}_\alpha, \tilde{\mu}_\beta, \tilde{\sigma}_\beta)$ must be specified. Like the HB-CRM and K-CRM-1-trial, the 1-CRM-K-trials design includes subgroup-specific, do-not-skip, and overdose-control rules.

Table 1 summarizes the 4 methods of the linear terms of their logistic models, the structural assumptions for the parameters, and the rules for trial conduct, ie, whether there is 1 trial or K separate trials that do not use each others' data to make decisions. While both the HB-CRM and K-CRM-1-trial are based on models that borrow strength between subgroups, the key diference is that the HB-CRM model has a hierarchial prior structure on $\alpha_1, \cdots, \alpha_K$, while the model used by the K-CRM-1-trial does not.

2.4 | Prior specification and numerical methods

Recall that, for the hierarchical model, to establish the prior the 5 fixed hyperparmeters $(\tilde{\mu}_\beta, \tilde{\sigma}_\beta)$ and $(\mu_{\alpha,\phi}, \sigma_{\alpha,\phi}, U_\phi)$ must be specified. To establish the prior for each of the 3 nonhierarchical models, the 4 fixed hyperparameters $(\tilde{\mu}_\alpha, \tilde{\sigma}_\alpha, \tilde{\mu}_\beta, \tilde{\sigma}_\beta)$ must be specified. We recommend minimally informative priors to allow the data to dominate the decisions in general and, in the present context, to ensure a fair comparison among the different methods. For the location parameters, either $(\tilde{\mu}_\alpha, \mu_{\alpha,\phi})$ for the hierarchical model or $(\tilde{\mu}_\alpha, \tilde{\mu}_\beta)$ for the nonhierarchical models, one can 2 elicited mean toxicity probabilities to solve for the fixed 2 fixed hyperprior means. Given the standardized doses, in each case the 2 fixed hyperprior means may be calculated by equating elicited values of toxicity probabilities at 2 different doses, say $x^{(1)}$ and $x^{(2)}$, with the corresponding formulas for $\pi(x^{(j)}, \theta)$, $j = 1,2$, replacing θ by its mean, and solving the 2 equations for the 2 unknown hyperparameters. This is illustrated below. Given these fixed location parameters, variance parameters may be determined for prior informativeness, quantified by prior effective sample size (ESS).[1,12] To speed up computation, one may use approximate ESS values[13] (computational details are given in the Appendix). One may set the values of the variance hyperparameters to control prior informativeness so that the per-subgroup ESS values are a small number, such as 1, 2, or 3.

CRM. For the CRM, first choose an overall ESS, and then divide it by K to obtain a common per-subgroup ESS value. As explained in Section 2.3, this method assumes $logit\{\pi(x_j, \alpha, \beta)\} = \alpha + \beta x_j$ (Equation 5), with $\alpha \sim N(\tilde{\mu}_\alpha, \tilde{\sigma}_\alpha^2)$ and $\beta \sim N(\tilde{\mu}_\beta, \tilde{\sigma}_\beta^2)$. To obtain the means $\tilde{\mu}_\alpha$ and $\tilde{\mu}_\beta$, if, for example, one elicits the 2 prior mean toxicity probabilities $E\{\pi(x_2, \theta)\} = 0.10$ and $E\{\pi(x_5, \theta)\} = 0.50$, then the 2 resulting equations $\tilde{\mu}_\alpha + \tilde{\mu}_\beta x_2 = \text{logit}(0.10)$ and $\tilde{\mu}_\alpha + \tilde{\mu}_\beta x_5 = \text{logit}(0.50)$ yield $\tilde{\mu}_\alpha = -1.23$ and $\tilde{\mu}_\beta = 2.40$. Assuming that $\tilde{\sigma}_\alpha^2 = \tilde{\sigma}_\beta^2$, one then may compute the approximate overall ESS values for a suitable range of $\tilde{\sigma}_\alpha^2 (= \tilde{\sigma}_\beta^2)$, eg, 0.01, 0.02, …, 10. Finally, one may choose a value of $\tilde{\sigma}_\alpha^2 (= \tilde{\sigma}_\beta^2)$ so that the overall ESS value is closest to $4 (= 1 \times 4)$, that is, the per-subgroup ESS value nearly equals 1, resulting in $\tilde{\sigma}_\alpha^2 = \tilde{\sigma}_\beta^2 = 1.25$.

1-CRM-K-trials. This method conducts K separate trials and assumes $logit\{\pi_k(x_j, \alpha_k, \beta_k)\} = \alpha_k + \beta_k x_j$ (Equation 6) with $\alpha_1, \cdots, \alpha_K \sim$ i.i.d. $N(\tilde{\mu}_\alpha, \tilde{\sigma}_\alpha^2)$ and $\beta_1, \cdots, \beta_K \sim$ i.i.d. $N(\tilde{\mu}_\beta, \tilde{\sigma}_\beta^2)$. Thus, one may perform the same calculations as done above for the CRM, but within each subgroup. Given $\tilde{\mu}_\alpha = -1.23$ and $\tilde{\mu}_\beta = 2.40$, one may choose $\tilde{\sigma}_\alpha^2 (= \tilde{\sigma}_\beta^2)$ so that the per-subgroup ESS takes a value close to 1, which in this case gives $\tilde{\sigma}_\alpha^2 = \tilde{\sigma}_\beta^2 = 5.92$.

K-CRM-1-trial. This method conducts a single trial and assumes $logit\{\pi_k(x_j, \alpha_k, \beta)\} = \alpha_k + \beta x_j$ (Equation 6) with priors $\alpha_1, \cdots, \alpha_K \sim$ i.i.d. $N(\tilde{\mu}_\alpha, \tilde{\sigma}_\alpha^2)$ and $\beta \sim N(\tilde{\mu}_\beta, \tilde{\sigma}_\beta^2)$. Since it is assumed that, a priori, $\alpha_1, \cdots, \alpha_K$ have the same mean $\tilde{\mu}_\alpha$, it is reasonable to derive a single value of $\tilde{\mu}_\alpha$. Thus, as above, $\tilde{\mu}_\alpha =$ -1.23 and $\tilde{\mu}_\beta = 2.40$ would be derived. In the evaluation of approximate ESS values with fixed $\tilde{\mu}_\alpha$ and $\tilde{\mu}_\beta$, one may compute the within-subgroup probability of toxicity $E\{\pi_k(x_j, \alpha_k, \beta)|\tilde{\mu}_\alpha, \tilde{\mu}_\beta, \tilde{\sigma}_\alpha^2, \tilde{\sigma}_\beta^2)\}$ for a suitable range of $\tilde{\sigma}_\alpha^2$ $(= \tilde{\sigma}_\beta^2)$. Then, choose a value of $\tilde{\sigma}_\alpha^2$ $(= \tilde{\sigma}_\beta^2)$ so that the per-subgroup ESS takes a value being closest to 1, resulting in $\tilde{\sigma}_\alpha^2 = \tilde{\sigma}_\beta^2 =$ 5.92. Although ESS is not necessarily additive over subgroups because β is the common slope parameter for patient subgroups in this method, we avoid more complicated ESS computations to facilitate practical application.

HB-CRM. As explained in Section 2.2, this method conducts a single trial assuming a hierarchical model with $logit\{\pi_k(x_j, \alpha_k, \beta)\} = \alpha_k + \beta x_j$ (Equation 1) for subgroup $k = 1, \ldots, K$. For the level 1 priors, it is assumed that $\alpha_1, \ldots, \alpha_K$ are i.i.d. $N(\tilde{\mu}_\alpha, \tilde{\sigma}_\alpha^2)$ and that $\beta \sim N(\tilde{\mu}_\beta, \tilde{\sigma}_\beta^2)$. For level 2 priors, it is assumed that $\tilde{\mu}_\alpha \sim N(\mu_{\alpha,\phi}, \sigma_{\alpha,\phi}^2)$ and $\tilde{\sigma}_\alpha \sim U(0.01, U_\phi)$. The location parameters in this method are $\tilde{\mu}_\beta$ in level 1 and $\mu_{\alpha,\phi}$ in level 2. Since $\mu_{\alpha,\phi}$ represents the overall mean of $\alpha_1, \ldots, \alpha_K$, it may be acceptable not to differentiate the prior levels of $\tilde{\mu}_\alpha$ and $\mu_{\alpha,\phi}$ to simplify the derivation of their estimates. Thus, similarly to the nonhierarchical methods, $\mu_{\alpha,\phi} =$ -1.23 and $\tilde{\mu}_\beta =$ 2.40 are obtained using the 2 elicited prior means $E\{\pi(d_2, \theta)\} = 0.10$ and $E\{\pi(d_5, \theta)\} = 0.50$. Next, 3 parameters $(\tilde{\sigma}_\beta^2, \sigma_{\alpha,\phi}^2, U_\phi)$ are specified using the ESS computation process. To simplify computation, we use a simplified algorithm to obtain these 3 parameters sequentially. First, use the same value of $\tilde{\sigma}_\beta^2$ specified in *K*-CRM-1-trial, which has a study design similar to that of HB-CRM, regardless of whether a hierarchical structure is assumed or not. Second, for U_ϕ, Morita et al[1] evaluate ESS at 2 values, $U_\phi = 2$ and $U_\phi = 5$. Then, for a suitable range of $\sigma_{\alpha,\phi}^2$, eg, 0.01, 0.02, …, 10, compute the within-subgroup prior mean probability of toxicity $E\{\pi_k(x_j, \alpha_k, \beta)|\mu_{\alpha,\phi}, \tilde{\mu}_\beta, \tilde{\sigma}_\beta^2, \sigma_{\alpha,\phi}^2, U_\phi\}$ to obtain approximate ESS values. Finally, choose a pair of values of $\sigma_{\alpha,\phi}^2$ and U_ϕ so that the per-subgroup ESS takes a value close to 1. If both values of U_ϕ in combination with some value of $\sigma_{\alpha,\phi}^2$ yield the per-subgroup ESS value 1, choose the smaller value, 2 for U_ϕ, thereby obtaining an appropriately informative hyperprior of $\tilde{\sigma}_\alpha$. It is expected that HB-CRM with a suitably informative hyperprior for $\tilde{\sigma}_\alpha$ that controls the between-subgroup variability in the intercepts $(\alpha_1, \cdots, \alpha_K)$ will do a better job of dealing with differences in toxicity probabilities by borrowing strength between subgroups.

Because these specifications of fixed prior para=meters involve some arbitrary choices, one should evaluate the operating characteristics of the design via simulation, and if necessary adjust the numerical hyperparameter values on that basis. To compute the posteriors, we use Markov chain Monte Carlo,[14] because the joint posterior distribution of the intercept and slope parameters is not readily available in closed form.

3 | SIMULATIONS AND CASE-BY-CASE EXAMPLES

3.1 | Simulation study design

We compared the 4 methods in several cases chosen to evaluate the advantages and disadvantages of HB-CRM and the 3 other designs of how accurately each estimates optimal doses (ODs).

To evaluate the performance of the 4 methods fairly, we used the same basic setup with respect to the dose levels $J = 6$ with $(d_1, \ldots, d_6) = (100, 200, 300, 400, 500, 600)$, starting dose (d_1), target toxicity level $\pi^* = .33$, and the number of subgroups $K = 4$. For the subgroups, we assumed 2 different distributions of population proportions: either $\boldsymbol{\xi} = (0.25, 0.25, 0.25, 0.25)$ or $\boldsymbol{\xi} = (0.40, 0.30, 0.20, 0.10)$, named "equal" and "different" prevalence patterns, respectively. In addition, we evaluated each design's operating characteristics using four maximum sample sizes ($N_{\max} = 48, 72, 96, 120$). We chose the minimum and maximum values of $N_{\max}$, 48 and 120, taking into account that their corresponding expected per-subgroup sample sizes under the equal population proportions, 12 and 30, often may be used in an ordinary phase 1 trial for each subgroup.

As explained in Section 2.4, we set up the priors of the 4 methods to ensure reasonably fair comparisons. The details are given in the Supporting Information. First, we used prior estimates of $E\{\pi(d_2, \boldsymbol{\theta})\}$ and $E\{\pi(d_5, \boldsymbol{\theta})\}$ to solve for the location parameters, $\tilde{\mu}_\alpha$ and $\tilde{\mu}_\beta$. With $E\{\pi(d_2, \boldsymbol{\theta})\} = 0.10$ and $E\{\pi(d_5, \boldsymbol{\theta})\} = 0.50$, the location parameters were specified as $\tilde{\mu}_\alpha = -1.23$ and $\tilde{\mu}_\beta = 2.40$ for all 4 methods (except for HB-CRM, which does not use a fixed $\tilde{\mu}_\alpha$). Given the location parameters, the scale parameters were specified as $\tilde{\sigma}_\alpha^2 = \tilde{\sigma}_\beta^2 = 1.25$ for the CRM so that the overall ESS value was close to $4 (= 1 \times 4)$, that is, the per-subgroup ESS values nearly equaled 1. For 1-CRM-*K*-trials and *K*-CRM-1-trial, the scale parameters were specified as $\tilde{\sigma}_\alpha^2 = \tilde{\sigma}_\beta^2 = 5.92$ to obtain per-subgroup ESS value close to 1. The priors and hyperpriors of HB-CRM were specified with $\tilde{\mu}_\beta = 2.40, \tilde{\sigma}_\beta^2 =$ 5.92, $\mu_{\alpha,\phi} = -1.23$, $\sigma_{\alpha,\phi}^2 = 4.85$, and $U_\phi = 2$. The location

parameters were specified on the basis of the prior estimates $E\{\pi(d_2, \boldsymbol{\theta})\} = 0.10$ and $E\{\pi(d_5, \boldsymbol{\theta})\} = 0.50$, and the dispersion parameters were specified according to prior ESS so that the per-subgroup ESS value was close to 1. For the design parameters of the overdose control rule, on the basis of preliminary investigation, we determined that the combinations of $\pi^{\text{odc}} = 0.50$ and $\psi^{\text{odc}} = 0.25$ gave designs with good operating characteristics.

We constructed 4 different dose-toxicity scenarios by specifying values of the true toxicity probabilities in each subgroup, shown in Figure 2 and Table S1. These scenarios were chosen to illustrate how the methods behave in a variety of settings for inter-subgroup difference of the dose-toxicity relationship. The scenarios are not based on any of the models. Each scenario is characterized by the true probabilities of toxicity, $\pi_{j,k}^{\text{true}}$, for the 6 dose levels $j = 1, \cdots, 6$ and 4 subgroups, $k = 1, \cdots, 4$. Thus, the assumed true dose-toxicity curve within subgroup k is characterized by the 6-dimensional vector $\boldsymbol{\pi}_k^{\text{true}} = (\pi_{1,k}^{true}, \cdots, \pi_{6,k}^{\text{true}})$. Under Scenario 1, given the target probability $\pi^* = 0.33$, d_4 is the OD in all the 4 subgroups. In contrast, Scenario 2 represents a case where the ODs differ between subgroups, with respective ODs d_4, d_6, d_1, and d_4, in the 4 subgroups, respectively. Scenarios 3 also has different ODs for the subgroups, but they are closer to each other than in Scenario 2. Scenario 4 is a difficult case where the 4 subgroup-specific dose-toxicity curves $\boldsymbol{\pi}_1^{\text{true}}, \boldsymbol{\pi}_2^{\text{true}}, \boldsymbol{\pi}_3^{\text{true}}$, and $\boldsymbol{\pi}_4^{\text{true}}$ differ substantially between subgroups, and moreover, each subgroup's curve has 2 dose levels with true toxicity probabilities equal to or close to the target $\pi^* = 0.33$. In this case, 2 doses both are good choices for each subgroup, but these 2 doses differ between subgroups.

To evaluate and compare the 4 designs under each of the dose-toxicity scenarios in the simulations, we use the following weighted average of the dose selection probabilities. Motivated by the idea that smaller values of $|\pi_{j,k}^{true} - \pi^*|$ correspond to d_j being more desirable in subgroup k, equivalently larger values of $1 - |\pi_{j,k}^{true} - \pi^*|$ are more desirable in that subgroup, we define the weights

$$w_{j,k} = \frac{1 - |\pi_{j,k}^{true} - \pi^*| - \min_{r=1,\cdots,J}\left\{1 - |\pi_{r,k}^{true} - \pi^*|\right\}}{\max_{r=1,\cdots,J}\left\{1 - |\pi_{j,k}^{true} - \pi^*|\right\} - \min_{r=1,\cdots,J}\left\{1 - |\pi_{r,k}^{true} - \pi^*|\right\}} \tag{7}$$

Then, we define the *subgroup-specific weighted probability of selection (WPS)*

$$\text{WPS}_k = \sum_{j=1}^{J} w_{j,k} \cdot \Pr(x_j \text{ is selected as the OD in subgroup } k), \tag{8}$$

for each $k = 1, \cdots, K$. We subtract the smallest value of $1 - |\pi_{r,k}^{true} - \pi^*|$ among the j doses in the numerator and denominator of Equation 7 so that $w_{j,k} = 0$ for the least desirable dose in that subgroup to give greater relative weights to the doses having $\pi_{j,k}^{\text{true}}$ closer to π^* in that subgroup. Particularly, Equation 7 gives weight 1 to the dose having $\pi_{j,k}^{\text{true}}$ closest to π^*. We do not take the alternative approach of using $|\pi_{j,k}^{\text{true}} - \pi^*|^{-1}$ as a

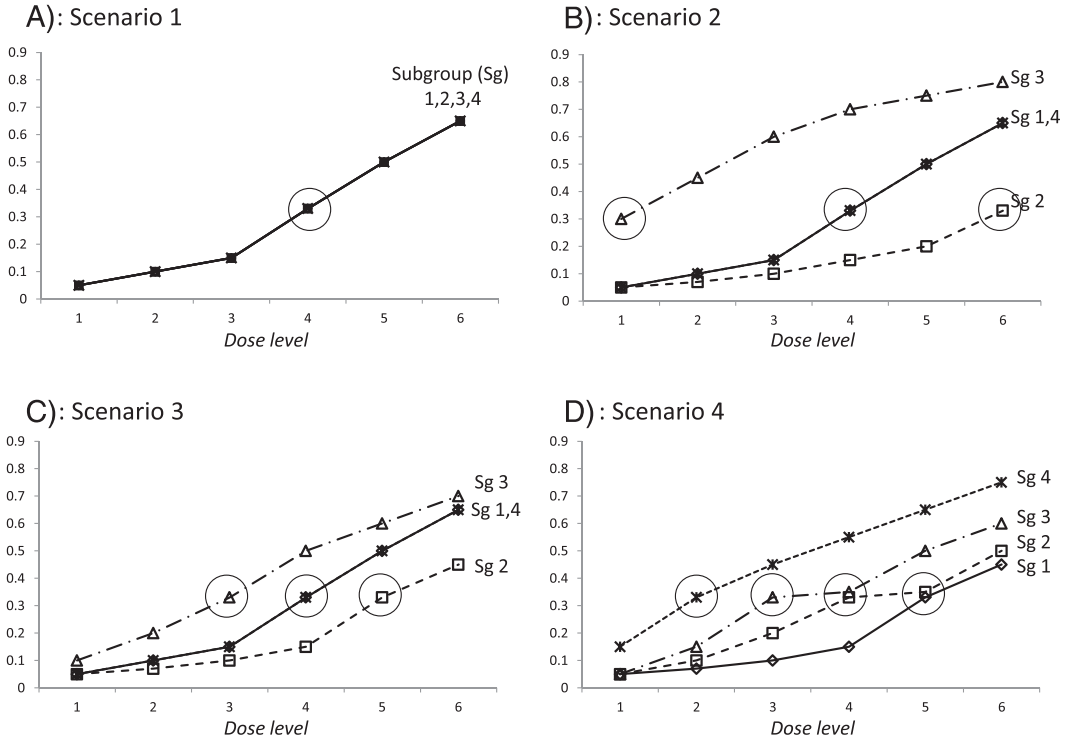

FIGURE 2 Subgroup-specific dose-toxicity curves assumed in the simulations, presented for the true dose-toxicity probabilities $\pi_{1,k}^{\text{true}}, \cdots, \pi_{6,k}^{\text{true}}$ for each subgroup (Sg): Sg 1, diamond and solid; Sg 2, square and dashed; Sg 3, triangle and dashed-dotted; Sg 4, star and dotted. Optimal doses are indicated by open circles. A, Scenario 1; B, Secnario 2; C, Scenario 3; D, Scenario 4

basis for constructing weights because this takes on the value ∞ if $\pi_{j,k}^{\text{true}} = \pi^*$. We also evaluate the statistic PCS_k, the probability of correct selection (PCS) of the dose that minimizes $|\pi_{j,k}^{\text{true}} - \pi^*|$ in subgroup k, which gives weight 1 to the dose having $\pi_{j,k}^{\text{true}}$ closest to the target and weight 0 to all other doses.

Under each scenario, we simulated the trial 1000 times using each method. The SAS program to implement HB-CRM is provided in the Supporting Information (SAS for Windows release 9.3; SAS Institute Inc., Cary, North Carolina).

3.2 | Simulation results

The operating characteristics for the 4 methods are summarized by toxicity scenarios, and the results are shown for the WPS and PCS only for $N_{\max} = 96$ and except for Scenario 1. Figures 3 and 4 show the results under the assumptions of equal and different subgroup proportions (0.25, 0.25, 0.25, 0.25) and (0.40, 0.30, 0.20, 0.10), respectively. The selection probabilities of x_j as the OD in subgroup k in Equation 8 were computed as the percentage of times that each of the methods selected x_j as the OD in each subgroup. More complete results are shown in Table S2.

Overall, the simulation study reconfirmed that ignoring subgroups resulted in undesirably low probabilities of selecting ODs, especially when the dose-toxicity relationships were largely different between subgroups. It also was reconfirmed that *K*-CRM-1-trial and 1-CRM-*K*-trials behaved about the same in a setting with more than 2 nonordered subgroups. The HB-CRM showed worse performance than *K*-CRM-1-trial and 1-CRM-*K*-trials in several cases. However, when the subgroup proportions were different, HB-CRM gave much better results in the subgroups with small proportions $\xi = 0.1$ or 0.2.

As shown in Figure 3, under the assumption of equal subgroup proportions, under Scenario 2, *K*-CRM-1-trial and 1-CRM-*K*-trials both performed best in subgroup 3, but in another subgroups the performance of HB-CRM was very similar to these 2 methods. As in Scenario 3, the performance of HB-CRM was quite favorable when the ODs were close

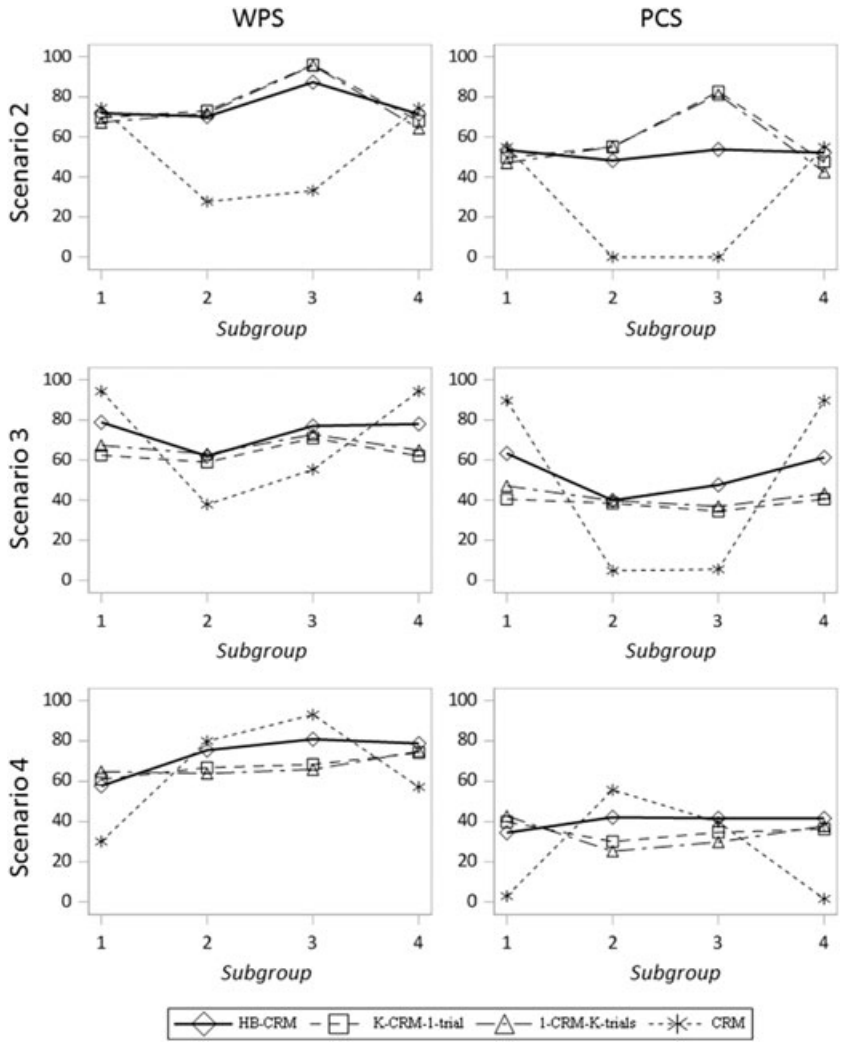

FIGURE 3 Subgroup-specific weighted probability of selection (WPS) of optimal dose and probability of correctly selecting (PCS) the optimal dose for HB-CRM (diamond and solid), *K*-CRM-1-trial (square and dashed), and 1-CRM-*K*-trials (triangle and dashed-dotted), and CRM (star and dotted) when the total sample size $N_{\max} = 96$ with assuming equal subgroup proportions $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_4) = (0.25, 0.25, 0.25, 0.25)$

to each other between patient subgroups. It appears that, in such a case, HB-CRM effectively borrows strength between subgroups through its hierarchical structure. Under Scenario 4, the WPS and PCS values of HB-CRM were higher in subgroups 2 and 3, for which 2 doses were good choices, compared to those of *K*-CRM-1-trial and 1-CRM-*K*-trials. Under the different subgroup proportions (Figure 4), HB-CRM yielded much higher WPS values in subgroup 4 under each of Scenarios 2 and 3 compared to *K*-CRM-1-trial and 1-CRM-*K*-trials. That is, the desirable effect of borrowing strength between subgroups in HB-CRM appeared to be more pronounced in subgroups with smaller numbers of patients. In other aspects, overall, the results were similar to those obtained in the case of equal subgroup proportions.

Figures 5A,B shows the WPS values for the 4 sample sizes by toxicity scenario under the equal and different subgroup proportions, respectively. The 4 columns correspond respectively to HB-CRM, *K*-CRM-1-trial, 1-CRM-*K*-trials, and CRM, starting from the left. Overall, the performance of HB-CRM improves as $N_{\max}$ increases much more than those of *K*-CRM-1-trial and 1-CRM-*K*-trials. Particularly, under Scenario 1, HB-CRM yielded high values of WPS even with $N_{\max} = 48$, and the WPS values of HB-CRM with $N_{\max} = 48$ were comparable to those of *K*-CRM-1-trial and 1-CRM-*K*-trials for $N_{\max} = 120$. This may be due to the way that HB-CRM borrows strength between subgroups. Under this scenario, it is not surprising that CRM provides the highest values of WPS, because the patients come from one population rather than multiple subpopulations.

Since similar conclusions were obtained from the simulation results in WPS and PCS, it might be acceptable to use the PCS to evaluate the operating characteristics of a dose-finding study design. However, because the toxicity probabilities for 1 or more other doses may be close to that of the optimal dose in 1 or more subgroups, the WPS may be a more suitable index to compare the performances between methods, especially in a difficult case like Scenario 4.

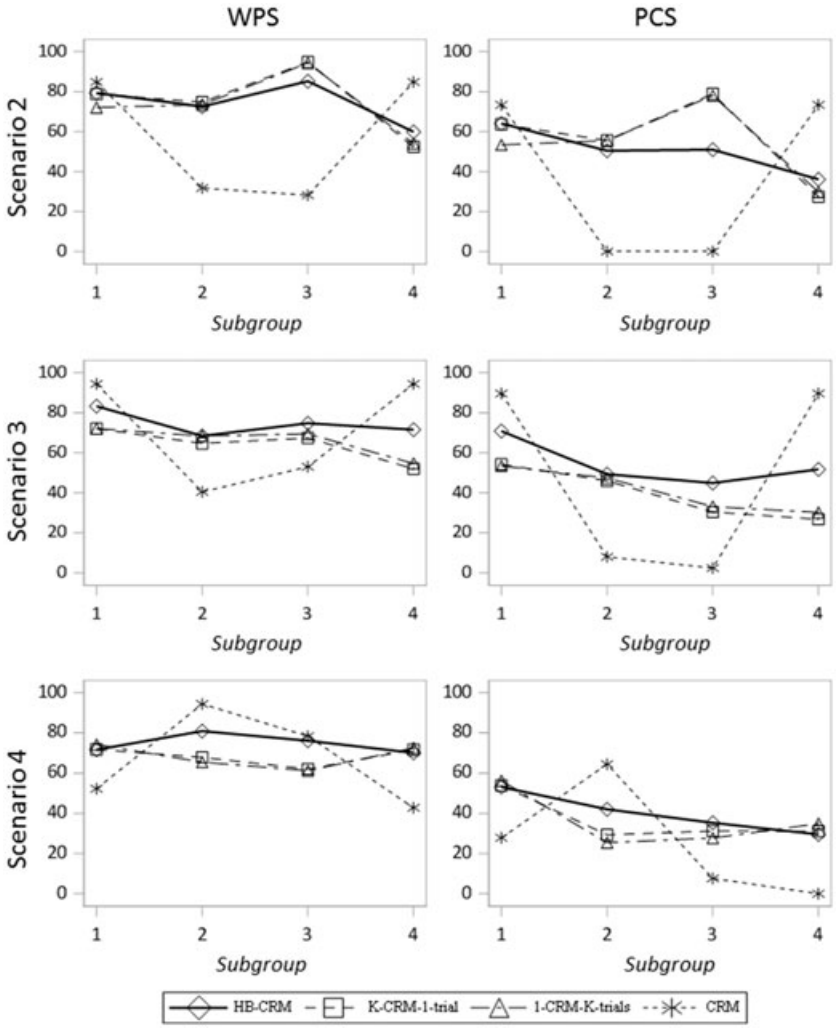

FIGURE 4 Subgroup-specific weighted probability of selection (WPS) of optimal dose and probability of correctly selecting (PCS) the optimal dose for *K*-CRM-1-trial (square and dashed), and 1-CRM-*K*-trials (triangle and dashed-dotted), and CRM (star and dotted) when the total sample size $N_{\max} = 96$ under the dose-toxicity Scenarios 2 to 4 with assuming different subgroup proportions $\boldsymbol{\xi} = (0.40, 0.30, 0.20, 0.10)$

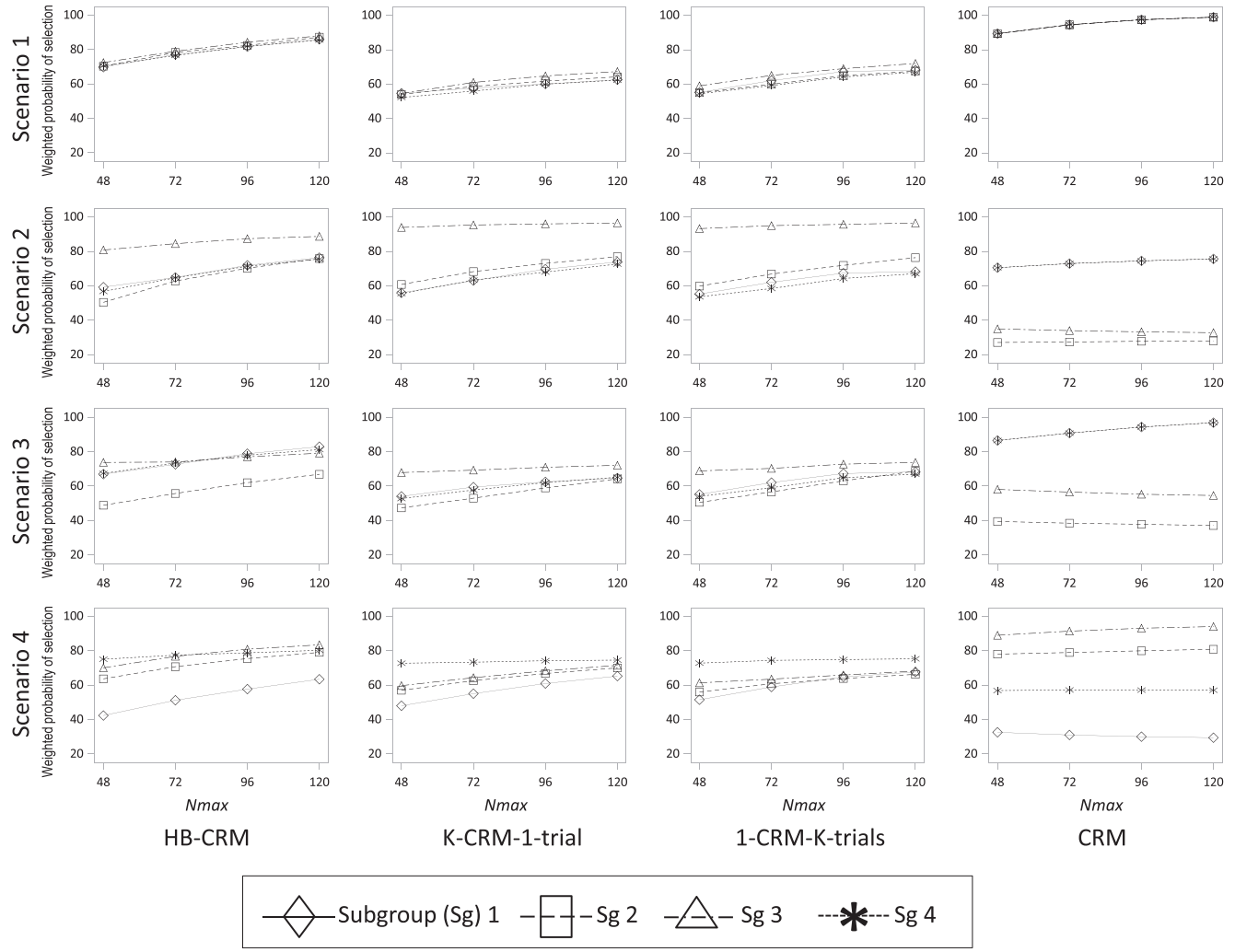

FIGURE 5A Subgroup-specific weighted probability of selection (WPS) of optimal dose for HB-CRM, K-CRM-1-trial, 1-CRM-K-trials, and CRM under the dose-toxicity scenarios 1 to 4 (from the first raw to the bottom), for maximum sample sizes $N_{\max} = 48, 72, 96, 120$, in subgroups 1: diamond and solid, 2: square and dashed, 3: triangle and dashed-dotted, 4: star and dotted. A, Assume equal subgroup proportions $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_4) = (0.25, 0.25, 0.25, 0.25)$. B, Assume different subgroup proportions $\boldsymbol{\xi} = (0.40, 0.30, 0.20, 0.10)$

4 | GUIDELINES FOR CONSTRUCTING DESIGNS

To construct a study design using the HB-CRM method, the following steps may be taken.

1. Determine the definition of toxicity, the target toxicity probability π^*, and the dose levels $(d_1, \cdots, d_J)$ to be tested.
2. Specify the patient subgroups $(1, \cdots, K)$, anticipated subgroup prevalences $\boldsymbol{\xi} = (\xi_1, \cdots, \xi_K)$, and patient accrual rates.
3. Set up the priors of the dose-finding model to be minimally informative for the prior ESS, following the approach described in Section 2.4.
4. Determine the total maximum sample size, $N_{\max}$, by running the computer program (provided in the supplementary materials) for a range of feasible values of $N_{\max}$ so that the study design has a sufficiently good performance for the WPS of the subgroups.

As a guide, in step 4 above, the values of $N_{\max}$ may range from $12 \times K$ to $30 \times K$ when the number of dose levels J is 4 to 6. If more dose levels are examined, one may consider increasing the per-subgroup sample size, as in an ordinary dose-finding trial. In some cases, some subgroups may be very small, that is, the corresponding values of $(\xi_1, \cdots, \xi_K)$ take very small values, eg, 0.05 or smaller. This is likely to occur if the number of subgroups is large, eg, $K = 10$ or 20. In such a case, due to the limited number of patients in a phase 1 trial, we strongly recommend reducing the number of subgroups, K, to a number that allows the proposed methodology to be applied in a practical way. In addition, one should take the patient accrual rates in subgroups into account when combining subgroups, to improve trial feasibility and simplify trial conduct, that is, to complete the trial within a realistic time frame. If different patient subgroups have different toxicity targets, $\pi_1^*, \cdots, \pi_K^*$, one can design the trial in the same way as a trial with a common target π^*, although one should

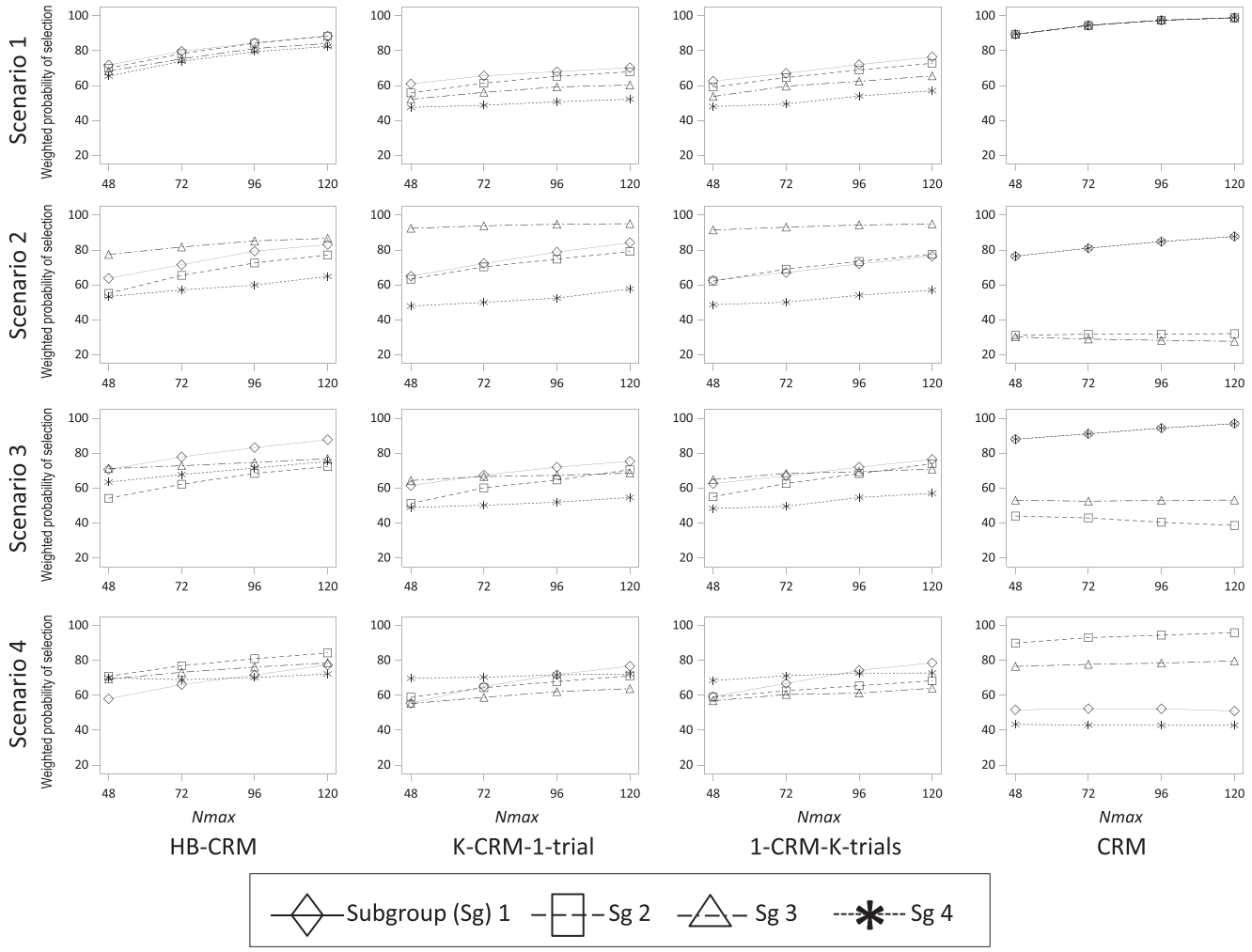

FIGURE 5B Continued

investigate the operating characteristics of the study design carefully (Figures 5).

5 | ILLUSTRATION

For illustration, on the basis of 2 real phase 1 trials, we show how HB-CRM and K-CRM-1-trial may work in practice, via simulations. To perform these simulations, we assumed true toxicity probabilities on the basis of the empirical data observed in the 2 clinical trials. The first example (Example 1) is a case where exchangeable toxicity probabilities may reasonably be assumed between patient subgroups, while exchangeability clearly should not be assumed in the second example (Example 2).

Example 1 is based on the work of Minami et al.[15] They examined 3 doses (400, 600, 800 *mg* once daily) of sonidegib (LDE225), a selective protein inhibitor, in $N = 45$ Asian patients with advanced solid tumors, including 2 racial subgroups. Subgroup 1 consisted of $N_1 = 21$ Japanese and subgroup 2 consisted $N_2 = 24$ of Hong Kong/Taiwanese, following the health authority's request. The maximum tolerated doses (MTDs) were reported to be 400 *mg* in both subgroups. For these patients, we consider it reasonable to assume, a priori, that the toxicity probabilities are exchangeable between the 2 racial subgroups. In this trial, dose-limiting toxicities (DLTs) were evaluated during the first treatment cycle. Table 2 shows the observed DLT data.

For Example 2, we use the data reported by Bendell et al.[16] They conducted a phase 1 trial to test 6 dose levels, 12.5, 25, 50, 80, 100, 150 *mg* once daily of BKM120, a pyrimidine-derived pan-PI3 K inhibitor with specific and potent activity against class I PI3Ks. A 100 *mg* was estimated to be the MTD in this trial. As a clinical background characteristic of the $N = 35$ patients enrolled in this trial, Bendell et al[16] reported the number of prior therapies for 2 subgroups, with >3 in subgroup 1, and $\leqslant 3$ in subgroup 2. The sizes of the 2 subgroups were nearly the same, with $N_1 = 18$ and $N_2 = 17$. Because, in general, patients who are more heavily pretreated are be more likely to experience toxicity, it is clearly not appropriate to assume that these 2 patient subgroups have exchangeable prior toxicity probabilities. Table 2 shows the DLTs observed during the first treatment cycle in this trial.

For each of the 2 examples, we simulated toxicity data for each of four hypothetical total sample sizes, $N = 35, 45, 70,$

TABLE 2 Illustration of HB-CRM and K-CRM-1-trial for 2 real phase 1 trials, the sonidegib trial[15] and the BKM120 trial[16]

Sonidegib trial			**Dose, mg**		
Subgroup			**400**	**600**	**800**
1 (Japanese)	No. of DLTs		2	5	0
	No. of patients		12	9	0
	Empirical Pr(DLT)		0.17	0.56	—
	Assumed true Pr(DLT)		0.15	0.55	0.65
2 (Taiwanese)	No. of DLTs		2	1	2
	No. of patients		12	8	4
	Empirical Pr(DLT)		0.17	0.13	0.50
	Assumed true Pr(DLT)		0.15	0.20	0.50
N	Method	Subgroup			
45	HB-CRM	1	**80.5**	19.5	0.0
		2	20.1	**71.8**	8.1
	K-CRM-1	1	**87.2**	12.8	0.0
		2	30.8	**60.5**	8.7
100	HB-CRM	1	**91.6**	8.4	0.0
		2	16.2	**80.6**	3.2
	K-CRM-1	1	**95.5**	4.5	0.0
		2	27.1	**69.6**	3.3

BKM120 trial			**Dose (mg)**					
			12.5	**25**	**50**	**80**	**100**	**150**
	No. of DLTs		0	0	0	1	4	1
	No. of patients		1	1	3	6	16	3
	Empirical Pr(DLT)		0.00	0.00	0.00	0.17	0.25	0.33
Subgroup								
1 (> 3 prior trts)	Assumed true Pr(DLT)		0.05	0.07	0.10	0.25	0.35	0.55
2 (⩽3 prior trts)	Assumed true Pr(DLT)		0.01	0.02	0.05	0.10	0.15	0.25
N	Method	Subgroup						
35	HB-CRM	1	0.9	8.1	39.0	**30.2**	18.0	3.8
		2	0.3	2.1	18.3	13.1	33.7	**32.5**
	K-CRM-1	1	1.2	12.6	33.1	**31.6**	17.9	3.6
		2	0.3	6.0	12.1	9.1	35.6	**36.9**
100	HB-CRM	1	0.2	3.8	25.1	**55.7**	14.2	1.0
		2	0.0	0.4	6.9	5.3	38.5	**48.9**
	K-CRM-1	1	0.3	7.4	25.0	**53.7**	12.7	0.9
		2	0.1	2.6	5.4	3.1	38.5	**50.3**

For each trial, the observed toxicity data and assumed true DLT probabilities used in the simulation first are summarized by subgroup, followed by simulation results for each method for each of 2 maximum sample sizes. Percentages of correct selection of the MTD within each subgroup are given in boldface.

and 100, with subgroup proportions $\xi = 0.467$ in Example 1 and $\xi = 0.514$ in Example 2, to mimic the proportions in the reported data. Since the implicit target range for the DLT probability was 0.16 to 0.33 to determine the MTD in both trials, we defined 0.25 as the target DLT probability for both examples. Table 2 shows the true DLT probabilities derived from the empirical data of the 2 clinical trials. In Example 1, where subgroups were determined by race, we assumed true DLT probabilities such that dose level 1 (400 *mg*) was the MTD in subgroup 1, while dose level 2 (600 *mg*) was the MTD in subgroup 2. For Example 2, where subgroups were determined by number of prior therapies, toxicity data within the patient subgroups were not given by Bendell et al.[16] Thus, we assumed true DLT probabilities so that dose levels 4 (80 *mg*) and 6 (150 *mg*) were the respective MTDs in subgroups 1 (> 3 prior therapies) and 2 (⩽ 3 prior therapies).

The simulation results for Examples 1 and 2 are summarized in Table 2 for 2 maximum sample sizes, $N = 45$ and 100 and $N = 35$ and 100, respectively. The table gives the percentages of times that each method selected each dose as the MTD in each subgroup. Correct selection percentages are given in boldface. Corresponding results for other maximum N values are summarized in the supplementary material.

In Example 1 (exchangeable case), HB-CRM performed better than K-CRM-1 overall in the 2 subgroups. In this example, Table 2 shows within-subgroup correct selection percentages of 71.8% and 80.5% for HB-CRM even with

N = 45, although these high values are due in part to there were only 3 the number of dose levels. Still, the example illustrates the ability of HB-CRM to reliably choose different optimal doses within subgroups. In Example 2 (nonexchangeable case), for N = 35, Table 2 shows that K-CRM-1 performed better than HB-CRM, especially in subgroup 2. With larger *N*, the performances of both methods improved, and the difference in performance between the 2 methods diminished. A key point is that, for phase 1 trials with patient heterogeneity and 6 or more doses levels, *N* should be larger than conventional values to obtain reliable subgroup-specific dose selections.

6 | DISCUSSION

Our simulation studies suggest that HB-CRM works well in situations where the dose-toxicity curves are expected to be similar or not largely different between multiple patient subgroups, and the exchangeability assumption is valid. This arises commonly in settings whether qualitatively different disease subgroups are included, and there is no prior knowledge about the comparative risks of toxicity in the subgroups. For overall performance quantified by the WPS or PCS, Figure 3 shows that the HB-CRM method (solid line) does well across all subgroups. In contrast, the nonhierarchical model-based methods *K*-CRM-1 and 1-CRM-*K* that account for subgroups may perform well for some subgroups but not as well for others. For example, in Scenario 2, the *K*-CRM-1 method outperforms the HB-CRM method slightly for WPS and markedly for subgroup 3 of PCS, but *K*-CRM-1 has the same or inferior performance compared to HB-CRM in Scenarios 3 and 4. Not surprisingly, the CRM that ignores subgroups has greatly inferior performance for several subgroups in each of Scenarios 2, 3, and 4. Figures 4A,B illustrate the extremely poor performance of the CRM for many subgroups in the presence of heterogeneity.

Since, in practice, one cannot know the true toxicity curves, there are 2 main messages. First, one certainly should account for known patient heterogeneity, because failure to do so is very likely to produce a selected dose that is far below optimal in some subgroups. Second, when the underlying assumptions are appropriate, the HB-CRM performs well consistently across a broad range of different dose-toxicity–subgroup scenarios, and it may be preferable to nonhierarchical model-based methods that choose subgroup-specific doses.

An important caveat is that the HB-CRM–based method is not appropriate when it is known that the subgroups are not exchangeable, with an important case being that where the risk of toxicity is known to be ordered by subgroup. If preclinical or clinical data identify multiple patient subgroups that are likely to have substantially different dose-toxicity relationships, but the HB-CRM is not appropriate, then one should use a nonhierarchical dose-finding method that chooses subgroup-specific doses.

ACKNOWLEDGMENTS

The authors thank Dr Peter Müller for his helpful advice on an earlier draft of this manuscript. We also thank the editor, associate editor, and 2 referees for their constructive comments and suggestions. Satoshi Morita research was partially supported by a Grant-in-Aid for Scientific ResearchC-24500345 from the Ministry of Health. Peter Thall research was supported by NIH/NCI grant RO1 CA 83932.

CONFLICT OF INTEREST

There are no conflicts of interest in this study.

REFERENCES

[1] S. Morita, P. F. Thall, P. Mueller, *Bayesian Anal.* **2012**, *7*, 591.
[2] J. O'Quigley, M. Pepe, L. Fisher, *Biometrics* **1990**, *46*, 33.
[3] J. O'Quigley, L. Z. Shen, A. Gamst, *J. Biopharm. Stat.* **1999**, *9*, 17.
[4] J. O'Quigley, X. Paoletti, *Biometrics* **2003**, *59*, 430.
[5] A. Ivanova, K. Wang, *Biometrics* **2006**, *25*, 2018.
[6] Z. Yuan, R. Chappell, *Clin. Trials* **2004**, *1*, 499.
[7] B. E. Storer, *Biometrics* **1989**, *45*, 925.
[8] D. H. Y. Leung, Y. G. Wang, *Control Clin. Trials* **2001**, *22*, 126.
[9] P. F. Thall, H. Nguyen, E. H. Estey, *Biometrics* **2008**, *64*, 1126.
[10] S. Liu, H. Pan, J. Xia, Q. Huang, Y. Yuan, *Stat. Med.* **2015**, *34*, 1681.
[11] A. Gelman, *Bayesian Anal.* **2006**, *1*, 515.
[12] S. Morita, P. F. Thall, P. Mueller, *Biometrics* **2008**, *64*, 595.
[13] P. F. Thall, RC Herrick, H. Nguyen, J. J. Venier, J. C. Norris,. *Clin. Trials* **2014**, *11*, 657.
[14] W. Gilks, S. Richardson, D. Spiegelhalter, *Markov Chain Monte Carlo in Practice*, Chapman & Hall, London **1996**.
[15] H. Minami, Y. Ando, B. B. Ma, J. H. Lee, H. Momota, Y. Fujiwara, L. Li, K. Fukino, K. Ito, T. Tajima, A. Mori, C. C. Lin, *Cancer Sci.* **2016**, doi:10.1111/cas.13022. [Epub ahead of print].
[16] J. C. Bendell, J. Rodon, H. A. Burris, M. de Jonge, J. Verweij, D. Birle, D. Demanse, S. S. De Buck, Q. C. Ru, M. Peters, M. Goldbrunner, J. Baselga, *J. Clin. Oncol.* **2012**, *30*, 282.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Morita S, Thall P, Takeda K. A simulation study of methods for selecting subgroup-specific doses in phase I trials. *Pharmaceut. Statist*. 2017;**16**:143–156. https://doi.org/10.1002/pst.1797

APPENDIX

The priors for the models underlying each of the 4 methods CRM, 1-CRM-*K*-trials, *K*-CRM-1-trial, HB-CRM may be constructed in the following 2 steps.

Step 1: location hyperparameters. Determine the location hyperparameters of the priors, $(\tilde{\mu}_\alpha, \mu_{\alpha,\phi})$ for the

HB-CRM and $(\tilde{\mu}_\alpha, \tilde{\mu}_\beta)$ for the 3 nonhierarchical methods, by first obtaining numerical values of the mean probability of toxicity at each of 2 doses. A convenient choice consists of the second lowest and second highest dose, denoted by $\pi(d_2, \boldsymbol{\theta})$ and $\pi(d_{J-1}, \boldsymbol{\theta})$, although other dose pairs may be used. These prior mean probabilities may be obtained by elicitation from the physicians, or on the basis of historical data. For the HB-CRM model, the prior elicitation process does not distinguish between patient subgroups, because the hierarchical model prior assumes that the toxicity probabilities of the patient subgroups are exchangeable. For each model, given the 2 elicited prior mean probabilities, the 2 equations are solved for the 2 location parameters.

Step 2: dispersion hyperparameters. Given the prior hypermeans, determine numerical values of the dispersion parameters controlling the informativeness by using prior ESS, as described in Section 2.4. These dispersion parameters are $(\tilde{\sigma}_\beta, \sigma_{\alpha,\phi}, U_\phi)$ for the HB-CRM model and $(\tilde{\sigma}_\alpha, \tilde{\sigma}_\beta)$ for each of the 3 other models. To speed up computation, compute approximate ESS values on the basis of a $\beta(a, b)$ distribution has ESS $= a + b$, by approximating the prior of any probability $\pi(\boldsymbol{\theta})$ by a $\beta(a, b)$ and matching the means and variances. One then solves the 2 equations

$$E\{\pi(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}})\} = a/(a+b)$$

and

$$var\{\pi(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}})\} = ab/\{(a+b)^2(a+b+1)\}$$

for ESS $\doteq a + b$.

Step 2a. To compute an approximate ESS within each subgroup, use the above approach to compute the β-approximated value ESS_j of the prior $\pi(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}})$ for each dose d_j, $j = 1, \ldots, J$, and use the mean of these J values as a per-subgroup approximate ESS.

Step 2b. Multiply the value of the per-subgroup ESS by the number of subgroups, K, to obtain an overall ESS.