Random-effects meta-analysis of Phase I dose-finding studies using stochastic process priors

Moreno Ursino, Christian Röver, Sarah Zohar, Tim Friede

The Annals of Applied Statistics
2021, Vol. 15, No. 1, 174–193
https://doi.org/10.1214/20-AOAS1390

RANDOM-EFFECTS META-ANALYSIS OF PHASE I DOSE-FINDING STUDIES USING STOCHASTIC PROCESS PRIORS

BY MORENO URSINO[1,*], CHRISTIAN RÖVER[2,‡], SARAH ZOHAR[**1,†] AND TIM FRIEDE[**2,§]

[1]*Centre de Recherche des Cordeliers, INSERM, Sorbonne Université, USPC, Université de Paris,* [*]*moreno.ursino@inserm.fr;* [†]*sarah.zohar@inserm.fr*

[2]*Department of Medical Statistics, University Medical Center Göttingen,* [‡]*christian.roever@med.uni-goettingen.de;* [§]*tim.friede@med.uni-goettingen.de*

Phase I dose-finding studies aim at identifying the maximum tolerated dose (MTD). Often, several dose-finding studies are conducted with some variation in the administration mode or dose panel. For instance, sorafenib (BAY 43-900) was used as monotherapy in 36 phase I trials, according to a recent clinicaltrials.gov search. Since the toxicity may not be directly related to the specific indication, synthesizing the information from several studies might be worthwhile. However, this is rarely done in practice and only a fixed-effect meta-analysis framework was proposed to date. We developed a Bayesian random-effects meta-analysis methodology to pool several phase I trials and suggest the MTD. A curve free hierarchical model on the logistic scale with random effects, accounting for between-trial heterogeneity, is used to model the probability of toxicity across the investigated doses. An Ornstein–Uhlenbeck Gaussian process is adopted for the random effects structure. Prior distributions for the curve-free model are based on a latent Gamma process. An extensive simulation study showed good performance of the proposed method also under model deviations. Sharing information between phase I studies can improve the precision of MTD selection, at least when the number of trials is reasonably large.

1. Introduction. Phase I dose-finding studies are conducted during early stages of the clinical development and aim at estimating the *maximum tolerated dose (MTD)* of a drug or a combination of molecules. The MTD is defined with reference to the occurrence of treatment-related adverse events, so-called *dose-limiting toxicities (DLTs)*. The MTD is reached once the rate of DLTs exceeds an acceptable level. Phase I studies usually involve small numbers of healthy volunteers, except in oncology, where, due to the potentially high toxicity of drugs, phase I trials are commonly performed on patients (Chevret (2006)).

In oncology, identifying the correct or reasonable dose or set of doses is a crucial objective in the drug development process: selecting too high a dose means exposing patients to an unacceptable toxicity profile, while selecting a dose of too low toxicity increases the likelihood that the treatment provides insufficient efficacy (Bretz, Pinheiro and Branson (2005)). The dose escalation paradigm in phase I (or I/II) trials thus generally aims to avoid recommending too toxic doses of an agent while maintaining an acceptable toxicity. Due to limited sample sizes, conventional statistical methods are often inaccurate so that adaptive sequential analyses have been proposed, as these can potentially find the MTD sooner and limit the number of exposed subjects (Le Tourneau, Lee and Siu (2009), Neuenschwander et al. (2015)).

Received August 2019; revised July 2020.
[**]Co-last authors as seniors.
Key words and phrases. Meta-analysis, dose finding, mixed-effect model.

Quoting Pretorius (2016), "Approximately 70% of Phase II trials are unsuccessful. [...] What is unexpected, however, is the percentage of "confirmatory" Phase III trials that fail—about 50%". The two main reasons for failure are toxicity and lack of efficacy (Harrison (2016)). An improper dose selection in the first stages of the clinical development plays an important role in the failure rate. Studies have shown that the wealth of data accumulated during clinical development is still poorly used, and several authors have pointed out the need to better use the existing data to improve the results (Arrowsmith (2011), Friede et al. (2018)).

When combining data across trials, two sources of potential heterogeneity need to be considered. First, there are differences in the outcomes of the control groups. In the context of dose-escalation studies, there might be differences in the (true) toxicity probabilities due to variations, for example, in the study populations or in the definition and assessment of toxicities. Second, the (true) treatment effects, even if defined on a relative scale, might vary across trials. In standard meta-analysis models, the former is addressed via stratification by study. In so-called random-effects meta-analyses, the latter is addressed by inclusion of random study-by-treatment interactions. In fixed-effect or common-effect meta-analysis, a homogeneous treatment effect across trials is assumed. For a recent discussion of the various statistical models, we refer here to Jackson et al. (2018). As evidenced by large-scale empirical investigations, some level of between-study heterogeneity is not unlikely to occur (Turner et al. (2012)). However, estimation of the corresponding variance component and accounting appropriately for the uncertainty in estimation in inference of relevant model parameters can be challenging, if the number of studies included in the meta-analysis is small (Friede et al. (2017)). In the context of meta-analyses of dose-escalation trials, we still lack an understanding as well as empirical evidence how best to account for the various forms of between-trial heterogeneity.

Zohar, Katsahian and O'Quigley (2011) proposed a meta-analysis approach for phase I clinical trials in oncology. Phase I data were pooled while accounting for the sequential nature of such trials to better estimate the overall MTD. However, this method did not deal with several important characteristics associated with phase I features. First, data were pooled under several administration schedules which may imply different toxicity profiles. Second, between-trial heterogeneity was not taken into account which may lead to inaccurate inference.

Thomas, Sweeney and Somayaji (2014) reported the results of a meta-analysis based on dose-response studies conducted by a large pharmaceutical company between 1998 and 2009. Data collection targeted efficacy endpoints, but safety data were not extracted. The goal of this meta-analysis was to identify consistent quantitative patterns in dose response across different compounds and diseases. The meta-analysis excluded oncology trials, as these have different dosing objectives and methods.

Kim et al. (2017) proposed a random-effects meta-analysis of phase I oncology clinical trials, where the exchangeability assumption is relaxed and rare events and missing data are investigated. However, efficacy rate was the primary criterion; drugs were grouped by type (cytotoxic or targeted agents), and each trial was confined to a single dose only.

In this manuscript we develop a novel meta-analysis approach for phase I clinical trials in oncology which takes into account the different features described above to better suit the requirements in estimating MTDs. We generalized the binomial-normal hierarchical model (BNHM) that is most commonly used in the literature for meta-analysis of studies involving a single dose. In the following section two motivating examples are described. In Section 3 the methodology is presented, along with prior distributions and variations of MTD definitions. In Section 4 we describe model variations and simulation settings that we used to test the developed method and its sensitivity to varying circumstances. Finally, in Section 5 the new methodology is applied to the two motivating examples; some limitations are discussed in Section 6.

TABLE 1

The results of 14 *studies on sorafenib monotherapy. For each dose considered in each trial, the numbers of patients experiencing DLT events and the total numbers of exposed patients are given*

	Dose (mg)						
Study	100	200	300	400	600	800	1000
Clark et al. (2005)	0/3	0/3		1/4	1/6	3/3	
Awada et al. (2005)	0/4	0/3	1/5	1/10	7/12	1/3	
Moore et al. (2005)	0/3	1/6		0/8	3/7		
Strumberg et al. (2005)	1/5	1/6		0/15	4/14	2/7	
Minami et al. (2008)	0/3	1/12		0/6	1/6		
Miller et al. (2009)		8/34		6/20			
Nabors et al. (2011)		0/3		1/6	0/3	1/5	3/3
Chen et al. (2007)		0/3		1/16			
Jia et al. (2013)				3/4			
Borthakur et al. (2011)-1		0/3		0/15	2/8		
Borthakur et al. (2011)-2		0/3		1/7	2/6		
Crump et al. (2010)-1	0/4	1/6	0/6	1/6			
Crump et al. (2010)-2	0/3	1/6		0/3	2/6		
Furuse et al. (2008)		0/12		1/14			

2. Two motivating examples. Meta-analyses have largely focused on late-stage trials. As phase I studies usually have small sample sizes and are mostly algorithm-based and only lately started using model-based designs, methodologists have given less attention to pooling them.

The first illustration concerns sorafenib (BAY 43-9006) which is a kinase inhibitor approved for the treatment of advanced renal cell carcinoma, hepatocellular carcinoma and radioactive iodine resistant advanced thyroid carcinoma. A search of the clinicaltrials.gov registry of clinical trials at the end of June 2019 revealed 833 studies using sorafenib (at any recruitment stage and type of study) of which 248 studies were labeled as "phase I" or "phase I/II", and 99 studies were labeled as "phase III" or "phase II/III". Of the 248 phase I or phase I/II studies using sorafenib, 36 used it in phase I as monotherapy (median sample size 22, range two to 158).

Motivated by the large number of trials found on clinicaltrials.gov, a search on PubMed found 12 manuscripts that reported on 14 trials. Their results are summarized in Table 1. These 14 trials tested a total of seven doses (100, 200, 300, 400, 600, 800 and 1000 mg), with most of these studies targeting solid tumors or leukemia. DLT definitions were comparable, and most of sorafenib schedules followed a 28-day cycle or similar. Not all of these trials originated from the clinicaltrials.gov search. Details are given in the Supplementary Material (Ursino et al. (2021a)). Today, the dose recommended by the European Medicines Agency (EMA) is 400 milligrams (mg) twice a day.

Applying the common-effect approach proposed by Zohar, Katsahian and O'Quigley (2011) (in the following referred to as the ZKO approach) to the sorafenib data (Table 1) and using (0.05, 0.1, 0.2, 0.3, 0.45, 0.6, 0.65) as initial guesses of the toxicity probabilities for the doses, also called "skeleton" (with an appropriate shape as in O'Quigley and Zohar (2010) and Zohar, Katsahian and O'Quigley (2011)), we obtained the following estimated toxicity probabilities: (0.012, 0.033, 0.093, 0.169, 0.308, 0.471, 0.530). Assuming a toxicity threshold of 0.33 and defining the MTD as the dose in the panel with an estimated probability of toxicity closest to the threshold, a dose of 600 mg is estimated as MTD, while for a threshold of 0.2, the MTD is at 400 mg.

TABLE 2

The results of 10 *studies on combination therapy of irinotecan and S-*1 (*tegafur/gimeracil/oteracil*). *For each dose considered in each trial, the numbers of patients experiencing DLT events and the total numbers of exposed patients are given*

	Dose (mg/m^2)									
Study	40	50	60	70	80	90	100	120	125	150
Ogata et al. (2009)	0/3	0/3	3/4							
Inokuchi et al. (2006)				0/3	10/42	0/3	2/3			
Goya et al. (2012)				0/3	0/3	3/5				
Takiuchi et al. (2005)	1/6		0/3		0/4		3/6			
Ishimoto et al. (2009)		0/3	0/3	0/3	2/4					
Kusaba et al. (2010)					0/6		2/3			
Nakafusa et al. (2008)			7/39		2/3					
Shiozawa et al. (2009)					1/6		2/6	2/6		2/3
Yoda et al. (2011)			0/3		3/6					
Komatsu et al. (2010)							1/9		1/9	0/3

The second example concerns a combination therapy of irinotecan and S-1 (S-1 refers to a combination of three pharmacological compounds, namely, tegafur, gimeracil and oteracil potassium). Irinotecan is a topoisomerase 1 inhibitor. It has proven effective in combination with 5-fluorouracil (5-FU) but was associated with many adverse events. Thus, the combination of irinotecan and S-1 was used to treat advanced colorectal and gastric cancer in a Japanese population. Ten trials (Table 2) evaluated a total of 10 doses, from 40 to 150 mg/m^2.

Applying the ZKO method on the irinotecan and S-1 data and using (0.005, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.65, 0.70) as skeleton, we obtained the following estimated toxicity probabilities: (0.002, 0.026, 0.061, 0.141, 0.231, 0.328, 0.430, 0.537, 0.592, 0.648). Assuming a toxicity threshold of 0.33, 90 mg/m^2 is estimated as MTD, while for a threshold of 0.2, the MTD is at 80 mg/m^2. To our knowledge, the combination of irinotecan and S-1 has not received market authorization.

In these two examples, the doses, dose ranges and sample sizes varied among trials. For estimating the overall MTD, the ZKO method of pooling adaptive sequential phase I data sets amounts to a fixed-effect meta-analysis method. The next section details our proposal that takes into account these specificities as well as inter- and intra-trial heterogeneity by developing a nonparametric random-effects approach.

3. Methods.

3.1. *The dose-response model.* For studies concerned with only a single dose, the binomial-normal hierarchical model (BNHM) or an approximation is most commonly used (Günhan, Röver and Friede (2020), Jackson et al. (2018)). When moving to several doses in the same study, we propose an extension of the BNHM that is adapted to the dose-finding context and that is able to also account for the ordering and spacing among doses.

Let $k \in \{1, \ldots, K\}$ be the study index and $i \in \{1, \ldots, I\} = \mathcal{I}$ be the dose level index, where all doses d_i used in any of the K trials are indexed in increasing order. The model must account for differences between the separation of adjacent d_i and the separation of consecutive doses in the individual trials (as occurs in Table 2). We define $\delta_{i,j}$ as the metric specifying the proximity or distance between doses. This may simply be defined as the linear difference ($\delta_{i,j} = d_i - d_j$). However, in many cases it may make sense to rather consider *relative* differences between dose levels on the logarithmic scale ($\delta_{i,j} = \log(d_i) - \log(d_j) =$

$\log(\frac{d_i}{d_j})$). Another option may be to assume unit increments for neighbouring doses ($\delta_{i,j} = i - j$).

The number of patients in study k allocated to dose i is given by n_{ik}, and X_{ik} is the number of patients experiencing a DLT. We propose the following model:

$$X_{ik} \sim \text{Binomial}(n_{ik}, p_{ik}), \tag{3.1}$$

$$\text{logit}(p_{ik}) = \sum_{j \le i} \mu_j + b_{ik}, \tag{3.2}$$

where p_{ik} is the probability of toxicity of dose i in the kth study. The probabilities p_{ik} are modelled on the logit scale. We assume that if the dose i is not used in the kth study, then $n_{ik} = X_{ik} = 0$. Therefore, as per convention $0^0 = 1$, it will not contribute to the likelihood.

The *fixed effects* $\mu_1 \in \mathbb{R}$ and $\mu_i \in \mathbb{R}^+$ (for $i > 1$) are common across all studies; the summation in (3.2) ensures strictly increasing overall mean probabilities of toxicity with increasing dose. The *random effects* accounting for between-study heterogeneity are represented by the (study-specific) vectors $\mathbf{b}_k \sim N(\mathbf{0}, \Sigma)$, where $\mathbf{0}$ represents the zero vector of dimension I and $\Sigma = \{\sigma^2_{i,j}\}_{i,j=1,\dots,I}$ the variance-covariance matrix. To meaningfully generalize from the BNHM for a single dose to a joint model for multiple doses, we specify the fixed and random effects accounting for the corresponding dose levels (d_i) and their ordering and proximity.

3.2. *Gaussian process for the random effects.* For the random effects we specify a model that accounts for the position of dose d_i on the dose continuum. We do not impose monotonicity on the elements of $\mathbf{b}_k$, and we rely on a relatively simple class of Gaussian processes. The model encompasses two interesting special cases, namely, *independent* and *identical* residuals at all doses. Between these extremes we use a stationary *Ornstein–Uhlenbeck process (OUP)* with covariance

$$\sigma^2_{i,j} = \sigma^2_{\text{m}} \exp\left(-\frac{|\delta_{i,j}|}{\ell}\right), \tag{3.3}$$

where σ^2_{m} is the marginal variance and $\ell > 0$ is a smoothness parameter determining how quickly the autocorrelation decays and residuals become less dependent, depending on the distance $\delta_{i,j}$ between doses (Neal (1999), Uhlenbeck and Ornstein (1930)). On small scales (relative to ℓ), the OUP behaves like a Wiener process (or Brownian motion); this nicely corresponds with the notion that *if* we know the residual at a certain dose, we know less about the neighbouring residual the further we move away from that dose, where increments behave (approximately) additively, as for the fixed effects model introduced below. For the limiting cases of $\ell \to 0$ and $\ell \to \infty$, it yields independent or identical residuals across doses, respectively (Doob (1942), Uhlenbeck and Ornstein (1930)). Prior distributions for the random effect's marginal variance σ^2_{m} and the OUP's distance scale ℓ need to be specified.

3.3. *Gamma process for fixed-effects prior distributions.* The definition of the common effect via a sum of unknown increments in (3.2) places the model in the class of stochastic processes that are commonly used as nonparametric models for unkown functions (Gelman et al. (2014), Chapter 21). Therefore, the prior distributions on the unknown increments may be inspired by a stochastic process. A natural and convenient class of models is defined via *infinitely divisible* probability distributions (Steutel (1979)); that means that we stay within the same distribution class for the increments (i.e., if we sum two increments, the sum's distribution again is in the same distribution class) which results in an overall consistent model. In the present case we consider strictly positive increments for increasing doses, so the Gamma process is an obvious choice (Lawless and Crowder (2004)).

The Gamma distribution is defined through two parameters, namely, the *shape* $s > 0$ and the *scale* $\theta > 0$ with mean $s\theta$ and variance $s\theta^2$. Choosing the first dose (d_1) as the *reference dose*, we can specify the prior distributions as a Gamma process with

$$\mu_1 \sim \text{Normal}(\mu^*, \sigma^*), \tag{3.4}$$

$$\mu_i \sim \text{Gamma}\big(s = \delta^*_{i,i-1}\kappa, \theta\big) \quad \text{for } i > 1, \tag{3.5}$$

where $\delta^*_{i,i-1}$ is the dose increment from dose d_{i-1} to d_i. δ^* can be equal to δ (used in the specification of the random effects) or it can use another underlying metric. The parameter μ_1 serves as an "intercept" term, and hyperparameters μ^* and σ^* (standard deviation) then need to be specified with reference to the expected toxicity at the reference dose. The Gamma process hyperparameters κ and θ also need to be prespecified. For a sensible choice, it is convenient to consider their effect on the conditional distribution for a unit increment,

$$\text{E}\big[\mu_i | \delta^*_{i,i-1} = 1\big] = \kappa\theta, \tag{3.6}$$

$$\text{Var}\big(\mu_i | \delta^*_{i,i-1} = 1\big) = \kappa\theta^2 \tag{3.7}$$

that suggests a reparametrisation in terms of

$$\text{slope } a = \kappa\theta \quad \text{and} \tag{3.8}$$

$$\text{coefficient of variation } c = \frac{1}{\sqrt{\kappa}}. \tag{3.9}$$

From this, we can see that, for small c, the (logit-) toxicity behaves approximately linearly, while larger c values allow for departures from linearity. In the limiting case of linearity, the model simplifies to a logistic model, which, in the special case of dose increments defined on the logarithmic scale as suggested above, again is a special case of the *Emax* model (Schwinghammer and Kroboth (1988)).

3.4. *Prior effective sample sizes for fixed effects.* To assess how informative certain choices of priors and hyperprior parameters for the fixed effect are, the notion of the *effective sample size (ESS)* can be used for the final calibration of the prior distributions and/or hyperprior parameters (Morita, Thall and Müller (2008)). In the present case we suggest to compute the approximate ESS as follows: (i) set the desired hyperparameters, (ii) simulate from the resulting set of prior distributions, (iii) for each simulated vector value, compute each p_i using (3.2) without random effects, (iv) approximate each p_i's distribution by a Beta(a_i, b_i), via a maximum likelihood approach or nonlinear least-squares method and (v) compute the approximate ESS as $\frac{1}{I}\sum_i (a_i + b_i)$, that is, the average of the ESS at each dose level.

3.5. *MTD estimation.* A range of rules have been proposed for estimating MTDs; several examples are given in the following. The most popular way uses the posterior mean estimates of the parameters in (3.2) and selects the MTD as the dose whose estimated DLT probability is closest to the pre-specified target $\tau \in [0, 1]$ (Cheung (2011)). In the meta-analysis context we may focus on the overall fixed effect; inverting from (3.2), we hence define

$$\pi_i = \text{logit}^{-1}\left(\sum_{j=1}^{i} \mu_j\right), \tag{3.10}$$

where the inverse logit is given by $\text{logit}^{-1}(x) = (1 + \exp(-x))^{-1}$. From this, we may then derive

$$\text{MTD} = d_j, \quad \text{where} \quad j = \arg\min_i \big|\text{E}[\pi_i | y] - \tau\big| \tag{3.11}$$

TABLE 3
Settings and parameters in the nine simulation scenarios. ℓ was chosen equal to 1 for all scenarios

Scenario	Fixed effect true $\mathbf{p}^\star$	Random effect	Study designs
1	(a)	OUP, $\sigma_m = 0.3$	CRM and $3+3$
2	(b)	OUP, $\sigma_m = 0.3$	CRM and $3+3$
3	(c)	OUP, $\sigma_m = 0.3$	CRM and $3+3$
4	(d)	OUP, $\sigma_m = 0.3$	CRM and $3+3$
5	(b)	OUP, $\sigma_m = 0.6$	CRM and $3+3$
6	(b)	$\Sigma = [\exp(-\frac{\lvert\delta_{i,j}\rvert}{\ell})\sigma_i\sigma_j]$ and $\boldsymbol{\sigma} = (0.1, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6)$	CRM and $3+3$
7	(c)	$\Sigma = [\exp(-\frac{\delta_{i,j}^2}{2\ell^2})\sigma_m^2]$, $\sigma_m = 0.3$	CRM and $3+3$
8	(c)	OUP, $\sigma_m = 0.3$	only $3+3$ design
9	(c)	OUP, $\sigma_m = 0.3$	only CRM design

OUP: Ornstein–Uhlenbeck process; CRM: continual reassessment method; $3+3$: $3+3$ algorithm design.

and where $\mathrm{E}[\pi_i \mid y]$ denotes the posterior expectation of π_i. Hence, the MTD is defined as the dose with estimated overall mean response closest to the targeted one. Alternatively, the posterior median may also be used instead of the mean in (3.11) (Ursino et al. (2019)).

In situations where investigators are particularly interested in overdose control, the classical *escalation with overdose control (EWOC)* principle may also be applied so that the MTD d_i is chosen as the largest dose satisfying

$$\mathrm{P}(\pi_i \geq \tau \mid y) < \tau_o, \tag{3.12}$$

that is, the dose whose posterior probability of exceeding the toxicity threshold τ is less than a prespecified threshold τ_o (Babb, Rogatko and Zacks (1998), Neuenschwander et al. (2015)). More complex rules involving loss functions, such as the one applied for the Bayesian Logistic Regression Model, can be also used (Neuenschwander, Branson and Gsponer (2008)).

4. Simulations. We performed an extensive simulation study to evaluate the operating characteristics of the proposed method. The aim was to compare the percentages of correct MTD selection to the ones of the ZKO method in several scenarios. A total of nine scenarios are proposed, detailed in Table 3, with variations in the position of the MTD, the heterogeneity structure and/or the design of the simulated trial. Details are given in Section 4.1. Then, we performed a sensitivity analysis aiming at checking the impact of prior distribution/hyperparameter choices and of random-effects model misspecification; details are shown in Sections 4.2 and 4.3.

All simulations have been done using the R software software, and R scripts are available as Supplementary Material (Ursino et al. (2021b)) and at the corresponding author's GitHub repository (Ursino (2020)).

4.1. *Data generation scenarios.* For each scenario, we simulated 1000 sets of completed trials that were subsequently meta-analyzed. Motivated by the sorafenib example (see Table 1), overall seven doses between $d_1 = 100$ mg and $d_7 = 1000$ mg were used. We first set the true probabilities of toxicity of the scenario for each of the $I = 7$ doses involved, $\mathbf{p}^\star = (p_1^\star, \ldots, p_I^\star)$. Four sets of $\mathbf{p}^\star$ were considered in total; these are illustrated in Figure 1. Then, the between-trial heterogeneity was added on the probability-transformed scale to set the probabilities of toxicity used to generate each single trial. The probit function was chosen for this data generation. Therefore, for the kth trial of the jth meta-analysis run, we first

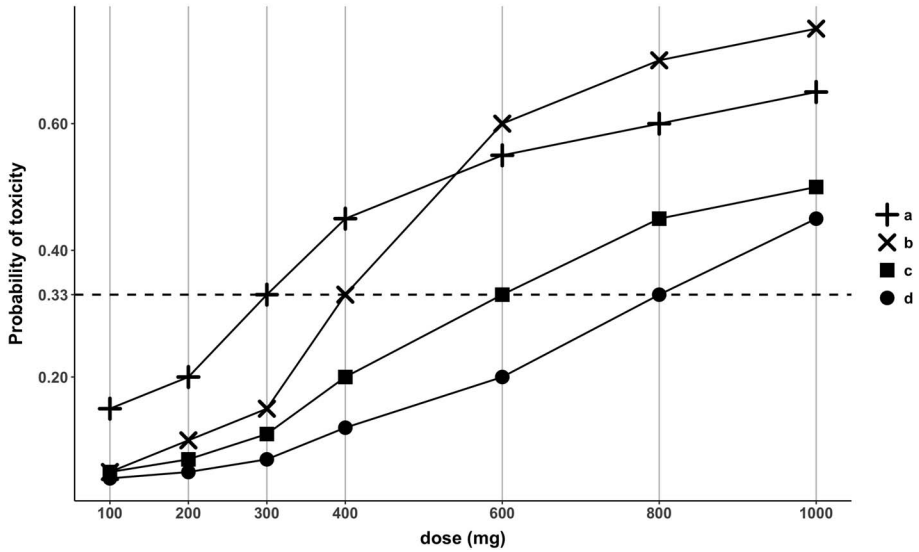

FIG. 1. *Four sets of probabilities* $\mathbf{p}^\star$ *used to set the fixed effects in the data generation scenarios*. *a*: $\mathbf{p}^\star = (0.15, 0.20, 0.33, 0{,}45, 0.55, 0.60, 0.65)$; *b*: $\mathbf{p}^\star = (0.05, 0.10, 0.15, 0.33, 0.60, 0.70, 0.75)$; *c*: $\mathbf{p}^\star = (0.05, 0.07, 0.11, 0.20, 0.33, 0.45, 0.50)$; *d*: $\mathbf{p}^\star = (0.04, 0.05, 0.07, 0.12, 0.20, 0.33, 0.45)$.

generated $\mathbf{p}_{kj}^{tr} = \mathcal{N}((\Phi^{-1}(p_1^\star), \ldots, \Phi^{-1}(p_I^\star)), \Sigma)$, where $\Phi(\cdot)$ represents the cumulative distribution function of the standard normal distribution. Then, we computed the probabilities as $\mathbf{p}_{kj} = (\Phi(p_{1,kj}^{tr}), \ldots, \Phi(p_{I,kj}^{tr}))$. We used the same autocovariance structure as in the estimation model (3.3) for all scenarios, allowing for a different σ_{m} value, except for scenario 6, where $\Sigma = [\exp(-\frac{|\delta_{i,j}|}{\ell})\sigma_i\sigma_j]$, and scenario 7, where $\Sigma = [\exp(-\frac{\delta_{i,j}^2}{2\ell^2})\sigma_m^2]$. For all scenarios, we set $\ell = 1$ and $\delta_{i,j} = \frac{d_i - d_j}{I^{-1}\sum_{n=1}^{I} d_n}$, while $\delta_{i,j}^* = \frac{d_i - d_j}{100\mathrm{mg}}$. This means that we used two related scales for δ and δ^* and that we use 100 mg as the measure unit for the fixed effect.

The number of doses used for each trial-panel is a random integer between 3 and 7 (sampled according to a uniform discrete distribution). Then, the true MTD (the dose in $\mathcal{I}$ whose probability of toxicity equals the target of $\tau = 0.33$ in the scenario) was assigned to the trial panel, and the other doses were selected randomly from the set $\{\mathcal{I} - \mathrm{MTD}\}$. Patients' responses are drawn from a Binomial distribution (3.1), according to the probability of toxicity associated to the dose where they were allocated to.

Depending on the scenarios and on the total number of trials used in the meta-analysis, some of the trials followed a *continual reassessment method* (CRM) design while others used the traditional algorithm *3 + 3* design (O'Quigley, Pepe and Fisher (1990), Le Tourneau, Lee and Siu (2009)).

The CRM is a model-based design that assumes a smooth parametric form for the dose-toxicity curve. Parameter estimation is updated after each patient, and the next patients are allocated to the dose whose posterior probability of toxicity is closest to the target. We used the `dfcrm` package, with the "empiric" model option and cohort size of one Cheung (2013). The 3 + 3 is an algorithm design which assumes cohorts of three patients and escalates, de-escalates or stops the trial according to the number of toxicities seen at the previous cohort. We used the function `sim3p3` of the `UBCRM` package to perform 3 + 3 simulation.

When 10 trials are included in each meta-analysis run and both designs are planned, five CRM and five 3 + 3 designs are used in the simulation. When only five trials are used in the meta-analysis, three CRM and two 3 + 3 are simulated. For the CRM trials the maximum sample size per study was sampled as an integer between 18 and 24 patients, and the number

of patients at each cohort between two and three (then, the maximum number of patients is automatically adjusted).

The (estimated) MTD is defined as the dose whose probability of toxicity is closest to the target of $\tau = 0.33$, and we adopted the posterior median variant of (3.11) as the estimation rule. The skeleton, that is, the prior guesses, was chosen to be (0.01, 0.05, 0.1, 0.15, 0.25, 0.38, 0.45), where only the probabilities linked to the doses in the trial panel are used and we selected the empirical working model. Finally, the CRM trials adopted the "no skipping" rule, that is, a higher dose is proposed to the next cohort only if all previous dose levels have already been given, while no stopping criteria were set.

In Scenarios 1–4 the true MTD is shifted from dose level 3 to dose level 6 while keeping the same $\sigma_m = 0.3$. This allows us to test the impact of the number of doses and MTD position in the meta-analysis run. Scenarios 5 and 6 have the same $\mathbf{p}^\star$ of Scenario 2, but we double the heterogeneity parameter in Scenario 5; we allow for dose-specific heterogeneity in Scenario 6. Then, Scenario 7 was added to check the impact of generating data under another Gaussian process. We evaluated the performance of the proposed model in case of 10 trials (made by five CRM and five $3 + 3$) and five trials (three CRM and two $3 + 3$) at each meta-analysis run. In the last two scenarios, that is, Scenarios 8 and 9, we evaluate the results given if all studies used an algorithm design (i.e., $3 + 3$) or model based design (i.e., the CRM), respectively. The simulation scenarios are summarised in Table 3.

4.2. *Prior settings.* To complete the method setting, prior distributions need to be set for random-effects parameters σ_{m} and ℓ, while fixed-effects hyperparameter values μ^*, σ^*, a and c need to be chosen. Regarding the first part, a half-Normal distribution was chosen as prior distribution for σ_{m} and an inverse Gamma distribution with shape and scale equal to 1 for ℓ.

Fixed effects hyperparameters play a crucial role in denoting the prior MTD, defined as the MTD estimated via prior distribution. In general, we suggest to select parameter values to set the prior MTD at the beginning of the second half part of the dose panel $\mathcal{I}$. μ^* and σ^* account for the probability of toxicity of the first dose which will impact the subsequent Gamma process. Once chosen the prior probability of toxicity desired for the first dose and having set the corresponding hyperparameters, a and c can be selected to have the prior MTD at a prespecified distance from the first dose. We suggest to first assign a value to c. Values not higher than 0.5 are are usually appropriate, since in our simulation experience (results not shown here) it gave more stable inferences. Last, find a to match the desired MTD position. It can be done via an optimization function that involves Monte Carlo steps (e.g., see the corresponding author's GitHub repository (Ursino (2020))).

When running a single meta-analysis, the user knows in advance the number of doses in the analysis and can select prior distributions and parameters, as described above. However, during simulations, depending on the scenarios, the number of doses in the panel, the dose-spacing and the related number of increments between doses can vary considerably. Therefore, we proposed an adaptive method to select the prior parameters of the Gamma prior process, considering the number of dose increments in each meta-analysis run.

We used an empirical Bayes approach here; specifically, we compute the empirical probability of toxicity of each dose by summing all DLTs reported on all studies at the same dose level and dividing it by the total number of patients treated at this dose level (in all studies). A pool adjacent violators algorithm (PAVA) was then used to perform a linear order isotonic regression to assure the nondecreasing behaviour of the dose-toxicity curve (Barlow et al. (1972)). Finally the empirical MTD was selected as the dose whose empirical probability of toxicity is closest to the target, set as 0.33 in this simulation study.

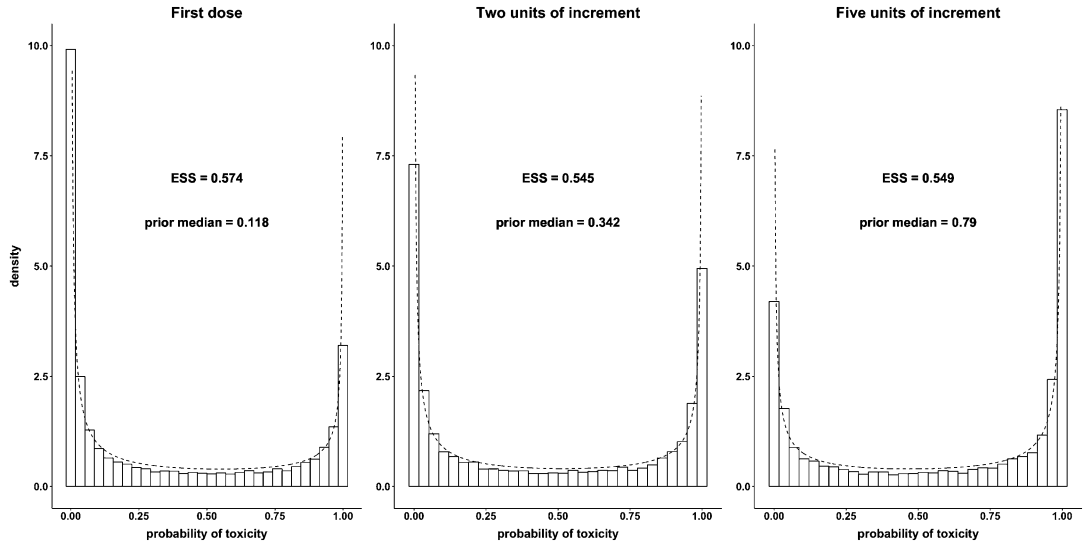

FIG. 2. *Induced prior probability of toxicities shown for the first dose of the panel and possible doses distant two and five units of increment from the first, respectively, when* $\mu^* = -2$, $\sigma^* = 5$, $a = 0.667$ *and* $c = 0.5$. *The histogram of the sampled distribution is plotted along with the fitted Beta distribution, whose parameters give the approximated effective sample size* (*ESS*).

Finally, we decided which set of prior parameters, among two prespecified ones, to select for the actual meta-analysis run. More specifically, the set of parameters was chosen considering the distance difference between the selected MTD and the first dose in the panel: if the difference is less or equal to two increment units, we select $\mu^* = -2$, $\sigma^* = 5$, $a = 0.667$ and $c = 0.5$ (set-1) that gives the induced prior probability of toxicities shown in Figure 2; otherwise, we select $\mu^* = -4$, $\sigma^* = 3.5$, $a = 0.642$ and $c = 0.5$ (set-2) that gives the induced prior probability of toxicities shown in Figure 3.

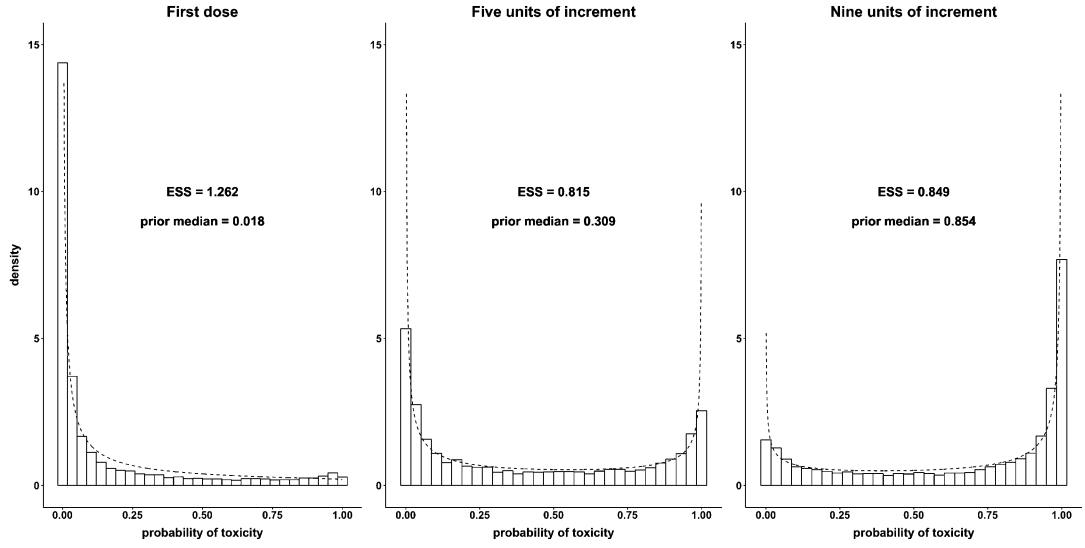

FIG. 3. *Induced prior probability of toxicities shown for the first dose of the panel and possible doses distant five and nine units of increment from the first, respectively, when* $\mu^* = -4$, $\sigma^* = 3.5$, $a = 0.642$ *and* $c = 0.5$. *The histogram of the sampled distribution is plotted along with the fitted Beta distribution, whose parameters give the approximated effective sample size* (*ESS*).

These values were chosen to have a good trade-off between ESS (lower numbers are desirable to have weakly informative prior) and the prior MTD placed at second and fifth dose increment from the first dose, respectively. Henceforth we refer to the resulting model as MADF (Meta-Analysis Dose-Finding).

4.3. *Sensitivity analyses*. We performed sensitivity analyses to check the impact of prior distributions and/or random-effects model misspecification. We considered four model modifications, changing the prior distribution for the fixed effect or changing the correlation structure for the random effects (or both).

Let MADF1 denote the model MADF where (3.5) is replaced by

$$\mu_i \sim \text{Gamma}(s = \kappa, \theta), \quad i > 1,$$

that is, the process assumes identical dose increments and all $\mu_{i>1}$ have the same prior distribution. In particular, we chose $\kappa = 3$ and $\theta = 2$ (that implies $a = 6$ and $c = \sqrt{3}^{-1}$) that led to very "pessimistic" prior probabilities of toxicities, that is, the prior probabilities of toxicities tends to be close to 1 for all doses larger than the first one.

In the second model we evaluate a more common Σ matrix structure and simpler prior for fixed effects. In detail, MADF2 denotes the model MADF with Σ as the variance-covariance of a heterogeneous first order autoregressive process, that is,

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 & \rho^2\sigma_1\sigma_3 & \dots & \rho^{I-1}\sigma_1\sigma_I \\ \rho\sigma_1\sigma_2 & \sigma_2^2 & \rho\sigma_2\sigma_3 & \dots & \rho^{I-2}\sigma_2\sigma_I \\ \rho^2\sigma_1\sigma_3 & \rho\sigma_2\sigma_3 & \sigma_3^2 & \dots & \rho^{I-3}\sigma_3\sigma_I \\ \dots & \dots & \dots & \dots & \dots \\ \rho^{I-1}\sigma_1\sigma_I & \dots & \dots & \dots & \sigma_I^2 \end{bmatrix},$$

along with a half-normal distribution with scale 1 as prior distribution for each σ_i and a uniform distribution across the interval $[0, 1]$ for ρ. Moreover, a Normal distribution and truncated Normal distributions were assigned as prior distribution to μ_1 and $\mu_i, i > 1$, respectively.

In the last two methods we evaluated uncorrelated random-effects and/or wider prior distributions. In MADF3 we have the same setting of MADF, but $\Sigma = \sigma_m^2 \mathbb{I}$, where $\mathbb{I}$ represents the identity matrix of $I \times I$ dimensions. In this case, random effects are uncorrelated. Again, a half-normal prior with scale 1 is assumed for σ_m.

Finally, MADF4 shares the same settings of MADF3, except for the Gamma prior distribution which is $\mu^* = -2$, $\sigma^* = 7$, $a = 3$ and $c = 0.5$ if the increment is less or equal to two units; otherwise, $\mu^* = -4$, $\sigma^* = 10$, $a = 3$ and $c = 2$. In this way we evaluated two sets of prior distributions with higher variance. All methods are summarized in Table 4.

4.4. *Results*. Table 5 shows the results in terms of proportion of final MTD selection, with the proportion of correct selection (PCS) in bold, of the proposed method, MADF, vs. the ZKO, when 10 studies are included in each meta-analysis run. MADF has higher PCS (interpreted also as percentage from now on), ranging from 61% to 92%, while ZKO PCS are in the range from 50% to 72%. The lower PCS performances of ZKO could be expected, since ZKO does not take into account heterogeneity between trials. ZKO tends to select overdoses more often than MADF, for example, in Scenario 1 where the MTD is at dose level 3; MADF suggests 34% over toxic doses vs. 41% for ZKO. PCS percentages decrease as σ increases, as in Scenario 5, and are stable also for random-effects misspecification, as in Scenarios 6 and 7.

These percentages decrease when only five studies are incorporated in the meta-analysis. The results are shown in Table 6, where MADF has still higher PCS, ranging from 50% to 82%, while ZKO has performance ranging from 38% to 61%.

TABLE 4
Summary of methods' parameters, priors and hyperpriors settings

Method	Fixed effects priors	Fixed effects hyperpriors	Random effects structure	Random effects hyperprior
MADF	$\mu_1 \sim \text{Normal}(\mu^*, \sigma^*)$; $\mu_i \sim \text{Gamma}(s = \delta^*_{i,i-1}\kappa, \theta)\ i > 1$	$\mu^* = -2, \sigma^* = 5$; $a = 0.667, c = 0.5$; or $\mu^* = -4, \sigma^* = 3.5$; $a = 0.642, c = 0.5$	$\mathbf{b}_k \sim N(\mathbf{0}, \Sigma)$; $\sigma^2_{i,j} = \sigma^2_{\text{m}} \exp(-\frac{\lvert\delta_{i,j}\rvert}{\ell})$	$\sigma_m \sim \text{Halfnormal}(1)$; $\ell \sim \text{Inv-Gamma}(1, 1)$
MADF1	$\mu_1 \sim \text{Normal}(\mu^*, \sigma^*)$; $\mu_i \sim \text{Gamma}(s = \kappa, \theta) i > 1$	$\mu^* = -2, \sigma^* = 5$; $a = 6, c = \sqrt{3}^{-1}$	$\mathbf{b}_k \sim N(\mathbf{0}, \Sigma)$; $\sigma^2_{i,j} = \sigma^2_{\text{m}} \exp(-\frac{\lvert\delta_{i,j}\rvert}{\ell})$	$\sigma_m \sim \text{Halfnormal}(1)$; $\ell \sim \text{Inv-Gamma}(1, 1)$
MADF2	$\mu_1 \sim \text{Normal}(\mu^*, \sigma^*)$; $\mu_i \sim T\ \text{Normal}(\mu^{**}, \sigma^{**})\ i > 1$	$\mu^* = -2, \sigma^* = 5$; $\mu^{**} = 3, \sigma^{**} = 5$	$\mathbf{b}_k \sim N(\mathbf{0}, \Sigma)$; $\sigma^2_{i,j} = \rho^{\lvert i-j-1\rvert}\sigma_i\sigma_j$	$\sigma_i \sim \text{Halfnormal}(1)$; $\rho \sim \text{Uniform}(-1, 1)$
MADF3	$\mu_1 \sim \text{Normal}(\mu^*, \sigma^*)$; $\mu_i \sim \text{Gamma}(s = \delta^*_{i,i-1}\kappa, \theta)\ i > 1$	$\mu^* = -2, \sigma^* = 5$; $a = 0.667, c = 0.5$; or $\mu^* = -4, \sigma^* = 3.5$; $a = 0.642, c = 0.5$	$\mathbf{b}_k \sim N(\mathbf{0}, \Sigma)$; $\Sigma = \sigma^2_m \mathbb{I}$	$\sigma_m \sim \text{Halfnormal}(1)$
MADF4	$\mu_1 \sim \text{Normal}(\mu^*, \sigma^*)$; $\mu_i \sim \text{Gamma}(s = \delta^*_{i,i-1}\kappa, \theta)\ i > 1$	$\mu^* = -2, \sigma^* = 7$; $a = 3, c = 0.5$; or $\mu^* = -4, \sigma^* = 10$; $a = 3, c = 2$	$\mathbf{b}_k \sim N(\mathbf{0}, \Sigma)$; $\Sigma = \sigma^2_m \mathbb{I}$	$\sigma_m \sim \text{Halfnormal}(1)$

TABLE 5

Proportion of dose selection using 10 *studies in each meta-analysis. The proportion of correct MTD selection (PCS) in each scenario is written in bold*

	Dose levels						
	1	2	3	4	5	6	7
Scenario 1							
MADF	0.000	0.082	**0.612**	0.305	0.001	0.000	0.000
ZKO	0.022	0.190	**0.496**	0.253	0.034	0.002	0.003
#patients	31 (23, 41)	31 (23, 41)	54 (43, 65)	15 (9, 23)	6 (3, 12)	2 (0, 6)	0 (0, 3)
Scenario 2							
MADF	0.000	0.000	0.032	**0.920**	0.048	0.000	0.000
ZKO	0.000	0.002	0.052	**0.695**	0.233	0.013	0.005
#patients	22 (18, 26)	26 (20, 32)	29 (23, 37)	59 (50, 68)	14 (9, 20)	5 (0, 9)	0 (0, 3)
Scenario 3							
MADF	0.000	0.000	0.000	0.084	**0.834**	0.082	0.000
ZKO	0.000	0.000	0.002	0.075	**0.676**	0.216	0.031
#patients	22 (17, 26)	23 (19, 29)	26 (20, 33)	29 (22, 38)	45 (36, 54)	11.5 (6, 18)	6 (2, 12)
Scenario 4							
MADF	0.000	0.000	0.000	0.000	0.228	**0.758**	0.014
ZKO	0.000	0.000	0.000	0.001	0.131	**0.680**	0.188
#patients	43 (37, 51)	43 (37, 51)	24 (19, 31)	26 (20, 34)	26 (20, 33)	40 (32, 48)	11 (6, 18)
Scenario 5							
MADF	0.000	0.000	0.085	**0.781**	0.134	0.000	0.000
ZKO	0.004	0.037	0.162	**0.561**	0.215	0.017	0.004
#patients	24 (19, 31)	27 (20, 35)	28 (21, 37)	51 (41, 59)	13 (8, 20)	6 (2, 12)	0 (0, 6)
Scenario 6							
MADF	0.000	0.000	0.019	**0.882**	0.099	0.000	0.000
ZKO	0.000	0.000	0.022	**0.665**	0.287	0.015	0.011
#patients	21 (17, 26)	25 (20, 30)	30 (23, 37)	61 (53, 69)	14 (8, 20)	5 (0, 9)	0 (0, 4)
Scenario 7							
MADF	0.000	0.000	0.000	0.069	**0.830**	0.101	0.000
ZKO	0.000	0.000	0.002	0.075	**0.653**	0.245	0.025
#patients	22 (18, 26)	22.5 (18, 28)	27 (20, 33.25)	30 (23, 38)	45 (36, 54)	12 (6, 18)	6 (3, 12)
Scenario 8							
MADF	0.000	0.000	0.000	0.150	**0.773**	0.077	0.000
ZKO	0.000	0.000	0.002	0.078	**0.591**	0.295	0.034
#patients	24 (18, 27)	24 (18, 27)	24 (18, 27)	24 (18, 27)	30 (24, 36)	9 (3, 12)	3 (0, 6)
Scenario 9							
MADF	0.000	0.000	0.000	0.064	**0.837**	0.099	0.000
ZKO	0.000	0.000	0.001	0.076	**0.715**	0.194	0.014
#patients	20 (16, 25)	24 (18, 31)	30 (23, 39)	37 (27, 47)	60 (49, 71)	15 (9, 23)	10 (4, 16)

MADF: Proposed method; ZKO: Zohar, Katsahian and O'Quigley (2011) method; #patients: median (first quartile—third quartile) number of patients allocated to each dose.

Figure 4 summarizes the results of the sensitivity analysis in terms of percentage of correct selection when 10 studies are included in each analysis. MADF1 has the best performance in Scenarios 1 and 6, while MADF3 is the best method in Scenario 4. MADF4 has the lowest PCS in all scenarios. Full results are given in Table 1 in the Supplementary Material (Ursino et al. (2021a)). We can see the same trend for five studies, except in Scenario 4 where MADF1 gets the lowest PCS (full results given in Table 2 in the Supplementary Material (Ursino et al. (2021a))).

TABLE 6

Proportion of dose selection using five studies in each meta-analysis. The proportion of correct MTD selection (PCS) in each scenario is written in bold

	Dose levels						
	1	2	3	4	5	6	7
Scenario 1							
MADF	0.007	0.153	**0.498**	0.324	0.018	0.000	0.000
ZKO	0.032	0.177	**0.377**	0.292	0.089	0.026	0.007
#patients	15 (10, 23)	16 (10, 23)	29 (21, 37)	8 (3, 14)	3 (0, 6)	0 (0, 3)	0 (0, 0)
Scenario 2							
MADF	0.000	0.002	0.068	**0.826**	0.103	0.001	0.000
ZKO	0.005	0.007	0.060	**0.490**	0.367	0.055	0.016
#patients	11 (8, 14)	12 (9, 17)	15 (9, 21)	32 (26, 38)	6 (3, 12)	0 (0, 6)	0 (0, 0)
Scenario 3							
MADF	0.000	0.000	0.003	0.169	**0.683**	0.144	0.001
ZKO	0.000	0.000	0.013	0.133	**0.544**	0.261	0.049
#patients	11 (8, 14)	11 (8, 15)	13 (9, 18)	15 (9, 21)	24 (18, 30)	6 (2, 10)	2 (0, 6)
Scenario 4							
MADF	0.000	0.000	0.000	0.003	0.320	**0.622**	0.055
ZKO	0.000	0.000	0.001	0.013	0.179	**0.610**	0.197
#patients	21 (17, 26)	21 (17, 26)	12 (8, 16)	12 (8, 18)	12.5 (9, 18)	20 (14, 27)	6 (0, 11)
Scenario 5							
MADF	0.000	0.017	0.153	**0.622**	0.200	0.008	0.000
ZKO	0.015	0.045	0.117	**0.436**	0.299	0.076	0.012
#patients	11 (8, 16)	13 (9, 19)	14 (9, 20)	27 (20, 34)	6 (3, 12)	3 (0, 6)	0 (0, 3)
Scenario 6							
MADF	0.000	0.000	0.059	**0.802**	0.137	0.002	0.000
ZKO	0.000	0.002	0.04	**0.449**	0.412	0.065	0.032
#patients	11 (8, 14)	12 (9, 16)	15 (10, 21)	33 (27, 39)	6 (3, 12)	2 (0, 6)	0 (0, 2)
Scenario 7							
MADF	0.000	0.000	0.001	0.152	**0.692**	0.155	0.000
ZKO	0.001	0.000	0.007	0.106	**0.546**	0.271	0.069
#patients	11 (8, 14)	11 (8, 15)	13 (9, 18)	15 (9, 21)	24 (17, 30)	6 (2, 11)	3 (0, 6)
Scenario 8							
MADF	0.000	0.001	0.009	0.203	**0.668**	0.116	0.003
ZKO	0.000	0.001	0.017	0.134	**0.489**	0.282	0.077
#patients	12 (9, 15)	12 (9, 15)	12 (9, 15)	12 (9, 15)	15 (12, 18)	3 (0, 6)	0 (0, 3)
Scenario 9							
MADF	0.000	0.000	0.002	0.143	**0.713**	0.141	0.001
ZKO	0.000	0.000	0.005	0.125	**0.557**	0.269	0.044
#patients	10 (7, 13)	11 (8, 16)	14 (9, 21)	18 (11.75, 25)	30 (22, 38)	7 (2, 13)	3 (0, 10)

MADF: Proposed method; ZKO: Zohar, Katsahian and O'Quigley (2011) method; #patients: median (first quartile—third quartile) number of patients allocated to each dose.

5. Application to the two examples. We applied the MADF method with the same setting and prior distributions, as described in the previous section, to the two examples introduced in Section 2.

5.1. *The sorafenib example*. Figure 5 shows the posterior distribution obtained for the probability of toxicity associated to each dose panel level. Using the posterior median variant of (3.11), we obtain the following estimates (0.032, 0.058, 0.085, 0.123, 0.307, 0.556, 0.834). This leads to selecting dose 600 mg as MTD if the target is $\tau = 0.33$ or $\tau = 0.25$, while

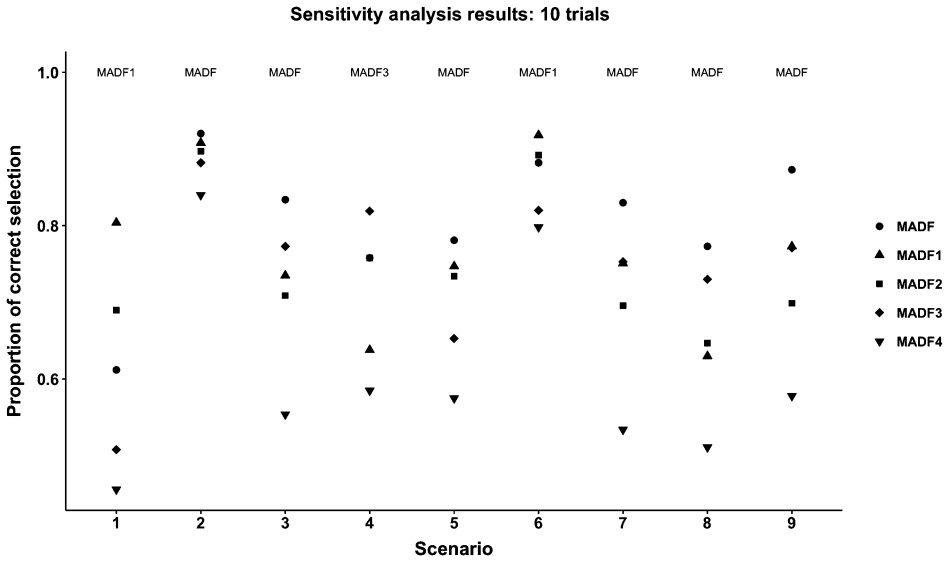

FIG. 4. *Results in terms of proportion of correct selection when* 10 *studies are included in each analysis. In each scenario the name of the method with the highest proportion of correct selection is shown. Standard errors, computed as* $\sqrt{p(1-p)/1000}$, *where* p *represents the proportion of correct selections, are in the range of* 0.009 *to* 0.016.

400 mg is chosen when $\tau = 0.20$. Adopting the EWOC rules as in (3.12), that is, computing $\mathrm{P}(\pi_i \geq \tau | y)$, we obtain (0, 0, 0, 0, 0.369, 0.991, 1), (0, 0, 0, 0.002, 0.832, 1, 1) and (0, 0, 0.001, 0.016, 0.964, 1, 1) for $\tau = 0.33$, $\tau = 0.25$ and $\tau = 0.20$, respectively. Setting $\tau_o = 0.25$, we select dose 400 mg in all cases.

In the light of these results, the EMA-recommended dose of 400 mg (twice a day) is safer whatever the model and criterion used. However, according to some models and criteria, the higher dose of 600 mg can also be considered safe.

5.2. *The irinotecan + S-1 example.* Results of the irinotecan + S-1 example are shown in Figure 6. Here, $\delta^*_{i,j} = \frac{d_i - d_j}{10\mathrm{mgm}^{-2}}$, while δ has the same specification, as described in Section 4.1. Again, using the posterior median variant of (3.11), we obtain the following estimates (0.022, 0.039, 0.070, 0.114, 0.194, 0.292, 0.413, 0.625, 0.678, 0.884). This leads to

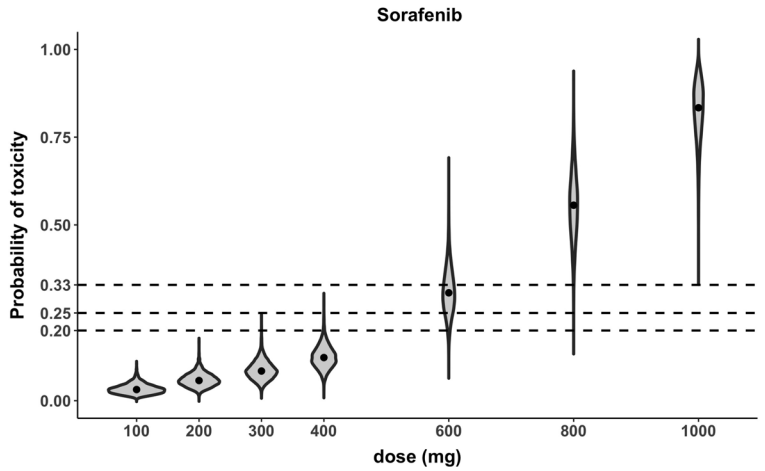

FIG. 5. *Posterior distribution for probability of toxicity for each dose level in the sorafenib example.*

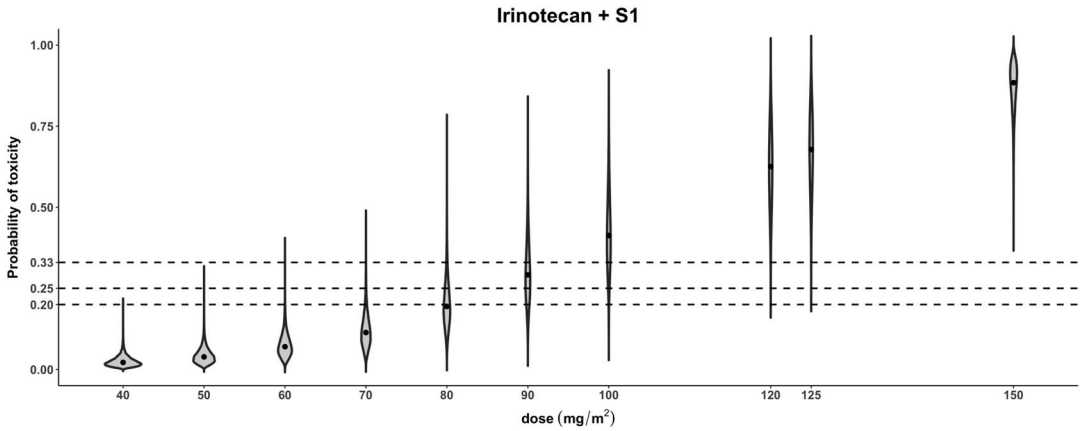

FIG. 6. *Posterior distribution for probability of toxicity for each dose level in the irinotecan + S-1 example.*

selecting dose 90 mg/m^2 as MTD for a target of $\tau = 0.33$ or $\tau = 0.25$, while 80 mg/m^2 is chosen when $\tau = 0.20$. Adopting the EWOC rules, as in (3.12), we obtain (0, 0, 0, 0.004, 0.061, 0.349, 0.773, 0.990, 0.996, 1), (0, 0, 0.002, 0.027, 0.238, 0.677, 0.944, 0.998, 1, 1) and (0, 0.001, 0.008, 0.082, 0.466, 0.866, 0.984, 1, 1, 1) for $\tau = 0.33$, $\tau = 0.25$ and $\tau = 0.20$, respectively. Setting $\tau_o = 0.25$, we select dose 80 mg/m^2 in the first two cases and 70 mg/m^2 in the last one.

6. Discussion. We proposed a new methodology for random-effects meta-analysis of phase I dose-finding trials, based on a Gaussian process for the random effect structure and on a Gamma process as a prior distribution for the fixed effects. The Gaussian process permits to share more information when doses are closer and less information when they are distant. In this way, for example, regarding a dose panel, dose levels 3 and 4 are more correlated than dose levels 1 and 4. The amount of correlation depends on the distance which seems more logical than assuming a constant value for the correlation. The Gamma prior process preserves the monotonicity assumption of toxicity. We do not suggest to add the full process to be estimated, since, in our experience, even if in meta-analysis more data are available than a single dose-finding trial, data are still not sufficient for a good estimation of the process parameters (results not shown in the paper). We focused on modelling toxicities exactly at the doses d_i that had also been investigated in the analysed trials. In general, in case of rich data and when the estimation of the underlying Gamma process is feasible, the full model actually also allows to interpolate or extrapolate across the continuum of doses. In this case, guidance on how to set the prior distributions can be found in Gelman et al. (2008).

The choice of the metrics δ^* and δ is mostly related to the computation stability and/or to the ease of performing experts' elicitation, enabling a more flexible model. However, when the two metrics are linked to each other via linear transformation, as in our case, one can also consider to use the same metric and scale the prior distributions accordingly.

With the above model specifications we have generalized the BNHM, as an obvious approach for the single-dose case, to the case of several adjacent doses. Note that, for the special case of a single dose, we actually again recover the BNHM with the parameters μ_1 and σ_{m} corresponding to the overall mean and heterogeneity parameters.

We chose the OUP process because it is the only nontrivial Gaussian process that includes independent and identical residuals as special cases and that allows linear transformations of the variables. Alternatively, one might also use processes from the more general class parameterized in terms of gamma-exponential autocovariance functions (Rasmussen and Williams (2006)).

In our results, ZKO had lower PCS performance. This is expected, since this method does not take into account the heterogeneity between trials. Also as expected, PCSs decrease when heterogeneity increases and when only 3 + 3 dose-finding trials are incorporated in meta-analysis (scenarios 5 and 8, respectively). MADF showed to be stable to model misspecification, as we can see in the results of scenarios 6 and 7 compared to scenarios 2 and 3, respectively. On the other hand, prior specification and simpler models, as MADF3 and MADF4, can give different operating characteristics. A conservative Gamma process prior, as MADF1, has better PCS when the MTD is located at the beginning of the dose panel. This behaviour is shared also by MADF2, as shown in three additional scenarios in the Supplementary Material (Ursino et al. (2021a)). Actually, this situation is not very realistic, since it would imply that subsequent trials have not adapted the dose range accordingly to previous results. However, if clinicians face a meta-analysis involving only few doses, a simpler method, as MADF2, can be a good solution.

In our examples, dose regimens were considered to be the same. It seems to have been a good approximation of the reality, as it is shown in the Supplementary Material (Ursino et al. (2021a)). However, our proposed methods can be used when doses were tested with different regimens (e.g., 800 mg once a day in one study and 400 mg twice a day in another study). It is possible to differentiate regimens with the same total dose passing to the exposition level, using pharmacokinetic (PK) principles. For example, to integrate varying dosing regimens, a latent pseudo-PK variable can be used to obtain an ordered set of regimens (Günhan et al. (2020)). Then, our methods can be applied.

A limitation of this work is that we rely on published data. Along with the well-known publication bias, the issue of poor reporting in phase I cancer trials was already raised by Zohar et al. (2008) and Comets and Zohar (2009). In articles describing early phase clinical trials, reporting is sometimes not complete in terms of DLTs and doses. Adverse events and DLTs are often mixed up together in tables or DLTs are reported by type and not by patient. Following the checklist proposed by Zohar et al. (2008) would improve the reporting of the results and would allow the use of meta-analytic methods, as the one proposed in this manuscript.

Acknowledgements. We thank the anonymous reviewer, the Editor and, especially, the anonymous Associate Editor for their careful reading of our manuscript and their many insightful comments and suggestions.

The first author is also affiliated with F-CRIN PARTNERS platform, AP-HP, Université de Paris.

SUPPLEMENTARY MATERIAL

Supplement to "Searches details and further scenarios results" (DOI: 10.1214/20-AOAS1390SUPPA; .pdf). Supplementary information.

Supplement to "R scripts used to generate manuscript results" (DOI: 10.1214/20-AOAS1390SUPPB; .zip). Supplementary information.

REFERENCES

ARROWSMITH, J. (2011). Trial watch: Phase III and submission failures: 2007–2010. *Nat. Rev. Drug. Discov.* **10** 87. https://doi.org/10.1038/nrd3375

AWADA, A., HENDLISZ, A., GIL, T., BARTHOLOMEUS, S., MANO, M., DE VALERIOLA, D., STRUMBERG, D., BRENDEL, E., HAASE, C. G. 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.

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

BARLOW, R. E., BARTHOLOMEW, D. J., BREMNER, J. M. and BRUNK, H. D. (1972). *Statistical Inference Under Order Restrictions. The Theory and Application of Isotonic Regression*. Wiley, London-Sydney. MR0326887

BORTHAKUR, G., KANTARJIAN, H., RAVANDI, F., ZHANG, W., KONOPLEVA, M., WRIGHT, J. J., FADERL, S., VERSTOVSEK, S., MATHEWS, S. et al. (2011). Phase I study of sorafenib in patients with refractory or relapsed acute leukemias. *Haematologica* **96** 62–68. https://doi.org/10.3324/haematol.2010.030452

BRETZ, F., PINHEIRO, J. C. and BRANSON, M. (2005). Combining multiple comparisons and modeling techniques in dose-response studies. *Biometrics* **61** 738–748. MR2196162 https://doi.org/10.1111/j.1541-0420.2005.00344.x

CHEN, Y.-B., LI, S., LANE, A. A., CONNOLLY, C., DEL RIO, C., VALLES, B., CURTIS, M., BALLEN, K., CUTLER, C., DEY, B. R. et al. (2014). Phase I trial of maintenance sorafenib after allogeneic hematopoietic stem cell transplantation for fms-like tyrosine kinase 3 internal tandem duplication acute myeloid leukemia. *Biology of Blood and Marrow Transplantation* **20** 2042–2048.

CHEUNG, Y. K. (2011). *Dose Finding by the Continual Reassessment Method*. Chapman and Hall/CRC, Boca Raton.

CHEUNG, K. (2013). dfcrm: Dose-finding by the continual reassessment method. R package version 0.2-2.

CHEVRET, S., ed. (2006). *Statistical Methods for Dose-Finding Experiments*. *Statistics in Practice*. Wiley, Chichester. MR2274231 https://doi.org/10.1002/0470861258

CLARK, J. W., EDER, J. P., RYAN, D., LATHIA, C. and LENZ, H.-J. (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.

COMETS, E. and ZOHAR, S. (2009). A survey of the way pharmacokinetics are reported in published phase I clinical trials, with an emphasis on oncology. *Clinical Pharmacokinetics* **48** 387–395.

CRUMP, M., HEDLEY, D., KAMEL-REID, S., LEBER, B., WELLS, R., BRANDWEIN, J., BUCKSTEIN, R., KASSIS, J., MINDEN, M. et al. (2010). A randomized phase I clinical and biologic study of two schedules of sorafenib in patients with myelodysplastic syndrome or acute myeloid leukemia: A NCIC (National Cancer Institute of Canada) Clinical Trials Group Study. *Leukemia & Lymphoma* **51** 252–260.

DOOB, J. L. (1942). The Brownian movement and stochastic equations. *Ann. of Math.* (2) **43** 351–369. MR0006634 https://doi.org/10.2307/1968873

FRIEDE, T., RÖVER, C., WANDEL, S. and NEUENSCHWANDER, B. (2017). Meta-analysis of few small studies in orphan diseases. *Res. Synth. Methods* **8** 79–91. https://doi.org/10.1002/jrsm.1217

FRIEDE, T., POSCH, M., ZOHAR, S., ALBERTI, C., BENDA, N., COMETS, E., DAY, S., DMITRIENKO, A., GRAF, A. et al. (2018). Recent advances in methodology for clinical trials in small populations: The InSPiRe project. *Orphanet Journal of Rare Diseases* **13** 186.

FURUSE, J., ISHII, H., NAKACHI, K., SUZUKI, E., SHIMIZU, S. and NAKAJIMA, K. (2008). Phase I study of sorafenib in Japanese patients with hepatocellular carcinoma. *Cancer Science* **99** 159–165.

GELMAN, A., JAKULIN, A., PITTAU, M. G. and SU, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *Ann. Appl. Stat.* **2** 1360–1383. MR2655663 https://doi.org/10.1214/08-AOAS191

GELMAN, A., CARLIN, J. B., STERN, H. S., DUNSON, D. B., VEHTARI, A. and RUBIN, D. B. (2014). *Bayesian Data Analysis*, 3rd ed. *Texts in Statistical Science Series*. CRC Press, Boca Raton, FL. MR3235677

GOYA, H., KURAISHI, H., KOYAMA, S., ICHIYAMA, T., YOSHIIKE, F., HIRAI, K., AGATSUMA, T., TATEISHI, K., KANDA, S. et al. (2012). Phase I/II study of S-1 combined with biweekly irinotecan chemotherapy in previously treated advanced non-small cell lung cancer. *Cancer Chemotherapy and Pharmacology* **70** 691–697.

GÜNHAN, B. K., RÖVER, S. and FRIEDE, T. (2020). Meta-analysis of few studies involving rare events. *Res. Synth. Methods* **11** 74–90.

GÜNHAN, B. K., WEBER, S. and FRIEDE, T. (2020). A Bayesian time-to-event pharmacokinetic model for phase I dose-escalation trials with multiple schedules. *Stat. Med.* **39** 3986–4000. https://doi.org/10.1002/sim.8703

HARRISON, R. K. (2016). Phase II and phase III failures: 2013–2015. *Nat. Rev. Drug. Discov.* **15** 817–818. https://doi.org/10.1038/nrd.2016.184

INOKUCHI, M., YAMASHITA, T., YAMADA, H., KOJIMA, K., ICHIKAWA, W., NIHEI, Z., KAWANO, T. and SUGIHARA, K. (2006). Phase I/II study of S-1 combined with irinotecan for metastatic advanced gastric cancer. *Br. J. Cancer* **94** 1130.

ISHIMOTO, O., ISHIDA, T., HONDA, Y., MUNAKATA, M. and SUGAWARA, S. (2009). Phase I study of daily S-1 combined with weekly irinotecan in patients with advanced non-small cell lung cancer. *International Journal of Clinical Oncology* **14** 43–47.

JACKSON, D., LAW, M., STIJNEN, T., VIECHTBAUER, W. and WHITE, I. R. (2018). A comparison of seven random-effects models for meta-analyses that estimate the summary odds ratio. *Stat. Med.* **37** 1059–1085. MR3777959 https://doi.org/10.1002/sim.7588

JIA, N., LIOU, I., HALLDORSON, J., CARITHERS, R., PERKINS, J., REYES, J., YEH, M., STOHR, E., RAO, S. et al. (2013). Phase I adjuvant trial of sorafenib in patients with hepatocellular carcinoma after orthotopic liver transplantation. *Anticancer Research* **33** 2797–2800.

KIM, M.-O., WANG, X., LIU, C., DORRIS, K., FOULADI, M. and SONG, S. (2017). Random-effects meta-analysis for systematic reviews of phase I clinical trials: Rare events and missing data. *Res. Synth. Methods* **8** 124–135.

KOMATSU, Y., YUKI, S., FUSE, N., KATO, T., MIYAGISHIMA, T., KUDO, M., KUNIEDA, Y., TATEYAMA, M., WAKAHAMA, O. et al. (2010). Phase 1/2 clinical study of irinotecan and oral S-1 (IRIS) in patients with advanced gastric cancer. *Advances in Therapy* **27** 483–492.

KUSABA, H., ESAKI, T., FUTAMI, K., TANAKA, S., FUJISHIMA, H., MITSUGI, K., SAKAI, K., ARIYAMA, H., TANAKA, R. et al. (2010). Phase I/II study of a 3-week cycle of irinotecan and S-1 in patients with advanced colorectal cancer. *Cancer Science* **101** 2591–2595.

LAWLESS, J. and CROWDER, M. (2004). Covariates and random effects in a gamma process model with application to degradation and failure. *Lifetime Data Anal.* **10** 213–227. MR2086957 https://doi.org/10.1023/B:LIDA.0000036389.14073.dd

LE TOURNEAU, C., LEE, J. J. and SIU, L. L. (2009). Dose escalation methods in phase I cancer clinical trials. *J. Natl. Cancer Inst.* **101** 708–720.

MILLER, A. A., MURRY, D. J., OWZAR, K., HOLLIS, D. R., KENNEDY, E. B., ABOU-ALFA, G., DESAI, A., HWANG, J., VILLALONA-CALERO, M. A. et al. (2009). Phase I and pharmacokinetic study of sorafenib in patients with hepatic or renal dysfunction: CALGB 60301. *J. Clin. Oncol.* **27** 1800–1805. https://doi.org/10.1200/JCO.2008.20.0931

MINAMI, H., KAWADA, K., EBI, H., KITAGAWA, K., KIM, Y.-I., ARAKI, K., MUKAI, H., TAHARA, M., NAKAJIMA, H. et al. (2008). Phase I and pharmacokinetic study of sorafenib, an oral multikinase inhibitor, in Japanese patients with advanced refractory solid tumors. *Cancer Science* **99** 1492–1498.

MOORE, M., HIRTE, H., SIU, L., OZA, A., HOTTE, S., PETRENCIUC, O., CIHON, F., LATHIA, C. and SCHWARTZ, B. (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., 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

NABORS, L. B., SUPKO, J. G., ROSENFELD, M., CHAMBERLAIN, M., PHUPHANICH, S., BATCHELOR, T., DESIDERI, S., YE, X., WRIGHT, J. et al. (2011). Phase I trial of sorafenib in patients with recurrent or progressive malignant glioma. *Neuro-Oncol.* **13** 1324–1330. https://doi.org/10.1093/neuonc/nor145

NAKAFUSA, Y., TANAKA, M., OHTSUKA, T., MIYOSHI, A., KOHYA, N., KITAJIMA, Y., SATO, S., MOCHINAGA, S., DOHI, S. et al. (2008). Phase I/II study of combination therapy with S-1 and CPT-11 for metastatic colorectal cancer. *Molecular Medicine Reports* **1** 925–930.

NEAL, R. M. (1999). Regression and classification using Gaussian process priors. In *Bayesian Statistics*, 6 (*Alcoceber*, 1998) (J. M. Bernardo et al., eds.) 475–501. Oxford Univ. Press, New York. MR1723510

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., MATANO, A., TANG, Z., ROYCHOUDHURY, S., WANDEL, S. and BAILEY, S. (2015). A Bayesian industry approach to phase I combination trials in oncology. In *Statistical Methods in Drug Combination Studies* (W. Zhao and H. Yang, eds.) **6**, 95–135. Chapman & Hall/CRC, Boca Raton.

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

O'QUIGLEY, J. and ZOHAR, S. (2010). Retrospective robustness of the continual reassessment method. *J. Biopharm. Statist.* **20** 1013–1025. MR2756534 https://doi.org/10.1080/10543400903315732

OGATA, Y., SASATOMI, T., AKAGI, Y., ISHIBASHI, N., MORI, S. and SHIROUZU, K. (2009). Dosage escalation study of S-1 and irinotecan in metronomic chemotherapy against advanced colorectal cancer. *Kurume Med. J.* **56** 1–7. https://doi.org/10.2739/kurumemedj.56.1

PRETORIUS, A. G. P. S. (2016). Phase III trial failures: Costly, but preventable. *Applied Clinical Trials* **25**.

RASMUSSEN, C. E. and WILLIAMS, C. K. I. (2006). *Gaussian Processes for Machine Learning*. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR2514435

SCHWINGHAMMER, T. L. and KROBOTH, P. D. (1988). Basic concepts in pharmacodynamic modeling. *J. Clin. Pharmacol.* **28** 388–394.

SHIOZAWA, M., SUGANO, N., TSUCHIDA, K., MORINAGA, S., AKAIKE, M. and SUGIMASA, Y. (2009). A phase I study of combination therapy with S-1 and irinotecan (CPT-11) in patients with advanced colorectal cancer. *Journal of Cancer Research and Clinical Oncology* **135** 365–370.

STEUTEL, F. W. (1979). Infinite divisibility in theory and practice. *Scand. J. Stat.* **6** 57–64. MR0538596

STRUMBERG, D., RICHLY, H., HILGER, R. A., SCHLEUCHER, N., KORFEE, S., TEWES, M., FAGHIH, M., BRENDEL, E., VOLIOTIS, D. 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.

TAKIUCHI, H., NARAHARA, H., TSUJINAKA, T., GOTOH, M., KAWABE, S.-I., KATSU, K.-I., IISHI, H., TATSUTA, M., FUJITANI, K. et al. (2005). Phase I study of S-1 combined with irinotecan (CPT-11) in patients with advanced gastric cancer (OGSG 0002). *Japanese Journal of Clinical Oncology* **35** 520–525.

THOMAS, N., SWEENEY, K. and SOMAYAJI, V. (2014). Meta-analysis of clinical dose-response in a large drug development portfolio. *Stat. Biopharm. Res.* **6** 302–317.

TURNER, R. M., DAVEY, J., CLARKE, M. J., THOMPSON, S. G. and HIGGINS, J. P. T. (2012). Predicting the extent of heterogeneity in meta-analysis, using empirical data from the cochrane database of systematic reviews. *Int. J. Epidemiol.* **41** 818–827.

UHLENBECK, G. E. and ORNSTEIN, L. S. (1930). On the theory of Brownian motion. *Phys. Rev.* **36** 823–841.

URSINO, M. (2020). MADF. Available at https://github.com/moreno-ursino/MADF/.

URSINO, M., YUAN, Y., ALBERTI, C., COMETS, E., FAVRAIS, G., FRIEDE, T., LENTZ, F., STALLARD, N. and ZOHAR, S. (2019). A dose finding design for seizure reduction in neonates. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **68** 427–444. MR3903002 https://doi.org/10.1111/rssc.12289

URSINO, M., RÖVER, C., ZOHAR, S. and FRIEDE, T. (2021a). Supplement to "Random-effects meta-analysis of Phase I dose-finding studies using stochastic process priors." https://doi.org/10.1214/20-AOAS1390SUPPA.

URSINO, M., RÖVER, C., ZOHAR, S. and FRIEDE, T. (2021b). Supplement to "Random-effects meta-analysis of Phase I dose-finding studies using stochastic process priors." https://doi.org/10.1214/20-AOAS1390SUPPB.

YODA, S., SOEJIMA, K., YASUDA, H., NAOKI, K., KAWADA, I., WATANABE, H., NAKACHI, I., SATOMI, R., NAKAYAMA, S. et al. (2011). A phase I study of S-1 and irinotecan combination therapy in previously treated advanced non-small cell lung cancer patients. *Cancer Chemotherapy and Pharmacology* **67** 717–722.

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

ZOHAR, S., LIAN, Q., LEVY, V., CHEUNG, K., IVANOVA, A. and CHEVRET, S. (2008). Quality assessment of phase I dose-finding cancer trials: Proposal of a checklist. *Clin. Trials* **5** 478–485.