RESEARCH ARTICLE

# A straightforward meta-analysis approach for oncology phase I dose-finding studies

## Christian Röver[1] | Moreno Ursino[2,3,4] | Tim Friede[1] | Sarah Zohar[3,4]

[1]Department of Medical Statistics, University Medical Center Göttingen, Göttingen, Germany

[2]Unit of Clinical Epidemiology, AP-HP, CHU Robert Debré, Université Paris Cité, Inserm CIC-EC 1426, Paris, France

[3]Inserm, Centre de Recherche des Cordeliers, Université Paris Cité, Sorbonne Université, Paris, France

[4]HeKA, Inria Paris, Paris, France

**Correspondence**
Christian Röver, Department of Medical Statistics, University Medical Center Göttingen, Göttingen, Germany.
Email: christian.roever@med.uni-goettingen.de

Phase I early-phase clinical studies aim at investigating the safety and the underlying *dose-toxicity* relationship of a drug or combination. While little may still be known about the compound's properties, it is crucial to consider quantitative information available from any studies that may have been conducted previously on the same drug. A meta-analytic approach has the advantages of being able to properly account for between-study heterogeneity, and it may be readily extended to prediction or shrinkage applications. Here we propose a simple and robust two-stage approach for the estimation of maximum tolerated dose(s) utilizing penalized logistic regression and Bayesian random-effects meta-analysis methodology. Implementation is facilitated using standard R packages. The properties of the proposed methods are investigated in Monte Carlo simulations. The investigations are motivated and illustrated by two examples from oncology.

**KEYWORDS**
Bayesian statistics, dose-escalation trial, random-effects meta-analysis, shrinkage estimation

## 1 | INTRODUCTION

Phase I dose-finding studies are the first-in-human studies of the clinical development which aim at estimating the maximum tolerated dose (MTD) of a drug or a combination of molecules. The MTD is usually defined as the dose level associated with a probability of occurrence of treatment-related adverse events, so-called dose-limiting toxicities (DLTs), at a prespecified level ($0 < \pi^\star < 1$; commonly: $\pi^\star = 33\%$ or $\pi^\star = 25\%$). DLTs are usually composite endpoints comprising a range of adverse events according their severity grade.

These exploratory early-phase studies are typically composed by a small sample of healthy volunteers, or patients (eg, in oncology because of the potential high toxicity of drugs, or in pediatrics because of the vulnerable population).[1,2] Due to the limited sample size and ethical concerns, simple randomized designs are usually not applied, but rather response-adaptive sequential designs, including some methods that can potentially find the MTD sooner and limit the number of observed DLTs.[3,4] These sequential designs can be divided in two main groups: (1) *algorithm-based* and (2)

---

*model-based* methods. In the first group, we can find the classical *3+3 design*, which uses prespecified fixed rules and only information of the patients at one dose (the one given to the previous cohort of patients) to allocate the next cohort to a dose level.[5,6] On the other hand, model-based designs, such as the *continual reassessment method* (CRM)[7] or the *Bayesian logistic regression model* (BLRM),[8] use an underlying statistical working model to describe the dose-toxicity relationship and a prespecified toxicity rate (the *target toxicity*), to allocate the next cohort of patients. While the former are easier to implement, the latter have better operating characteristics, that is, a higher percentage of correct MTD selection and greater chances of promoting effective drugs to subsequent development phases.[9-11] As a compromise between these two approaches, *model-assisted* designs have been proposed, which allow to combine the simplicity of rule-based designs with the estimation of model-based designs.[12,13] Like the model-based approaches, these use a statistical model to develop the design, but fixed dose allocation rules are determined at the preplanning stage, which makes implementation about as straightforward as for rule-based designs.

In all medical fields, and above all in oncology, identifying the appropriate therapeutical dose range is crucial to avoid exposing patients to an unacceptable toxicity profile or to insufficient efficacy.[14] Recently, Harrison[15] reported that phase II and phase III study failures were often due to lack of efficacy (52%) and safety (24%). They failed owing to inadequate patient selection, study design, biomarkers, and schedules, as well as data analysis, but the most evident reason was wrong estimation of the dose-response relationship during exploratory clinical trials. Thus, an improper dose selection at the first stages of clinical development may lead either to abandonment of promising drugs, or to the focusing on a wrong dose (too toxic or lacking efficacy). Therefore, using all available data, including other exploratory trials, in order to improve the estimation of the MTD is of utmost importance and the use of meta-analysis methods could be helpful. To cope with the specifics of phase I trials, that is, small sample sizes and the consideration of several dose levels, only few meta-analytic methods have been proposed so far. Zohar et al[16] proposed a common-effect (fixed-effect) method based on the retrospective CRM. Ursino et al[17] developed a more sophisticated one-stage random-effects approach based on stochastic process priors. Other authors, instead, focused directly on efficacy endpoints, such as Kim et al[18] who developed random-effects meta-analysis methodology to deal with a relaxed exchangeability assumption and rare events.

Our take was to propose the adaptation and the use of simpler approaches developed for regular meta-analysis, and to check their suitability in realistic phase I scenarios. We considered estimates of the MTD along with their standard errors, which are readily derived from the outcomes of a dose-finding trial based on a simple logistic model. In order to account for the small sample sizes typically encountered in early-phase studies, we suggest the use of a penalization approach correcting for bias in the logistic regression stage.[19,20] We then utilize a univariate two-stage meta-analysis model to synthesize the different estimates. Implementation is based on the `logistf` and `bayesmeta` R packages.[21,22] The results of the proposed meta-analysis methods may be utilized not only for an overall summary estimate (eg, for consideration by regulatory agencies), but also for the design of a prospective trial,[23,24] for the selection of the dose panel, for sample size determination,[25] or as a meta-analytic predictive (MAP) prior for Bayesian inference.[26,27]

The organization of this article is as follows; in Section 2, we motivate the problem by considering two example datasets; in both cases, a literature search on `clinicaltrials.gov` and PubMed yielded a number of phase I studies investigating a novel drug at a range of dose levels. In Section 3, we introduce the statistical methods for MTD estimation and meta-analysis used in the following. In Section 4, we investigate the proposed methods' performance using Monte Carlo simulations considering the combined estimate as well as shrinkage and prediction. In Section 5, we analyze the example data, and Section 6 finally closes with some conclusions.

## 2 | ILLUSTRATION

### 2.1 | Sorafenib example

The first illustrating example concerns Sorafenib (BAY 43-9006), a kinase inhibitor for the treatment of advanced renal cell carcinoma, hepatocellular carcinoma, and radioactive iodine resistant advanced thyroid carcinoma.[28] Thirteen trials with published results, described in 11 manuscripts, were identified in a literature search. The doses used in each trial, the numbers of patients allocated, and the total numbers of DLTs at each dose are summarized in Table 1.

**TABLE 1** The results of 13 studies on Sorafenib monotherapy

| | | | Dose (mg) | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Publication | Year | Country | 100 | 200 | 300 | 400 | 600 | 800 | 1000 |
| Awada et al[29] | 2005 | Belgium | 0/4 | 0/3 | 1/5 | 1/10 | 7/12 | 1/3 | |
| Clark et al[30] | 2005 | USA | 0/3 | 0/3 | | 1/4 | 1/6 | 3/3 | |
| Moore et al[31] | 2005 | Canada | 0/3 | 1/6 | | 0/8 | 3/7 | | |
| Strumberg et al[32] | 2005 | Germany | 1/5 | 1/6 | | 0/15 | 4/14 | 2/7 | |
| Furuse et al[33] | 2008 | Japan | | 0/12 | | 1/14 | | | |
| Minami et al[34] | 2008 | Japan | 0/3 | 1/12 | | 0/6 | 1/6 | | |
| Miller et al[35] | 2009 | USA | | 8/34 | | 6/20 | | | |
| Crump et al (A)[36] | 2010 | Canada | 0/4 | 1/6 | 0/6 | 1/6 | | | |
| Crump et al (B)[36] | 2010 | Canada | 0/3 | 1/6 | | 0/3 | 2/6 | | |
| Borthakur et al (A)[37] | 2011 | USA | | 0/3 | | 0/15 | 2/8 | | |
| Borthakur et al (B)[37] | 2011 | USA | | 0/3 | | 1/7 | 2/6 | | |
| Nabors et al[38] | 2011 | USA | | 0/3 | | 1/6 | 0/3 | 1/5 | 3/3 |
| Chen et al[39] | 2014 | USA | | 0/3 | | 1/16 | | | |

*Note*: For each dose considered in each trial, the numbers of patients experiencing DLT events, and the total numbers of exposed patients are given.

Seven doses (100, 200, 300, 400, 600, 800, and 1000 mg) were investigated across all studies. DLT definitions and Sorafenib administration schedules (about a 28-day cycle) were comparable across studies. The total sample sizes vary from 16 up to 54 patients, the maximum was achieved in one of the studies where Sorafenib was tested in 9 different patient subgroups defined by organ dysfunctions.[35] As expected for dose-escalation trials, observed DLT frequencies generally tend to increase with increasing dose, with zero or a low event rates at the lower doses placed at the beginning of the dose panel.

The majority of studies were performed in European or North American populations, while only two studies were done in Japanese patients.[33,34] Indeed, the Japanese Pharmaceuticals and Medical Devices Agency (PMDA) recommends a treatment re-evaluation in case insufficient data are available from a Japanese population.[40] As it is unclear to what extent evidence from the differing (Western and Japanese) populations is directly transferrable to one another,[41,42] we can consider a cautious *dynamical borrowing* approach in order to bridge between the two study subsets rather than pooling the data "naïvely".[43]

## 2.2 | Irinotecan/S-1 example

The second illustrating example concerns a combination therapy of Irinotecan (a topoisomerase 1 inhibitor) and S-1 (a combination of three pharmacological compounds, namely, tegafur, gimeracil, and oteracil potassium) that was tested in advanced colorectal and gastric cancer in a Japanese population.[44,45] Data extracted from 12 studies are shown in Table 2.

A total of 10 doses, ranging from 40 up to 150 mg/m$^2$, were evaluated among all trials. Two or more infusions of Irinotecan were planned in all trials, except for Yamada et al[46] and Yoshioka et al[53] with only one infusion at the first day cycle. The sample sizes range from 6, in agreement with the well-known 3+3 design, to 51, since, in Inokuchi et al,[48] DLTs were still recorded at the phase II stage of the study.

**TABLE 2** The results of 12 Japanese studies on combination therapy of Irinotecan and S-1

| Publication | Year | Dose (mg/m²) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 40 | 50 | 60 | 70 | 80 | 90 | 100 | 120 | 125 | 150 |
| Yamada et al[46] | 2003 | | | | | | | 0/3 | | 0/3 | 1/6 |
| Takiuchi et al[47] | 2005 | 1/6 | | 0/3 | | 0/4 | | 3/6 | | | |
| Inokuchi et al[48] | 2006 | | | | 0/3 | 10/42 | 0/3 | 2/3 | | | |
| Nakafusa et al[49] | 2008 | | | 7/39 | | 2/3 | | | | | |
| Ishimoto et al[50] | 2009 | | 0/3 | 0/3 | 0/3 | 2/4 | | | | | |
| Ogata et al[51] | 2009 | 0/3 | 0/3 | 3/4 | | | | | | | |
| Shiozawa et al[52] | 2009 | | | | | 1/6 | | 2/6 | 2/6 | | 2/3 |
| Yoshioka et al[53] | 2009 | | | | | | | 0/3 | | 1/6 | 0/3 |
| Komatsu et al[54] | 2010 | | | | | | | 1/9 | | 1/9 | 0/3 |
| Kusaba et al[55] | 2010 | | | | | 0/6 | | 2/3 | | | |
| Yoda et al[56] | 2011 | | | 0/3 | | 3/6 | | | | | |
| Goya et al[57] | 2012 | | | | 0/3 | 0/3 | 3/5 | | | | |

*Note*: For each dose considered in each trial, the numbers of patients experiencing DLT events, and the total numbers of exposed patients are given.

# 3 | METHODS

## 3.1 | Dose-finding experiments: Assumptions

Within a dose-finding experiment, we have a discrete set of covariable levels $x_j$, usually denoted as *doses* or *exposures*, or transformations thereof, which are indexed in increasing order. The response is an event count $r_j$ among a total number $n_j$ of patients that have been exposed to the $j$th dose.

We assume the DLT count $r_j$ among the total of $n_j$ patients exposed at level $x_j$ to follow a binomial distribution,

$$r_j \sim \text{Binomial}(p_j, \; n_j). \tag{1}$$

MTD estimation then aims at determining the dose $x^\star$ for which the DLT probability reaches a certain threshold $\pi^\star$ (or sometimes also the largest dose with DLT probability $p_j \leq \pi^\star$). Depending on the approach taken, the dose level $x^\star$ may be assumed to be among the set of experimentally considered dose levels $(x_1, \dots, x_J)$ or also in between or beyond the investigated doses.

## 3.2 | Logistic regression model

We apply logistic regression models with a logit link for each study. The DLT counts are modeled via a binomial distribution as in (1). The logistic regression model provides a joint parameterization of the DLT probabilities $p_j$ as a parametric function of the dose covariable $x_j$:

$$\text{logit}(p_j) = \beta_0 + \beta_1 x_j, \tag{2}$$

where the logit link function and its inverse are defined as

$$\text{logit}(x) = \log\left(\frac{x}{1-x}\right) \quad \text{and} \quad \text{logit}^{-1}(y) = \frac{\exp(y)}{1 + \exp(y)}. \tag{3}$$
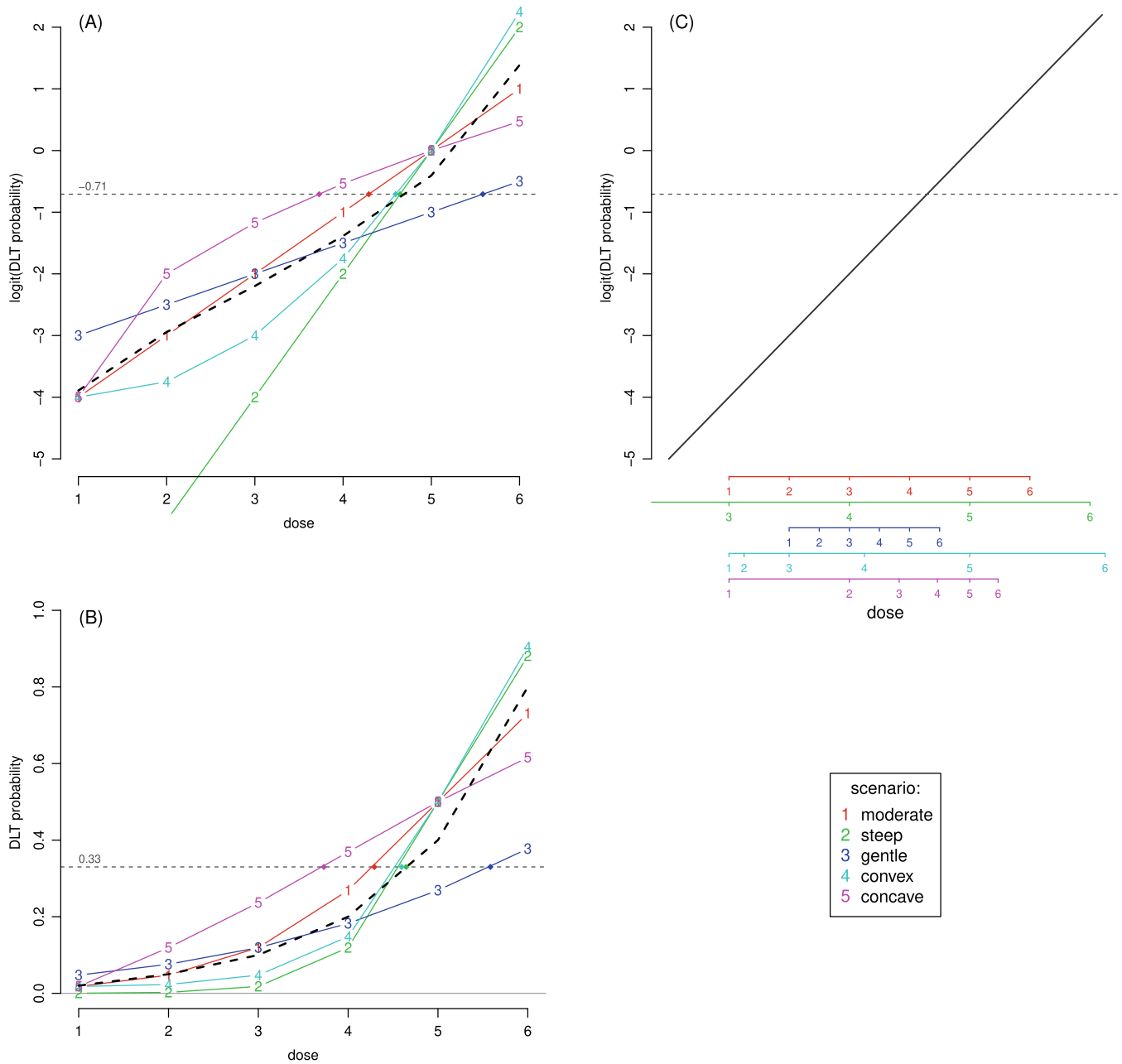
**FIGURE 1** The five simulation scenarios (dose-response curves) on the probability (A) and logit scales (B). The dashed line indicates the "skeleton" (a prior guess of the curve, which needs to be specified for CRM and BLRM methods). The corresponding MTDs, the doses at which a probability of 0.33 is attained (where logit(0.33) = −0.71), are marked by a dot. See also Table 3 for the actual numbers. (C) The five scenarios may also be motivated via a common underlying dose-response curve, but differing sets of dose levels

Note that the *logits* of the probabilities corresponds to the *logarithmic odds*, where the *odds* corresponding to a probability $p \in [0, 1]$ are given by $\frac{p}{1-p} \in [0, \infty]$. The linearity assumption constrains the DLT probabilities and reduces dimensionality of the problem: only two unknown parameters, the intercept $\beta_0$ and slope $\beta_1$ now determine the dose-response relationship. In addition, it facilitates interpolation or extrapolation beyond the discrete set of covariable ($x_j$) values. Introduction of a parametric model also means that monotonic transformations of the dose levels ($x_i$) may affect the model fit (see also Figure 1).

Within the logistic model, the MTD directly results from the regression parameters ($\beta_0$ and $\beta_1$). Inverting Equation (2) to derive the dose level $x^\star$ for which the prespecified toxicity of $\pi^\star$ is attained (ie, so that $\beta_0 + \beta_1 x^\star = \text{logit}(\pi^\star)$) leads to

$$x^\star = \frac{\text{logit}(\pi^\star) - \beta_0}{\beta_1} \tag{4}$$

as the corresponding MTD.

The logistic regression model as a very common procedure is a simple and pragmatic model choice here; a number of other models or parameterizations are also commonly used in dose-response modeling.[14,58] The logistic model also includes the so-called *Emax model* as a special case (see Appendix C for details).

### 3.3 | Parameter estimation in the logistic model

Simple maximum-likelihood (ML) estimation within the logistic regression model is sometimes problematic, in particular in case of smaller sample sizes. ML estimators are affected by certain biases, and *separation* issues may arise.[59] Separation refers to "pathological" data constellations where the predictor allows to perfectly split the data into cases and non-cases, which in the estimation stage leads to problems (a likelihood that does not attain a definite maximum at finite parameter values, which in practice often leads to numerical problems, resulting in large estimates as well as associated standard errors).

The use of the *Firth correction* has been suggested in order to reduce bias and to ensure finite parameter estimates.[19,60] Heinze and colleagues[61,62] showed that the use of the Firth correction avoids separation issues and may improve properties also in small samples. In the present context, our focus is not only on contrast estimates (odds ratios) but also hinges on the accuracy absolute (DLT) event probability estimates. Puhr et al[20] introduced *Firth's logistic regression with added covariate* (FLAC) as a method to further reduce bias in event probability estimates; the FLAC method has been recommended due to improved performance as well as its invariance properties.[20]

The logistic model falls within the class of *generalized linear models* (GLMs), and as such, implementations are readily available. Within R, one may utilize the "`glm()`" function for simple logistic regression, while the "`MASS`" library also provides the functionality to retrieve estimates and standard errors for the MTD.[63] Firth logistic regression and FLAC are implemented in the "`logistf`" R package.[21] Point estimates and their associated variance-covariance may again be utilized to derive corresponding MTD estimates and standard errors.

### 3.4 | Deriving MTD dose estimates

#### 3.4.1 | General remarks

Note that the MTD $x^\star$ does not necessarily fall within the range of investigated doses $(x_1, \ldots, x_j)$; the resulting estimate hence is usually an interpolation, or even an extrapolation. The actual task is to use the inverted equation (4) based on *estimates* of the coefficients $\beta_0$ and $\beta_1$ in order to derive an estimate of the MTD $x^\star$. A common approach is to assume a joint normal distribution for the ML estimates $\hat{\beta}_0$ and $\hat{\beta}_1$. The distribution of the estimated MTD then results as a ratio of (correlated) normal variates (see (4)). Such a ratio may often again be reasonably approximated by a normal distribution, however, in general the distribution may also be bimodal and heavy-tailed.[64] In particular, its first two moments (mean and variance) in general do not exist.

In the following, we will investigate a simple error propagation approach based on the *delta method*. Note that the eventual objective of this exercise is to provide estimates and standard errors to be passed on to the meta-analysis procedure at a subsequent stage. Estimates and standard errors are then utilized to specify the (approximately) normal likelihood (see Section 3.5). The aim hence is to provide an accurate (or reasonably conservative) normal approximation to the uncertainty in the MTD.[65]

#### 3.4.2 | The delta method

Consideration of standard errors (the variance-covariance matrix) of the regression coefficients $\beta_0$ and $\beta_1$ via the *delta method* yields a corresponding standard error for the MTD estimate $x^\star$.[66] This procedure is described in detail, including R code, by Venables and Ripley.[63] It should be noted, however, that this only works well in case of reasonably small standard

errors; in particular, the denominator in Equation (4), the regression slope, needs to be somewhat bounded away from zero (ie, it needs to have a small coefficient of variation).[64]

## 3.5 | Two-stage approach to meta-analysis

### 3.5.1 | The normal-normal hierarchical model

With MTD estimates that are provided along with their standard errors, the meta-analysis problem falls into the generic class of problems that may be addressed using the normal-normal hierarchical model (NNHM).[67-72] For each study $i$ ($i = 1, \ldots, k$), there is an underlying (unknown) true MTD value $\theta_i$. Analysis of the experimental data yields an MTD estimate $y_i$ that has a standard error $s_i$ associated. Utilizing a simple normal model, we assume that

$$y_i | \theta_i, s_i \sim \text{Normal}(\theta_i, s_i^2), \tag{5}$$

where the study-specific true values $\theta_i$ (the MTDs) are not necessarily identical across the different studies; these also have a certain amount of variability associated, which is implemented using another variance component via

$$\theta_i | \mu, \tau \sim \text{Normal}(\mu, \tau^2). \tag{6}$$

The heterogeneity parameter $\tau$ denotes the between-study variability, and $\mu$ is the overall mean. The model may also be expressed in its marginal form as

$$y_i | \mu, \tau, s_i \sim \text{Normal}(\mu, \ s_i^2 + \tau^2). \tag{7}$$

### 3.5.2 | Estimation within the NNHM

In the Bayesian framework, the estimates $y_i$ and their standard errors $s_i$ are utilized to define the (approximately) normal data likelihood; inference needs to consider the unknowns $\mu$ and $\tau$, and primary interest is usually in the overall mean $\mu$. For the design and analysis of a future dose-finding study, prediction (of a "new" MTD $\theta_{k+1}$) or shrinkage estimation (of one of the MTDs $\theta_i$) may also be relevant.[26,27] Prior distributions need to be specified for the overall mean ($\mu$) and the heterogeneity ($\tau$), expressing any a-priori information that may be available on these parameters. In the following, we will utilize uniform effect and heterogeneity priors, due to the mostly large sample sizes (numbers of studies $k$) in the examples. In general, or in particular when faced with only a small number of studies, the use of (weakly) informative priors may also be appropriate.[72,73]

## 4 | SIMULATIONS

### 4.1 | Aims

The aim is to first investigate the performance of the (first-stage) regression methods, and then also of the (second-stage) meta-analysis. For the logistic regression models, we will check bias, CI coverage probabilities and CI widths; for the meta-analyses, we focus on CI coverages and widths (for overall mean, prediction, shrinkage estimates). To this end, we will utilize Monte Carlo experiments using simulated data from a range of relevant scenarios.

Even when assuming that the logistic regression analyses yield relatively unbiased estimates of the regression coefficients, the eventual consideration of derived MTDs (a *nonlinear* transformation, see Equation (4)) means that we again should expect some bias. Considered as a function of the slope $\beta_1$, the transformation in (4) is a *convex* function, so if we assume that the MTD uncertainty was dominated by uncertainty in the slope, then (following Jensen's inequality) we may expect a negative bias here. Considered as a function of both intercept and slope, however, the MTD is neither convex nor concave.

In addition, it also remains to be seen how well a normal approximation works in a setup where moments may actually not be finite.[64]

## 4.2 | Simulation scenarios

The simulation scenarios considered are intended to be realistic in terms of the data that are commonly encountered in dose-escalation studies (as, eg, in the examples from Section 2; see also Appendix A). We will use a mixture of data generated by employing the common "3+3,"[5,6] "CRM,"[7,74] and "BLRM"[4,8,75] study designs (with equal probabilities). While the algorithmic 3+3 design used to be the most common in the past, alternative designs have recently gained popularity.[76,77] The different designs will then result in differing kinds of datasets, in terms of sample sizes, event rates, investigated dose ranges, and targeted DLT probabilities (as, eg, the 3+3 design is known to target a probability of 20%-25%).[78,79] Implementation details regarding the three designs are provided in Appendix D.

*Toxicity profiles*: We will use 5 dose-response curves that are shown in Figure 1 and Table 3. All of these are based on 6 dose levels ($x_i \in \{1, 2, \ldots, 6\}$), and they are labeled as "moderate," "steep," "gentle," "convex," or "concave." The first three are linear, while the latter two deviate from the logit model assumed in the analyses. Defining the targeted DLT probability as $\pi^\star = 0.33$, the MTD is attained within the dose range (between 1 and 6) in all 5 cases; the MTDs are also listed in Table 3.

*Design criteria*: The differing dose-escalation designs require the specification of certain free parameters. The 3+3 design is initiated at the lowest dose, a cohort size of 3 is used, there is no skipping of doses, and no dose escalation following a DLT event. CRM and BLRM designs utilize the "skeleton" (a prior guess of the response curve) shown in Figure 1 and Table 3. For CRM and BLRM, the maximum sample size was chosen randomly as a multiple of 3 between 15 and 30.

*Heterogeneity*: Heterogeneity in MTDs will be implemented via variability in the *intercept* term of the dose-response curve. In Figure 1, one can see that shifting the curves in the vertical direction will also increase or decrease the associated MTD. For a logistic dose-response (as in Equation 2), adding a random offset with variance $(\tau\beta_1)^2$ will imply a heterogeneity variance of $\tau^2$ in the associated MTD. For the nonlinear scenarios, the relationship is not as simple, but instead of the slope $\beta_1$, the slope *at the MTD* (as in Table 3) is simply used instead as an approximation. Heterogeneity levels of $\tau \in \{0.0, 0.5, 1.0\}$ will be considered.

*Numbers of studies*: At the meta-analysis stage, totals of $k \in \{5, 10, 20\}$ studies will be considered.

It should be noted that the five different dose-response curves may also be related to one another by considering a "common" curve with differing dose placement; see also the illustration in Figure 1. Figure 1C shows how the five scenarios may result from considering different sets of investigated doses. A denser or wider spacing affects the resulting slope, and constancy of distances between neighboring doses affects overall linearity. Differences between scenarios do not only relate to the underlying dose-response mechanism, but also to choices made at the experimental design stage. For example, the third ("gentle") scenario may be viewed as a scenario in which only a very narrow dose range is

**TABLE 3** The DLT probabilities in the five example scenarios, and the associated true MTDs

| | | Scenario | | | | | |
|---|---|---|---|---|---|---|---|
| | | **Moderate** | **Steep** | **Gentle** | **Convex** | **Concave** | **Skeleton** |
| Dose | 1 | 0.018 | 0.00034 | 0.047 | 0.018 | 0.018 | 0.02 |
| | 2 | 0.047 | 0.0025 | 0.076 | 0.023 | 0.119 | 0.05 |
| | 3 | 0.119 | 0.018 | 0.119 | 0.047 | 0.237 | 0.10 |
| | 4 | 0.269 | 0.119 | 0.182 | 0.148 | 0.369 | 0.20 |
| | 5 | 0.500 | 0.500 | 0.269 | 0.500 | 0.500 | 0.40 |
| | 6 | 0.731 | 0.881 | 0.378 | 0.905 | 0.616 | 0.80 |
| MTD | | 4.29 | 4.65 | 5.58 | 4.60 | 3.73 | |

*Note*: See also Figure 1.

**TABLE 4** Mean numbers of doses/patients/events resulting from the different data scenarios and designs

| Design | Scenario | | | | |
| | Moderate | Steep | Gentle | Convex | Concave |
| --- | --- | --- | --- | --- | --- |
| 3+3 | 4.3/16.1/2.8 | 5.0/17.1/2.7 | 4.6/17.4/2.7 | 4.9/17.2/2.8 | 3.7/13.9/2.7 |
| CRM | 4.5/22.3/4.6 | 5.0/22.5/4.8 | 4.6/22.6/3.5 | 4.8/22.5/4.7 | 4.0/22.5/5.3 |
| BLRM | 4.5/22.6/4.8 | 5.0/22.6/4.0 | 4.9/22.8/3.8 | 4.9/22.5/4.2 | 4.0/22.4/5.4 |
| Overall | 4.5/20.3/4.1 | 5.0/20.7/3.9 | 4.7/20.9/3.3 | 4.9/20.7/3.9 | 3.9/19.6/4.5 |

*Note*: For CRM and BLRM, the maximum sample size is also random here (between 15 and 30).

investigated—which eventually (and maybe as expected) makes it harder to identify the exact MTD, while at the same time problems appear amplified since uncertainty in the MTD in terms of a certain absolute dose range then quickly translates into uncertainty spanning several dose levels. This highlights the importance of the careful choice of the dose (or *exposure*) scale and of possible transformations (eg, when considering log-transformed doses). It also emphasizes the difficulties arising in the comparison of results (in particular: standard errors and CI widths quoted in units of *doses*) across the different simulation scenarios. Seemingly "steep" and "gentle" shapes may simply arise from considering wider or narrower dose ranges.

## 4.3 | Simulation scenario properties

We briefly investigated the kinds of data that are returned when implementing the above model assumptions. Table 4 characterizes the data in terms of the mean numbers of doses used, the numbers of patients recruited and the numbers of DLT events under the three designs and overall (on average across designs).

Most notably, CRM and BLRM yield very similar data, while 3+3 differs slightly from the other two. The 3+3 design generally yields a smaller total sample size (numbers of patients *and* events). In the following, we will not distinguish between data generating models, but will consider a mixture of the three. The data characteristics are also roughly similar to those encountered in the Sorafenib and Irinotecan examples above (see also Tables 1, 2, A1, and A2).

## 4.4 | Results

### 4.4.1 | Comparing standard and penalized logistic regression methods

Figure 2 illustrates the estimation performance of the three investigated methods ("plain" logistic regression within the GLM framework, Firth correction, and FLAC). Figure 2A shows the error in the estimated MTD; within each scenario, all three methods perform comparably, while FLAC appears to be overall least biased. Overall there appears to be a tendency to underestimate the MTD, that is, the errors are on the conservative side. In particular the third scenario, characterized by a very gentle slope, leads to relatively large offsets (in terms of dose steps) and a negative bias. This may in fact be expected, as this is the scenario in which the slope estimate should tend to be particularly uncertain.

Figure 2B again illustrates the offset in the MTD estimate, this time in terms of the DLT probability at the estimated MTD. On this scale, differences between scenarios do not appear quite as dramatic. The probabilities are centered at their aimed for value of $\pi^\star = 0.33$ (marked by a red line), and depending on the scenario, they exhibit more or less variability. The FLAC estimate again appears to be closest to the target on average.

Figure 2C shows the uncertainties (estimated standard errors) associated with the DLT estimates. The Firth and FLAC estimates behave similarly, while the simple GLM estimate yields more extreme (either very small or very large) errors. Note that the standard error is in units of doses here, that is, a standard error > 1 means that the associated CI will span across several doses.

Figure 2D eventually shows the 95% CIs' coverage probabilities for the different combinations; in general we see some undercoverage, which is not actually satisfactory, while in most scenarios (in particular in those where the linearity assumption is met), the FLAC method again tends to perform best.

Overall, the regularized (Firth or FLAC) estimates clearly outperform the "plain" logistic regression for MTD estimation based on relatively small sample sizes. Also, while "plain" regression would fail in about 4% of cases, the other two always yield finite parameter estimates. Firth and FLAC estimates seem to behave more similarly, with a slight advantage apparent for the FLAC method. In the following, we will hence focus on the FLAC method for deriving MTD estimates (and associated standard errors) in the first stage of the analysis.

### 4.4.2 | Investigating MA performance

After checking the performance of MTD estimators for the first stage of the analysis, we will investigate the performance of meta-analyses based on estimated MTDs. In the meta-analysis setup, we will vary the five dose-response scenarios, as well as the amount of heterogeneity and the number of studies considered. As in the previous section, we will be considering the aspects of estimation error (in terms of doses and in terms of DLT probability), as well as CI width and coverage probability.
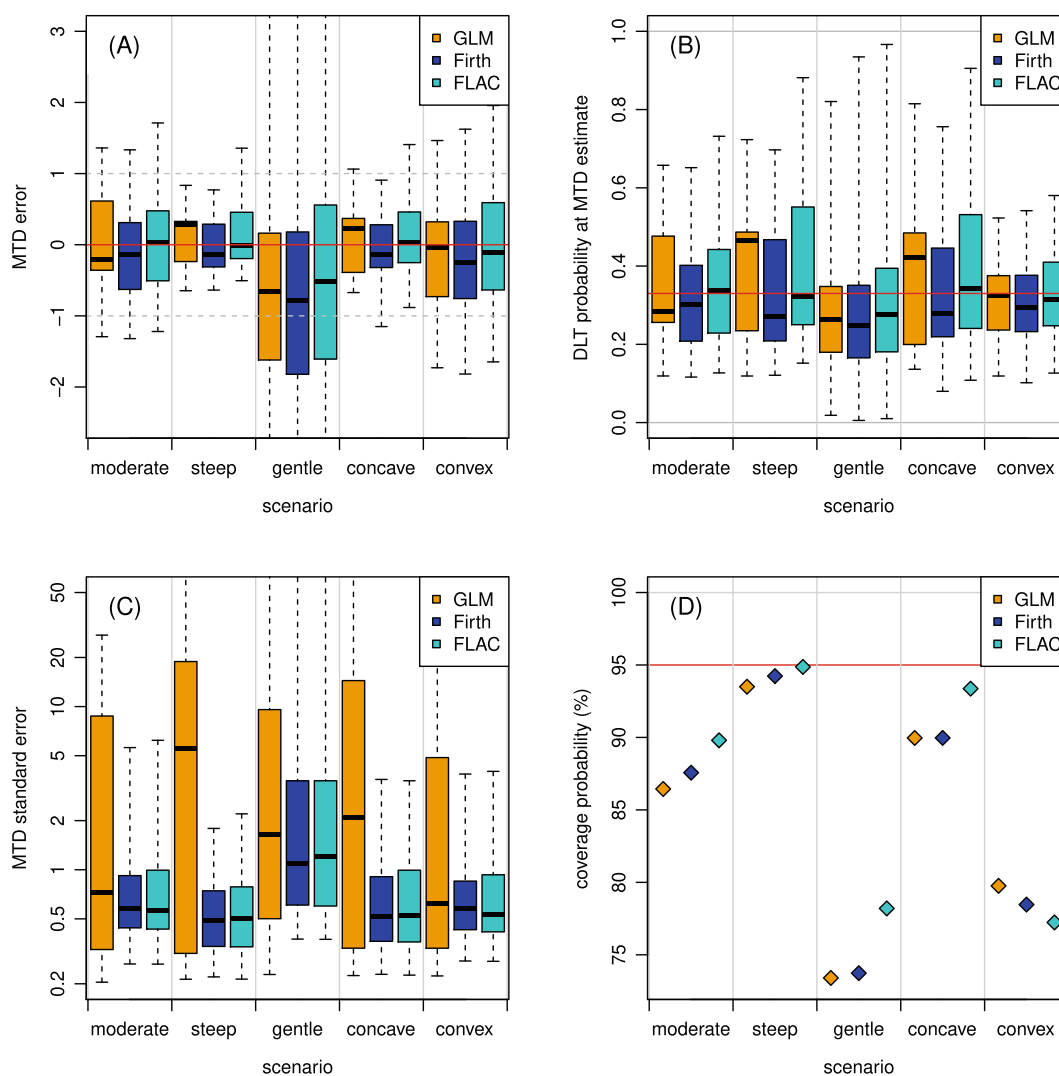


**FIGURE 2** Performance of the first-stage logistic regression methods. (A) The errors in MTD estimates in terms of the offset in *doses*. (B) The error in terms of the *DLT probability* at the estimated MTD; the horizontal reference line indicates the target of 33%. (C) The standard errors associated with the differing estimates, and (D) the coverage probabilities of derived 95% confidence intervals. Boxplots indicate the three quartiles and the central 90% range

Figure 3 illustrates the errors in MTD estimation, both on the dose scale as well as the DLT probability scale. It becomes evident that negative biases already apparent in the previous simulations propagate through to the meta-analysis results. Especially in the "gentle" scenario, the MTD gets underestimated, and an increasing number of studies ($k$) reinforces the biased estimate. In this scenario, the MTD estimate is about one dose off, corresponding to a DLT probability of roughly 20%-25% instead of the aimed for $\pi^\star$ of 33% (see also Table 3). For all the other scenarios, however, and in particular those where the linearity assumption is met, the estimates appear to behave reasonably, irrespective of the amount of heterogeneity ($\tau$) or the number of studies ($k$).

Figure 4 illustrates the widths and coverage probabilities of credible intervals for the MTD. The length of CIs behaves as expected; an increasing number ($k$) of studies considered leads to shorter intervals, while increasing the amount of heterogeneity ($\tau$) makes intervals slightly wider. Coverage probabilities of 95% intervals however are substantially off in some cases, which makes sense given the previously considered simulation results; when estimation bias is substantial and a large number of studies is included, so that bias is of the order of the CI width, then the coverage deteriorates. In the majority of scenarios considered, however, coverage probabilities are at reasonable levels.

As outlined in Section 3.5.2, instead of the overall mean $\mu$, in some applications it may also make sense to consider estimation of one of the study-specific means $\theta_i$ (shrinkage estimation), or of a "new" instance $\theta_{k+1}$ (prediction). For analogous results for prediction and shrinkage intervals, see Appendix E. In brief, one can see that the biases observed previously appear to propagate through, however, the coverage probabilities are substantially closer to their nominal levels.

# 5 | APPLICATIONS

## 5.1 | Sorafenib example

### 5.1.1 | MTD estimation

Now consider MTD estimation in the context of the Sorafenib example (see Table 1), again aiming for a DLT probability of $\pi^\star = 0.33$. Considering the empirical DLT frequencies encountered, note that a number of studies never actually reach an empirical DLT frequency of 33%, which also makes MTD estimation a bit of an extrapolation exercise here. Also, *separation* occurs, for example, in the study by Furuse et al,[33] where only two doses are investigated and events only were observed in one of the two groups. We will consider doses on their logarithmic scale for the analysis. Table A1 (in Appendix A) also shows the resulting MTD estimates and their standard errors.

The joint analysis of the 13 estimates is illustrated in Figure 5. Note that the data from some of the studies barely allow to constrain the MTD, which is reflected in correspondingly large standard errors associated with the estimates. The eventual analysis is hence supported by certain studies, while others may not contribute substantially. The contribution of studies to the resulting estimate may be expressed in terms of (percentage) *weights*, which are also shown in the figure.[80] The eventual estimate of the overall mean MTD is at 608.1 [470.5, 795.6]. In terms of dose *selection* (among the investigated doses; see Table 1), this range only includes the dose of 600 mg. This is in agreement with the result previously found by Ursino et al,[17] who worked using discrete dose levels. Therefore, the EMA-recommended dose of 400 mg (twice a day) is still considered a safer choice taking into account these new findings.[28]

When the meta-analysis is performed at the design stage of a new study, it may be more relevant to look at a prediction interval; this may be useful to help designing the study, or to formally include the "historical" information in the eventual analysis.[25-27] The prediction interval is wider, as it includes the estimated between-study heterogeneity, and here it ranges from 363.3 to 1044.8 mg.

For any of the 13 included studies, we may also investigate the study-specific MTDs $\theta_i$ in the light of the combined data via shrinkage estimation; for example the shrinkage estimate for the most recent study by Chen et al,[39] which contributed only little information (a rather vague estimate of 3149.5 [0.0, 2 146 367 310.5]), is at 607.0 [364.6, 1046.8]. As expected, this is very close to the prediction, but also substantially shorter than the original interval.

The heterogeneity ($\tau$) is estimated at 0.13 [0.00, 0.45], which seems to be at a reasonable level for a log-transformed MTD.[73,81]

Given the weights that are shown in Figure 5, an obvious question is to what extent the studies with "small" weights are actually relevant to the analysis. The weights are closely related to the standard errors ($\sigma_i$) associated with each estimate,[80] and it seems sensible that those studies with very large uncertainties associated may only help very little in constraining
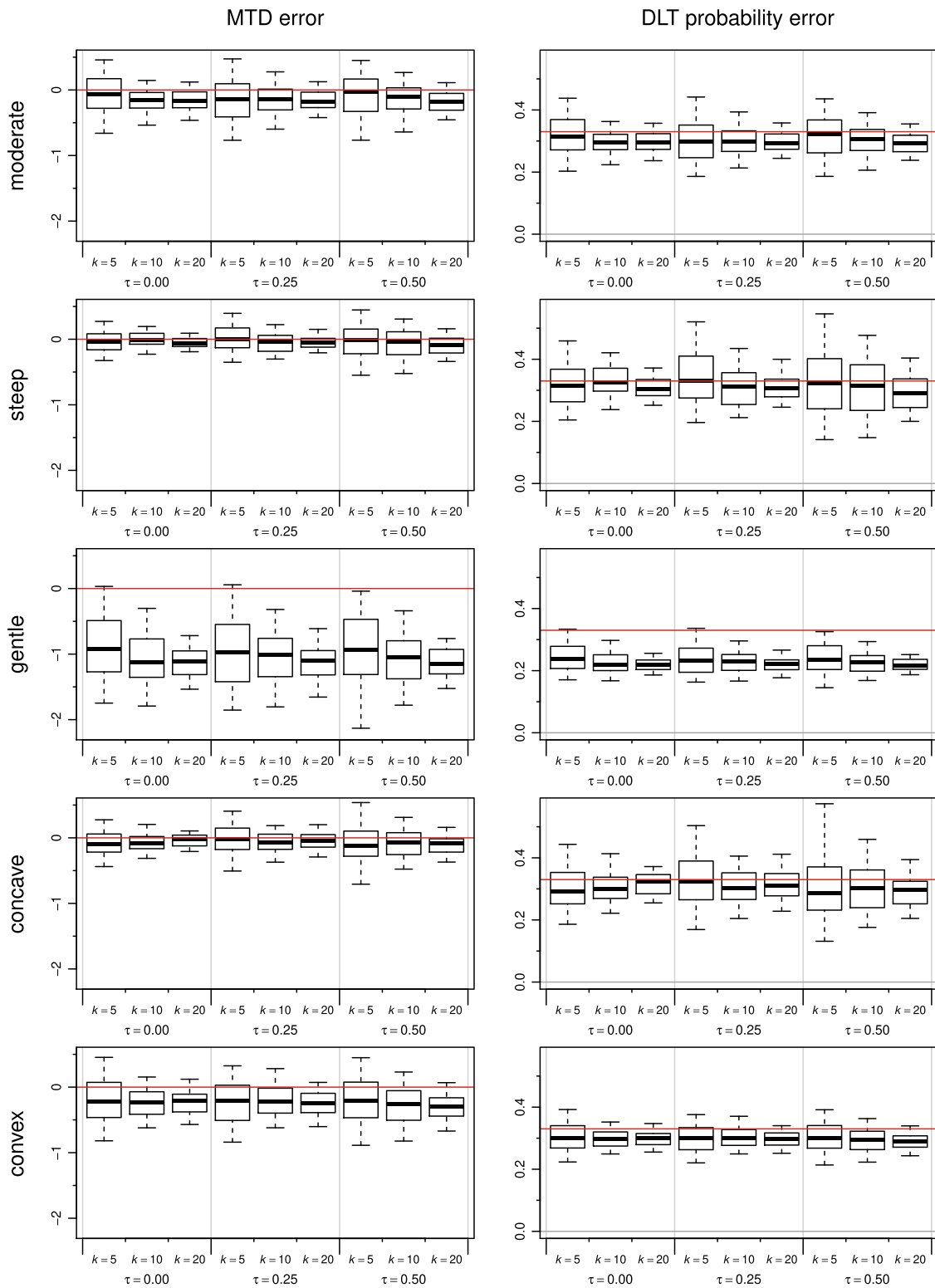
**FIGURE 3**  Performance of meta-analysis (based on FLAC estimates) in different scenarios, and for different numbers of studies and amounts of heterogeneity. The left column shows the offset in estimated MTD in terms of *doses*, and the right column shows the DLT probabilities at the estimated MTDs. Rows correspond to the different dose-response scenarios. Boxplots indicate the three quartiles and the central 90% range

**FIGURE 4** Widths and coverage probabilities of credible intervals for the (mean) MTD based on the meta-analyses. Boxplots indicate the three quartiles and the central 90% range

■ quoted estimate   ◆ shrinkage estimate

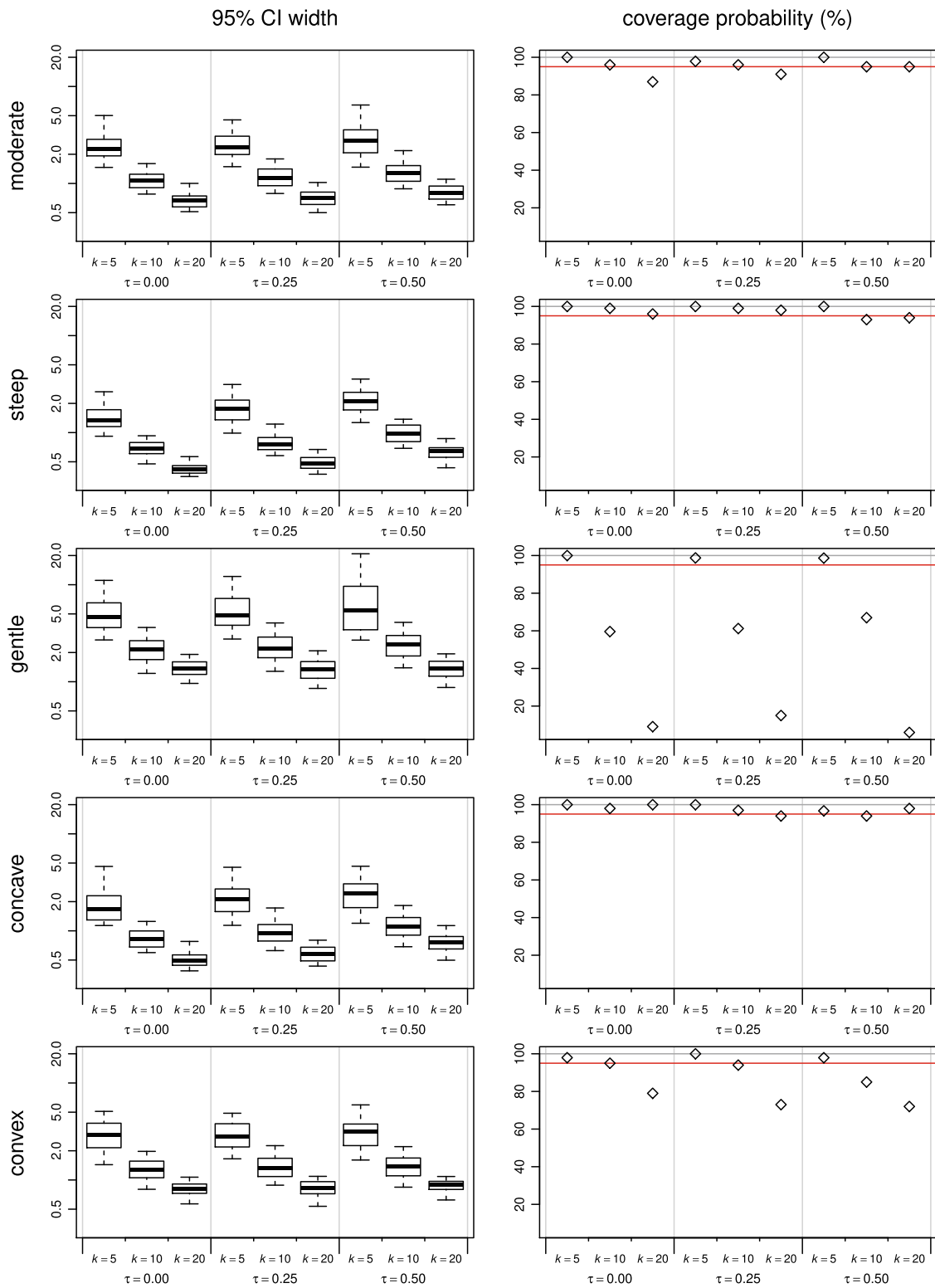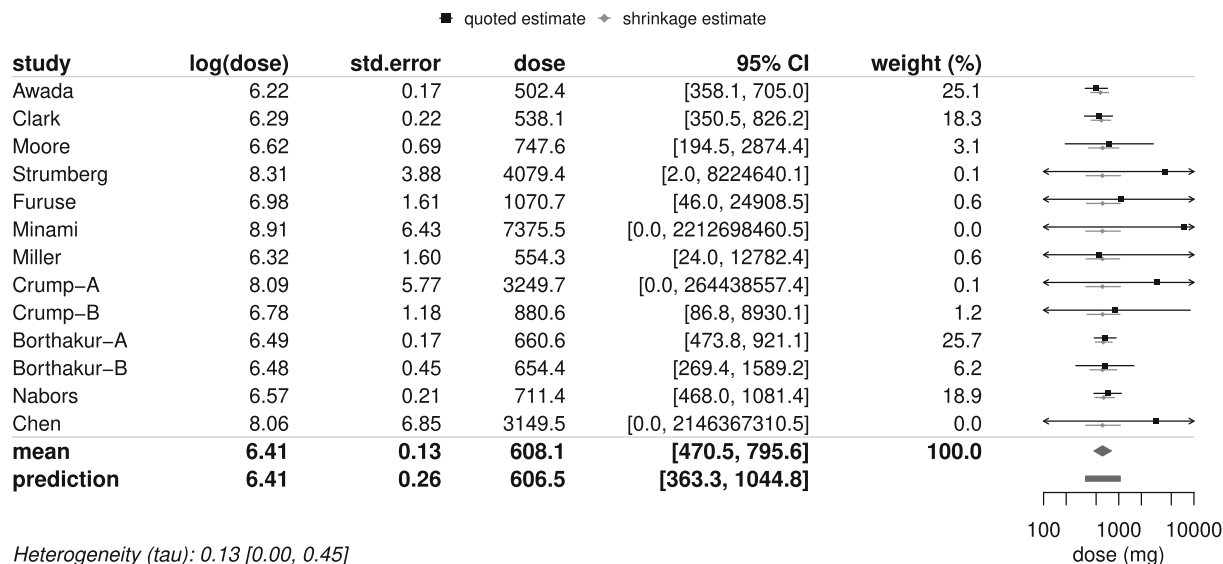| study | log(dose) | std.error | dose | 95% CI | weight (%) | |
|-------|-----------|-----------|------|--------|------------|---|
| Awada | 6.22 | 0.17 | 502.4 | [358.1, 705.0] | 25.1 | |
| Clark | 6.29 | 0.22 | 538.1 | [350.5, 826.2] | 18.3 | |
| Moore | 6.62 | 0.69 | 747.6 | [194.5, 2874.4] | 3.1 | |
| Strumberg | 8.31 | 3.88 | 4079.4 | [2.0, 8224640.1] | 0.1 | |
| Furuse | 6.98 | 1.61 | 1070.7 | [46.0, 24908.5] | 0.6 | |
| Minami | 8.91 | 6.43 | 7375.5 | [0.0, 2212698460.5] | 0.0 | |
| Miller | 6.32 | 1.60 | 554.3 | [24.0, 12782.4] | 0.6 | |
| Crump–A | 8.09 | 5.77 | 3249.7 | [0.0, 264438557.4] | 0.1 | |
| Crump–B | 6.78 | 1.18 | 880.6 | [86.8, 8930.1] | 1.2 | |
| Borthakur–A | 6.49 | 0.17 | 660.6 | [473.8, 921.1] | 25.7 | |
| Borthakur–B | 6.48 | 0.45 | 654.4 | [269.4, 1589.2] | 6.2 | |
| Nabors | 6.57 | 0.21 | 711.4 | [468.0, 1081.4] | 18.9 | |
| Chen | 8.06 | 6.85 | 3149.5 | [0.0, 2146367310.5] | 0.0 | |
| **mean** | **6.41** | **0.13** | **608.1** | **[470.5, 795.6]** | **100.0** | |
| **prediction** | **6.41** | **0.26** | **606.5** | **[363.3, 1044.8]** | | |

100   1000  10000
dose (mg)

*Heterogeneity (tau): 0.13 [0.00, 0.45]*

**FIGURE 5**   A forest plot illustrating the meta-analysis results for the Sorafenib example dataset. MTD estimates are based on penalized logistic regression (FLAC). The estimates are shown in terms of logarithmic dose as well as on their original scale. The "weight" column indicates each study's contribution to the overall estimate

the MTD. If we simply omit those studies with $\sigma_i > 1$ (on the logarithmic scale this means that lower and upper 95% interval bound are more than a factor of 50 apart), we are left with 6 studies, and the overall analysis results remain almost the same; see Figure B1 in Appendix B. We may also check back with the original data (see Tables 1 and A1) and check why these particular studies only seem to provide little evidence; these are all studies either involving only few DLT events, or studies that (empirically) exhibited a gently-sloping response, so that MTD estimates are large and at the same time associated with correspondingly large uncertainties.

## 5.1.2 | Meta-analytic approach to bridging

Suppose one was interested in quantifying the MTD in Japanese patients. Strictly speaking, only two studies are available that investigated this figure[33,34] (see also Table 2). A meta-analysis of this pair of studies is readily performed and only requires the additional specification of a proper heterogeneity prior—unlike in the case of many studies (large $k$), where an improper uniform heterogeneity prior may be considered appropriate, the case of only $k = 2$ studies requires a proper prior here.[72,73,82] For example, one may argue that a half-normal prior with scale 0.2 may be appropriate here, implying that while one anticipates some differences in study-specific MTDs, these are expected to vary around their common mean by a factor ranging mostly between $\frac{2}{3}$ and $\frac{3}{2}$.[73(Tab.3)] However, due to the low precision that the two Japanese studies provide, the meta-analysis also remains rather inconclusive with an estimated MTD of 1200 mg and a CI ranging from 56 mg up to 26 000 mg. With this, the estimate covers *all* of the potential doses (from 100 to 1000 mg) that had been considered in any of the experiments.

While evidence from the remaining European/North American studies may not be *directly* transferable to the Japanese context, it seems obvious that still these are of some relevance here.[41,42] A way to formally consider the external data for estimating the Japanese MTD in a dynamic fashion works via *shrinkage estimation*.[43] Technically, this may be implemented via several meta-analyses. Two summary estimates resulting from analyzing the groups of Japanese and Western studies separately are then combined in a third meta-analysis. At this second stage, the focus is not on an "overall mean" parameter, but rather on an updated *shrinkage* estimate (here: the Japanese MTD).[43] The assumption being implemented here is that there is heterogeneity within the Japanese and non-Japanese studies, as well as between the two groups of studies, resulting in an additional hierarchical model stage, and a total of three variance parameters.[83] The second-stage meta-analysis again requires a proper heterogeneity prior; as in the above example, we will use a half-normal prior with scale 0.2. The results are shown in Table 5. Excluding the two Japanese studies from the analysis yields an MTD estimate for the "Western" studies similar to the overall result shown in Table 5. The two Japanese studies alone on the other hand

**TABLE 5** Meta analysis aiming for an estimate of the MTD for a Japanese population

| | Log-MTD | | MTD (mg) | |
|---|---|---|---|---|
| Analysis | Mean | SD | Median | 95% CI |
| Western (mean) | 6.41 | 0.14 | 606 | [467, 794] |
| Japanese (mean) | 7.09 | 1.57 | 1199 | [56, 25 574] |
| Japanese (shrinkage) | 6.43 | 0.30 | 618 | [337, 1179] |

*Note*: Assuming heterogeneity among Western and Japanese studies, as well as between the two groups of studies, first two separate analyses combining the Japanese studies[33,34] as well as the remaining Western studies are performed. Subsequently the two resulting estimates are analyzed jointly, aiming for a shrinkage estimate of the Japanese MTD.

yield only a very vague estimate; the standard error (on the logarithmic scale) is roughly 10 times as wide as the "Western" standard error. Performing the meta-analysis at the second stage then yields a shrinkage estimate closer to the Western estimate and with a substantially reduced standard error.

The new shrinkage estimate of 617 mg and the associated interval ranging from 337 up to 1179 mg do not allow to pinpoint a single dose, yet they allow to narrow down the range of likely and plausible doses considerably based on explicit and transparent model assumptions regarding the relationships between Japanese and external data. Again, we may also determine the *weights* of the Western and Japanese studies as these contribute to the resulting shrinkage estimate;[80] the Japanese contribution here amounts to only 3.6%, once more highlighting the potential gain from considering the external data.

## 5.2 | Irinotecan/S-1 example

Although there is one study less in the second example, as well as fewer patients and fewer events involved, all studies saw an empirical DLT rate of at least 11%, and there are eventually more studies that are able to contribute substantial information on the MTD (see also Table 2). The resulting DLT estimates are shown in Table A2 in Appendix A; only two studies yield (log-) DLT estimates with standard errors $\sigma_i > 1$. One example of an ambiguous data setup is given by Yoshioka et al;[53] the data do not even clearly suggest an increasing or decreasing toxicity with increasing dose (see Table 2), and consequently the resulting MTD estimate ends up with a huge standard error. Note also the noticeable differences between MTD estimates based on simple logistic regression compared to the regularized ones.

Figure 6 shows the results of a meta-analysis of all 12 MTD estimates. The eventual MTD estimate is at 80.3 [67.4, 97.3], which in terms of dose *selection*, would include the doses of 70, 80, or 90 mg/m$^2$. To our knowledge, the combination of Irinotecan and S-1 has not yet received market authorization, but these figures are in agreement with the range determined by Ursino et al[17] for the MTD. The prediction interval, denoting the expected MTD in a new study, ranges from 47.6 up to 138.1 mg/m$^2$.

We may again also consider shrinkage estimation; for example, for the most recent study,[57] which contributed a pretty precise estimate of 85.9 [78.1, 94.6] on its own already, considering the remaining studies in addition does not add much information and the shrinkage estimate is very similar at 85.6 [77.9, 94.0]. Another interesting example is the one by Yoshioka et al,[53] where one DLT was observed in 12 patients allocated to three doses (100, 125, and 150 mg/m$^2$; see Table 2). The authors eventually determined the highest dose as the recommended one, which seems appropriate considering only the data of the trial. We note, however, that the shrinkage interval for this study, [47.6, 138.1], excludes the dose of 150 (see also Figure 6).

## 6 | DISCUSSION

The aim of this article is to propose a simple meta-analysis approach for the estimation of the MTD from multiple phase I dose-finding studies. The proposed two-stage approach is easier to implement than a one-stage approach that would require additional assumptions on the model, such as the control group response or the implementation of pooling and stratification schemes.[84]
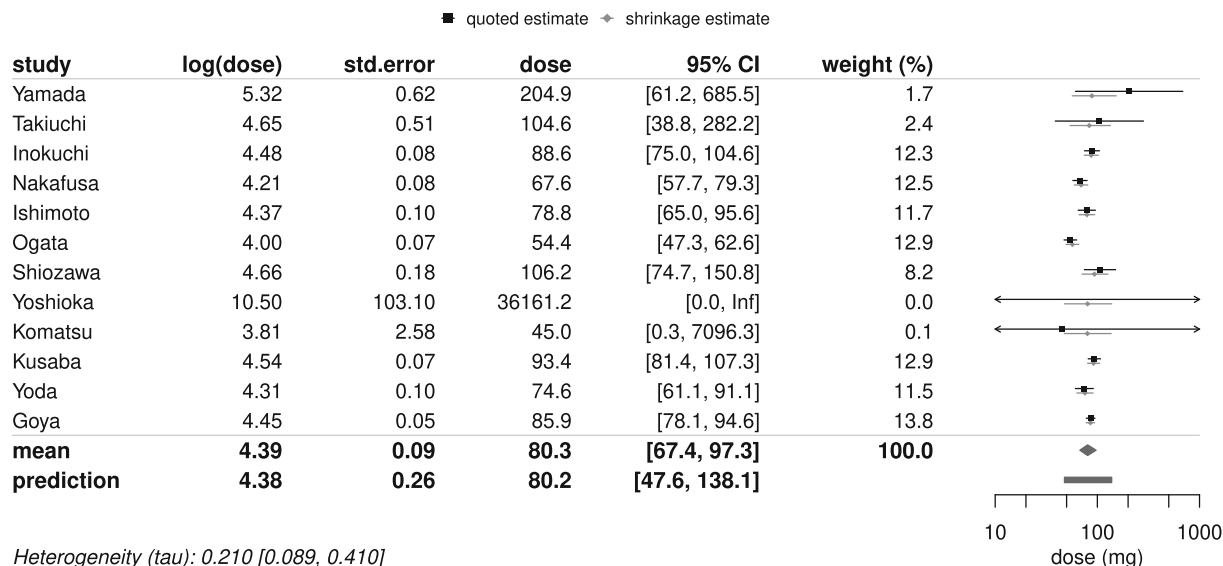
■ quoted estimate ◆ shrinkage estimate

| study | log(dose) | std.error | dose | 95% CI | weight (%) | |
|---|---|---|---|---|---|---|
| Yamada | 5.32 | 0.62 | 204.9 | [61.2, 685.5] | 1.7 | |
| Takiuchi | 4.65 | 0.51 | 104.6 | [38.8, 282.2] | 2.4 | |
| Inokuchi | 4.48 | 0.08 | 88.6 | [75.0, 104.6] | 12.3 | |
| Nakafusa | 4.21 | 0.08 | 67.6 | [57.7, 79.3] | 12.5 | |
| Ishimoto | 4.37 | 0.10 | 78.8 | [65.0, 95.6] | 11.7 | |
| Ogata | 4.00 | 0.07 | 54.4 | [47.3, 62.6] | 12.9 | |
| Shiozawa | 4.66 | 0.18 | 106.2 | [74.7, 150.8] | 8.2 | |
| Yoshioka | 10.50 | 103.10 | 36161.2 | [0.0, Inf] | 0.0 | |
| Komatsu | 3.81 | 2.58 | 45.0 | [0.3, 7096.3] | 0.1 | |
| Kusaba | 4.54 | 0.07 | 93.4 | [81.4, 107.3] | 12.9 | |
| Yoda | 4.31 | 0.10 | 74.6 | [61.1, 91.1] | 11.5 | |
| Goya | 4.45 | 0.05 | 85.9 | [78.1, 94.6] | 13.8 | |
| **mean** | **4.39** | **0.09** | **80.3** | **[67.4, 97.3]** | **100.0** | |
| **prediction** | **4.38** | **0.26** | **80.2** | **[47.6, 138.1]** | | |

10    100    1000
dose (mg)

*Heterogeneity (tau): 0.210 [0.089, 0.410]*

**FIGURE 6** A forest plot illustrating the meta-analysis for the Irinotecan/S-1 example data. MTD estimates are given on the logarithmic as well as on the dose scale

Based on our extensive simulation study, this simple approach gave on average good operating characteristics, in terms of MTD standard error (in the first stage) and MTD correct dose selection associated with DLT probability estimated at MTD (at the second stage). Only in the "gentle" scenario the proposed approach led to an underestimation of the selected MTD. Unless the therapeutic window of the tested drug is very narrow (ie, if the minimum effective dose, MED, is close to the MTD), an underestimation of the MTD could be still considered acceptable, since a nontoxic, yet efficacious dose is selected. Nevertheless, the "gentle" scenario reflects a case where an inappropriate dose range has been used in the original study. Indeed, if the toxicity probabilities associated with the investigated dose are similar, the first-stage regression (based on small samples) is likely to yield very imprecise MTD estimates, entailing problems with the delta method being applied to a nonlinear response curve. The second-stage meta-analysis then may also not improve upon this. The selection of an appropriate set of doses in phase I dose-finding trials is hence important in order to be able to derive useful MTD estimates.

We also investigated the case of bridging studies where information from one population is used to extrapolate knowledge about another population. In the Sorafenib illustration, we considered dynamic shrinkage estimation when using Western data for the estimation of the Japanese MTD. Ollier et al[85] have proposed a complementary approach to evaluate the similarity between two population distributions of a parameter of interest (MTD, dose-response model parameter(s), etc.) and thus help in deciding whether or not to use and to tailor the available information.

A limitation of this two-stage approach is related to the bias and coverage issues that seem to originate from the first-stage logistic regressions. A one-stage approach might avoid this, but would come at the cost of additional modeling assumptions and a more complicated analysis, as recalled at the beginning of this section. However, the heterogeneity parameter might 'absorb' some of the erratic small-sample-size behavior at the first stage, which might explain why prediction and shrinkage estimates are not as badly affected.

Bayesian approaches involve the specification of prior distributions, which is relevant for the meta-analysis at the second stage of our proposed approach. Here we relied on "uninformative" (uniform) priors for the overall mean and heterogeneity parameters as a default. In particular in cases involving few studies only, the use of (weakly) informative priors is recommended.[72,73] In practice, some prior information may come from earlier phases or from PK/PD considerations or preclinical data,[86] and it may be worth trying to include these in terms of informative priors.[72,73]

Another point is associated with the way dose-finding designs are constructed. In the interest of participants, early-phase dose-finding studies are designed with the aim to minimize the number of DLTs occurring, in particular when volunteers are involved. This is why these methods are often sequential, avoiding allocating patients to toxic doses and targeting the MTD. However, such schemes might result in large uncertainty in the eventual MTD estimate; from the estimation perspective, it may be more desirable to also investigate wider ranges of (not necessarily more toxic) doses in order to better probe the actual dose-response curve.[87]

Finally, our proposed simple meta-analysis method may be transferable to related estimation problems, for example, for median lethal doses (LD$_{50}$), estimation of the MED coming from phase II dose-ranging studies, etc. User-friendly R code is provided in the supplement to help clinical trials statisticians and stakeholders to implement the proposed method.

## DATA AVAILABILITY STATEMENT

The data that supports the findings of this study are available in the supplementary material of this article.

## ORCID

*Christian Röver* https://orcid.org/0000-0002-6911-698X
*Moreno Ursino* https://orcid.org/0000-0002-5709-4322
*Tim Friede* https://orcid.org/0000-0001-5347-7441
*Sarah Zohar* https://orcid.org/0000-0002-8429-2340

## REFERENCES

1. Chevret S. *Statistical Methods for Dose-Finding Experiments*. Chichester, West Sussex, England: John Wiley & Sons; 2006.
2. Daimon T, Hirakawa A, Matsui S. *Dose-Finding Designs for Early-Phase Cancer Clinical Trials*. Tokyo: Springer; 2019.
3. Le Tourneau C, Lee JJ, Siu LL. Dose escalation methods in phase I cancer clinical trials. *J Natl Cancer Inst*. 2009;101(10):708-720.
4. Neuenschwander B, Matano A, Tang Z, Roychoudhury S, Wandel S, Bailey S. A Bayesian industry approach to phase I combination trials in oncology. In: Zhao W, Yang H, eds. *Statistical Methods in Drug Combination Studies*. Boca Raton: Chapman & Hall / CRC; 2015: 95-135.
5. Reiner E, Paoletti X, O'Quigley J. Operating characteristics of the standard phase I clinical trial design. *Comput Stat Data Anal*. 1999;30(3):303-315.
6. Paoletti X, Ezzalfani M, Le Tourneau C. Statistical controversies in clinical research: requiem for the 3+3 design for phase I trials. *Ann Oncol*. 2015;26(9):1808-1812.
7. O'Quigley J, Pepe M, Fisher L. Continual reassessment method: a practical design for phase I clinical trials in cancer. *Biometrics*. 1990;46(1):33-48.
8. Neuenschwander B, Branson M, Gsponer T. Critical aspects of the Bayesian approach to phase I cancer trials. *Stat Med*. 2008;27(13):2420-2439.
9. Onar A, Kocak M, Boyett JM. Continual reassessment method vs. traditional empirically based design: modifications motivated by phase I trials in pediatric oncology by the pediatric brain tumor consortium. *J Biopharm Stat*. 2009;19(3):437-455.
10. Conaway MR, Petroni GR. The impact of early-phase trial design in the drug development process. *Clin Cancer Res*. 2019;25(2):819-827.
11. Wei L, Pan X, Fernandez S. Practical considerations for the implementation of adaptive designs for oncology phase I dose-finding trials. *Future Drug Discov*. 2019;1(2). doi:10.4155/fdd-2019-0021
12. Zhou H, Murray TA, Pan H, Yuan Y. Comparative review of novel model-assisted designs for phase I clinical trials. *Stat Med*. 2018;37(14):2208-2222.
13. Yuan Y, Lee JJ, Hilsenbeck SG. Model-assisted designs for early-phase clinical trials: simplicity meets superiority. *JCO Precis Oncol*. 2019;3:1-12.
14. Bretz F, Pinheiro JC, Branson M. Combining multiple comparisons and modeling techniques in dose-response studies. *Biometrics*. 2005;61(3):738-748.
15. Harrison RK. Phase II and phase III failures: 2013–2015. *Nat Rev Drug Discov*. 2016;15:817-818.
16. Zohar S, Katsahian S, O'Quigley J. An approach to meta-analysis of dose-finding studies. *Stat Med*. 2011;30(17):2109-2116.
17. Ursino M, Röver C, Zohar S, Friede T. Random-effects meta-analysis of phase I dose-finding studies using stochastic process priors. *Ann Appl Stat*. 2021;15(1):174-193.
18. Kim M-O, Wang X, Liu C, Dorris K, Fouladi M, Song S. Random-effects meta-analysis for systematic reviews of phase I clinical trials: rare events and missing data. *Res Synth Methods*. 2017;8(2):124-135.
19. Kosmidis I, Firth D. Jeffreys-prior penalty, finiteness and shrinkage in binomial-response generalized linear models. *Biometrika*. 2021;108(1):71-82.
20. Puhr R, Heinze G, Nold M, Lusa L, Geroldinger A. Firth's logistic regression with rare events: accurate effect estimates and predictions? *Stat Med*. 2017;36(14):2302-2317.
21. Heinze G, Ploner M, Jiricka L. logistf: Firth's bias-reduced logistic regression; 2020. R package version 1.24.
22. Röver C. bayesmeta: Bayesian random-effects meta analysis R package; 2015. http://cran.r-project.org/package=bayesmeta
23. DerSimonian R. Meta-analysis in the design and monitoring of clinical trials. *Stat Med*. 1996;15(12):1237-1248.
24. Goudie AC, Sutton AJ, Jones DR, Donald A. Empirical assessment suggests that existing evidence could be used more fully in designing randomized controlled trials. *J Clin Epidemiol*. 2010;63(9):983-991.

25. De Santis F. Using historical data for Bayesian sample size determination. *J Royal Stat Soc A*. 2007;170(1):95-113.

26. Schmidli H, Gsteiger S, Roychoudhury S, O'Hagan A, Spiegelhalter D, Neuenschwander B. Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics*. 2014;70(4):1023-1032.

27. Wandel S, Neuenschwander B, Röver C, Friede T. Using phase II data for the analysis of phase III studies: an application in rare diseases. *Clin Trials*. 2017;14(3):277-285.

28. European Medicines Agency (EMA). Nexavar (sorafenib) EPAR Summary. https://www.ema.europa.eu/en/medicines/human/EPAR/nexavar. Accessed September 25, 2021.

29. Awada A, Hendlisz A, Gil T, et al. 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*. 2005;92(10):1855.

30. Clark JW, Eder JP, Ryan D, Lathia C, Lenz H-J. 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*. 2005;11(15): 5472-5480.

31. Moore M, Hirte HW, Siu L, et al. 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*. 2005;16(10):1688-1694.

32. Strumberg D, Richly H, Hilger RA, et al. 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*. 2005;23(5): 965-972.

33. Furuse J, Ishii H, Nakachi K, Suzuki E, Shimizu S, Nakajima K. Phase I study of sorafenib in Japanese patients with hepatocellular carcinoma. *Cancer Sci*. 2008;99(1):159-165.

34. Minami H, Kawada K, Ebi H, et al. Phase I and pharmacokinetic study of sorafenib, an oral multikinase inhibitor, in Japanese patients with advanced refractory solid tumors. *Cancer Sci*. 2008;99(7):1492-1498.

35. Miller AA, Murry DJ, Owzar K, et al. Phase I and pharmacokinetic study of sorafenib in patients with hepatic or renal dysfunction: CALGB 60301. *J Clin Oncol*. 2009;27(11):1800.

36. Crump M, Hedley D, Kamel-Reid S, et al. 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. *Leuk Lymphoma*. 2010;51(2):252-260.

37. Borthakur G, Kantarjian H, Ravandi F, et al. Phase I study of sorafenib in patients with refractory or relapsed acute Leukemias. *Haematologica*. 2011;96(1):62-68.

38. Nabors LB, Supko JG, Rosenfeld M, et al. Phase I trial of sorafenib in patients with recurrent or progressive malignant glioma. *Neuro-Oncol*. 2011;13(12):1324-1330.

39. Chen Y-B, Li S, Lane AA, et al. Phase I trial of maintenance sorafenib after allogeneic hematopoietic stem cell transplantation for FMS-like tyrosine kinase 3 internal tandem duplication acute myeloid leukemia. *Biol Blood Marrow Transpl*. 2014;20(12):2042-2048.

40. Pharmaceuticals and Medical Devices Agency (PMDA). Basic principles for conducting phase I trials in the Japanese population prior to global clinical trials; 2015. https://www.pmda.go.jp/files/000157777.pdf. Accessed September 6, 2021.

41. ICH. Ethnic factors in the acceptability of foreign clinical data. ICH Harmonized Tripartite Guideline E5(R1): international Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use; 1998.

42. European Medicines Agency (EMA). Concept paper on extrapolation of efficacy and safety in medicine development. EMA/129698/2012; 2013.

43. Röver C, Friede T. Dynamically borrowing strength from another study. *Stat Methods Med Res*. 2020;29(1):293-308.

44. European Medicines Agency (EMA). Onivyde pegylated liposomal (irinotecan hydrochloride trihydrate) EPAR summary; 2021. https://www.ema.europa.eu/en/medicines/human/EPAR/onivyde-pegylated-liposomal. Accessed September 25, 2021.

45. European Medicines Agency (EMA). Teysuno (tegafur/gimeracil/oteracil) EPAR summary; 2020. https://www.ema.europa.eu/en/medicines/human/EPAR/teysuno. Accessed September 25, 2021.

46. Yamada Y, Yasui H, Goto A, et al. Phase I study of irinotecan and S-1 combination therapy in patients with metastatic gastric cancer. *Int J Clin Oncol*. 2003;8(6):374-380.

47. Takiuchi H, Narahara H, Tsujinaka T, et al. Phase I study of S-1 combined with irinotecan (CPT-11) in patients with advanced gastric cancer (OGSG 0002). *Jpn J Clin Oncol*. 2005;35(9):520-525.

48. Inokuchi M, Yamashita T, Yamada H, et al. Phase I/II study of S-1 combined with irinotecan for metastatic advanced gastric cancer. *Br J Cancer*. 2006;94(8):1130.

49. Nakafusa Y, Tanaka M, Ohtsuka T, et al. Phase I/II study of combination therapy with S-1 and CPT-11 for metastatic colorectal cancer. *Mol Med Rep*. 2008;1(6):925-930.

50. Ishimoto O, Ishida T, Honda Y, Munakata M, Sugawara S. Phase I study of daily S-1 combined with weekly irinotecan in patients with advanced non-small cell lung cancer. *Int J Clin Oncol*. 2009;14(1):43-47.

51. Ogata Y, Sasatomi T, Akagi Y, Ishibashi N, Mori S, Shirouzu K. Dosage escalation study of S-1 and irinotecan in metronomic chemotherapy against advanced colorectal cancer. *Kurume Med J*. 2009;56(1+ 2):1-7.

52. Shiozawa M, Sugano N, Tsuchida K, Morinaga S, Akaike M, Sugimasa Y. A phase I study of combination therapy with S-1 and irinotecan (CPT-11) in patients with advanced colorectal cancer. *J Cancer Res Clin Oncol*. 2009;135(3):365-370.

53. Yoshioka T, Kato S, Gamoh M, et al. Phase I/II study of sequential therapy with irinotecan and S-1 for metastatic colorectal cancer. *Br J Cancer*. 2009;101:1972-1977.

54. Komatsu Y, Yuki S, Fuse N, et al. Phase 1/2 clinical study of irinotecan and oral S-1 (IRIS) in patients with advanced gastric cancer. *Adv Ther*. 2010;27(7):483-492.

55. Kusaba H, Esaki T, Futami K, et al. Phase I/II study of a 3-week cycle of irinotecan and S-1 in patients with advanced colorectal cancer. *Cancer Sci*. 2010;101(12):2591-2595.

56. Yoda S, Soejima K, Yasuda H, et al. A phase I study of S-1 and irinotecan combination therapy in previously treated advanced non-small cell lung cancer patients. *Cancer Chemother Pharmacol*. 2011;67(3):717-722.

57. Goya H, Kuraishi H, Koyama S, et al. Phase I/II study of S-1 combined with biweekly irinotecan chemotherapy in previously treated advanced non-small cell lung cancer. *Cancer Chemother Pharmacol*. 2012;70(5):691-697.

58. Pinheiro J, Bornkamp B, Glimm E, Bretz F. Model-based dose finding under model uncertainty using general parametric models. *Stat Med*. 2014;33(10):1646-1661.

59. Albert A, Anderson JA. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*. 1984;71(1):1-10.

60. Firth D. Bias reduction of maximum likelihood estimates. *Biometrika*. 1993;80(1):27-38.

61. Heinze G, Schemper M. A solution to the problem of separation in logistic regression. *Stat Med*. 2002;21(16):2409-2419.

62. Heinze G, Puhr R. Bias-reduced and separation-proof conditional logistic regression with small or sparse data sets. *Stat Med*. 2010;29(7-8):770-777.

63. Venables WN, Ripley BD. *Modern Applied Statistics: Generalized Linear Mixed Models*. 4th ed. New York, NY: Springer-Verlag; 2002:183-210.

64. Pham-Gia T, Turkkan N, Marchand E. Density of the ratio of two normal random variables and applications. *Commun Stat Theory Methods*. 2006;35(9):1569-1591.

65. Bretthorst GL. The near-irrelevance of sampling frequency distributions. In: Linden W, ed. *Maximum Entropy and Bayesian Methods*. Dordrecht, Netherlands: Kluwer Academic Publishers; 1999:21-46.

66. Cox C. Delta method. In: Armitage P, Colton T, eds. *Encyclopedia of Biostatistics*. 2nd ed. Chichester, UK: Wiley & Sons; 2005.

67. Hedges LV, Olkin I. *Statistical Methods for Meta-Analysis*. San Diego, CA: Academic Press; 1985.

68. Hartung J, Knapp G, Sinha BK. *Statistical Meta-Analysis with Applications*. Hoboken, NJ: John Wiley & Sons; 2008.

69. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. *Introduction to Meta-Analysis*. Chichester, UK: John Wiley & Sons; 2009.

70. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Res Synth Methods*. 2010;1(2):97-111.

71. Viechtbauer W. Conducting meta-analyses in R with the metafor package. *J Stat Softw*. 2010;36(3):1-48.

72. Röver C. Bayesian random-effects meta-analysis using the Bayes meta R package. *J Stat Softw*. 2020;93(6):1-51.

73. Röver C, Bender R, Dias S, et al. On weakly informative prior distributions for the heterogeneity parameter in Bayesian random-effects meta-analysis. *Res Synth Methods*. 2021;12(4):448-474.

74. Garrett-Mayer E. The continual reassessment method for dose-finding studies: a tutorial. *Clin Trials*. 2006;3(1):57-71.

75. Sweeting M, Mander A, Sabin T. bcrm: Bayesian continual reassessment method designs for phase I dose-finding trials. *J Stat Softw*. 2013;54(13):1-26.

76. Riviere M-K, Le Tourneau C, Paoletti X, Dubois F, Zohar S. Designs of drug-combination phase I trials in oncology: a systematic review of the literature. *Ann Oncol*. 2015;26(4):669-674.

77. Brummelen EMJ, Huitema ADR, Werkhoven E, Beijnen JH, Schellens JHM. The performance of model-based versus rule-based phase I clinical trials in oncology. *J Pharmacokinet Pharmacodyn*. 2016;43:235-242.

78. Lin Y, Shih WJ. Statistical properties of the traditional algorithm-based designs for phase I cancer clinical trials. *Biostatistics*. 2001;2(2):203-215.

79. Kang S-H, Ahn CW. An investigation of the traditional algorithm-based designs for phase I cancer clinical trials. *Drug Inf J*. 2002;36(4):865-873.

80. Röver C, Friede T. Bounds for the weight of external data in shrinkage estimation. *Biom J*. 2021;65(5):1131-1143.

81. Spiegelhalter DJ, Abrams KR, Myles JP. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. Chichester, UK: John Wiley & Sons; 2004.

82. Friede T, Röver C, Wandel S, Neuenschwander B. Meta-analysis of two studies in the presence of heterogeneity with applications in rare diseases. *Biom J*. 2017;59(4):658-671.

83. Morris TP, Fisher DJ, Kenward MG, Carpenter JR. Meta-analysis of Gaussian individual patient data: two-stage or not two-stage? *Stat Med*. 2018:37(9):1419-1438.

84. Günhan BK, Weber S, Seroutou A, Friede T. Phase I dose-escalation oncology trials with sequential multiple schedules. *BMC Med Res Methodol*. 2021;21:69.

85. Ollier A, Zohar S, Morita S, Ursino M. Estimating similarity of dose-response relationships in phase I clinical trials — Case study in bridging data package. *Int J Environ Res Public Health*. 2021;18(4):1639.

86. Zheng H, Hampson LV, Wandel S. A robust Bayesian meta-analytic approach to incorporate animal data into phase I oncology trials. *Stat Methods Med Res*. 2020;29(1):94-110.

87. Dette H, Bretz F, Pepelyshev A, Pinheiro J. Optimal designs for dose-finding studies. *J Am Stat Assoc*. 2008;103(483):1225-1237.

88. Schwinghammer TL, Kroboth PD. Basic concepts in pharmacodynamic modeling. *J Clin Pharmacol*. 1988;28(5):388-394.

89. Esterni B, Mane B. UBCRM: functions to simulate and conduct dose-escalation phase I studies; 2015. R package version 1.0.1.

90. Cheung K, Duong J. dfcrm: dose-finding by the continual reassessment method; 2019. R package version 0.2-2.1.
91. Sweeting M, Wheeler G. bcrm: Bayesian continual reassessment method for phase I dose-escalation trials; 2019. R package version 0.5.4.

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

## APPENDIX A. EXAMPLE DATA

Tables A1 and A2 below characterize the datasets introduced in Section 2 in terms of the numbers of doses, patients and events, and show the slightly differing MTD estimates (on the logarithmic scale) from "plain" logistic regression, regression using Firth correction, and Firth's logistic regression with added covariate (FLAC) approaches.

**TABLE A1** Characteristics of the Sorafenib example introduced in Section 2 (see also Table 1); the numbers of doses, patients, and DLT events are given for each study, as well as the MTD estimates (on the logarithmic scale) from plain logistic regression, regression using Firth correction, and the FLAC approach

| Publication | Year | Doses | Patients | Events | Plain logistic | | Firth | | FLAC | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Estimate | SE | Estimate | SE | Estimate | SE |
| Awada et al[29] | 2005 | 6 | 37 | 10 | 6.22 | 0.15 | 6.19 | 0.17 | 6.22 | 0.17 |
| Clark et al[30] | 2005 | 5 | 19 | 5 | 6.33 | 0.15 | 6.24 | 0.22 | 6.29 | 0.22 |
| Moore et al[31] | 2005 | 4 | 24 | 4 | 6.47 | 0.46 | 6.50 | 0.67 | 6.62 | 0.69 |
| Strumberg et al[32] | 2005 | 5 | 47 | 8 | 8.01 | 3.16 | 8.53 | 5.35 | 8.31 | 3.88 |
| Furuse et al[33] | 2008 | 2 | 26 | 1 | 6.06 | 20.33 | 7.00 | 2.10 | 6.98 | 1.61 |
| Minami et al[34] | 2008 | 4 | 27 | 2 | 8.01 | 3.89 | 8.27 | 5.25 | 8.91 | 6.43 |
| Miller et al[35] | 2009 | 2 | 54 | 14 | 6.28 | 1.49 | 6.19 | 1.30 | 6.32 | 1.60 |
| Crump et al (A)[36] | 2010 | 4 | 22 | 2 | 7.21 | 3.14 | 8.59 | 10.96 | 8.09 | 5.77 |
| Crump et al (B)[36] | 2010 | 4 | 18 | 3 | 6.57 | 0.80 | 6.56 | 1.08 | 6.78 | 1.18 |
| Borthakur et al (A)[37] | 2011 | 3 | 26 | 2 | 6.40 | 2.15 | 6.44 | 0.15 | 6.49 | 0.17 |
| Borthakur et al (B)[37] | 2011 | 3 | 16 | 3 | 6.37 | 0.25 | 6.38 | 0.43 | 6.48 | 0.45 |
| Nabors et al[38] | 2011 | 5 | 20 | 5 | 6.57 | 0.17 | 6.52 | 0.21 | 6.57 | 0.21 |
| Chen et al[39] | 2014 | 2 | 19 | 1 | 6.07 | 30.28 | 3.10 | 13.91 | 8.06 | 6.85 |

**TABLE A2** Characteristics of the Irinotecan/S-1 example introduced in Section 2 (see also Table 2); the numbers of doses, patients, and DLT events are given for each study, as well as the MTD estimates (on the logarithmic scale) from plain logistic regression, regression using Firth correction, and the FLAC approach

| Publication | Year | Doses | Patients | Events | Plain logistic | | Firth | | FLAC | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Estimate | SE | Estimate | SE | Estimate | SE |
| Yamada et al[46] | 2003 | 3 | 12 | 1 | 5.02 | 2.63 | 5.28 | 0.81 | 5.32 | 0.62 |
| Takiuchi et al[47] | 2005 | 4 | 19 | 4 | 4.59 | 0.38 | 4.56 | 0.48 | 4.65 | 0.51 |
| Inokuchi et al[48] | 2006 | 4 | 51 | 12 | 4.47 | 0.07 | 4.47 | 0.07 | 4.48 | 0.08 |
| Nakafusa et al[49] | 2008 | 2 | 42 | 9 | 4.20 | 0.07 | 4.20 | 0.08 | 4.21 | 0.08 |
| Ishimoto et al[50] | 2009 | 4 | 13 | 2 | 4.38 | 1.26 | 4.33 | 0.09 | 4.37 | 0.10 |
| Ogata et al[51] | 2009 | 3 | 10 | 3 | 4.08 | 5.51 | 3.98 | 0.07 | 4.00 | 0.07 |
| Shiozawa et al[52] | 2009 | 4 | 21 | 7 | 4.67 | 0.16 | 4.64 | 0.18 | 4.66 | 0.18 |
| Yoshioka et al[53] | 2009 | 3 | 12 | 1 | 7.93 | 41.94 | -48.06 | 14805.91 | 10.50 | 103.10 |
| Komatsu et al[54] | 2010 | 3 | 21 | 2 | 4.07 | 1.56 | 3.86 | 3.02 | 3.81 | 2.58 |
| Kusaba et al[55] | 2010 | 2 | 9 | 2 | 4.59 | 5.01 | 4.52 | 0.07 | 4.54 | 0.07 |
| Yoda et al[56] | 2011 | 2 | 9 | 3 | 4.37 | 3.30 | 4.28 | 0.12 | 4.31 | 0.10 |
| Goya et al[57] | 2012 | 3 | 11 | 3 | 4.49 | 1.65 | 4.44 | 0.05 | 4.45 | 0.05 |

## APPENDIX B. SENSITIVITY ANALYSIS

Figure B1 illustrates a sensitivity analysis based on the Sorafenib data (see also Figure 5). A number of studies (those with "large" standard errors associated) contribute little to the analysis and hence have low weights associated. Omitting studies with low weights yields very similar, consistent results.
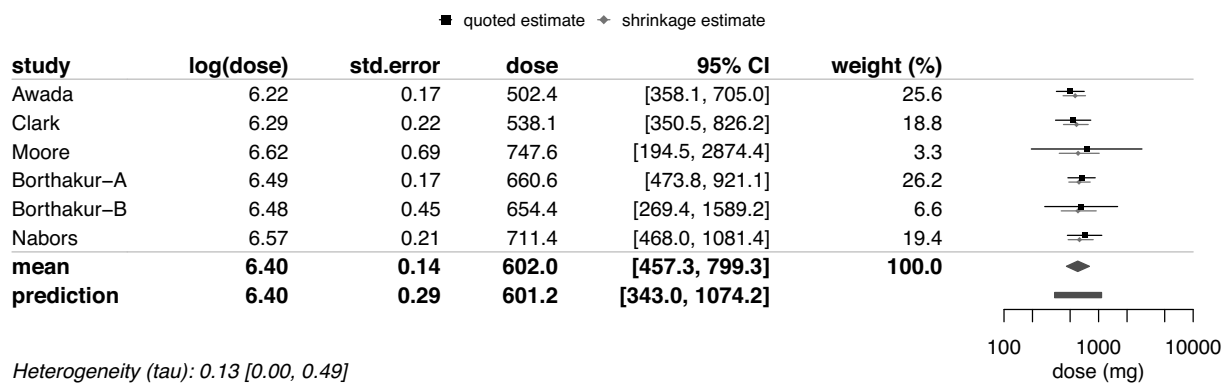


| study | log(dose) | std.error | dose | 95% CI | weight (%) |
|---|---|---|---|---|---|
| Awada | 6.22 | 0.17 | 502.4 | [358.1, 705.0] | 25.6 |
| Clark | 6.29 | 0.22 | 538.1 | [350.5, 826.2] | 18.8 |
| Moore | 6.62 | 0.69 | 747.6 | [194.5, 2874.4] | 3.3 |
| Borthakur–A | 6.49 | 0.17 | 660.6 | [473.8, 921.1] | 26.2 |
| Borthakur–B | 6.48 | 0.45 | 654.4 | [269.4, 1589.2] | 6.6 |
| Nabors | 6.57 | 0.21 | 711.4 | [468.0, 1081.4] | 19.4 |
| **mean** | **6.40** | **0.14** | **602.0** | **[457.3, 799.3]** | **100.0** |
| **prediction** | **6.40** | **0.29** | **601.2** | **[343.0, 1074.2]** | |

*Heterogeneity (tau): 0.13 [0.00, 0.49]*

**FIGURE B1** A sensitivity analysis for the Sorafenib example data. MTD estimates are based on "FLAC" regressions. Omitting studies with "large" standard errors ($\sigma_i > 1$), which do not contribute much information on the overall mean estimate yields very similar results (see also Figure 5)

## APPENDIX C. CORRESPONDENCE OF LOGISTIC AND EMAX MODELS

In case of a logarithmic dose as covariable ($x_j = \log(d_j)$, and $d_j > 0$), the logistic model from Section 3 is equivalent to the *Emax model*. The Emax model is defined as

$$p_j = \text{emax}(d_j) = E_{\max} \frac{d_j^n}{E_{50}^n + d^n}, \tag{C1}$$

where $E_{max} > 0$ denotes the maximal possible response, and $E_{50} > 0$ denotes the dose level yielding half the maximum response.[88] In case of $E_{max} = 1$, the logistic and Emax models are equivalent; the Emax model's exponent $n$ corresponds to the slope parameter $\beta_1$, and the $E_{50}$ parameter is related to the intercept $\beta_0$ as

$$\exp(-\beta_0) = E_{50}^n \quad \Leftrightarrow \quad \beta_0 = -n\log(E_{50}) \quad \Leftrightarrow \quad E_{50} = \exp\left(-\frac{\beta_0}{n}\right). \tag{C2}$$

## APPENDIX D. DATA GENERATION DETAILS

The simulation of data from common dose-finding designs (see Section 4) was facilitated by utilizing existing R libraries. The "3+3" data were generated using the UBCRM package's "sim3p3()" function.[89] The 3+3 simulation only requires specification of the (true) toxicity profile; the dose range is probed starting from the lower end, patients are recruited in cohorts of three, and decision on increasing, decreasing or proceeding with the current dose is done using fixed rules based on the number of DLTs observed in the previous cohort.[5,6]

The "CRM" data generation was based on the dfcrm package's "crmsim()" function,[90] using a cohort size of 3, the first dose as the starting dose, the skeleton (the "prior" argument in crmsim()) as shown in Figure 1 and Table 3, and a targeted DLT probability of $\pi^\star = 0.33$. The default settings include a "no-skipping rule," an "empiric" working model, as well as prior settings for the model parameter. The sample size was drawn uniformly (among multiples of 3) between 15 and 30. Patients are again recruited in small cohorts, and decision on the dose to be utilized for the next cohort is based on a parametric (logistic regression) model fitted to the past data.[7,74]

"BLRM" data utilized the bcrm package's "bcrm()" function[91] using settings as above, and in addition using an EWOC criterion of 0.25 and vague independent lognormal priors for the logistic regression parameters (intercept, slope) with means logit(0.01) and 0.0 and variances 4 and 1, respectively. The BLRM works similarly to the CRM, but is based on a Bayesian model and includes consideration of the (posterior) probability (here: 25%) of exposing patients to overly toxic doses.[4,8,75]

Table D1 illustrates how the different designs employed in the simulations probe the dose-response curve in terms of the mean numbers of patients assigned to each of the six doses (see Section 4 for more details on the scenarios). Simulations here are based on "fixed" dose-response curves (no heterogeneity).

**TABLE D1** Mean numbers of patients assigned to each dose in the different simulation scenarios, and based on the different experimental designs

| Scenario | Method | Dose | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 |
| Moderate | 3+3 | 3.15 | 3.35 | 3.79 | 3.69 | 1.82 | 0.27 |
| | CRM | 3.17 | 3.49 | 4.45 | 7.25 | 3.84 | 0.14 |
| | BLRM | 3.00 | 3.15 | 4.61 | 8.87 | 2.78 | 0.19 |
| Steep | 3+3 | 3.00 | 3.02 | 3.14 | 3.90 | 3.53 | 0.46 |
| | CRM | 3.00 | 3.03 | 3.17 | 5.23 | 7.75 | 0.28 |
| | BLRM | 3.00 | 3.00 | 3.05 | 7.48 | 5.74 | 0.33 |
| Gentle | 3+3 | 3.44 | 3.48 | 3.42 | 3.19 | 2.47 | 1.40 |
| | CRM | 3.50 | 3.78 | 4.39 | 5.69 | 4.63 | 0.65 |
| | BLRM | 3.03 | 3.22 | 4.22 | 6.19 | 4.26 | 1.83 |
| Convex | 3+3 | 3.13 | 3.21 | 3.33 | 3.89 | 3.26 | 0.43 |
| | CRM | 3.16 | 3.25 | 3.50 | 5.77 | 6.59 | 0.22 |
| | BLRM | 3.00 | 3.02 | 3.29 | 7.69 | 5.17 | 0.28 |
| Concave | 3+3 | 3.17 | 3.77 | 3.60 | 2.39 | 0.82 | 0.14 |
| | CRM | 3.18 | 4.58 | 6.96 | 6.43 | 1.33 | 0.05 |
| | BLRM | 3.05 | 4.03 | 7.54 | 6.39 | 1.33 | 0.09 |

## APPENDIX E. PREDICTION AND SHRINKAGE RESULTS

The four figures below illustrate the meta-analysis performance for prediction and shrinkage estimation. Figures E1 and E2 show the absolute offset in terms of *dose levels* as well as *DLT probability* in analogy to Figure 3. Figures E3 and E4 illustrate the credible interval lengths and their coverage probabilities analogously to Figure 4.
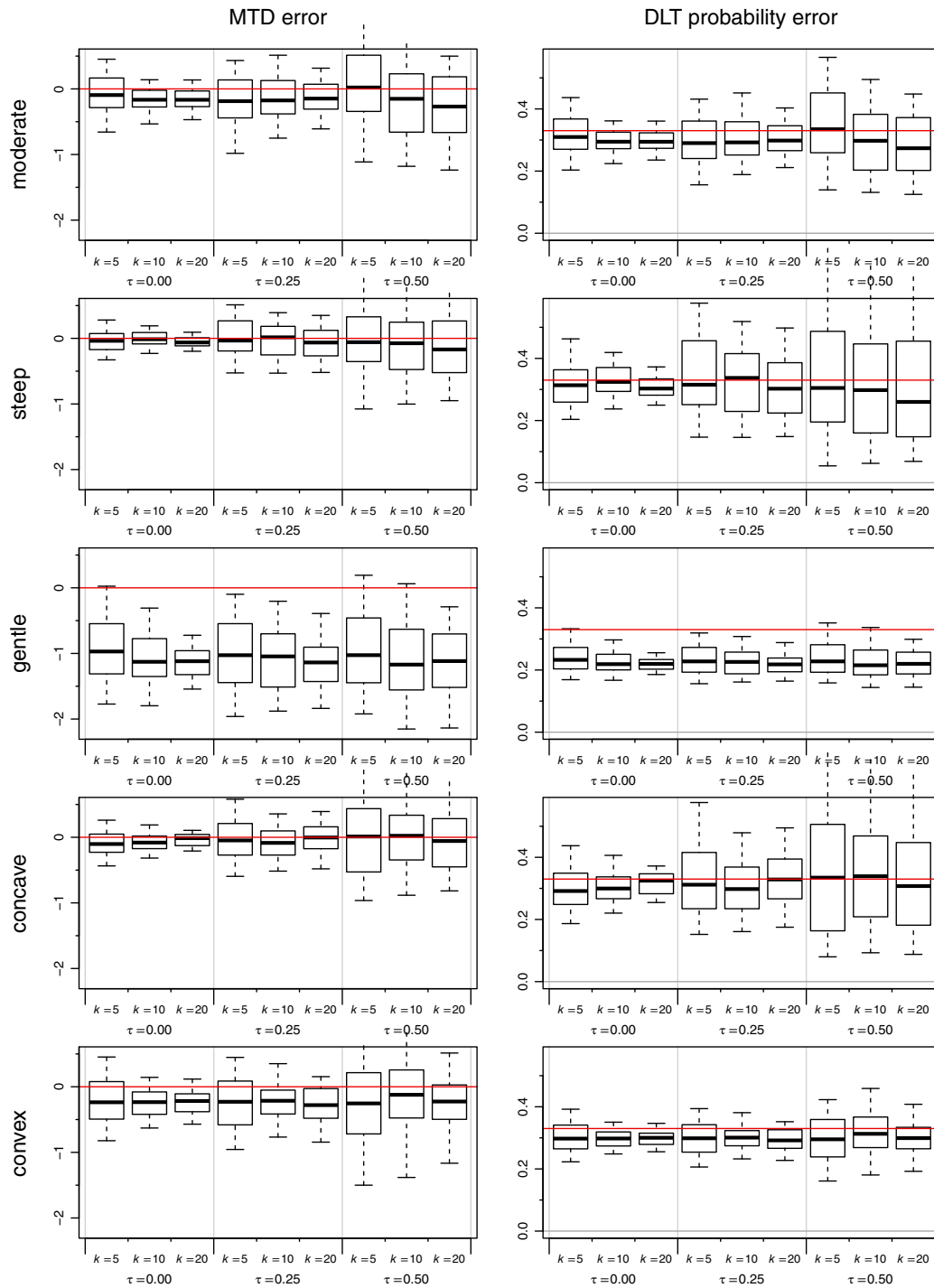


**FIGURE E1** Prediction performance of meta-analysis based on FLAC estimates in different scenarios, and for different numbers of studies and amounts of heterogeneity. The left column shows the offset in estimated MTD in terms of *doses*, and the right column shows the DLT probabilities at the estimated MTDs (see also Figure 3)
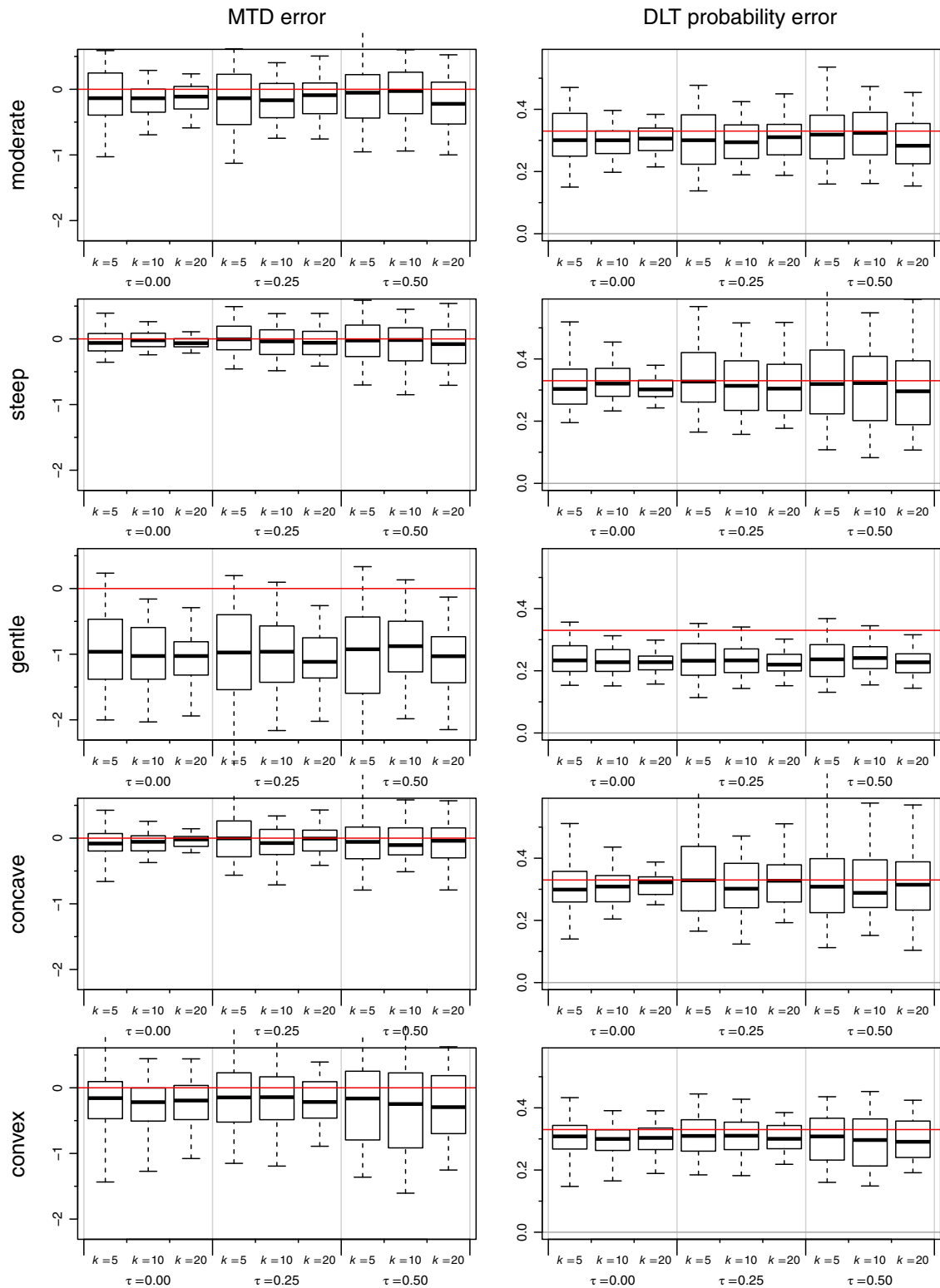
**FIGURE E2** Shrinkage performance of meta-analysis based on FLAC estimates in different scenarios, and for different numbers of studies and amounts of heterogeneity. The left column shows the offset in estimated MTD in terms of *doses*, and the right column shows the DLT probabilities at the estimated MTDs (see also Figure 3)
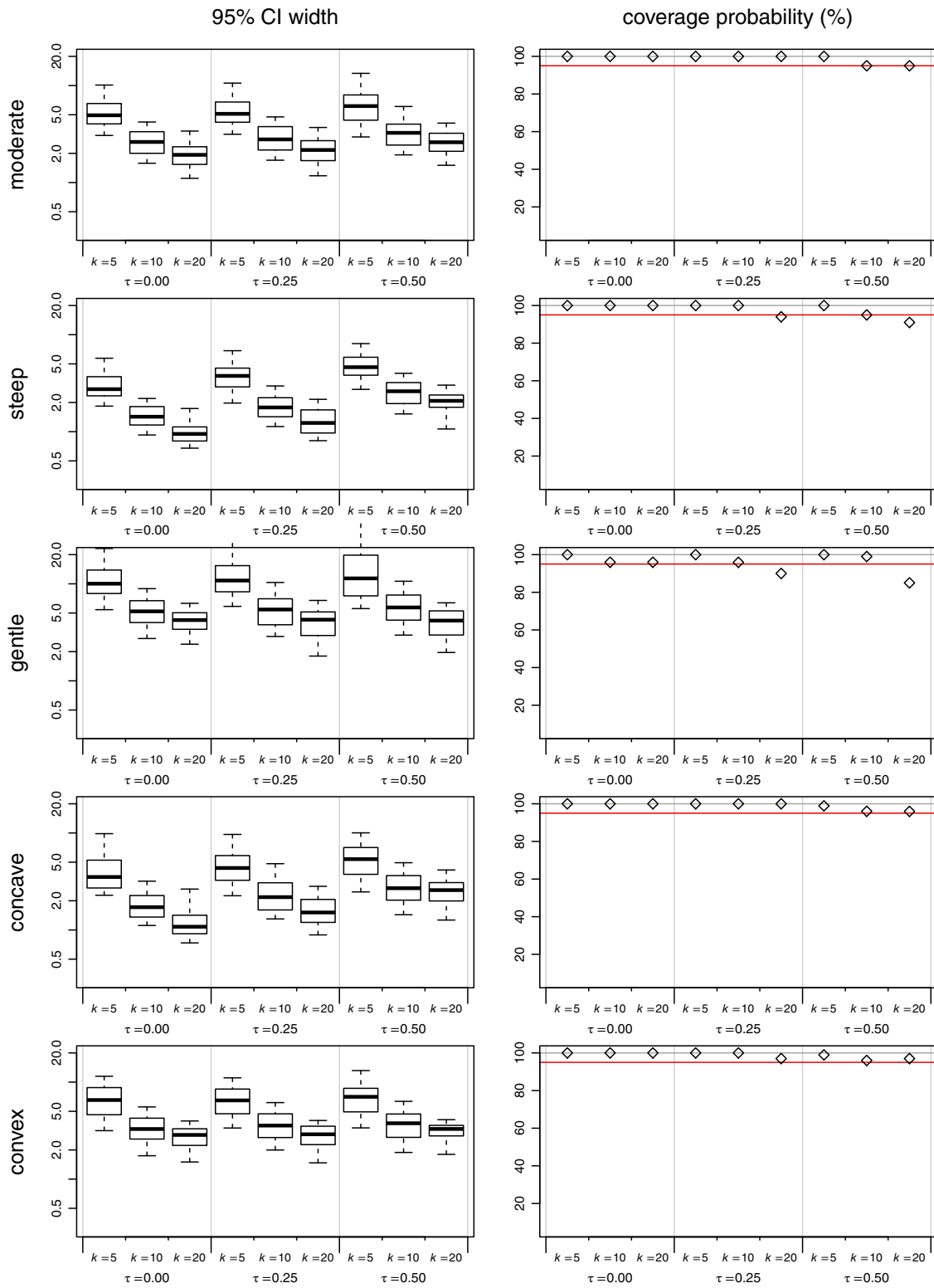
**F I G U R E   E3**   Widths and coverage probabilities of prediction intervals for the MTD based on the meta-analyses (see also Figure 4)
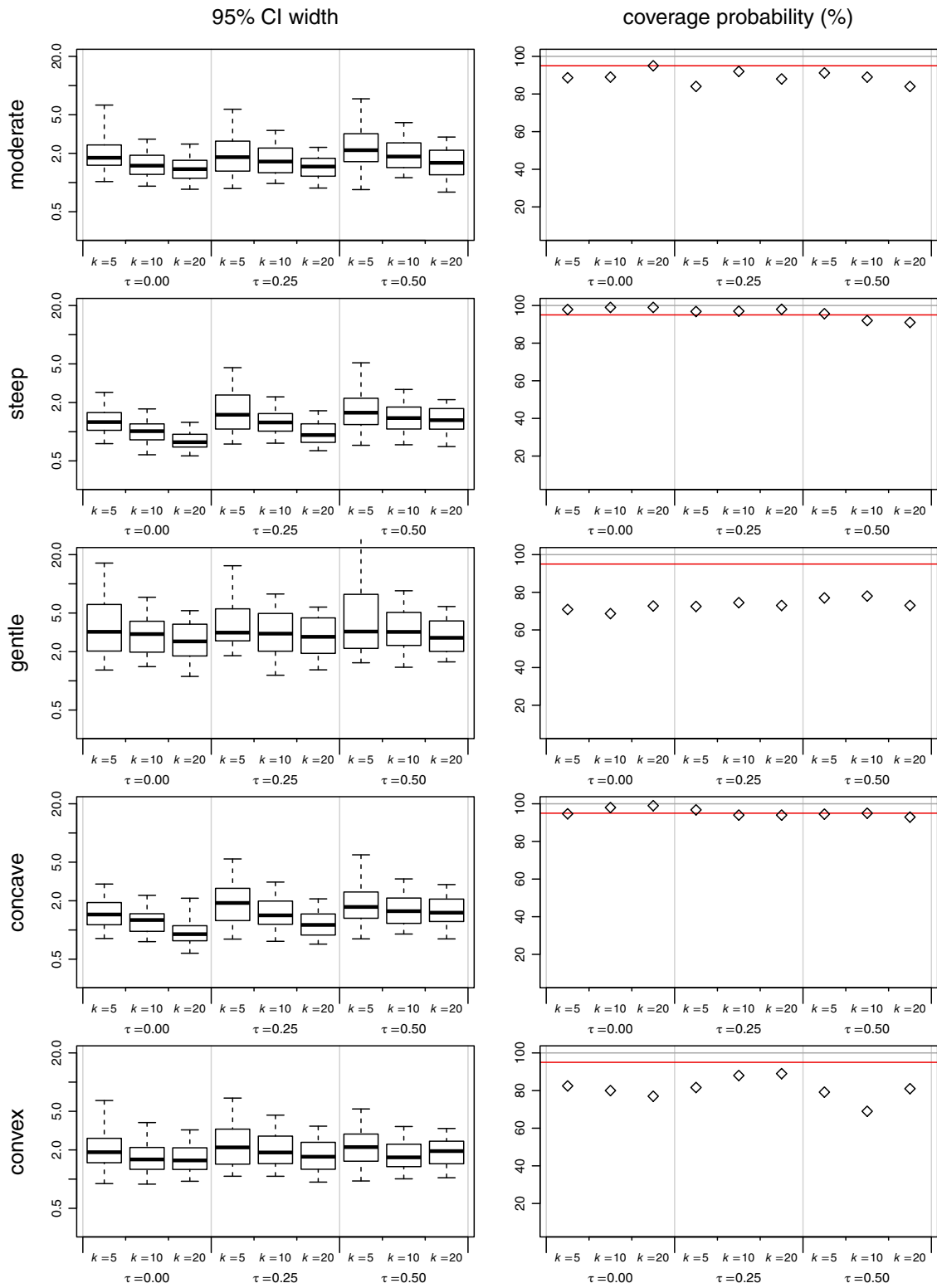
**FIGURE E4** Widths and coverage probabilities of shrinkage credible intervals for the MTD based on the meta-analyses (see also Figure 4)