The Annals of Applied Statistics
2022, Vol. 16, No. 4, 2481–2504
https://doi.org/10.1214/22-AOAS1600

BAYESIAN HIERARCHICAL RANDOM-EFFECTS META-ANALYSIS AND DESIGN OF PHASE I CLINICAL TRIALS

BY RUITAO LIN[1,a], HAOLUN SHI[2,d], GUOSHENG YIN[3,e], PETER F. THALL[1,b], YING YUAN[1,c] AND CHRISTOPHER R. FLOWERS[4,f]

[1]*Department of Biostatistics, The University of Texas MD Anderson Cancer Center,* [a]*rlin@mdanderson.org,* [b]*rex@mdanderson.org,* [c]*yyuan@mdanderson.org*

[2]*Department of Statistics and Actuarial Science, Simon Fraser University,* [d]*haoluns@sfu.ca*

[3]*Department of Statistics and Actuarial Science, The University of Hong Kong,* [e]*gyin@hku.hk*

[4]*Department of Lymphoma/Myeloma, The University of Texas MD Anderson Cancer Center,* [f]*crflowers@mdanderson.org*

We propose a curve-free random-effects meta-analysis approach to combining data from multiple phase I clinical trials to identify an optimal dose. Our method accounts for between-study heterogeneity that may stem from different study designs, patient populations, or tumor types. We also develop a meta-analytic-predictive (MAP) method, based on a power prior, that incorporates data from multiple historical studies into the design and conduct of a new phase I trial. Performances of the proposed methods for data analysis and trial design are evaluated by extensive simulation studies. The proposed random-effects meta-analysis method provides more reliable dose selection than comparators that rely on parametric assumptions. The MAP-based dose-finding designs are generally more efficient than those that do not borrow information, especially when the current and historical studies are similar. The proposed methodologies are illustrated by a meta-analysis of five historical phase I studies of Sorafenib and design of a new phase I trial.

1. Introduction. The aims of a phase I clinical trial of a new drug are to investigate its dose-limiting toxicity (DLT) and identify a recommended phase II dose, such as the maximum tolerated dose (MTD) from a set of candidate doses. We define the MTD as the dose having probability of DLT closest to a prespecified fixed target. Existing dose-finding designs include various versions of the algorithmic 3 + 3 design (Storer (1989)), the up-and-down design (Gezmu and Flournoy (2006)), the model-based continual reassessment method (CRM) (O'Quigley, Pepe and Fisher (1990)), escalation with overdose control (EWOC) (Babb, Rogatko and Zacks (1998)), the Bayesian logistic regression method (BLRM) (Neuenschwander, Branson and Gsponer (2008)), the Bayesian optimal interval design (BOIN) (Liu and Yuan (2015), Yuan et al. (2016)), and a nonparametric overdose control (NOC) design (Lin and Yin (2017)), among others. A major practical problem is that differences between designs, biological mechanisms of the agent, or objectives may produce different choices of an MTD.

Because the sample size of a phase I trial is typically small, the probability of correctly identifying the MTD tends to be low. To address this problem, one may take advantage of the fact that, for a new agent, quite often several clinical centers conduct independent phase I trials (Zohar, Katsahian and O'Quigley (2011)). A meta-analysis of data from multiple independent phase I trials of the same agent may improve estimation of the dose–toxicity curve and thus identification of the MTD (García et al. (2014)). This must be done carefully, however, due to between-study variability arising from the use of different dose-finding designs

Received June 2021; revised November 2021.
Key words and phrases. Bayesian adaptive method, meta-analysis, phase I clinical trials, power prior, random-effects model.

as well as differences in patient prognostic characteristics, tumor types, supportive care, and treatment administration schedules.

In most settings, compared with inferences that rely on a single study, a meta-analysis can achieve higher reliability, more accurate estimation, and greater reproducibility while also accounting for between-study variability. Research on methods for meta-analysis of phase I dose-finding studies is quite limited. Zohar, Katsahian and O'Quigley (2011) proposed a parametric CRM-based fixed-effect meta-analysis approach, but it requires the strong assumption that the selected studies are homogeneous. This is at odds with the fact that substantial between-study heterogeneity should be expected in multiple phase I studies of the same agent, due to the sources noted above. Another limitation of this parametric model based approach is that its performance is sensitive to model misspecification. For example, antineoplastic agents typically have a steep dose–toxicity curve (García et al. (2014)), which the power model used by the CRM often cannot reliably quantify. This may lead to poor performance of the meta-analysis. In other meta-analysis approaches, random-effects models are employed routinely to account for between-study heterogeneity (Higgins, Thompson and Spiegelhalter (2009)). We take this approach for meta-analysis of multiple phase I trials by assuming a random-effects model which provides more reliable results than using a model that assumes homogeneity. Recently, Ursino et al. (2021) also proposed a Bayesian meta-analysis method where the random-effects structure was modeled by the Ornstein–Uhlenbeck Gaussian process. Our approach differs from their method in the model structure, prior distribution, and the amount of information shared across dose levels.

To address the pervasive problem that the small sample sizes of most phase I studies lead to a low probability of correctly identifying the MTD, we also consider the problem of how to use available historical information when designing a new dose-finding trial. If the new and historical trials are similar, then it is natural to borrow information from the historical studies to achieve more efficient interim adaptive dose selection and greater accuracy in MTD identification. A naïve approach would adopt a "one-size-fits-all" model that incorporates all of the historical information. However, this approach is problematic if the difference between the current and historical trials is substantial (Huang and Temple (2008), Yasuda, Zhang and Huang (2008)). To address this issue, several adaptive information-borrowing designs have been proposed, particularly for bridging trials, which evaluate the effect of a treatment in a certain population, for example, pediatric patients, by using data from historical trials of the same treatment in a different population, for example, adults. For example, Morita (2011) used informative priors to incorporate historical data into a bridging study based on the CRM. Liu et al. (2015) introduced a bridging CRM to facilitate dose finding for follow-up bridging trials. More generally, several authors have proposed bridging methods for incorporating historical data in a variety of settings when planning or conducting a phase II or phase III clinical trial (Chen et al. (2011), Chow et al. (2012), Gould et al. (2012), Hobbs, Carlin and Sargent (2013), Schmidli et al. (2014)).

In practice, two important differences between a current trial and historical trials should be considered. On one hand, the information observed from the current trial may be inconsistent (i.e., nonexchangeable) with that from the historical trials. This might be caused by differences in disease types, treatment schedules, or other factors. When these differences are large, the data are not exchangeable between trials, and it is best to not borrow information from the historical trials. Some existing methods, such as the bridging CRM of Liu et al. (2015), do not take potential inconsistency into consideration appropriately, and thus their performance tends to be unsatisfactory. On the other hand, even if the new and historical trials are similar, one still must account for the intrinsic heterogeneity between studies. The methods of Morita (2011) and Liu et al. (2015), while useful, are limited by the fact that they only borrow information from one historical study, and they do not account for heterogeneity between multiple

historical studies. Existing methods also fail to agree on how much information to borrow from historical data. Ideally, a method that does this should strike a compromise between simply combining the historical and current trial data and completely ignoring the historical data.

In this paper we propose a Bayesian hierarchical random-effects meta-analysis approach to exploiting data from multiple phase I dose-finding studies when estimating the MTD in a new phase I trial. We also apply the proposed meta-analysis method to the problem of designing a new dose-finding trial. Our contributions are twofold:

(i) In a Bayesian hierarchical framework, we extend the standard *logistic-normal distribution* for the probability of DLT as a curve-free monotonic function of the dose level. Under the proposed model, independent and identically distributed (i.i.d.) study-specific parameters are assumed to follow a normal distribution, mimicking a standard random-effects meta-analysis. Our curve-free approach accounts for heterogeneity while also flexibly accommodating a wide range of possible shapes for the dose–toxicity curves.

(ii) We facilitate more informative dose-finding by assuming a power prior in the context of the proposed model and develop meta-analytic-predictive (MAP) versions of both the CRM and BOIN (Schmidli et al. (2014)). In this sense we combine standard dose-finding methods with MAP approaches. Our model may be considered an extension of the meta-analytic-predictive prior (Neuenschwander et al. (2010), Schmidli et al. (2014), Spiegelhalter, Abrams and Myles (2004)) to phase I trials. The proposed MAP-based dose-finding methods account for heterogeneity across studies and the difference between the current and historical studies.

Extensive simulation studies, reported below, show that, when current and historical trials are similar, the proposed methods are more efficient than existing dose-finding methods. When the trials are different, the proposed methods quickly move to a no-information-borrowing mode with negligible deterioration in performance. Our approach is conceptually similar to that of Ibrahim et al. (2012), with the key differences that we consider more complicated dose-finding trials with smaller sample sizes and monotone dose-response constraints. Major advantages of the proposed designs are that they adaptively borrow historical information to reduce sample sizes and achieve higher efficiency in selecting the MTD, and they accommodate different numbers and sample sizes of multiple historical trials.

The remainder of the paper is organized as follows. In Section 2 we present the real meta-analysis that motivated the proposed method. In Section 3 we develop a curve-free random-effects meta-analysis method to synthesize multiple heterogeneous dose-finding studies. In Section 4 we develop two MAP dose-finding designs based on the proposed random-effects meta-analysis model using a power prior. As an illustration, we apply the proposed methods to the motivating example in Section 5. The performance of our methods is examined by extensive simulation studies which are summarized in Sections 6 and 7. We conclude with a brief discussion in Section 8.

2. Motivating example. Sorafenib is an orally administered multikinase inhibitor that slows tumor growth by disrupting tumor microvasculature through antiproliferative, antiangiogenic, and/or proapoptotic effects. Several phase I and pharmacokinetic studies of Sorafenib have been conducted in patients with various advanced solid tumors. Zohar, Katsahian and O'Quigley (2011) analyzed five dose-finding trials of Sorafenib. Table 1 summarizes the numbers of patients treated and numbers of observed DLTs at each dose level in these trials. Combining the five trials, a total of 154 patients were treated at one of six doses: 100, 200, 300, 400, 600, or 800 mg. The trials differed in study design, patient population, tumor type, and other factors. Based on these data, Figure 1 illustrates study-specific and pooled estimates

TABLE 1

Historical data from five phase I dose-finding studies of Sorafenib. Each entry is y_{kj}/n_{kj}, where y_{kj} is the number of DLTs observed and n_{kj} is the number of patients treated at dose level j in study k

No.	Study	Dose (mg)					
		100	200	300	400	600	800
1	Clark et al.	0/3	0/3	–	1/4	1/6	3/3
2	Awada et al.	0/4	0/3	1/5	1/10	7/12	1/3
3	Moore et al.	0/3	1/6	–	0/8	3/7	–
4	Strumberg et al.	1/5	1/6	–	0/15	4/14	2/7
5	Minami et al.	0/3	1/12	–	0/6	1/6	–
Overall data		1/18	3/30	1/5	2/43	16/45	6/13
Empirical Pr(toxicity)		0.06	0.10	0.20	0.05	0.36	0.46

of dose–response curves, obtained by isotonic regression (Bril et al. (1984)). Methodological details of the isotonic regression are described in the Appendix.

The five trials have two important features. First, the estimated dose-toxicity curves have a variety of different shapes (upper panel of Figure 1) which strongly suggests that it is not appropriate to assume a fixed-effect model. Second, the estimated curves have irregular patterns. Based on the pooled data, the empirical toxicity rates for the six doses are 0.06, 0.10, 0.20, 0.05, 0.36, and 0.46. The pooled toxicity probability estimate at dose level 4 is anomalous in that it is lower than the estimates at the lower dose levels 1, 2, and 3 which violates the assumption that the dose–toxicity curves are monotone increasing. As shown in the lower panel of Figure 1, when monotonicity is imposed on the toxicity estimates, using isotonic regression (Bril et al. (1984), Yuan and Chappell (2004)), the resulting estimated dose–response curve is flat from dose level 2 to dose level 4, followed by a steep increase at dose level 5. These results suggest that parametric models, such as a simple logistic model (dashed line), might not be suitable to capture such irregularities. This suggests that a more flexible model is needed to account for heterogeneity between studies and steep or irregular dose-toxicity curves.

After completion of these five studies, several more dose-finding trials of Sorafenib were conducted with different objectives or types of patients. Some of the newer trials had estimated dose-toxicity curves similar to those of the five historical trials, while others did not. For example, Borthakur et al. (2011) conducted a phase I study of Sorafenib in 16 patients with refractory or relapsed acute leukemia and studied three doses: 200, 400, and 600 mg. The empirical toxicity rates at these doses were $0/3 = 0.00$, $1/7 = 0.14$, and $3/6 = 0.50$. Using 0.33 as the target toxicity probability, 400 mg was declared as the MTD. Due to the small sample size, this conclusion is unreliable. For example, if one less DLT had been observed at 600 mg, with 2/6 rather than 3/6 DLTs, then 600 mg would have been declared the MTD, so the final conclusion would be over turned on the outcome of a single patient. From a Bayesian viewpoint, if a Beta(0.5, 0.5) prior is assumed for the probability of toxicity at 400 mg, then the posterior would be Beta(1.5, 6.5) with 95% posterior credible interval [0.016, 0.501]. A similar computation shows that a 95% posterior credible interval for the probability of toxicity at 600 mg is [0.167, 0.833]. Since the overlap between these two posterior credible intervals is [0.167, 0.501], very little is known inferentially about the dose-toxicity curve or whether the MTD using any fixed target between 0.20 and 0.40 should be 400 mg or 600 mg. These numerical results illustrate the general problem that sample sizes of conventional phase I trials are far too small to obtain reliable inferences. When analyzing the results obtained by Borthakur et al. (2011), if historical data from the previous five trials had been incorporated in a meta-analysis, it might have led to more reliable conclusions.

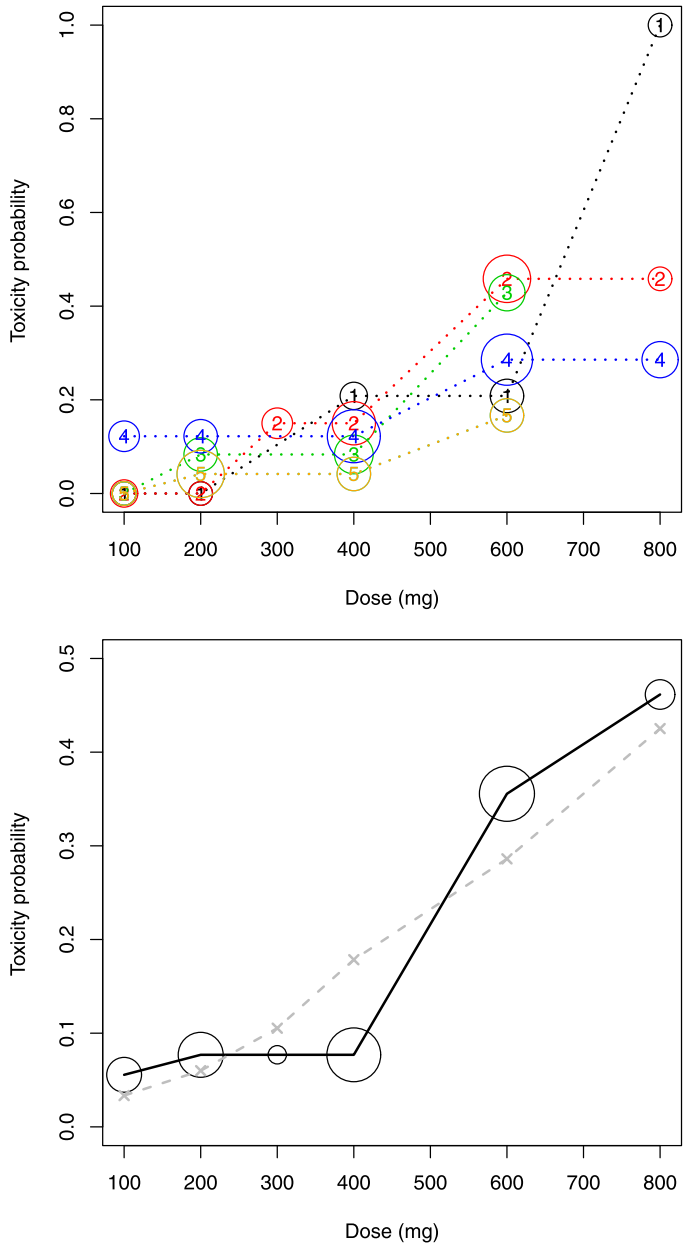

FIG. 1. *Estimated dose–toxicity curves based on the historical Sorafenib clinical trial data. The upper panel gives individual estimates of the study-specific dose–response curves, based on isotonic regression to force each estimated curve to be nondecreasing with the dose level. The lower panel gives a pooled estimate of the dose–response curve based on isotonic regression (solid line) and the estimate based on a logistic model (dashed line). The area of the circle at each dose level is proportional to the sample size.*

3. Combining multiple phase I studies.

3.1. *Probability models.* Suppose that a new agent has been studied in K historical phase I dose-finding trials with a total of J dose levels tested. In study k, let n_{kj} denote the number of patients treated at dose level j, and let y_{kj} denote the corresponding number of DLTs, for $k = 1, \ldots K$ and $j = 1, \ldots, J$. If some dose level j is not considered in trial k, we set $n_{kj} = 0$ and $y_{kj} = 0$. Denote $p_{kj} = \Pr(\text{DLT} \mid \text{dose level } j \text{ in study } k)$, with $\boldsymbol{p}_k = (p_{k1}, \ldots, p_{kJ})^T$. For each dose j the probabilities $\{p_{kj}, k = 1, \ldots, K\}$ may vary across

studies due to between-study heterogeneity. In most settings, sources of between-trial heterogeneity cannot be quantified accurately, fully observed, and often are unknown. The goal of a meta-analysis of K phase I studies is to obtain a reliable estimate of the average toxicity probability $\tilde{\boldsymbol{p}} = (\tilde{p}_1, \ldots, \tilde{p}_J)$ across the studies for use to identify a more accurate MTD.

A random-effects model can be described as a Bayesian hierarchical model, where an assumed common distribution of random effects associated with trials is used to characterize between-trial heterogeneity (Stangl and Berry (2000)). For each dose j and study k, we assume that the DLT counts follow Binomial distributions with $y_{kj} \mid p_{kj} \sim \text{Bin}(n_{kj}, p_{kj})$. To ensure monotone increasing dose-toxicity curves, with $p_{k1} < p_{k2} < \cdots < p_{kJ}$, and to allow the observed data to be shared dynamically across dose levels for each study, we reparameterize p_{kj} as

$$p_{kj} = \frac{\sum_{i=1}^{j} \exp(\phi_{ki})}{1 + \sum_{i=1}^{j} \exp(\phi_{ki})}, \tag{1}$$

where each ϕ_{ki} is real-valued. The vectors $\boldsymbol{\phi}_k = (\phi_{k1}, \ldots, \phi_{kJ})^T$ and $\boldsymbol{p}_k = (p_{k1}, \ldots, p_{kJ})^T$ have a one-to-one correspondence, given by the equations

$$\phi_{k1} = \log\left(\frac{p_{k1}}{1 - p_{k1}}\right) \quad \text{and} \quad \phi_{kj} = \log\left(\frac{p_{kj}}{1 - p_{kj}} - \frac{p_{k(j-1)}}{1 - p_{k(j-1)}}\right), \quad j = 2, \ldots, J. \tag{2}$$

Thus, $\boldsymbol{\phi}_k$ determines $\boldsymbol{p}_k$ and vice versa. For a single dose-finding study, similar dynamic models have been considered by Gasparini and Eisele (2000), Yin, Li and Ji (2006), and Liu and Johnson (2016). The reparameterization (1) facilitates exchanging information across dose levels, and it also provides a flexible curve-free model for the dose-toxicity relationship, since no parametric assumption is imposed on the toxicity probability p_{kj} as a function of the dose level j. Because the J real-valued parameters $\boldsymbol{\phi}_k$ determine $\boldsymbol{p}_k$, the curve-free model (2) is very flexible and can accommodate a wide range of dose–response relationships.

To account for heterogeneity between trials, we assume the following hierarchical model:

$$\begin{aligned}
&\text{Level 1}: \quad y_{kj} \mid \boldsymbol{\phi}_k \sim \text{Bin}\left(n_{kj}, \frac{\sum_{i=1}^{j} \exp(\phi_{ki})}{1 + \sum_{i=1}^{j} \exp(\phi_{ki})}\right),\\
&\text{Level 2}: \quad \boldsymbol{\phi}_k \mid \tilde{\boldsymbol{\phi}}, \sigma^2 \sim \mathbf{N}_J(\boldsymbol{\phi}_k \mid \tilde{\boldsymbol{\phi}}, \sigma^2 \mathbf{I}_J),\\
&\text{Level 3}: \quad \tilde{\boldsymbol{\phi}} \sim \pi(\tilde{\boldsymbol{\phi}}), \quad \sigma^2 \sim \pi(\sigma^2),
\end{aligned} \tag{3}$$

for each $k = 1, \ldots, K$ and $j = 1, \ldots, J$. Level 1 gives the likelihood, Level 2 gives the priors on the $\boldsymbol{\phi}_k$'s, and Level 3 specifies the hyperpriors on the prior mean vector $\tilde{\boldsymbol{\phi}} = (\tilde{\phi}_1, \ldots, \tilde{\phi}_J)^T$ and variance σ^2. Here, $\mathbf{N}_J(\boldsymbol{\phi}_k \mid \tilde{\boldsymbol{\phi}}, \sigma^2 \mathbf{I}_J)$ denotes a J-variate normal distribution with mean vector $\tilde{\boldsymbol{\phi}}$ and $J \times J$ covariance matrix $\boldsymbol{\Sigma}_J = \sigma^2 \mathbf{I}_J$, where $\mathbf{I}_J$ is a $J \times J$ identity matrix. We assume that $\tilde{\boldsymbol{\phi}}$ and σ^2 are independent, with σ^2 following a half-t prior distribution with one degree of freedom (i.e., a half-Cauchy distribution) and scale parameter 25 (Gelman (2006)), and that $\tilde{\boldsymbol{\phi}}$ follows a prior that is the product of J independent normal distributions with zero means and large variances, such as 10.

This hierarchical model has the following desirable features:

(i) The study-specific $\boldsymbol{\phi}_k$'s are modeled as random effects sampled from a common distribution which is a standard approach in random-effects meta-analysis. In the J-variate normal distribution for $\boldsymbol{\phi}_k$, the parameter σ^2 accounts for heterogeneity between the K trials. If $\sigma^2 = 0$, the model reduces to a fixed-effect model in which all study-specific toxicity probabilities at each dose are homogeneous. The random-effects model (3) borrows information across all trials, facilitates estimation of trial-specific dose-toxicity curves, and provides an average toxicity probability $\tilde{p}_j$ over the K trials, obtained by plugging $\tilde{\boldsymbol{\phi}}$ into equation (1).

(ii) When $J = 1$, the distribution of p_{k1} based on the hierarchical model (3) reduces to the so-called *logistic-normal distribution* which is a popular choice for dealing with hierarchically exchangeable data (Aitchison and Shen (1980), Lenk (1988)). Our hierarchical random-effects model thus extends the logistic-normal distribution for a single-dose DLT probability to accommodate a vector of J monotonically increasing DLT probabilities.

Here, we use a simple structure of the covariance matrix for $\boldsymbol{\Sigma}_J$ (i.e., $\boldsymbol{\Sigma}_J = \sigma^2 \mathbf{I}_J$) so that the latent variables $\phi_{k1}, \ldots, \phi_{kJ}$ are modeled independently. Ursino et al. (2021) proposed to use the Ornstein–Uhlenbeck process for $\boldsymbol{\Sigma}_J$ which allows more information sharing when doses are closer. Such an Ornstein–Uhlenbeck process can also be applied to our hierarchical model. Since the sample sizes for phase I clinical trials are typically limited, however, we find that this more sophisticated covariance structure does not offer much performance gain to our approach.

Let $\mathcal{D}_K$ denote the observed data from the K historical studies. Denoting the JK-dimensional real-valued vector $\boldsymbol{\Phi} = (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_K)$, the full likelihood is

$$L(\boldsymbol{\Phi} \mid \mathcal{D}_K) \propto \prod_{k=1}^{K} \prod_{j=1}^{J} p_{kj}^{y_{kj}} (1 - p_{kj})^{n_{kj} - y_{kj}} = \prod_{k=1}^{K} \prod_{j=1}^{J} \frac{\{\sum_{i=1}^{j} \exp(\phi_{ki})\}^{y_{kj}}}{\{1 + \sum_{i=1}^{j} \exp(\phi_{ki})\}^{n_{kj}}}.$$

The joint posterior distribution based on $\mathcal{D}_K$ is

$$\pi\big(\boldsymbol{\Phi}, \tilde{\boldsymbol{\phi}}, \sigma^2 \mid \mathcal{D}_K\big) \propto L(\boldsymbol{\Phi} \mid \mathcal{D}_K) f\big(\boldsymbol{\Phi} \mid \tilde{\boldsymbol{\phi}}, \sigma^2\big) \pi(\tilde{\boldsymbol{\phi}}) \pi\big(\sigma^2\big), \tag{4}$$

where $f(\boldsymbol{\Phi} \mid \tilde{\boldsymbol{\phi}}, \sigma^2)$ is the density of the multivariate normal distribution given in (3). We compute the posterior distribution (4) using Markov chain Monte Carlo with a Gibbs sampler.

3.2. *Incorporating data from an ongoing trial.* During conduct of an ongoing trial of the agent, our framework adaptively incorporates newly collected data for making real-time statistical inferences about the dose–toxicity curve. Indexing the current trial by $k = 0$, let $\mathcal{D}_0 = \{(y_{01}, n_{01}), \ldots, (y_{0J}, n_{0J})\}$ denote the current data, and let $\boldsymbol{p}_0 = (p_{01}, \ldots, p_{0J})^T$ denote the dose–toxicity probabilities of the current trial. We reparameterize $\boldsymbol{p}_0$ through $\boldsymbol{\phi}_0 = (\phi_{01}, \ldots, \phi_{0J})^T$ using equation (1).

We exploit the historical data $\mathcal{D}_K$ to obtain more efficient inferences for $\boldsymbol{\phi}_0$. The degree to which the historical data are informative for an ongoing trial depends on the similarity of the design and study characteristics between the current and historical trials. This is determined by the degree to which the historical data and the current trial data agree. If the historical and current trial data are not exchangeable, then borrowing too much from the historical data may lead to an inaccurate MTD estimate. In contrast, if there is good agreement between the historical and current trial data, then information borrowing in real time can lead to a much more efficient trial. To quantify and estimate the degree of agreement between the historical and current trials, we use the ideas of a power prior (Ibrahim and Chen (2000)), a commensurate prior (Hobbs et al. (2011)), and a meta-analytic-predictive prior (Schmidli et al. (2014)). For hierarchical modeling, Chen and Ibrahim (2006) established a formal analytic connection between the power parameter and the variance component of the hierarchical model. Intuitively, the power prior adaptively inflates the variances of the historical studies. To apply this, we replace the Level 2 and Level 3 components in the hierarchical model (3) as follows:

$$\text{Level 2}: \quad \boldsymbol{\phi}_k \mid \alpha, \tilde{\boldsymbol{\phi}}, \sigma^2 \sim \begin{cases} \pi_0(\boldsymbol{\phi}_0) \mathbf{N}_J\big(\boldsymbol{\phi}_0 \mid \tilde{\boldsymbol{\phi}}, \alpha\sigma^2 \mathbf{I}_J\big), & k = 0, \\ \mathbf{N}_J\big(\boldsymbol{\phi}_k \mid \tilde{\boldsymbol{\phi}}, \sigma^2 \mathbf{I}_J\big), & k = 1, \ldots, K, \end{cases}$$

$$\text{Level 3}: \quad \tilde{\boldsymbol{\phi}} \sim \pi(\tilde{\boldsymbol{\phi}}), \qquad \sigma^2 \sim \pi\big(\sigma^2\big), \qquad \alpha \sim \pi(\alpha),$$

where $\pi_0(\boldsymbol{\phi}_0)$ denotes a noninformative prior on $\boldsymbol{\phi}_0$ and $\alpha \in [1, \infty)$ is the power (or commensurate) parameter that controls the degree of borrowing (Hobbs et al. (2011), Ibrahim and Chen (2000)). A key property of this model is that α corresponds to a multiplicative rescaling by $\sqrt{\alpha}$, that is, the variance of $\boldsymbol{\phi}_0$ in the multivariate normal prior distribution is inflated by the power parameter α. In Level 3 we use independent priors for $(\alpha, \tilde{\boldsymbol{\phi}}, \sigma^2)$. For brevity, we abuse the notation slightly by using π generically to denote the marginal priors of the elements of $(\alpha, \tilde{\boldsymbol{\phi}}, \sigma^2)$. The Level 3 priors $\pi(\tilde{\boldsymbol{\phi}})$ and $\pi(\sigma^2)$ were described in Section 3.1, and we will describe the prior $\pi(\alpha)$ below in Section 4.3.

To show how the model connects the historical and current trial data, given a value of α, the marginal power prior distribution for $\boldsymbol{\phi}_0$, based on the historical data $\mathcal{D}_K$, can be computed as

$$\pi(\boldsymbol{\phi}_0 \mid \mathcal{D}_K, \alpha) \propto \int \pi_0(\boldsymbol{\phi}_0) f\big(\boldsymbol{\phi}_0 \mid \tilde{\boldsymbol{\phi}}, \alpha\sigma^2 \mathbf{I}_J\big) \pi\big(\tilde{\boldsymbol{\phi}}, \sigma^2 \mid \mathcal{D}_K\big)\, \mathrm{d}\tilde{\boldsymbol{\phi}}\, \mathrm{d}\sigma^2,$$

where $f(\boldsymbol{\phi}_0 \mid \tilde{\boldsymbol{\phi}}, \alpha\sigma^2 \mathbf{I}_J)$ is the multivariate normal distribution of $\boldsymbol{\phi}_0$ and $\pi(\tilde{\boldsymbol{\phi}}, \sigma^2 \mid \mathcal{D}_K)$ is the posterior distribution of $(\tilde{\boldsymbol{\phi}}, \sigma^2)$ based on the historical data $\mathcal{D}_K$. The posterior distribution of $\boldsymbol{\phi}_0$ based on the combined datasets $\mathcal{D}_0 \cup \mathcal{D}_K$ can be represented as

$$\begin{aligned}
\pi(\boldsymbol{\phi}_0 \mid \mathcal{D}_0, \mathcal{D}_K) &\propto \int \pi\big(\boldsymbol{\phi}_0, \alpha, \tilde{\boldsymbol{\phi}}, \sigma^2 \mid \mathcal{D}_0, \mathcal{D}_K\big)\, \mathrm{d}\alpha\, \mathrm{d}\tilde{\boldsymbol{\phi}}\, \mathrm{d}\sigma^2 \\
&\propto \int L(\boldsymbol{\phi}_0 \mid \mathcal{D}_0) L(\boldsymbol{\Phi} \mid \mathcal{D}_K) \pi_0(\boldsymbol{\phi}_0) f\big(\boldsymbol{\phi}_0 \mid \tilde{\boldsymbol{\phi}}, \alpha\sigma^2 \mathbf{I}_J\big) \\
&\quad \times f\big(\boldsymbol{\Phi} \mid \tilde{\boldsymbol{\phi}}, \sigma^2 \mathbf{I}_J\big) \pi\big(\alpha, \tilde{\boldsymbol{\phi}}, \sigma^2\big)\, \mathrm{d}\alpha\, \mathrm{d}\tilde{\boldsymbol{\phi}}\, \mathrm{d}\sigma^2,
\end{aligned}$$

where $L(\boldsymbol{\phi}_0 \mid \mathcal{D}_0)$ is the likelihood function based on the current trial data.

Although the power prior parameter originally proposed by Ibrahim and Chen (2000) has domain [0, 1], here we use the parameterization where $\alpha \in [1, \infty)$ is the inverse of the original power parameter. For example, for the five numerical values $\{5, 25, 45, 65, 85\}$ of α, the corresponding power parameters, based on the definition of Ibrahim and Chen (2000), are the inverses $\{0.20,\ 0.04,\ 0.022,\ 0.015,\ 0.011\}$, and the scale inflation factors are the square roots $\{2.2,\ 5.0,\ 6.7,\ 8.1,\ 9.2\}$. Note that $\pi_0(\boldsymbol{\phi}_0)$ is noninformative, with $\alpha = 1$ corresponding to full information borrowing by simply pooling the samples. In this case the current trial data fall into the random-effects structure of the historical data, and the power prior distribution of $\boldsymbol{\phi}_0$ is determined mainly by the distribution of $\boldsymbol{\phi}_0$ based on $\mathcal{D}_K$. The amount of cross-study borrowing decreases as α increases, and this corresponds to greater heterogeneity between the historical and current data. When the power parameter α is sufficiently large, the historical data will have negligible impact on $\boldsymbol{\phi}_0$. The impact of the historical data on the current trial also depends on the posterior of the heterogeneity parameter σ^2. If there are large differences among historical studies, when σ^2 is large, then the degree of borrowing from the historical data is small. If the differences are small, then more information will be borrowed. While we consider a single α for the entire model, it is also possible to specify a different α for each historical trial. However, this is feasible only if the sample sizes and the number of historical studies are large.

In the prior specification, α can be assumed to be either known or unknown, depending on whether the exchangeability between the historical and current trials can be defined clearly based on the available information or input from clinical investigators. For example, if the historical and current trials have exactly the same study characteristics and patient population, then α can be assumed to equal 1. For random α a prior distribution $\pi(\alpha)$ can be used which is independent of $\tilde{\boldsymbol{\phi}}$ and σ^2, and in this case the data can adaptively choose the amount of information borrowed from the historical studies. However, a *caveat* with this approach is that α may be identified weakly in the model due to overly sparse data in phase I clinical trials

so that estimation is sensitive to the choice of the prior for the proposed models, as noted by Hobbs, Sargent and Carlin (2012). This instability can be mitigated by assuming a discrete uniform prior with finite support for $\alpha \in [1, \infty)$ (Yin and Ibrahim (2005)). For example, one may specify M fixed values $1 \le \alpha_1 < \alpha_2 < \cdots < \alpha_M$ and assume $\Pr(\alpha = \alpha_m) = 1/M$, a priori for $m = 1, \ldots, M$. This prior may help to compensate for a small amount of empirical data from the current phase I trial and may be constructed to optimize the operating characteristics of the planned dose-finding design. We will discuss the choice of $\pi(\alpha)$ in more detail in Section 4.

4. Meta-analytic-predictive (MAP) dose finding. In this section, we explain how the random-effects meta-analysis model given in Section 3 and its predictive framework can be incorporated into an existing model-based dose-finding design. We extend the CRM and BOIN designs, calling these extensions the meta-analytic-predictive CRM (MAP-CRM) and meta-analytic-predictive BOIN (MAP-BOIN). Both MAP-based dose-finding designs are capable of adaptively borrowing information. If the data from the current and historical trials disagree, the current trial data play a dominant role in adaptively choosing the dose levels. If the data are similar, then the MAP-based design adaptively borrows strength from the historical studies for sequential decision making.

4.1. *An MAP-CRM design.* The basic idea of the CRM is that a dose–toxicity model is fit repeatedly to the accumulating data with each new patient cohort assigned the dose that has the estimated toxicity probability, based on the most recent data, closest to a prespecified fixed target probability, θ. In practice, this is typically done for successive cohorts of three or possibly two patients. Let $\mathcal{D}_0(n)$ denote the data from the first n patients in the ongoing trial. Under our model the posterior mean $\hat{\boldsymbol{p}}_0 = (\hat{p}_{01}, \ldots, \hat{p}_{0J})$ of $\boldsymbol{p}_0$ is computed by averaging over the posterior of $\boldsymbol{\phi}_0$,

$$\hat{p}_{0j} = \int \frac{\sum_{i=1}^{j} \exp(\phi_{0i})}{1 + \sum_{i=1}^{j} \exp(\phi_{0i})} \pi\big(\boldsymbol{\phi}_0 \mid \mathcal{D}_0(n), \mathcal{D}_K\big)\, \mathrm{d}\boldsymbol{\phi}_0, \quad j = 1, \ldots, J. \tag{5}$$

Thus, the MAP-CRM estimator $\hat{\boldsymbol{p}}_0$ adaptively incorporates the current and historical trial data across all dose levels. A trial using the MAP-CRM design is conducted as follows:

(i) For safety, the trial starts by treating the first cohort one dose level below the MTD identified by a preliminary meta-analysis using the historical data.

(ii) At any point in the trial, let j^{curr} denote the current dose level. To choose a dose level for the next cohort, calculate $\hat{\boldsymbol{p}}_0$ and identify the dose level j^* having estimated toxicity probability closest to θ,

$$j^* = \underset{j \in \{1, \ldots, J\}}{\arg\min} \, |\hat{p}_{0j} - \theta|.$$

If $j^{\text{curr}} > j^*$, the dose level is deescalated to $j^{\text{curr}} - 1$; if $j^{\text{curr}} < j^*$, the dose level is escalated to $j^{\text{curr}} + 1$; otherwise, the dose level j^{curr} is retained to treat the next cohort.

(iii) The trial is terminated early, due to excessive toxicity, with no MTD chosen, if the lowest dose level is unsafe, formally if

$$\Pr\big(p_{01} > \theta \mid \mathcal{D}_0(n), \mathcal{D}_K\big) = \int_{-\infty}^{\log\{\theta/(1-\theta)\}} \pi\big(\phi_{01} \mid \mathcal{D}_0(n), \mathcal{D}_K\big)\, \mathrm{d}\phi_{01} > c_U,$$

where

$$\pi\big(\phi_{01} \mid \mathcal{D}_0(n), \mathcal{D}_K\big) = \int \pi\big(\boldsymbol{\phi}_0 \mid \mathcal{D}_0(n), \mathcal{D}_K\big)\, \mathrm{d}\phi_{02} \ldots \mathrm{d}\phi_{0J},$$

and c_U is a fixed cut-off probability.

(iv) If the trial is not stopped early and the maximum sample size is reached, then the dose level j^* with estimated toxicity probability closest to θ is selected as the MTD.

In practice, c_U is a tuning parameter chosen via preliminary simulations to obtain a design that stops with high probability if the lowest dose is truly unsafe with typical values in the range $0.80 \le c_U \le 0.95$.

Cunanan and Koopmeiners (2018) also considered a hierarchical modeling approach for phase I dose-finding studies. Their approach applies to the case of multiple studies conducted simultaneously, however, where it is difficult for a hierarchical model to detect heterogeneity early when the data are sparse. In contrast, our approach has only one ongoing trial and multiple completed historical trials, providing rich information about heterogeneity at the start of the new trial.

4.2. *An MAP-BOIN design.* BOIN is a model-assisted design that has simple and transparent dose-finding rules yet provides satisfactory operating characteristics comparable to other model-based designs, such as the CRM. The dose-assignment decision of BOIN is guided by comparing the empirical toxicity probability y_{0j}/n_{0j} at the current dose level with a pair of predetermined optimal dose escalation and deescalation boundaries (λ_1, λ_2). Thus, BOIN is myopic in that it uses only the data at the current dose level for decision making, while ignoring all other data.

Let $\mathcal{D}_0^j(n) = (y_{0j}, n_{0j})$ denote the current observed data at dose level j from the first n patients. To combine the proposed meta-analysis model with BOIN, we estimate p_{0j} using the data $\mathcal{D}_0^j(n)$ as well as information adaptively borrowed from $\mathcal{D}_K$. The estimator is

$$\bar{p}_{0j} = \int \frac{\sum_{i=1}^{j} \exp(\phi_{0i})}{1 + \sum_{i=1}^{j} \exp(\phi_{0i})} \pi\big(\boldsymbol{\phi}_0 \mid \mathcal{D}_0^j(n), \mathcal{D}_K\big)\, \mathrm{d}\boldsymbol{\phi}_0, \quad j = 1, \ldots, J.$$

The amount of information borrowed from the historical studies by the MAP-BOIN estimator $\bar{\boldsymbol{p}}_0 = (\bar{p}_{01}, \ldots, \bar{p}_{0J})$ reflects the degree of heterogeneity between $\mathcal{D}_0^j(n)$ and $\mathcal{D}_K$. The rationale behind this estimator is that, when there is a conflict between $\mathcal{D}_0^j(n)$ and $\mathcal{D}_K$, the posterior estimator $\bar{p}_{0j}$ should be close to the empirical estimate y_{0j}/n_{0j}. Thus, if little information is borrowed from $\mathcal{D}_K$, the proposed MAP-BOIN design tends to coincide with the standard BOIN design. On the other hand, if there is little conflict, then more information from the historical data is borrowed, and the final posterior estimator will be more efficient. Given current dose level j^{curr}, to choose a dose for the next cohort we first calculate the posterior mean estimate $\bar{p}_{0j^{\text{curr}}}$. If $\bar{p}_{0j^{\text{curr}}} \le \lambda_1$, the dose level is escalated to $j^{\text{curr}} + 1$; if $\bar{p}_{0j^{\text{curr}}} \ge \lambda_2$, the dose level is deescalated to $j^{\text{curr}} - 1$; otherwise, j^{curr} is retained for the next cohort. The remaining rules are the same as those of the MAP-CRM design. At the end of the trial, the MTD is estimated based on (5) by adaptively pooling the data from the current and historical trials.

The MAP-BOIN design inherits the practical advantage of BOIN that dose escalation and deescalation rules can be predetermined prior to the trial conduct and summarized in a decision-making table (e.g., Table 2). This can be done because the decision of BOIN only depends on the current dose level's data. Enumerating all possible outcome combinations $\mathcal{D}_0^j(n)$ and computing $\bar{p}_{0j}$ allows one to determine the possible decisions. Unlike the BOIN design, which applies the same dose escalation/de-escalation boundaries to all dose levels, the decision boundaries for MAP-BOIN vary across dose levels, since information is borrowed from the historical data.

TABLE 2

Dose escalation and de-escalation boundaries of the BOIN and MAP-BOIN designs in the Sorafenib trial. The boundaries of the MAP-BOIN design are calculated based on the historical data of the five studies in Table 1, *which vary with dose levels*

	Escalation boundary							De-escalation boundary						
		MAP-BOIN (Dose level)							MAP-BOIN (Dose level)					
n_j	BOIN	1	2	3	4	5	6	BOIN	1	2	3	4	5	6
3	0	0	0	0	0	0	0	2	2	2	2	2	1	1
6	1	1	1	0	0	0	0	3	3	3	3	3	3	2
9	2	2	2	2	2	1	1	4	4	4	4	4	4	4
12	3	3	3	3	3	2	2	5	6	6	6	6	5	5
15	3	4	4	4	4	3	2	6	7	7	7	7	6	6
18	4	5	5	5	4	4	3	8	8	8	8	8	8	7
21	5	5	5	5	5	4	4	9	9	9	9	9	9	8

If the number of DLTs at the current dose level is less than or equal to the escalation boundary, then the dose for the new cohort is escalated to the next higher dose level; if the number of DLTs at the current dose level is greater than or equal to the de-escalation boundary, then the dose for the new cohort is de-escalated to the next lower dose level; otherwise, the dose stays at the same level for the new cohort.

4.3. *Calibrating the power prior parameter.* Because the amount of information borrowed from the historical data is determined by the power parameter α, the choice of the prior $\pi(\alpha)$ plays a critical role in the performance of the MAP dose-finding designs. It is important to calibrate a suitable prior $\pi(\alpha)$ to strike a balance between "fully borrowing" ($\alpha = 1$) and "no borrowing" ($\alpha = \infty$) from the historical data. If α is close to 1, then fully borrowing the historical data may lead to inappropriate dose-assignments if there is disagreement between the historical data and current trial data. If α is very large, then there is very little borrowing from the historical data which may cause a loss of efficiency.

To formalize prior specification for α, we define a set of M_α equidistant values having equal prior probability $1/M_\alpha$. Let α_0 denote the minimum value in the discrete support and d_α the distance between the values in the support so that the support for α is the set $\{\alpha_0, \alpha_0 + d_\alpha, \ldots, \alpha_0 + (M_\alpha - 1)d_\alpha\}$. To form a candidate set of prior distributions, $\mathcal{A}(\alpha)$, we first fix M_α and investigate various values of α_0 and d_α.

Our motivation for optimizing the prior of α is twofold: (a) to improve accuracy in selection of the MTD and (b) to have a low probability of allocating patients to overly toxic dose levels. To do this, we consider two metrics: the correct selection (CS) percentage of the MTD and the overdose allocation (OA) percentage, defined as the percentage of patients allocated to dose levels higher than the underlying true MTD. As a criterion for selecting the optimal candidate prior $\pi(\alpha)$, we use the rank of the weighted operating characteristics (RWOC). The RWOC is based on a weighted statistic computed from J simulations, where the jth simulation consists of a large number of random scenarios in which dose level j is the MTD among J prespecified dose levels. The procedure for computing the RWOC requires a total of J simulations, carried out as follows:

1. In the jth simulation, randomly generate a total of T (say, $T = 10,000$) toxicity probability vectors, each satisfying the constraint that dose level j is the MTD. To generate a toxicity probability vector satisfying this constraint, draw a vector of J i.i.d. uniform random values from [0, 1] as the toxicity probabilities, sort the vector from smallest to largest, and identify the MTD (i.e., the dose has probability of DLT closest to θ) among the J values. If the MTD is located at level j, accept the toxicity probability vector; otherwise, discard the vector and repeat the process until the MTD is located at the jth dose level.

2. Given a candidate prior $\pi(\alpha)$ and M underlying dose–response probabilities, compute the proportion of trials, denoted by CS_j^π, which correctly select level j as the MTD, and the proportion of patients allocated to dose levels higher than the true MTD, denoted by OA_j^π.

3. Repeat this process for each $j = 1, \ldots, J$, and compute the weighted averages of the correct selection percentages, $\overline{\text{CS}}^\pi = \sum_{j=1}^J w_j \text{CS}_j^\pi$, and overdose allocation percentages, $\overline{\text{OA}}^\pi = \sum_{j=1}^J w_j \text{OA}_j^\pi$, where $w_j > 0$ is the weight for $j = 1, \ldots, J$, with $\sum_{j=1}^J w_j = 1$.

4. Repeat the above computations for each candidate prior, $\pi(\alpha) \in \mathcal{A}(\alpha)$, and compute its RWOC^π as $R(\overline{\text{CS}}^\pi) + R(1 - \overline{\text{OA}}^\pi)$, where $R(\cdot)$ is the rank among the candidate priors.

5. The optimal prior is chosen to minimize RWOC^π, formally $\pi^*(\alpha) = \arg\min_{\pi(\alpha)\in\mathcal{A}(\alpha)}\{\text{RWOC}^\pi\}$.

A key aspect of this approach is that $\overline{\text{CS}}^\pi$ and $\overline{\text{OA}}^\pi$ are average values obtained from a large number of randomly chosen scenarios rather than a few scenarios that may have been cherry picked to make a prior appear to have good properties. In practice, the weights $\{w_1, \ldots, w_J\}$ may be elicited from the clinical investigators. For example, let j^* denote the estimated MTD from the meta-analysis of historical data. If the clinicians, based on their understanding and judgment of the similarity between the historical and current trials, have an a priori belief that there is, at least, an 80% chance that the MTD should not differ from the historically estimated MTD j^* by more than one level then, for example, one may assign weights of 0.15, 0.50, and 0.15 to the simulation values $j^* - 1$, j^*, and $j^* + 1$, respectively, and equal weights summing to 0.20 for the rest of the simulations.

When evaluating each candidate prior, its expected sample size (ESS) (Lee et al. (2015), Morita, Thall and Müller (2008), Neuenschwander et al. (2020)) may be used to avoid assuming a prior having an ESS that is unacceptably large, compared with the trial sample size. The ESS of $\pi(\alpha)$ may be approximated as follows. First, assume a vague prior on $\boldsymbol{\phi}_0$, using an improper prior, $\phi_{0j} \overset{\text{i.i.d.}}{\propto} 1$, $j = 1, \ldots, J$, since the performance of the proposed methods is not sensitive to the choice of $\pi(\boldsymbol{\phi}_0)$, provided that it is noninformative. Next, specify the prior distribution for $(\tilde{\boldsymbol{\phi}}, \sigma^2)$ which is the same as that used in the meta-analysis.

After specifying all priors, simulate 10,000 or more pseudo samples of $\boldsymbol{p}_0 = (p_{01}, \ldots, p_{0J})$, and use the method of moments to approximate the distribution of the simulated samples of p_{0j} with a Beta(a_j, b_j) distribution, for $j = 1, \ldots, J$. Under a Beta-Binomial model, a Beta(a_j, b_j) distribution can be thought of as the posterior from an experiment in which, in a sample of size $a_j + b_j$, one observes a_j successes and b_j failures, after assuming a very vague Beta(c_j, d_j) prior distribution with $0 < c_j, d_j < \epsilon$ for arbitrarily small ϵ. Since the Beta$(a_j + c_j, b_j + d_j)$ posterior is arbitrarily close to Beta(a_j, b_j) for small ϵ, one may think of $a_j + b_j$ as the ESS (Morita, Thall and Müller (2008)). Consequently, for our dose-finding designs at dose level j the approximate prior ESS is $a_j + b_j$. To approximate the ESS for the overall prior across the J dose levels, we compute the average ESS $= \sum_{j=1}^J (a_j + b_j)/J$. This may be used to ensure that the prior ESS is not overly large, and it should be substantially less than the sample size of the current trial.

5. Applications.

5.1. *A meta-analysis of five sorafinib trials.* For illustration, we apply the proposed meta-analysis methods to the motivating example given in Section 2. For dose-finding studies where a meta-analysis is practical, the number of historical trials and sample sizes are likely to be moderate to relatively large (e.g., $\geq$ three historical trials, and ≥ 10 patients in each trial), and thus vague priors for $(\tilde{\boldsymbol{\phi}}, \sigma^2)$ are suitable. The historical Sorafenib trials have sample sizes 18, 30, 5, 43, 45, and 13, respectively. We assume i.i.d. normal distributions

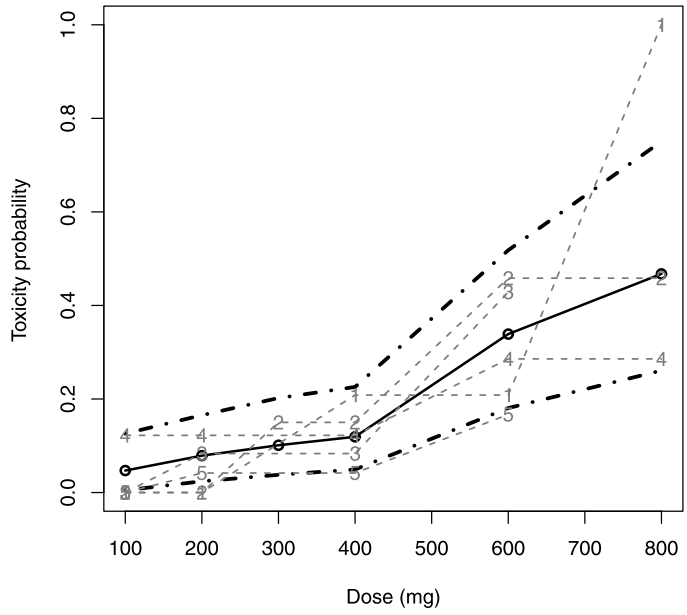

FIG. 2. *Posterior mean* (*solid line*) *and* 95% *credible bands* (*dot–dashed lines*) *of the average dose–response curve in the retrospective analysis of the Sorafenib clinical trials. The gray dotted lines display the isotonic regression estimates of the study-specific dose–response curves, with the numbers identifying the studies.*

with mean 0 and variance 10 as the priors for $\phi_1, \ldots, \phi_J$ and a half-t distribution with one degree of freedom and scale parameter 25 as the prior for σ^2. Figure 2 displays the posterior mean estimates and 95% credible intervals for the dose-toxicity curves, together with isotonic regression estimates of the study-specific dose-toxicity curves. It shows that the average dose-toxicity curve is relatively flat, for the first four doses, and then increases sharply at dose level 5. This pattern is also observed in the study-specific curves of studies 2–5. The shape of the average dose-toxicity curve thus reflects the individual curves. Most of the isotonically adjusted estimates of the toxicity probabilities lie inside the 95% credible intervals of the average dose-toxicity curve, indicating that the model offers sufficient flexibility to accommodate the data from the individual studies.

The posterior means of the average toxicity probabilities $(\tilde{p}_1, \ldots, \tilde{p}_6)$ are (0.05, 0.08, 0.10, 0.12, 0.34, 0.47), based on the historical data in Table 1. Using the target toxicity probability $\theta = 0.33$, dose level 5 (600 mg) is chosen as the MTD in this retrospective analysis. In contrast, using the same target, the method of Zohar, Katsahian and O'Quigley (2011) chooses dose level 4 (400 mg) as the MTD. However, the observed and estimated toxicity rates of dose level 4 are 0.05 and 0.12, respectively, both of which are far below the target 0.33, so the recommended dose level 4 appears to be suboptimal for this dataset. Another advantage of our Bayesian hierarchical meta-analysis approach is that the credible interval bands for the average dose–response curve can be readily computed based on the posterior samples. In contrast, under the frequentist paradigm the confidence band for the dose-response curve can only be obtained by applying a complicated asymptotic approximation.

5.2. *Designing a new trial.* Next, we apply the proposed method to designing a new dose-finding trial for Sorafinib, using a target toxicity rate $\theta = 0.33$, seven patient cohorts with three patients per cohort, and $J = 6$ dose levels. The starting dose of the new trial is level 4 which is one level below the estimated MTD $j^* = 5$ from the meta-analysis of the historical data.

The prior specification is the same as in the meta-analysis model. For the prior of α, we calibrate its value to maximize the RWOC under randomly chosen scenarios. Supposing that the clinician believes that there is an 80% chance that the actual MTD should not differ from the historically estimated MTD $j^* = 5$ by more than one level, we use the weight vector (0.067, 0.067, 0.067, 0.15, 0.50, 0.15).

The search range for the smallest value α_0 in the α's discrete support is $\{1, 5, 10, 15, 20, 30\}$. The search range for the distance d_α between points on the support is $\{5, 10, 20, 30, 40, 50\}$. The value of M_α should not be too small, since this would give an inflexible prior; nor should it be too large, since this would easily give an overly dispersed prior. We found that setting M_α in the range of five to 10 leads to reasonably good simulation performance, and we set $M_\alpha = 5$. A more refined search is possible, but this simple search is sufficient for the purpose of illustration, and it should work well in most applications.

We apply the procedure for both the MAP-CRM and MAP-BOIN designs. The optimal parameters are chosen as $\alpha_0 = 5$ and $d_\alpha = 20$, that is, $\mathcal{A} = \{5, 25, 45, 65, 85\}$ with equal prior probabilities 0.2, for both MAP-CRM and MAP-BOIN, corresponding to a prior expected sample size of roughly one observation per dose level. Other values of α_0 and d_α may separately yield a better RWOC for the MAP-CRM and MAP-BOIN designs. To simplify the presentation, we choose the same α_0 and d_α parameters for both designs. These parameters give nearly identical operating characteristics as those selected separately as optimal for the MAP-CRM and MAP-BOIN designs.

5.3. *Illustration of the MAP-CRM in the homogeneous case.* We first illustrate how MAP-CRM may be applied in order to design and conduct a dose-finding trial when the historical and the current trial data agree, that is, the homogeneous case. For this case, we simulated data for a single illustrative trial using the estimated dose-response curve based on the historical data. Figure 3 (upper panel) presents the entire dose-assignment history of one trial conducted using MAP-CRM.

The trial starts by treating the first cohort at dose level 4, and none of the three patients experience DLT. According to the updated estimate of $\boldsymbol{p}_0$, using (5), dose level 5 has the estimated toxicity probability closest to $\theta = 0.33$, and thus the second cohort is treated at dose level 5. In the second cohort, one patient experiences DLT, and MAP-CRM recommends the same dose level for the third cohort. For the fourth cohort, the observed data in the trial are $(y_{01}, \ldots, y_{06}) = (0, 0, 0, 0, 1, 0)$ and $(n_{01}, \ldots, n_{06}) = (0, 0, 0, 3, 6, 0)$. Although the observed toxicity data at dose level 5 are $(y_{05}, n_{05}) = (1, 6)$, which is smaller than $\theta = 0.33$, because it borrows strength from the historical data the MAP-CRM design yields the estimates $(\hat{p}_{01}, \ldots, \hat{p}_{06}) = (0.04, 0.08, 0.10, 0.11, 0.24, 0.43)$ which leads to the fourth cohort being assigned to dose level 5. In contrast, both the CRM using a noninformative prior and the BOIN design would escalate to dose level 6 for the fourth cohort. In this case the MAP-CRM finds the current and the historical data to be partially similar, and thus it borrows some historical information in the decision making. The fourth cohort has no DLT, so the observed data at dose level 5 are $(y_{05}, n_{05}) = (1, 9)$. At this point the MAP-CRM design identifies a difference between the current and historical trials and makes adaptive decisions on that basis, that is, dose escalation for the fifth cohort. The subsequent treatment assignments in this trial are displayed in Figure 3 (upper panel). At the end of the trial, the observed data are $(y_{01}, \ldots, y_{06}) = (0, 0, 0, 0, 4, 2)$ and $(n_{01}, \ldots, n_{06}) = (0, 0, 0, 3, 15, 3)$, leading to the posterior estimates $(\hat{p}_{01}, \ldots, \hat{p}_{06}) = (0.05, 0.09, 0.11, 0.13, 0.30, 0.49)$. As a result, the MAP-CRM design selects dose level 5 as the MTD.

5.4. *Illustration of the MAP-CRM in the heterogeneous case.* We next consider the MAP-CRM in the case of heterogeneous trials. For illustration we mimic the trial conducted

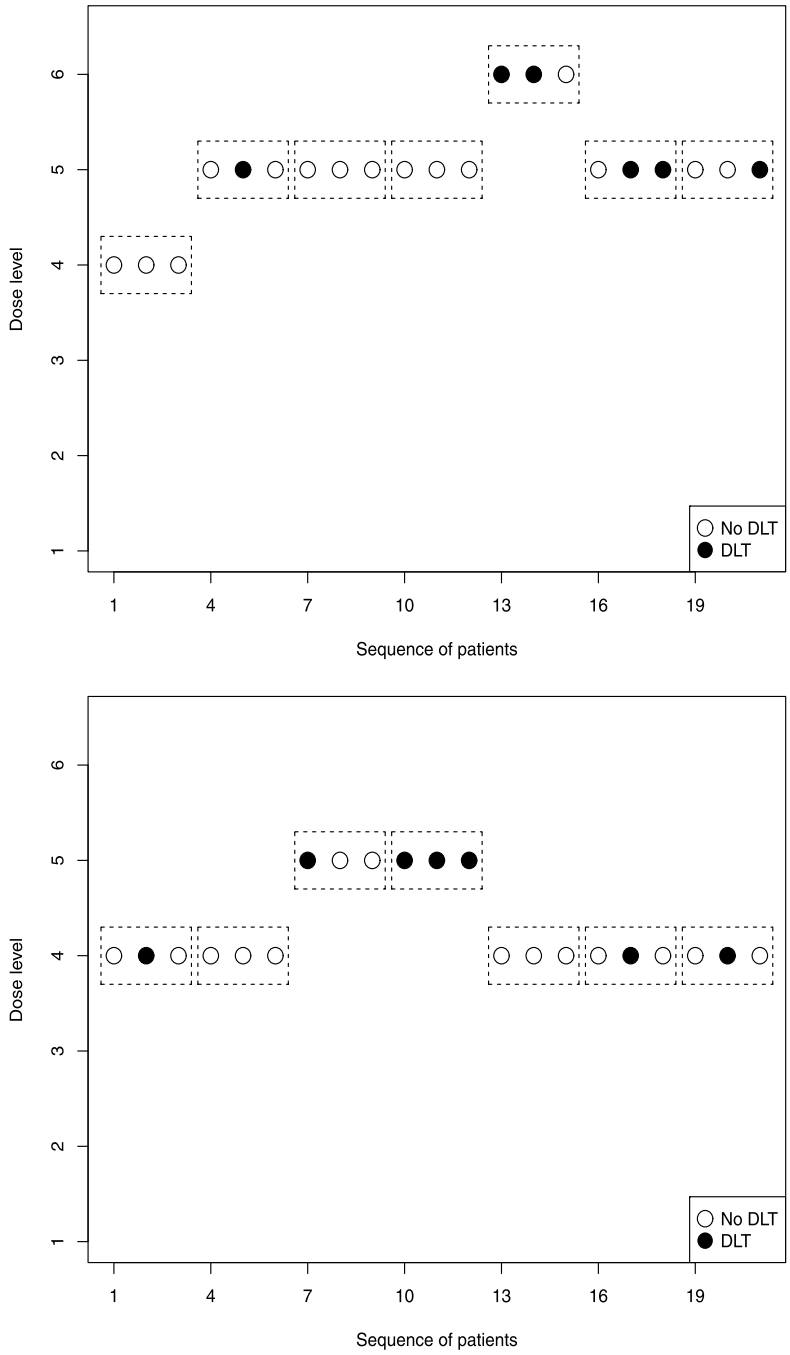

FIG. 3. *Single trial illustrations of the proposed MAP-CRM design. The upper panel shows the case where the current and historical trials are similar, while the lower panel shows the case of heterogeneity between the current and historical trials. Open circles indicate patients without toxicity, and solid circles indicate patients with toxicity. A dashed box represents a patient cohort.*

by Borthakur et al. (2011), introduced in Section 2, with the three doses 200, 400, and 600 mg studied in the new trial. For the simulations we embed these three doses in the larger set {100, 200, 300, 400, 600, 800}, as shown in Table 1, indexed by $1, \ldots, 6$. Thus, while only three doses are studied in the new trial, by using historical information, data from all six doses are used for decision making by MAP-CRM. The dose–toxicity data were simulated based on

the observed toxicity rates reported by Borthakur et al. (2011) with the dose escalation/deescalation decided by MAP-CRM, as shown in Figure 3 (lower panel).

In the new trial the first cohort is treated at 400 mg, and one patient experiences DLT. The MAP-CRM design gives the estimates $(\hat{p}_{01}, \ldots, \hat{p}_{06}) = (0.12, 0.19, 0.26, 0.29, 0.53, 0.62)$, and thus it retains dose level 4 (400 mg) for the second cohort. In contrast, if the information from the historical trials were borrowed fully, the dose would be escalated to dose level 5 for the second cohort which would be overly aggressive. Since only one patient in the first two cohorts experiences DLT at dose level 4, the dose is escalated to level 5 for the third cohort. One DLT is observed at this dose level, and thus this level is retained for the fourth cohort. However, all three patients in the fourth cohort have DLTs at dose level 5, so the next cohort's dose is deescalated to dose level 4. The MAP-CRM assigns all remaining patients to dose level 4, and it identifies dose level 4 (400 mg) as the MTD at the end of the trial. This differs from the historical trials for which the MTD is dose level 5 (600 mg).

5.5. *Illustration of the MAP-BOIN design.* We apply MAP-BOIN to designing a new trial of Sorafenib. As previously discussed, the dose escalation and deescalation rules of MAP-BOIN can be pretabulated before the trial, similarly to algorithm-based designs. We summarize the decision rules of MAP-BOIN in Table 2 which shows that the interval boundaries of the MAP-BOIN design vary with the dose levels. Because the historical data show substantial evidence that the target dose is level 5, the escalation boundaries of the MAP-BOIN design at dose levels 1–3 for $n_j = 15$ or 18 are greater than those of the BOIN design. In other words, when the dose levels for cohorts 5 or 6 are 1, 2, or 3, MAP-BOIN has a larger chance of escalating than the BOIN design. On the other hand, at dose level 5, MAP-BOIN tends to prevent dose escalation because the historical data indicate this dose as the MTD; thus, its escalation boundaries are no greater than those of BOIN for all n_j's. Similarly, because dose level 6 is overly toxic according to the historical data, the MAP-BOIN has a larger probability of deescalating from dose level 6 than the BOIN. In summary, the dose-specific escalation and deescalation boundaries of the MAP-BOIN design indicate that it can partly reflect the information contained in the historical data yet assign a dominant role to the current data in decision making. The dose-finding procedure of the new Sorafenib trial using the MAP-BOIN design follows exactly the same decisions in Table 2, so we omit implementation details.

6. Simulation studies of the meta-analysis methods.

6.1. *Comparison of models.* We conducted simulation studies to assess the finite-sample performance of our proposed random-effects meta-analysis model. We compared the proposed curve-free random-effects meta-analysis method (CF-RMA) with the CRM-based fixed-effect meta-analysis approach (CRM-FMA) of Zohar, Katsahian and O'Quigley (2011) and also with a parametric random-effects meta-analysis (P-RMA) method. Prior specification of the proposed method is the same as in Section 5. For the CRM-FMA we pool all the historical data together and adopt monotonicity assumptions, based on a power model (O'Quigley, Pepe and Fisher (1990)), $p_{kj} = a_j^{\exp(\beta)}$, where $(a_1, \ldots, a_6)$ denote prespecified fixed toxicity probabilities, that is, the model's *skeleton*, and β is an unknown parameter to be estimated. Applying the model calibration method of Lee and Cheung (2009), we obtain the skeleton values $(a_1, \ldots, a_6) = (0.02, 0.05, 0.12, 0.21, 0.33, 0.45)$.

For the P-RMA model we assume the power model but allow the parameter β to vary across trials, following a hierarchical structure, given by

$$p_{kj} = a_j^{\exp(\beta_k)}, \quad \beta_k \sim \mathrm{N}(\beta, \sigma^2), \beta, \sigma^2 \sim \pi(\beta, \sigma^2).$$

For the prior $\pi(\beta, \sigma^2)$, we assume that β and σ^2 are independent with $\beta \sim \mathrm{N}(0, 2)$ and σ^2 following a half-t distribution with one degree of freedom and scale parameter of 25 (Gelman (2006)).

6.2. *Configuration for fixed scenarios*. Following the setup of the motivating example, we consider six fixed dose-toxicity scenarios, each with six dose levels, and target $\theta = 0.33$. The goal in setting these scenarios is to examine the method's performance in terms of the random effects, location of the MTD, and the underlying data generation scheme. Table S.1 in the Supplementary Material (Lin et al. (2022)) summarizes the configuration of the simulation scenarios.

For each scenario, six true study-specific dose-response curves are simulated, leading to a total of six historical trials. To mimic reality, for each trial we randomly select four to six dose levels with equal probabilities from the six dose levels being considered. As a result, some dose levels are not included in some trials, and thus those trials have missing data for the absent dose levels. To generate data of the historical trials, we first specify a mean toxicity probability vector $\boldsymbol{p}_{\text{true}}$. A true study-specific latent vector for each historical trial is generated from a multivariate normal distribution with mean $\tilde{\boldsymbol{\phi}}_{\text{true}}$ and covariance matrix $\sigma^2_{\text{true}}\mathbf{I}_6$, and we back-solve for the value of $\tilde{\boldsymbol{\phi}}_{\text{true}}$ from $\boldsymbol{p}_{\text{true}}$, based on the link function, for example, in (2). For the kth historical trial, $k = 1, \ldots, 6$, given the study-specific latent vector $\boldsymbol{\phi}_{k,\text{true}}$ that is generated from $\mathbf{N}_6(\boldsymbol{\phi}_{k,\text{true}} \mid \tilde{\boldsymbol{\phi}}_{\text{true}}, \sigma^2_{\text{true}}\mathbf{I}_6)$, the toxicity probability $p_{kj,\text{true}}$ at dose j can be derived using a link function, for example, the logistic function in (2).

For each set of six simulated dose-response curves, we conduct a trial using the 3 + 3 design, up and down design, even allocation of patients to the dose levels, BOIN, CRM, and EWOC, respectively. The total sample size of the 3 + 3 design is random. For each of the other five designs, we randomly choose seven to 15 cohorts of size three each which gives the maximum sample size of 21 to 45. A total of 10,000 historical datasets are simulated under each scenario.

In scenarios 1 and 2, we assume $\boldsymbol{p}_{\text{true}} = (0.03, 0.07, 0.14, 0.23, 0.34, 0.47)$, so the MTD is dose level 5. The link function for back-solving for the true study-specific latent vector $\tilde{\boldsymbol{\phi}}_{\text{true}}$ from $\boldsymbol{p}_{\text{true}}$ is the same as (2). For the covariance matrix $\sigma^2_{\text{true}}\mathbf{I}_6$, we set $\sigma_{\text{true}} = 0$ for scenario 1, yielding a fixed-effect scenario. In scenario 2, we set $\sigma_{\text{true}} = 0.3$, leading to approximately 15% of the simulated study-specific dose–toxicity curves having an MTD different from dose level 5. In scenarios 3 and 4, the configurations are similar to those of scenarios 1 and 2. We take $\boldsymbol{p}_{\text{true}} = (0.05, 0.09, 0.32, 0.46, 0.59, 0.67)$, so dose level 3 is the MTD. We set $\sigma_{\text{true}} = 0$ in scenario 3, and $\sigma_{\text{true}} = 0.3$ in scenario 4.

In scenarios 5 and 6, we use a different link function in the data-generating procedures. We set $\boldsymbol{p}_{\text{true}} = (0.02, 0.03, 0.09, 0.33, 0.40, 0.42)$ with dose level 4 as the MTD and use a probit link function, $\tilde{\boldsymbol{\phi}}_{\text{true}} = \Phi^{-1}(\boldsymbol{p}_{\text{true}})$, where $\Phi^{-1}(\cdot)$ is the inverse cumulative distribution function (CDF) of the standard normal distribution. Under the probit link the true study-specific latent vector $\boldsymbol{\phi}_{k,\text{true}}$ is generated from a multivariate normal distribution with mean $\tilde{\boldsymbol{\phi}}_{\text{true}}$ and covariance matrix $\sigma^2_{\text{true}}\mathbf{I}_6$. The true study-specific dose-toxicity probabilities are calculated as $p_{kj,\text{true}} = 1 - \Phi(\phi_{k(j),\text{true}})$, where $\phi_{k(j),\text{true}}$ is the jth smallest element of $\boldsymbol{\phi}_{k,\text{true}}$. We set $\sigma_{\text{true}} = 0$ in scenario 5, and $\sigma_{\text{true}} = 0.2$ in scenario 6 which implies that approximately 35% of the simulated dose-toxicity curves have an MTD different from dose level 4.

6.3. *Configuration for random scenarios*. To avoid cherry-picking scenarios, we also consider scenarios where $\boldsymbol{p}_{\text{true}}$ is generated randomly with each of the six dose levels having an equal probability of being the MTD. For each simulated study we first randomly generate the average dose-toxicity curve $\boldsymbol{p}_{\text{true}}$ and then compute $\tilde{\boldsymbol{\phi}}_{\text{true}}$ using the logistic link function

in (2) or the probit link. The true study-specific latent vector $\boldsymbol{\phi}_{k,\text{true}}$ is generated from a multivariate normal distribution with mean $\tilde{\boldsymbol{\phi}}_{\text{true}}$ and covariance matrix $\sigma^2_{\text{true}}\mathbf{I}_6$. The study-specific dose-response curve is calculated from $\boldsymbol{\phi}_{k,\text{true}}$ based on the link function. Given $\boldsymbol{p}_{\text{true}}$, generation of the study-specific dose-response curves and the historical trial results are the same as in Section 6.2.

In scenario 7 we set $\sigma_{\text{true}} = 0$, so the study-specific dose–toxicity curves are the same across all studies. In scenario 8 we take $\sigma_{\text{true}} = 0.3$, thus introducing random effects into the studies. In scenario 9 the setup is the same as in scenario 8, but the link function in the data-generating procedure is the probit function.

6.4. *Results for fixed scenarios*. The simulation results for the fixed scenarios are summarized in Table S.2 in the Supplementary Material which gives the percentages of trials that correctly identify the MTD and that identify overly toxic doses. In scenarios 1 and 2 the P-RMA method performs best, essentially because the prior guess of the dose-toxicity curve is close to the truth. In scenarios 3–6 the CF-RMA performs best in terms of both the highest percentage of correct selection (PCS) and the lowest chance of selecting an overly toxic dose. Because the P-RMA method suffers from model misspecification, it has the worst performance among the three methods. In contrast, our CF-RMA method flexibly accommodates a wide variety of dose-response curves and true data-generating procedures. In most scenarios the advantages of CF-RMA over the parametric P-RMA and fixed-effect CRM-FMA are substantial. In terms of the percentage of overdose selection (POS), our proposed method is the safest in four of the six scenarios.

6.5. *Results for random scenarios*. In scenarios 7–9, where the true dose–response curves are randomly generated, CF-RMA is the best performer with the highest PCS and lowest POS. The CF-RMA has, on average, 10 and five higher percentage points of correct identification, compared with P-RMA and CRM-FMA, respectively. These may be considered substantial advantages, given the small sample sizes of the phase I studies. In terms of POS, our proposed method is the safest in all random scenarios. Overall, the simulations show that the proposed method is superior in both MTD identification and safety compared with the other two meta-analysis methods.

7. Simulation studies for dose-finding designs.

7.1. *Design comparisons*. We conducted a simulation study of the MAP-CRM and MAP-BOIN designs with existing methods included for comparison, including two no-information-borrowing designs and two information-borrowing designs. The no-information-borrowing designs are the CRM and BOIN. For the information-borrowing designs we examine the CRM using an informative prior (IP-CRM) (Morita (2011)), and the bridging CRM (B-CRM) (Liu and Yuan (2015)). To ensure a fair comparison, the starting dose for all dose-finding designs considered is dose level 4, one dose level below the estimated MTD from the historical data.

In the CRM design, the model with the skeleton $(a_1, \ldots, a_6) = (0.02, 0.05, 0.12, 0.21, 0.33, 0.45)$ is used, and the prior distribution of the unknown variable follows a normal distribution with a mean of 0 and a variance of 2. In the BOIN design, the default setting is used. In the IP-CRM design, a logistic regression model is used by fixing the intercept at 3 and treating the dose as the covariate that is obtained using the "backward fitting" procedure based on the historical data. The unknown slope parameter of the IP-CRM follows a Gamma(6, 6) prior. In the B-CRM design, the design parameters are set at default values. The IP-CRM and B-CRM designs cannot account for heterogeneity across multiple historical

studies. Instead, these two designs are implemented by pooling the multiple historical studies into a single study. For the proposed method the prior specifications are the same as those in Section 5.

7.2. *Configuration for fixed s*. Following the motivating example, 21 patients in cohorts of size three are enrolled and assigned adaptively to one of the six dose levels. The historical data include the information from the five dose-finding trials of Sorafenib.

The performance of each design is examined under six scenarios. The goal is to examine performance under varying degrees of similarity between the true dose-toxicity profile and the historical data, ranging from highly similar to vastly different, for example, whether the locations of MTDs are the same. Recall that dose level 5 was identified as the MTD by applying the proposed MAP-CRM meta-analysis to the historical data. Scenarios 1 and 2 assume toxicity profiles most commensurate to historical data, that is, the same location of the MTD and similar shape of the dose–response curve. Under scenarios 1 and 2 it is expected that the information-borrowing designs would perform better than the no-information-borrowing designs. On the other hand, Scenarios 5 and 6 have the most different toxicity profiles from the historical estimate, that is, the MTD locates far from the one estimated from the historical trial. Under scenarios 5 and 6, it is expected that the information-borrowing designs would not perform as well as the case where the toxicity profiles are highly similar to the historical estimates. Scenarios 3 and 4 have toxicity profiles where the MTD locations differ from the historically estimated location by one level. Figure S.1 in the Supplementary Material shows the dose-response curves under the six scenarios alongside the curve estimated from the historical trials for comparison.

7.3. *Configuration for random scenarios*. To establish random dose–response scenarios, we simulate two sets of dose–response curves. The first set, which we refer to as the homogeneous case, fixes the MTD at dose level 5, while the second set, which we refer to as the heterogeneous case, selects one of the dose levels, excluding dose level 5, with equal probabilities of being the MTD. The first set represents the case in which the current trial has the same MTD as the one in the historical trials, but the dose–response curve may take various shapes. The second set represents the case where there are substantial differences between the current and historical trials. It may be expected that the information-borrowing designs would perform better than the no-information-borrowing designs in the homogeneous and worse in the heterogeneous case.

7.4. *Results for fixed scenarios*. Table S.3 in the Supplementary Material summarizes the percentage of dose selections and the average number of patients treated at each dose based on 10,000 simulated trials for each scenario and method. In scenario 1 the specified dose-toxicity probabilities are very close to the estimates from the historical data, and thus the generated current trial data tend to be similar to the historical data. All of the information-borrowing designs outperform the no-information-borrowing designs in terms of PCS. Compared with CRM, MAP-CRM assigns approximately three more patients to the MTD. Compared with BOIN, MAP-BOIN assigns nearly one more patient to the MTD. Both the MAP-CRM and MAP-BOIN designs are less likely to select overly toxic dose levels than the other methods.

Scenario 2 is a case where the current trial MTD is the same as that in the historical trials, but the shape of the dose–toxicity curve is different. All information-borrowing designs outperform the no-information-borrowing designs. Compared with the other methods, the MAP-CRM and MAP-BOIN have better performances in terms of PCS.

In scenario 3 the ongoing trial's MTD is one level below the identified MTD from the historical trials, and MAP-CRM outperforms the no-information-borrowing designs. The PCS of

MAP-CRM is 3.8 percentage points higher than that of CRM. Patient allocation to the MTD under MAP-CRM and MAP-BOIN is also slightly better compared with other methods.

When the MTD is the highest dose level, as in scenario 4, the MAP-CRM and MAP-BOIN designs are relatively conservative, since the historical data indicate that the highest dose is likely to be excessively toxic. This information has a large impact on the decision of whether to escalate above dose level 5. As a result, the PCS of MAP-CRM is not as good as that of CRM, and for MAP-BOIN its PCS is slightly lower than that of the BOIN design.

In scenarios 5 and 6, the MTD locations and dose-response curves differ dramatically from those in the historical studies. As expected, because the MAP-CRM and MAP-BOIN designs adaptively examine the degree of agreement between the current and historical studies, this determines the amount of information they borrow from the historical data. The MAP-CRM and MAP-BOIN designs still achieve reasonably good performances, outperforming the no-information-borrowing designs. Since the IP-CRM and B-CRM do not account for heterogeneity between the current and historical data, their performances in scenarios 5 and 6 are unstable or undesirable.

In summary, these simulations reveal several appealing features of the MAP-CRM and MAP-BOIN designs. They reliably identify similarity between the current and historical studies and adaptively determine the amount of information borrowing. When the studies are similar, the proposed designs tend to be more efficient, and when the studies differ substantially, the proposed designs can quickly switch to the no-information-borrowing mode and still achieve a reasonably good performance.

Furthermore, we quantify the sample size saving of the MAP-CRM design, compared with the CRM design, by increasing the number of cohorts under the CRM design until its PCS is no less than that of the MAP-based methods. In the first three scenarios, the CRM design would, respectively, require additional five, seven, and two cohorts (or 15, 21, and six patients) to achieve similar performances of the MAP-CRM design. Considering that the original number of cohorts is only seven (or 21 patients), in this sense the MAP-CRM design achieves substantial sample size saving percentages of 41.7%, 50%, and 22.2% in comparison with the no-information-borrowing CRM design.

We illustrate the model's ability to determine the degree of information borrowing by evaluating the average posterior distribution of α under the six scenarios. As shown in Figure S.2 in the Supplementary Material, under scenarios 1 and 2, where the historical data tend to agree with the current toxicity estimates, the posterior distribution of α is concentrated on the left, indicating strong information borrowing. On the other hand, under scenarios 5 and 6, where the true toxicity probabilities differ sharply from those implied by the historical data, the posterior distribution of α is concentrated more to the right side. The posterior means of α in scenarios 1–6 are 33.6, 35.5, 40.6, 36.5, 49.0, and 50.2, respectively. The results indicate that model can reliably detect the degree of similarity between historical and current data and make adaptive decisions accordingly.

7.5. *Results for random scenarios*. Under the random scenarios, we evaluated the PCS and the percentage of overdose allocation (POA), based on 10,000 data replications. Figure S.3 in the Supplementary Material summarizes the performances of the designs for the homogeneous and heterogeneous cases. For each case, six distinct plotting symbols representing the six designs are aligned vertically, the horizontal position of the vertical line corresponds to the mean of the current set of six bars, and the vertical position of the plotting symbol represents the difference in percentage points from the center. For example, in the homogeneous case, the PCS is centered around 46%, and the PCS of the MAP-CRM design is roughly three percentage points higher than the center, that is, around 49%.

In terms of PCS, in both the homogeneous and heterogeneous cases, the MAP-CRM and MAP-BOIN designs are among the best performers. While the B-CRM attains the highest

PCS in the homogeneous case, it performs worst under the heterogeneous case. The CRM and IP-CRM designs achieve PCS similar to or slightly better than the proposed methods in the heterogeneous case but perform worse in the homogeneous case.

In terms of POA, in the homogeneous case the MAP-CRM and MAP-BOIN designs achieve the lowest percentages. In the heterogeneous case, the POA of the MAP-CRM and MAP-BOIN are neither best nor worst. The CRM design has the highest POA in both the homogeneous and heterogeneous cases. The center POA in the homogeneous case is much lower than that of the heterogeneous case. This is because dose level 5 is chosen as the MTD in the homogeneous case, that is, only one dose level is higher than the MTD, whereas in the heterogeneous case, for example, dose levels 1–4 chosen as the MTD, more dose levels are defined as overdoses.

8. Concluding remarks. We have proposed a novel random-effects model for the meta-analysis of multiple historical phase I dose-finding studies. The proposed model is an extension of the standard *logistic-normal distribution* for a single-dose DLT probability, and it leads to a monotonically increasing vector of DLT probabilities with the random effects quantifying between-trial heterogeneity. Our simulation studies show that the proposed random-effects meta-analysis method generally has superior performance, compared with a large set of other methods. We also incorporate the proposed random-effects model and power prior into several existing dose-finding methods to obtain new MAP-based dose-finding methods that adaptively borrow information from multiple previous studies. Their performance is shown to be superior by extensive simulation studies. In some senses, dose-finding studies with multiple dose levels may be treated as multiarm trials, where some arms may be subject to missingness in certain studies. Since network meta-analysis is an approach tailored for dealing with such a problem of mixed treatment comparisons (Lu and Ades (2004)), it might be interesting to adapt the meta-analysis of multiple dose-finding studies to the framework of network meta-analysis. Additionally, our approach can be framed as an extension of model-based meta-analysis (Mould (2012), Boucher and Bennetts (2015)) to analyze phase I clinical trials. This suggests that the meta-analytic-predictive approach might be applied to other problems in a more general setting under the framework of model-based meta-analysis.

APPENDIX

Illustration of isotonic regression: Isotonic regression is commonly adopted for estimating a dose-response curve when it is desired to enforce a monotonicity assumption of the dose-toxicity relationship, $p_1 \le p_2 \le \cdots \le p_J$ (Yuan and Chappell (2004)). For dose finding, when the observed DLT proportions at some dose levels do not follow a monotonically increasing pattern, isotonic regression replaces the empirical proportions of any adjacent levels that violate the monotonicity assumption by their weighted average. To illustrate the idea, suppose that the observed [number of DLTs]/[number of patients] at dose levels 1 and 2 are 1/3 and 0/6, respectively. Since the monotonicity assumption is violated in this case, a pooled estimate, to enforce monotonicity, for these two levels could be $(1/3 \times 3 + 0/6 \times 6)/(3+6) = 0.11$, that is, averaging the DLT proportions using the numbers of patients as the weights.

For more general problems where J doses are involved, isotonic regression can be done using the pool-adjacent-violators (PAVA) algorithm (Bril et al. (1984)) which repeatedly replaces the adjacent violators with their weighted average until a monotonic ordering is achieved. The resulting estimates $\tilde{p}_i$ are those that, among all the isotonic candidate estimates p_i', minimize the weighted sum of squares,

$$\sum_{j=1}^{J} w_j (p_i - p_i')^2,$$

where w_j is the weight, which, in the dose-finding problem, is specified as the number of patients at each dose $w_j = n_j$.

Acknowledgments. We would like to thank the Editor, the Associate Editor, and the reviewers for their valuable comments and suggestions with special thanks to the Associate Editor whose dedicated and meticulous effort has led to a much improved version of our paper.

RL and HS contributed equally to this work.

Funding. Lin's research was partially supported by grants from the National Cancer Institute (5P30CA016672 and 1R01CA261978).

Yin's research was partially supported by funding from the Research Grants Council of Hong Kong (17308420).

Thall's research was partially supported by grants from the National Cancer Institute (5P30CA016672 and 1R01CA261978).

Flowers's research was partially supported by grants from the Cancer Prevention & Research Institute of Texas (RR190079) and Burroughs Wellcome Fund (1016433.01).

SUPPLEMENTARY MATERIAL

Supplement to "Bayesian hierarchical random-effects meta-analysis and design of phase I clinical trials" (DOI: 10.1214/22-AOAS1600SUPP; .pdf). The Supplementary Materials include the simulation scenarios and results that are discussed in Section 7.

REFERENCES

AITCHISON, J. and SHEN, S. M. (1980). Logistic-normal distributions: Some properties and uses. *Biometrika* **67** 261–272. MR0581723 https://doi.org/10.2307/2335470

AWADA, A., HENDLISZ, A., GIL, T. et al. (2005). Phase I safety and pharmacokinetics of BAY 43-9006 administered for 21 days on/7 days off in patients with advanced, refractory solid tumours. *Br. J. Cancer* **92** 1855–1861.

BABB, J., ROGATKO, A. and ZACKS, S. (1998). Cancer phase I clinical trials: Efficient dose escalation with overdose control. *Stat. Med.* **17** 1103–1120.

BORTHAKUR, G., KANTARJIAN, H., RAVANDI, F. et al. (2011). Phase 1 study of sorafenib in patients with refractory or relapsed acute leukemias. *Haematol.* **96** 61–68.

BOUCHER, M. and BENNETTS, M. (2015). The many flavors of model-based meta-analysis: Part I—introduction and landmark data. *CPT*: *Pharmacometrics & Systems Pharmacology* **5** 54–64.

BRIL, G., DYKSTRA, R., PILLERS, C. and ROBERTSON, T. (1984). Isotonic regression in two independent variables. *Appl. Stat.* **33** 352–357.

CHEN, M.-H. and IBRAHIM, J. G. (2006). The relationship between the power prior and hierarchical models. *Bayesian Anal.* **1** 551–574. MR2221288 https://doi.org/10.1214/06-BA118

CHEN, M.-H., IBRAHIM, J. G., LAM, P., YU, A. and ZHANG, Y. (2011). Bayesian design of noninferiority trials for medical devices using historical data. *Biometrics* **67** 1163–1170. MR2829252 https://doi.org/10.1111/j.1541-0420.2011.01561.x

CHOW, S.-C., CHIANG, C., LIU, J. and HSIAO, C.-F. (2012). Statistical methods for bridging studies. *J. Biopharm. Statist.* **22** 903–915. MR2970563 https://doi.org/10.1080/10543406.2012.701578

CLARK, J. W., EDER, J. P., RYAN, D. et al. (2005). Safety and pharmacokinetics of the dual action Raf kinase and vascular endothelial growth factor receptor inhibitor, BAY 43-9006, in patients with advanced, refractory solid tumors. *Clin. Cancer Res.* **11** 5472–5480.

CUNANAN, K. M. and KOOPMEINERS, J. S. (2018). Hierarchical models for sharing information across populations in phase I dose-escalation studies. *Stat. Methods Med. Res.* **27** 3447–3459. MR3865285 https://doi.org/10.1177/0962280217703812

GARCÍA, V. M., OLMOS, D., GOMEZ-ROCA, C. et al. (2014). Dose-response relationship in phase I clinical trials: A European Drug Development Network (EDDN) collaboration study. *Clin. Cancer Res.* **20** 5663–5671.

GASPARINI, M. and EISELE, J. (2000). A curve-free method for phase I clinical trials. *Biometrics* **56** 609–615. MR1795024 https://doi.org/10.1111/j.0006-341X.2000.00609.x

GELMAN, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Anal.* **1** 515–533. MR2221284 https://doi.org/10.1214/06-BA117A

GEZMU, M. and FLOURNOY, N. (2006). Group up-and-down designs for dose-finding. *J. Statist. Plann. Inference* **136** 1749–1764. MR2255594 https://doi.org/10.1016/j.jspi.2005.08.002

GOULD, A. L., JIN, T., ZHANG, L. X. and WANG, W. W. B. (2012). A predictive Bayesian approach to the design and analysis of bridging studies. *J. Biopharm. Statist.* **22** 916–934. MR2970564 https://doi.org/10.1080/10543406.2012.701579

HIGGINS, J. P. T., THOMPSON, S. G. and SPIEGELHALTER, D. J. (2009). A re-evaluation of random-effects meta-analysis. *J. Roy. Statist. Soc. Ser. A* **172** 137–159. MR2655609 https://doi.org/10.1111/j.1467-985X.2008.00552.x

HOBBS, B. P., CARLIN, B. P. and SARGENT, D. J. (2013). Adaptive adjustment of the randomization ratio using historical control data. *Clin. Trials* **10** 430–440.

HOBBS, B. P., SARGENT, D. J. and CARLIN, B. P. (2012). Commensurate priors for incorporating historical information in clinical trials using general and generalized linear models. *Bayesian Anal.* **7** 639–673. MR2981631 https://doi.org/10.1214/12-BA722

HOBBS, B. P., CARLIN, B. P., MANDREKAR, S. J. and SARGENT, D. J. (2011). Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics* **67** 1047–1056. MR2829239 https://doi.org/10.1111/j.1541-0420.2011.01564.x

HUANG, S. M. and TEMPLE, R. (2008). Is this the drug or dose for you?: Impact and consideration of ethnic factors in global drug development, regulatory review, and clinical practice. *Clin. Pharmacol. Ther.* **84** 287–294.

IBRAHIM, J. G. and CHEN, M.-H. (2000). Power prior distributions for regression models. *Statist. Sci.* **15** 46–60. MR1842236 https://doi.org/10.1214/ss/1009212673

IBRAHIM, J. G., CHEN, M.-H., XIA, H. A. and LIU, T. (2012). Bayesian meta-experimental design: Evaluating cardiovascular risk in new antidiabetic therapies to treat type 2 diabetes. *Biometrics* **68** 578–586. MR2959625 https://doi.org/10.1111/j.1541-0420.2011.01679.x

LEE, S. M. and CHEUNG, Y. K. (2009). Model calibration in the continual reassessment method. *Clin. Trials* **6** 227–238.

LEE, J., THALL, P. F., JI, Y. and MÜLLER, P. (2015). Bayesian dose-finding in two treatment cycles based on the joint utility of efficacy and toxicity. *J. Amer. Statist. Assoc.* **110** 711–722. MR3367259 https://doi.org/10.1080/01621459.2014.926815

LENK, P. J. (1988). The logistic normal distribution for Bayesian, nonparametric, predictive densities. *J. Amer. Statist. Assoc.* **83** 509–516. MR0971380

LIN, R. and YIN, G. (2017). Nonparametric overdose control with late-onset toxicity in phase I clinical trials. *Biostatistics* **18** 180–194. MR3612282 https://doi.org/10.1093/biostatistics/kxw038

LIN, R., SHI, H., YIN, G., THALL, P. F., YUAN, Y. and FLOWERS, C. R. (2022). Supplement to "Bayesian hierarchical random-effects meta-analysis and design of phase I clinical trials." https://doi.org/10.1214/22-AOAS1600SUPP

LIU, S. and JOHNSON, V. E. (2016). A robust Bayesian dose-finding design for phase I/II clinical trials. *Biostatistics* **17** 249–263. MR3515998 https://doi.org/10.1093/biostatistics/kxv040

LIU, S. and YUAN, Y. (2015). Bayesian optimal interval designs for phase I clinical trials. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **64** 507–523. MR3325461 https://doi.org/10.1111/rssc.12089

LIU, S., PAN, H., XIA, J., HUANG, Q. and YUAN, Y. (2015). Bridging continual reassessment method for phase I clinical trials in different ethnic populations. *Stat. Med.* **34** 1681–1694. MR3334684 https://doi.org/10.1002/sim.6442

LU, G. and ADES, A. E. (2004). Combination of direct and indirect evidence in mixed treatment comparisons. *Stat. Med.* **23** 3105–3124.

MINAMI, H., KAWADA, K., EBI, H. et al. (2008). Phase I and pharmacokinetic study of sorafenib, an oral multikinase inhibitor, in Japanese patients with advanced refractory solid tumors. *Cancer Sci.* **99** 1492–1498.

MOORE, M., HIRTE, H. W., SIU, L. et al. (2005). Phase I study to determine the safety and pharmacokinetics of the novel Raf kinase and VEGFR inhibitor BAY 43-9006, administered for 28 days on/7 days off in patients with advanced, refractory solid tumors. *Ann. Oncol.* **16** 1688–1694.

MORITA, S. (2011). Application of the continual reassessment method to a phase I dose-finding trial in Japanese patients: East meets West. *Stat. Med.* **30** 2090–2097. MR2829159 https://doi.org/10.1002/sim.3999

MORITA, S., THALL, P. F. and MÜLLER, P. (2008). Determining the effective sample size of a parametric prior. *Biometrics* **64** 595–602, 669–670. MR2432433 https://doi.org/10.1111/j.1541-0420.2007.00888.x

MOULD, D. R. (2012). Model-based meta-analysis: An important tool for making quantitative decisions during drug development. *Clin. Pharmacol. Ther.* **92** 283–286.

NEUENSCHWANDER, B., BRANSON, M. and GSPONER, T. (2008). Critical aspects of the Bayesian approach to phase I cancer trials. *Stat. Med.* **27** 2420–2439. MR2432497 https://doi.org/10.1002/sim.3230

NEUENSCHWANDER, B., CAPKUN-NIGGLI, G., BRANSON, M. and SPIEGELHALTER, D. J. (2010). Summarizing historical information on controls in clinical trials. *Clin. Trials* **7** 5–18. https://doi.org/10.1177/1740774509356002

NEUENSCHWANDER, B., WEBER, S., SCHMIDLI, H. and O'HAGAN, A. (2020). Predictively consistent prior effective sample sizes. *Biometrics* **76** 578–587. https://doi.org/10.1111/biom.13252

O'QUIGLEY, J., PEPE, M. and FISHER, L. (1990). Continual reassessment method: A practical design for phase 1 clinical trials in cancer. *Biometrics* **46** 33–48. MR1059105 https://doi.org/10.2307/2531628

SCHMIDLI, H., GSTEIGER, S., ROYCHOUDHURY, S., O'HAGAN, A., SPIEGELHALTER, D. and NEUENSCHWANDER, B. (2014). Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics* **70** 1023–1032. MR3295763 https://doi.org/10.1111/biom.12242

SPIEGELHALTER, D. J., ABRAMS, K. R. and MYLES, J. P. (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. Wiley, Chichester.

STANGL, D. and BERRY, D. A. (2000). *Meta-Analysis in Medicine and Health Policy*. CRC Press, Boca Raton.

STORER, B. E. (1989). Design and analysis of phase I clinical trials. *Biometrics* **45** 925–937. MR1029610 https://doi.org/10.2307/2531693

STRUMBERG, D., RICHLY, H., HILGER, R. A. et al. (2005). Phase I clinical and pharmacokinetic study of the Novel Raf kinase and vascular endothelial growth factor receptor inhibitor BAY 43-9006 in patients with advanced refractory solid tumors. *J. Clin. Oncol.* **23** 965–972.

URSINO, M., RÖVER, C., ZOHAR, S. and FRIEDE, T. (2021). Random-effects meta-analysis of phase I dose-finding studies using stochastic process priors. *Ann. Appl. Stat.* **15** 174–193. MR4255273 https://doi.org/10.1214/20-aoas1390

YASUDA, S. U., ZHANG, L. and HUANG, S. M. (2008). The role of ethnicity in variability in response to drugs: Focus on clinical pharmacology studies. *Clin. Pharmacol. Ther.* **84** 417–423.

YIN, G. and IBRAHIM, J. G. (2005). Cure rate models: A unified approach. *Canad. J. Statist.* **33** 559–570. MR2232380 https://doi.org/10.1002/cjs.5550330407

YIN, G., LI, Y. and JI, Y. (2006). Bayesian dose-finding in phase I/II clinical trials using toxicity and efficacy odds ratios. *Biometrics* **62** 777–784. MR2247206 https://doi.org/10.1111/j.1541-0420.2006.00534.x

YUAN, Z. and CHAPPELL, R. (2004). Isotonic designs for phase I cancer clinical trials with multiple risk groups. *Clin. Trials* **1**, 499–508.

YUAN, Y., HESS, K. R., HILSENBECK, S. G. and GILBERT, M. R. (2016). Bayesian optimal interval design: A simple and well-performing design for phase I oncology trials. *Clin. Cancer Res.* **22** 4291–4301.

ZOHAR, S., KATSAHIAN, S. and O'QUIGLEY, J. (2011). An approach to meta-analysis of dose-finding studies. *Stat. Med.* **30** 2109–2116. MR2829161 https://doi.org/10.1002/sim.4121