Effective sample size for computing prior hyperparameters in Bayesian phase I–II dose-finding

Peter F Thall, Richard C Herrick, Hoang Q Nguyen, John J Venier and J Clift Norris

Clin Trials published online 1 September 2014
DOI: 10.1177/1740774514547397

Article

CLINICAL TRIALS

Clinical Trials
1–10

DOI: 10.1177/1740774514547397
ctj.sagepub.com
SAGE

Effective sample size for computing prior hyperparameters in Bayesian phase I–II dose-finding

Peter F Thall, Richard C Herrick, Hoang Q Nguyen, John J Venier and J Clift Norris

Abstract

Background: The efficacy–toxicity trade-off based design is a practical Bayesian phase I–II dose-finding methodology. Because the design's performance is very sensitive to prior hyperparameters and the shape of the target trade-off contour, specifying these two design elements properly is essential.

Purpose: The goals are to provide a method that uses elicited mean outcome probabilities to derive a prior that is neither overly informative nor overly disperse, and practical guidelines for specifying the target trade-off contour.

Methods: A general algorithm is presented that determines prior hyperparameters using least squares penalized by effective sample size. Guidelines for specifying the trade-off contour are provided. These methods are illustrated by a clinical trial in advanced prostate cancer. A new version of the efficacy–toxicity program is provided for implementation.

Results: Together, the algorithm and guidelines provide substantive improvements in the design's operating characteristics.

Limitations: The method requires a substantial number of elicited values and design parameters, and computer simulations are required to obtain an acceptable design.

Conclusion: The two key improvements greatly enhance the efficacy–toxicity design's practical usefulness and are straightforward to implement using the updated computer program. The algorithm for determining prior hyperparameters to ensure a specified level of informativeness is general, and may be applied to models other than that underlying the efficacy–toxicity method.

Keywords

Adaptive design, Bayesian design, clinical trial, dose-finding, phase I/II trial

Introduction

In early phase clinical trial design based on a Bayesian parametric model, the hyperparameter vector $\tilde{\boldsymbol{\theta}}$ that determines the prior $p(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}})$ on the model parameter vector $\boldsymbol{\theta}$ plays a key role. Determining $\tilde{\boldsymbol{\theta}}$ is very important because such designs must make adaptive decisions repeatedly based on small to moderate interim sample sizes. Consequently, the design's operating characteristics (OCs) are very sensitive to the numerical values of $\tilde{\boldsymbol{\theta}}$. In phase I or phase I–II dose-finding trials, the adaptive decisions include choosing doses for successive patient cohorts and deciding whether to stop the trial early, and the OCs include the selection probability and sample size for each dose under each of several dose–outcome scenarios. Discussions of this problem in particular settings, and various methods for determining $\tilde{\boldsymbol{\theta}}$, are given by Braun and Wang,[1] Cheung,[2] and Thall et al.[3] (Section 4.3).

The efficacy–toxicity (EffTox) trade-off based design[4,5] is a Bayesian sequentially adaptive procedure for determining an optimal dose in phase I–II clinical trials. This design is based on the bivariate binary outcome $\mathbf{Y} = (Y_E, Y_T)$, where Y_E indicates efficacy and Y_T indicates toxicity. Two key design elements are the prior and the target EffTox trade-off contour, $\mathcal{C}$, both of which are based on elicited information. Because the design's OCs are very sensitive to both $\tilde{\boldsymbol{\theta}}$ and $\mathcal{C}$, these must be specified carefully to obtain an acceptable

Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

Corresponding author:
Peter F Thall, Department of Biostatistics, The University of Texas MD Anderson Cancer Center, 1400 Pressler Street, Houston, TX 77030, USA.
Email: rex@mdanderson.org

design. If $p(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}})$ is either overly informative or overly disperse, it may lead to a design with undesirable properties. Denoting

$$\pi_{j,k}(\boldsymbol{\theta}) = \Pr(Y_j = 1|x_k, \boldsymbol{\theta})$$

for standardized doses $x_1 < \ldots < x_K$ and $j = E,T$, the prior $p(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}})$ induces priors on the $2K$ probabilities $\{\pi_{j,k}(\boldsymbol{\theta}),\ j = E,T,\ k = 1,\ldots,K\}$. These induced priors have substantive impacts on posterior quantities used for decision making. A practical difficulty is that it is not obvious precisely which $\tilde{\boldsymbol{\theta}}$ and $\mathcal{C}$ lead to either good or bad designs. In practice, $\tilde{\boldsymbol{\theta}}$ and $\mathcal{C}$ have been determined in a trial-and-error fashion by guessing values, simulating the trial, modifying $\tilde{\boldsymbol{\theta}}$ or $\mathcal{C}$ on that basis, and repeating this until a reasonable design is obtained. The purpose of this article is to provide a formal basis for determining $\tilde{\boldsymbol{\theta}}$ and $\mathcal{C}$ that is much less time-consuming.

We address the problem of determining $\tilde{\boldsymbol{\theta}}$ based on elicited mean outcome probabilities by exploiting the idea of prior effective sample size (ESS), introduced by Morita et al.[6,7] We present a general algorithm for computing $\tilde{\boldsymbol{\theta}}$ that uses the sum of squared differences between elicited and actual mean probabilities as an objective function of $\tilde{\boldsymbol{\theta}}$, penalized by the difference between the prior ESS and a specified target. The objective function is minimized by applying the Nelder–Mead algorithm.[8] While we focus on the EffTox model, the algorithm is quite general and, in principle, can be applied to calibrate the priors of other parametric Bayesian models used for either clinical trial design or data analysis. The general idea that one should elicit prior information about outcome probabilities, rather than about regression coefficient parameters, was established by Tsutakawa and Lin.[9] Formal methods using this idea to establish priors for generalized linear models were given by Bedrick et al.[10]

Denoting a generic pair of probabilities in the unit square $[0, 1]^2$ by $\boldsymbol{\pi} = (\pi_E,\pi_T)$, the target EffTox trade-off contour $\mathcal{C}$ of the EffTox design is obtained as a curve fit to several equally desirable $\boldsymbol{\pi}$ pairs. This is used to generate a family of contours partitioning $[0, 1]^2$, which in turn is used to define the desirability, $\delta(\boldsymbol{\pi})$, of any $\boldsymbol{\pi}$ in $[0, 1]^2$. For each successive cohort during the trial, the most desirable dose is determined by evaluating δ at the pair of posterior means $\mathrm{E}\{(\pi_{E,k}(\boldsymbol{\theta}),\pi_{T,k}(\boldsymbol{\theta}))|data)\}$ for each dose k. Recently, we discovered that certain shapes of $\mathcal{C}$ produce EffTox designs that are pathological in that they are not likely to escalate to higher doses having higher efficacy and acceptably low toxicity. We address this problem by providing simple, practical guidelines for specifying $\mathcal{C}$.

We will illustrate the improved EffTox methods with a phase I–II trial in advanced prostate cancer with goal to find the optimal dose of a novel targeted peptide to induce apoptosis in the cancer cells. Part of the rationale for the agent is that obesity is highly associated with aggressive prostate cancer, although the relationship is complex and not completely understood. Efficacy is defined as a loss of $\geq$10% in body mass index compared to baseline. Toxicity is defined as any grade 3 or higher regimen-related adverse event, or a rise in serum creatinine $\geq$ 40% from baseline. Both outcomes are evaluated within 28 days from the start of therapy. A lower limit of 0.50 on the probability of efficacy and an upper limit of 0.30 on the probability of toxicity were specified by the investigators.

The idea of representing EffTox trade-offs graphically and using them for trial design has been applied by Conaway and Petroni[11] to specify alternative hypotheses in phase II trials, Thall and Russell[12] for dose-finding with the trinary outcomes where efficacy and toxicity are disjoint, and Thall et al.[13] to compare treatments in a competing risks setting. Many phase I–II dose-finding designs have been proposed.[14–17] Reviews of dose-finding methods are given by Chevret[18] and Zohar and Chevret.[19]

The remainder of the article is organized as follows. We first review the EffTox design and prior ESS. The algorithm for establishing a prior is then presented. The methods are illustrated by application to a phase I–II clinical trial in advanced prostate cancer, including a simulation study. Practical step-by-step guidelines for constructing a design are presented, and we close with a brief discussion. To facilitate implementation, a new version of the EffTox program is available from the website https://biostatistics.mdanderson.org/SoftwareDownload

Review of the trade-off based design

Given raw doses $d_1 < d_2 < \ldots < d_K$, to stabilize numerical computations the standardized doses

$$x_k = \log(d_k) - K^{-1}\sum_{r=1}^{K}\log(d_r)$$

are used in the model. Denoting the parameter subvectors $\boldsymbol{\theta}_E = (\mu_E,\beta_{E,1},\beta_{E,2})$ and $\boldsymbol{\theta}_T = (\mu_T,\beta_{T,1})$, the linear terms

$$\eta_{E,k}(\boldsymbol{\theta}_E) = \mu_E + \beta_{E,1}x_k + \beta_{E,2}x_k^2$$

and

$$\eta_{T,k}(\boldsymbol{\theta}_T) = \mu_T + \beta_{T,1}x_k$$

determine the marginal probabilities as

$$\pi_{j,k}(\boldsymbol{\theta}_j) = \mathrm{logit}^{-1}\{\eta_{j,k}(\boldsymbol{\theta}_j)\} \quad \text{for} j = E, T$$

We require $\beta_{T,1} > 0$ to ensure that $\eta_{T,k}(\boldsymbol{\theta}_T)$ increases with dose, but the form of $\eta_{E,k}(\boldsymbol{\theta}_E)$ allows $\pi_{E,k}(\boldsymbol{\theta}_E)$ to be non-monotone in dose. The entire model parameter

vector is $\boldsymbol{\theta} = (\boldsymbol{\theta}_T, \boldsymbol{\theta}_E, \psi)$, where ψ is a real-valued association parameter. Denote the joint probabilities

$$\pi(a, b|x_k, \boldsymbol{\theta}) = \Pr(Y_E = a, Y_T = b|x_k, \boldsymbol{\theta})$$

for $a, b = 0$ or 1. Suppressing x_k and $\boldsymbol{\theta}$ for brevity, the joint distribution is

$$\pi(a, b) = \pi_E^a(1 - \pi_E)^{1-a}\pi_T^b(1 - \pi_T)^{1-b} + (-1)^{a+b}\left(\frac{e^{\psi} - 1}{e^{\psi} + 1}\right)\pi_E(1 - \pi_E)\pi_T(1 - \pi_T)$$

Denote the observed data after n patients by

$$\mathcal{D}_n = \{(\mathbf{Y}_i, x_{[i]}), \quad i = 1, \ldots, n\}$$

where $\mathbf{Y}_i = (Y_{i,E}, Y_{i,T})$ and $x_{[i]}$ is the dose given to patient i. The likelihood is

$$\mathcal{L}(\mathcal{D}_n|\boldsymbol{\theta}) = \prod_{i=1}^{n} \pi(\mathbf{Y}_i|x_{[i]}, \boldsymbol{\theta})$$

and the posterior is

$$p(\boldsymbol{\theta}|\mathcal{D}_n, \tilde{\boldsymbol{\theta}}) \propto \mathcal{L}(\mathcal{D}_n|\boldsymbol{\theta}) \times p(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}}).$$

Our notation makes explicit the fact that all posterior probabilities depend on both the data $\mathcal{D}_n$ and the fixed prior hyperparameters $\tilde{\boldsymbol{\theta}}$.

A fixed lower limit π_E^* on $\pi_E(x, \boldsymbol{\theta}_E)$ and a fixed upper limit π_T^* on $\pi_T(x, \boldsymbol{\theta}_T)$ must be elicited from the physician. A dose x_k *is acceptable* if

$$\Pr\{\pi_{E,k}(\boldsymbol{\theta}_E) > \pi_E^*|\mathcal{D}_n, \tilde{\boldsymbol{\theta}}\} \geq p_{E,L}$$

and

$$\Pr\{\pi_{T,k}(\boldsymbol{\theta}_T) < \pi_T^*|\mathcal{D}_n, \tilde{\boldsymbol{\theta}}\} \geq p_{T,L} \quad (1)$$

where $p_{E,L}$ and $p_{T,L}$ are fixed cutoffs, usually in the range 0.05–0.20, with $p_{E,L} = p_{T,L} = 0.10$ working well in most settings. The criteria in expression (1) say that, given the current data, it is not unlikely that x_k is efficacious and not unlikely that x_k has acceptably low toxicity. Since these criteria often seem non-intuitive with the cutoff 0.10, they may be given in the equivalent form that x_k *is unacceptable* if either

$$\Pr\{\pi_{E,k}(\boldsymbol{\theta}_E) < \pi_E^*|\mathcal{D}_n, \tilde{\boldsymbol{\theta}}\} > 0.90$$

or

$$\Pr\{\pi_{T,k}(\boldsymbol{\theta}_T) > \pi_T^*|\mathcal{D}_n, \tilde{\boldsymbol{\theta}}\} > 0.90$$

that is, x_k is unacceptable if it is likely to be either inefficacious or too toxic. Only acceptable doses are given, and if all doses are unacceptable, the trial is stopped with no dose selected.

As implemented in the EffTox program, the target contour $\mathcal{C}$ and corresponding family of contours partitioning $[0, 1]^2$ are defined as follows, although alternative constructions may be used.[2] Denote three equally desirable outcome probability pairs $\boldsymbol{\pi}_1^* = (\pi_{1,E}^*, 0)$, $\boldsymbol{\pi}_2^* = (1, \pi_{2,T}^*)$, and $\boldsymbol{\pi}_3^* = (\pi_{3,E}^*, \pi_{3,T}^*)$, with $\pi_{1,E}^* < \pi_{3,E}^*$ and $\pi_{3,T}^* < \pi_{2,T}^*$. While initially these values are elicited from the physicians, because in practice the practical consequences of a particular $\mathcal{C}$ often are not obvious, $\mathcal{C}$ must be calibrated to obtain a design with good OCs. In the section "Practical guidelines for trial design," we will explain how to do this quickly and effectively. The *desirability function* of $(\pi_E, \pi_T) = \boldsymbol{\pi} \in [0, 1]^2$ is defined to be

$$\delta(\pi_E, \pi_T) = 1 - \|(\pi_E, \pi_T) - (1, 0)\|_p = 1 - \left\{\left(\frac{\pi_E - 1}{\pi_{1,E}^* - 1}\right)^p + \left(\frac{\pi_T - 0}{\pi_{2,T}^* - 0}\right)^p\right\}^{1/p} \quad (2)$$

where $p > 0$. We solve $\delta(\pi_{E,3}^*, \pi_{T,3}^*) = 0$ for p using the bisection method, wherein intervals known to bracket the solution are successively refined. This gives $\delta(\boldsymbol{\pi}) = 0$ on $\mathcal{C}$ with $\delta(\boldsymbol{\pi})$ increasing as $\boldsymbol{\pi}$ moves along any straight line from a point in $[0, 1]^2$ to the ideal pair $(\pi_E, \pi_T) = (1,0)$. To use this structure for dose selection, denote the marginal hyperparameter vectors by $\tilde{\boldsymbol{\theta}}_E$ and $\tilde{\boldsymbol{\theta}}_T$ with

$$\mu_{j,k,n}(\tilde{\boldsymbol{\theta}}_j) = E\{\pi_j(x_k, \boldsymbol{\theta}_j)|\mathcal{D}_n, \tilde{\boldsymbol{\theta}}_j\}$$

for $j = E, T$ and $k = 1, \ldots, m$. The dose with largest desirability $\delta(\mu_{E,k,n}(\tilde{\boldsymbol{\theta}}_E), \mu_{T,k,n}(\tilde{\boldsymbol{\theta}}_T))$ among the acceptable doses is chosen.

In the next section, we describe a method for establishing the prior $p(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}})$ from $2K$ elicited mean outcome probabilities $\boldsymbol{\mu}_j^{(e)} = (\mu_{j,1}^{(e)}, \ldots, \mu_{j,K}^{(e)}) \quad j = E, T$.

Using ESS for penalized prior computation

As noted above, the following algorithm for establishing a prior that incorporates elicited prior means and has a given level of informativeness is quite general. In the following presentation, we present the algorithm in the context of the model underlying the EffTox method to make things concrete and also because this application is a central focus of the article.

Prior parameterization

Our method for determining a prior on $\boldsymbol{\theta}$ is to first make the practical simplifying assumptions that $\psi \sim N(0,1)$ and $\beta_{E,2} \sim N(0,0.20)$ and then use prior ESS to calibrate the priors of the remaining parameters. We assume $\mu_j \sim N(\tilde{\mu}_{j,\mu}, \tilde{\sigma}_{j,\mu}^2)$ and $\beta_{j,1} \sim N(\tilde{\mu}_{j,\beta}, \tilde{\sigma}_{j,\beta}^2)$ for

$j = E,T$, since we have found that these hyperparameter values and assumptions give priors that work well in all settings. Determining the priors of $(\mu_E,\beta_{E,1},\mu_T,\beta_{T,1})$ requires specifying numerical values of the four hyper-means and four hyper-standard deviations, denoted respectively by

$$\tilde{\boldsymbol{\mu}} = (\tilde{\mu}_{E,\mu}, \tilde{\mu}_{T,\mu}, \tilde{\mu}_{E,\beta}, \tilde{\mu}_{T,\beta})$$

and

$$\tilde{\boldsymbol{\sigma}} = (\tilde{\sigma}_{E,\mu}, \tilde{\sigma}_{E,\beta}, \tilde{\sigma}_{T,\mu}, \tilde{\sigma}_{T,\beta}).$$

We also will write these hyperparameters as vectors rearranged in terms of the outcomes

$$\tilde{\boldsymbol{\theta}}_j = (\tilde{\mu}_{j,\mu}, \tilde{\mu}_{j,\beta}, \tilde{\sigma}_{j,\mu}, \tilde{\sigma}_{j,\beta}) \quad \text{for} \quad j = E, T.$$

A key point is that, given $\tilde{\boldsymbol{\mu}}$, the hyper-standard deviations $\tilde{\boldsymbol{\sigma}}$ determine the informativeness of the prior, but the precise degree of informativeness cannot be known from their numerical values per se. For example, while $\tilde{\sigma}_{j,\mu} = \tilde{\sigma}_{j,\beta} = 1$ gives a more informative prior than $\tilde{\sigma}_{j,\mu} = \tilde{\sigma}_{j,\beta} = 10$ for $j = E,T$, it is not clear precisely how informative either prior may be, or whether, for given $\tilde{\boldsymbol{\mu}}$, it will yield a design with good properties. The algorithm given below provides a practical way to deal with this issue.

Prior ESS

Intuitively, the ESS of a prior $p(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}})$ is the number of observations that, starting with a vague version of the prior, gives a posterior very close to $p(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}})$. The definition of ESS for regression models given by Morita, Thall, and Müller (MTM)[6,7] may be summarized as follows. Let $f(Y|\boldsymbol{\theta},X)$ denote a parametric probability distribution for a random vector Y in a regression model with covariates X and parameter vector $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_d)$. MTM first define an "$\varepsilon$-information" prior $q_0(\boldsymbol{\theta})$ that is vague in a suitable sense, and then define the ESS to be the sample size m of outcomes $\mathbf{Y}_m = (Y_1,\ldots,Y_m)$ that, starting with $q_0(\boldsymbol{\theta})$, yields a posterior $q_m(\boldsymbol{\theta}|\mathbf{Y}_m)$ very close to $p(\boldsymbol{\theta})$. Formally, for sample $\mathbf{Y}_m = (Y_1,\ldots,Y_m)$ with covariates $\mathbf{X}_m = (X_1,\ldots,X_m)$, the likelihood is

$$f_m(\mathbf{Y}_m|\boldsymbol{\theta},\mathbf{X}_m) = \prod_{i=1}^{m} f(Y_i|\boldsymbol{\theta},X_i).$$

Let $g_m(\mathbf{X}_m|\boldsymbol{\xi})$ be a sampling model for the covariates with prior $r(\boldsymbol{\xi}\,|\,\tilde{\boldsymbol{\xi}})$. The ε-information prior, $q_0(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}}_0)$, is defined by requiring matching means, $\mathrm{E}_{q0}(\boldsymbol{\theta}) = \mathrm{E}_p(\boldsymbol{\theta})$, and correlations, $\mathrm{Corr}_{q0}(\theta_j,\theta_{j'}) = \mathrm{Corr}_p(\theta_j,\theta_{j'})$, $j \neq j'$, but inflated variances of the elements of $\boldsymbol{\theta}$. Given a sample $\mathbf{Y}_m$, the posterior is

$$q_m(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}}_0, \mathbf{Y}_m, \mathbf{X}_m) \propto q_0(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}}_0) f_m(\mathbf{Y}_m|\mathbf{X}_m,\boldsymbol{\theta}).$$

MTM define the prior-posterior distance as the difference between the traces or determinants of the information matrix of $p(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}})$ and the mean information matrix of $q_m(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}}_0, \mathbf{Y}_m, \mathbf{X}_m)$, with the mean computed with respect to the prior predictive distribution $f_m(\mathbf{Y}_m|\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\xi}})$. The ESS is defined as the interpolated value of m minimizing the prior-posterior distance. For a subvector $\boldsymbol{\theta}_r$ of $\boldsymbol{\theta}$, the ESS can be determined similarly in terms of the marginal prior $p(\boldsymbol{\theta}_r|\tilde{\boldsymbol{\theta}})$. In most settings, the expectation cannot be obtained analytically, and a simulation-based numerical approximation is used.

To speed up computation, we will rely on the following method for computing approximate ESS values that relies on the well-known fact that a *beta*(*a*,*b*) distribution has $ESS = a + b$. This is done by approximating the prior of any probability $\pi(\boldsymbol{\theta})$ by a *beta*(*a*,*b*), matching the means and variances, and solving the equations

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

and

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

which gives $a + b$ as an approximate ESS of the prior of $\pi(\boldsymbol{\theta})$ that is induced by $p(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}})$.

We apply this to the EffTox model to obtain approximate ESS values of the marginal priors of $(\mu_E,\beta_{E,1})$ and $(\mu_T,\beta_{T,1})$, since we have specified the priors of $\beta_{E,2}$ and ψ. To do this, we compute the beta-approximated value $ESS_{j,k}$ of the prior of $\pi_{j,k}(\mu_j,\beta_{j,1}|\tilde{\boldsymbol{\theta}}_j)$ for each j and dose k. This gives $2K$ ESS values, and we use the mean

$$\overline{ESS}_j(\tilde{\boldsymbol{\theta}}_j) = K^{-1} \sum_{k=1}^{K} ESS_{j,k}(\tilde{\boldsymbol{\theta}}_j)$$

over the K doses as an approximate ESS of the marginal prior $p(\mu_j,\beta_{j,1}|\tilde{\boldsymbol{\theta}}_j)$ for each $j = E,T$. The mean $\{\overline{ESS}_E(\tilde{\boldsymbol{\theta}}_E) + \overline{ESS}_T(\tilde{\boldsymbol{\theta}}_T)\}/2$ may be considered an average marginal ESS. However, it is not necessarily the case that the ESS of $p(\boldsymbol{\theta}_E,\boldsymbol{\theta}_T|\tilde{\boldsymbol{\theta}}_E,\tilde{\boldsymbol{\theta}}_T)$ equals $\overline{ESS}_E(\tilde{\boldsymbol{\theta}}_E) + \overline{ESS}_T(\tilde{\boldsymbol{\theta}}_T)$, since in general, ESS is not additive over marginals. Moreover, the ESS of the prior on the entire vector $\boldsymbol{\theta}$ has not been determined. We avoid this more complex derivation to speed up computation and facilitate practical application, since we have found that only $\overline{ESS}_E(\tilde{\boldsymbol{\theta}}_E)$ and $\overline{ESS}_T(\tilde{\boldsymbol{\theta}}_T)$ are needed to calibrate $\tilde{\boldsymbol{\sigma}}$.

Computational algorithm

We will determine $(\tilde{\boldsymbol{\mu}},\tilde{\boldsymbol{\sigma}})$ from the elicited mean outcome probabilities $(\boldsymbol{\mu}_E^{(e)},\boldsymbol{\mu}_T^{(e)})$ by exploiting two key ideas. The first is the fact that the prior means of the $\pi_{j,k}(\boldsymbol{\theta}_j)$s are functions of $(\tilde{\boldsymbol{\mu}},\tilde{\boldsymbol{\sigma}})$. The second is that ESS is an intuitively straightforward criterion for calibrating

the numerical values of $\tilde{\boldsymbol{\sigma}}$. An important technical point is that even for $K = 3$ dose levels, there are $2K = 6$ elicited means, a number larger than the $\dim(\tilde{\mu}) = 4$, even after setting $\tilde{\mu}_{\beta_{E,2}} = 0$, so any set of equations in $\boldsymbol{\mu}^{(e)}$ and $\tilde{\boldsymbol{\mu}}$ would be over-determined. While a simple approach would be to use only 4 of the $2K$ elicited values, this would require discarding elicited information.

An alternative approach to eliciting the prior means is to elicit probability intervals for the $\pi_{j,k}(\boldsymbol{\theta})$s, such as prior coverage intervals of the form $\Pr\{L < \pi_{j,k}(\boldsymbol{\theta}) < U\} = 0.90$, in order to obtain information about $\tilde{\boldsymbol{\sigma}}$. However, this approach might easily lead to overly informative priors that dominate the data, which could have disastrous consequences during trial conduct if the physician's prior optimism about toxicity turns out to be wrong. The consequences of an overly informative prior will be illustrated below.

For each $j = E,T$ and $k = 1,\ldots,K$, the prior mean outcome probability is

$$\mu_{j,k}(\tilde{\boldsymbol{\theta}}_j) = E\{\pi_{j,k}(\boldsymbol{\theta}_j)|\tilde{\boldsymbol{\theta}}_j\} = \int_{\boldsymbol{\theta}_j} \pi_{j,k}(\boldsymbol{\theta}_j) p(\boldsymbol{\theta}_j|\tilde{\boldsymbol{\theta}}_j) d\boldsymbol{\theta}_j.$$

Recall that the corresponding elicited mean is denoted by $\mu^{(e)}_{j,k}$. Let $ESS_{target,j}$ be a targeted prior ESS for $p(\boldsymbol{\theta}_j|\tilde{\boldsymbol{\theta}}_j)$. In practice, values between 0.3 and 1.5 usually work well, with $ESS_{target,E} = ESS_{target,T}$. An empirical justification for this range is given in the context of the illustrative application, below. For $j = E,T$ our objective functions are

$$\begin{aligned}\phi_j(\tilde{\boldsymbol{\theta}}_j) = &\sum_{k=1}^{K} \{\mu^{(e)}_{j,k} - \mu_{j,k}(\tilde{\boldsymbol{\theta}}_j)\}^2 \\ &+ 0.1\{\overline{ESS}_j(\tilde{\boldsymbol{\theta}}_j) - ESS_{target,j}\}^2 + 0.02\,(\tilde{\sigma}_{j,\mu} - \tilde{\sigma}_{j,\beta})^2. \end{aligned} \quad (3)$$

The $\tilde{\boldsymbol{\theta}}_j$ that minimizes (3) can be found quickly and easily using the Nelder–Mead algorithm.[8] To start the algorithm, we obtain initial $\tilde{\boldsymbol{\mu}}$ values by solving the linear equations

$$\text{logit}\{\mu^{(e)}_{j,k}\} = \tilde{\mu}_{j,\mu} + \tilde{\mu}_{j,\beta} x_k, \quad j = E,T$$

using the lowest and highest doses, x_1 and x_K. Applying the delta method, we determine the initial $\tilde{\boldsymbol{\sigma}}$ by solving the equations

$$\begin{aligned}\text{var}\{\pi_{j,k}(\boldsymbol{\theta}_j|\tilde{\boldsymbol{\theta}}_j)\} = &\left[\left\{\text{logit}^{-1}\left(\tilde{\mu}_{j,\mu} + \tilde{\mu}_{j,\beta} x_k\right)\right\}'\right]^2 \\ &\times \left(\tilde{\sigma}^2_{j,\mu} + \tilde{\sigma}^2_{j,\beta} x_k^2\right), \quad j = E,T.\end{aligned}$$

Applying the previously noted relationship between the ESS, $a + b$, and the mean and variance of a $beta(a,b)$, we compute the left-hand side of this equation as $\text{var}\{\pi_{j,k}(\boldsymbol{\theta}_j|\tilde{\boldsymbol{\theta}}_j)\} = \mu^{(e)}_{j,k}(1 - \mu^{(e)}_{j,k})/ESS_{target,j} + 1)$. We do this evaluation for x_1 and x_K and take the geometric mean. We obtain the right-hand side of the equation by doing the evaluation at x_1 and x_K, and taking the arithmetic mean. Finally, we equate the left- and right-hand sides and solve for $\tilde{\boldsymbol{\sigma}}$.

The two penalty terms are included essentially to stabilize the Nelder–Mead algorithm. The penalty term $\{\overline{ESS}_j(\tilde{\boldsymbol{\theta}}_j) - ESS_{target,j}\}^2$ in the objective function (3) ensures that the marginal of $(\mu_j,\beta_{j,1})$ has the desired degree of informativeness in terms of the specified $ESS_{target,j}$. The penalty term $(\tilde{\sigma}_{j,\mu} - \tilde{\sigma}_{j,\beta})^2$ is included to ensure that the prior standard deviations $\tilde{\sigma}_{j,\mu}$ and $\tilde{\sigma}_{j,\beta}$ do not have different orders of magnitude, which can destabilize the Nelder–Mead algorithm and cause it to fail to converge. As the Nelder–Mead algorithm approaches its minimum, both penalty terms converge to 0. The tuning parameters in (3) were determined empirically. The value 0.1 that multiplies the penalty term $\{\overline{ESS}_j(\tilde{\boldsymbol{\theta}}_j) - ESS_{target,j}\}^2$ was determined by examining a range of values with the two aims (1) to obtain results that are robust to the numerical values of elicited means and starting values of the hyperparameters $\tilde{\boldsymbol{\theta}}_j$ and (2) to reduce the number of iterations to obtain convergence. Similarly, if, for example, the value 0.02 that multiplies the penalty term $(\tilde{\sigma}_{j,\mu} - \tilde{\sigma}_{j,\beta})^2$ is replaced by 1, in many cases, the Nelder–Mead algorithm takes longer to converge. The small value 0.02 works well in practice because when $\phi_j(\tilde{\boldsymbol{\theta}}_j)$ is far from its minimum, this value avoids problems caused by the term $(\tilde{\sigma}_{j,\mu} - \tilde{\sigma}_{j,\beta})^2$ dominating $\phi_j(\tilde{\boldsymbol{\theta}}_j)$, whereas if $\phi_j(\tilde{\boldsymbol{\theta}}_j)$ is near its minimum, $(\tilde{\sigma}_{j,\mu} - \tilde{\sigma}_{j,\beta})^2$ converges to 0 and thus its tuning parameter no longer matters.

Illustrative application

The doses of the targeted peptide in the advanced prostate cancer trial are $d = 1, 2, 4, 6.6, 10$ mcL/kg. The elicited prior marginal mean outcome probabilities for these dose levels were $\boldsymbol{\mu}^{(e)}_E = (0.20, 0.40, 0.60, 0.80, 0.90)$ and $\boldsymbol{\mu}^{(e)}_T = (0.02, 0.04, 0.06, 0.08, 0.10)$. The maximum sample size was $N = 39$ with cohorts of size 3, the first cohort was treated at $d = 1$, the upper limit on $\pi_{T,k}(\boldsymbol{\theta})$ was $\pi^*_T = 0.30$, the lower limit on $\pi_{E,k}(\boldsymbol{\theta})$ was $\pi^*_E = 0.50$, and the acceptability criterion lower probability cutoffs were $p_{E,L} = p_{T,L} = 0.10$. These design elements were chosen by the investigators who organized the trial.

The values $ESS_{target,E} = ESS_{target,T} = 0.9$ were used to solve for the prior hyperparameters, which were $(\tilde{\mu}_{E,\mu}, \tilde{\mu}_{E,\beta_1}, \tilde{\mu}_{T,\mu}, \tilde{\mu}_{T,\beta_1}) = (0.74, 3.42, -7.96, 1.55)$ and $(\tilde{\sigma}_{E,\mu}, \tilde{\sigma}_{E,\beta}, \tilde{\sigma}_{T,\mu}, \tilde{\sigma}_{T,\beta}) = (2.54, 2.44, 3.55, 3.50)$. Note that while these particular numerical hyperparameter values have no intuitive meaning, it is easy to understand the prior ESS 0.9 and elicited prior mean outcome probabilities.

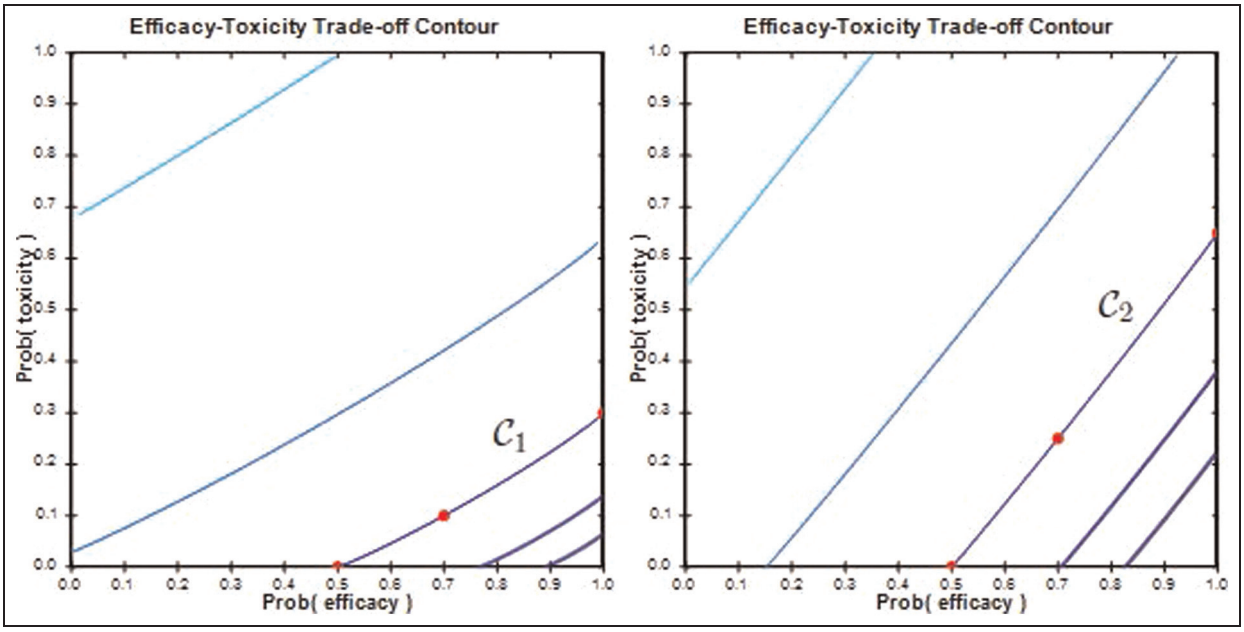

Figure 1. Two sets of efficacy–toxicity trade-off contours for the phase I–II prostate cancer trial. The contours in the left-hand figure generated by target $\mathcal{C}_1$ are insufficiently steep, and produce a design with poor properties. The contours in the right-hand figure generated by target $\mathcal{C}_2$ produce a design with more desirable properties, as shown in Table 1.

A design initially considered for this trial used the trade-off contour $\mathcal{C}_1$ based on the probability pairs (0.50,0), (1, 0.30), (0.70, 0.10), illustrated in the left-hand plot in Figure 1. This contour produces a design that fails to escalate reliably in scenarios where higher doses have higher $\pi_{E,k}(\boldsymbol{\theta})$ and acceptable $\pi_{T,k}(\boldsymbol{\theta})$. This is because, as shown in the left-hand plot of Figure 1, the contours determined by $\mathcal{C}_1$ are not sufficiently steep to give substantially higher values of $\delta(\pi_{E,k},\pi_{T,k})$ if $\pi_{E,k}$ increases substantially with dose k while $\pi_{T,k}$ increases very little. The contours in the right-hand plot in Figure 1, determined by target $\mathcal{C}_2$ based on the trade-off probability pairs (0.50,0), (1,0.65), (0.70,0.25), give a design that does not have this pathological property. These contours are sufficiently steep so that they give a higher desirability to higher doses having higher $\pi_{E,k}(\boldsymbol{\theta})$ with acceptable $\pi_{T,k}(\boldsymbol{\theta})$, so the algorithm is much more likely to escalate to more desirable higher doses.

Table 1 illustrates this effect in terms of the OCs of two designs for the prostate cancer trial that differ only in terms of their trade-off contours. The first design uses $\mathcal{C}_1$, and the second uses the steeper target $\mathcal{C}_2$. Recall that the desirability function is defined so that $\delta(\boldsymbol{\pi}) = 0$ on the target contour, so this function may take on positive or negative values. To facilitate interpretation, the values $\delta^+(\boldsymbol{\pi}) = 100\, e^{\delta(\boldsymbol{\pi})}$ are given in the tables. The four scenarios in Table 1 were chosen to illustrate how the method behaves in settings of the type described above. In all cases, $\pi_{T,5}^{true}$ at the highest dose $k = 5$ is within 0.10 of the limit $\pi_T^* = 0.30$, with $0.20 \leq \pi_{T,5}^{true} \leq 0.40$, and $\pi_{E,k}$ increases monotonically in k to high values at the two highest dose levels, $k = 4,5$. In these scenarios, the contour $\mathcal{C}_1$ gives a design that tends to get stuck at dose levels 3 or 4, while $\mathcal{C}_2$ roughly doubles the selection percentage for the most desirable dose level 5 in Scenarios 1–3 and increases it from 21% to 35% in Scenario 4. Note that in Scenario 4, while $\pi_{T,5}^{true} = 0.40$ is slightly above the specified upper limit $\pi_T^* = 0.30$, with such small sample sizes, the difference $0.40 - 0.30 = 0.10$ is extremely difficult to detect using any statistical method. In this case, the trade-off function $\mathcal{C}_2$ says that $\delta^+(\pi_{T,4},\pi_{E,4})^{true} = \delta^+(0.20,0.50) = 73$ is slightly smaller than $\delta^+(\pi_{T,5},\pi_{E,5})^{true} = \delta^+(0.40,0.70) = 79$. That is, with $\mathcal{C}_2$ increasing π_E^{true} from 0.50 to 0.70 and π_T^{true} from 0.20 to 0.40 gives a small increase in δ^+, so this is a desirable trade-off. If this is not what is wanted, then a contour less steep than $\mathcal{C}_2$ should be used. However, if one were to use a contour less steep with slope somewhere between those of $\mathcal{C}_2$ and $\mathcal{C}_1$, the price would be smaller selection percentages for dose levels 4 and 5 in each Scenario compared to the design with contour $\mathcal{C}_2$.

A sensitivity analysis of prior informativeness in terms of ESS shows why using $\tilde{\sigma}_{j,\mu}$ or $\tilde{\sigma}_{j,\beta}$ values that are either too large or too small can produce designs with very undesirable properties. This is quantified in Table 2, which summarizes the design's OCs in four scenarios using three different priors. An essential point is that the prior affects the stopping criterion in terms of the marginal posterior of $\pi(x_k, \boldsymbol{\theta}_j)$ for each $j = E,T$ in (1). The first prior in Table 2 has target $ESS = 10$ and, as might be expected, this prior turns out to be overly informative. The second prior is that which was derived

Table 1. Dose selection percentages for the prostate cancer trial design using either the problematic target contour $\mathcal{C}_1$ or the sufficiently steep contour $\mathcal{C}_2$.

Scenario	Contour		Dose level					None
			1	2	3	4	5	
1		$(\pi_{T,k},\pi_{E,k})^{true}$	(0.05,0.20)	(0.10,0.40)	(0.15,0.60)	(0.20,0.80)	(0.40,0.90)	
	$\mathcal{C}_1$	Desirability	44	54	68	87	55	
		% Selected	1	11	56	25	7	1
	$\mathcal{C}_2$	Desirability	51	69	96	132	118	
		% Selected	0	3	31	49	16	1
2		$(\pi_{T,k},\pi_{E,k})^{true}$	(0.05,0.20)	(0.10,0.40)	(0.15,0.60)	(0.20,0.80)	(0.30,0.90)	
	$\mathcal{C}_1$	Desirability	44	54	68	87	77	
		% Selected	1	11	53	23	11	1
	$\mathcal{C}_2$	Desirability	51	69	96	132	139	
		% Selected	0	3	31	39	25	1
3		$(\pi_{T,k},\pi_{E,k})^{true}$	(0.05,0.20)	(0.10,0.40)	(0.15,0.60)	(0.18,0.80)	(0.20,0.90)	
	$\mathcal{C}_1$	Desirability	44	54	68	93	108	
		% Selected	1	11	49	21	18	1
	$\mathcal{C}_2$	Desirability	51	69	96	136	162	
		% Selected	0	3	28	32	36	1
4		$(\pi_{T,k},\pi_{E,k})^{true}$	(0.05,0.10)	(0.10,0.15)	(0.15,0.20)	(0.20,0.50)	(0.40,0.70)	
	$\mathcal{C}_1$	Desirability	36	32	29	45	34	
		% Selected	0	0	3	59	21	17
	$\mathcal{C}_2$	Desirability	41	42	43	73	79	
		% Selected	0	0	1	47	35	17

Table 2. Dose selection percentages for the prostate cancer trial design using a prior with *ESS* = 10, 0.9, or 0.02.

Scenario	*ESS*		Dose level					None
			1	2	3	4	5	
1		$(\pi_{T,k},\pi_{E,k})^{true}$	(0.02,0.20)	(0.04,0.40)	(0.06,0.60)	(0.08,0.80)	(0.10,0.90)	
		Desirability	53	76	111	160	190	
	10	% Selected	0	0	0	4	96	0
	0.9	% Selected	0	1	13	38	48	0
	0.02	% Selected	1	4	15	9	68	3
2		$(\pi_{T,k},\pi_{E,k})^{true}$	(0.05,0.10)	(0.10,0.30)	(0.15,0.60)	(0.30,0.62)	(0.45,0.65)	
		Desirability	41	57	96	79	66	
	10	% Selected	0	0	15	68	17	0
	0.9	% Selected	0	1	45	38	11	5
	0.02	% Selected	1	4	26	18	34	17
3		$(\pi_{T,k},\pi_{E,k})^{true}$	(0.20,0.60)	(0.40,0.62)	(0.55,0.65)	(0.70,0.70)	(0.85,0.75)	
		Desirability	89	67	57	50	44	
	10	% Selected	14	61	12	0	0	13
	0.9	% Selected	56	27	2	0	0	15
	0.02	% Selected	44	25	1	0	0	30
4		$(\pi_{T,k},\pi_{E,k})^{true}$	(0.30,0.10)	(0.40,0.30)	(0.55,0.60)	(0.60,0.62)	(0.65,0.65)	
		Desirability	28	35	51	49	48	
	10	% Selected	0	7	24	2	0	67
	0.9	% Selected	0	6	6	1	0	87
	0.02	% Selected	0	0	0	0	0	100

ESS: effective sample size.

above to have target $ESS = 0.9$, used for the designs evaluated in Table 1. In Table 2, under Scenario 1, dose 5 is best based on the trade-off δ value, and $\boldsymbol{\pi}^{true} = \boldsymbol{\mu}^{(e)}$, the corresponding elicited means. Under Scenario 1, the design with prior $ESS = 10$ chooses the best dose 5 with very high probability, essentially because the prior means equal $\boldsymbol{\pi}^{true}$. In Scenario 2, where dose 3 is the best, the prior with $ESS = 10$ degrades performance substantially by biasing the selections toward dose 5. It chooses the best dose 15% of the time and the two higher suboptimal doses 68% + 17% = 85% of the time compared, respectively, to 45% and 49% for the design with prior $ESS = 0.9$. The same effect is seen in Scenario 3, where dose 1 is best but the design

with prior $ESS = 10$ has much larger selection percentages for the higher suboptimal doses 2 and 3. The greatest danger of an overly informative optimistic prior is that it will cause the posterior to underestimate toxicity or overestimate efficacy, and thus delay shutting down a trial in a case where all doses are inefficacious or too toxic. This is illustrated by Scenario 4, where all doses are unacceptable and the early stopping percentage drops from 87% for the design with $ESS = 0.90$ to 67% for the design with $ESS = 10$. In particular, the design with $ESS = 10$ has a 24% chance of selecting dose 3, which has $\pi_{T,3}^{true} = 0.55$, compared to a 6% selection rate for the design with $ESS = 0.9$.

The prior with $ESS = 10$ has $(\tilde{\sigma}_{E,\mu}, \tilde{\sigma}_{E,\beta}, \tilde{\sigma}_{T,\mu}, \tilde{\sigma}_{T,\beta}) = (0.63, 0.58, 0.02, 0.99)$. These numerical hyperparameter values have no specific intuitive meaning, other than that they are smaller than the corresponding values (2.54,2.44,3.55,3.50) for the prior with $ESS = 0.90$. However, it should not be surprising that a prior with $ESS = 10$ is overly informative for a trial of 39 patients. Thus, the ESS values give a simple, easily understandable way to quantify prior informativeness that cannot be obtained from the $\tilde{\boldsymbol{\sigma}}$ values.

It might seem that one may deal with the problem of prior specification by simply using large $\tilde{\boldsymbol{\sigma}}$ values, equivalently a small value of ESS, to obtain a putatively "non-informative" prior. This is a mistake that has undesirable consequences when using Bayesian designs for early phase trials, since early adaptive decisions are based on very small interim sample sizes. The numerical integration routine used to compute posteriors by the EffTox program does not converge for the very large values of $\tilde{\sigma}_{j,\mu}$ and $\tilde{\sigma}_{j,\beta}$ corresponding to very small ESS. Thus, in order to illustrate this case, we used a different computer program that was tailored to apply a new extension of EffTox, called the "late onset EffTox" method, that deals with delayed outcomes.[20] This program uses Markov chain Monte Carlo to compute posteriors. We used this program to implement the EffTox method for very small ESS values by matching all elements of the design with those given above, except prior informativeness, and setting $\tilde{\sigma}_{j,\mu} = \tilde{\sigma}_{j,\beta} = 100$ with domain $-300 \leq \mu_j, \beta_{j,1} \leq 300, j = E,T$, which gives $ESS = 0.02$. The OCs in Table 2 illustrate what can happen if one assumes a prior with very large hypervariances and thus ESS value that is too small. This prior causes the pathological behavior that the design is very likely to stop the trial early in cases where it should not. Specifically, the design with $ESS = 0.02$ has extremely high false-negative rates in scenarios where there are acceptable doses, with 17% and 30% of trials stopped early incorrectly in Scenarios 2 and 3, compared to 5% and 15% for the design with $ESS = 0.90$. In terms of correct selection rates, compared to the design with $ESS = 0.90$, the design with prior $ESS = 0.02$ has a larger probability of selecting the best dose level 5 in Scenario 1, but smaller probabilities of selecting the best doses in Scenarios 2 and 3. This example illustrates the fact that in this setting, the notion that a prior with very large variances is "non-informative" in the sense that it has little effect on inferences is inappropriate.

Setting $\tilde{\sigma}_{j,\mu} = \tilde{\sigma}_{j,\beta} = 100$ causes the problem of incorrectly stopping early because the induced priors on the $\pi_{j,k}(\boldsymbol{\theta}_j|\tilde{\boldsymbol{\theta}}_j)$s have most of their mass very near 0 or 1. This, in turn, causes the posteriors used by the early stopping rules (3) to fall below the cutoff 0.10 for very small early samples. For example, suppose that the outcomes for the first three patients treated at the lowest dose $k = 1$ all are $(Y_E, Y_T) = (0,0)$, and the data for the next cohort treated at $k = 2$ consist of two with outcome $(Y_E, Y_T) = (0,0)$ and one with $(Y_E, Y_T) = (1,1)$. Based on these six outcomes, the design with $ESS = 0.02$ gives the following posterior dose acceptability criteria probabilities. At the five doses, the respective efficacy acceptability probabilities $\Pr\{\pi_{T,k}(\boldsymbol{\theta}_E) > \pi_E^*|\mathcal{D}_n\}$ are (0.01,0.09,0.84,0.92,0.94), and the respective toxicity acceptability probabilities $\Pr\{\pi_{T,k}(\boldsymbol{\theta}_T) < \pi_T^*|\mathcal{D}_n\}$ are (0.99,0.68,0.09,0.06,0.04). With the early stopping cutoffs 0.10, these probabilities cause the design to conclude that doses $k = 1$ and 2 are inefficacious and doses $k = 3$, 4, and 5 are too toxic, thus stopping the trial. In contrast, the design with prior $ESS = 0.9$ concludes that dose $k = 1$ is inefficacious but that all other acceptability criteria are met, and escalates to dose level $k = 3$.

Figure 2 provides a graphical illustration of how the probabilities of correct selection and of incorrectly stopping the trial early vary with ESS, under Scenario 2 of

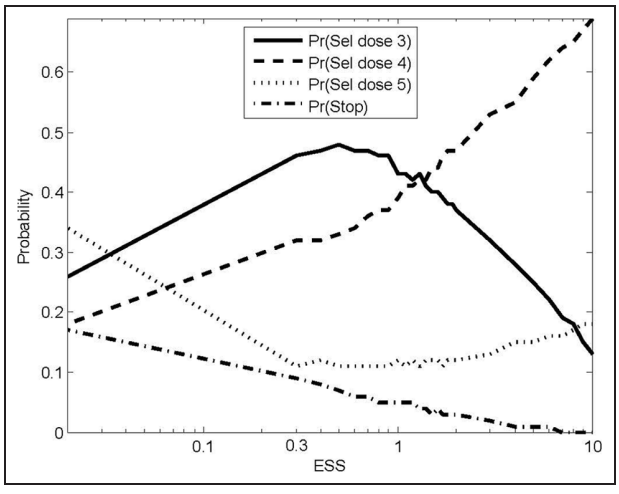

Figure 2. Probabilities of selecting doses 3, 4, or 5, or stopping the trial early, under Scenario 2 of Table 2, where doses 1, 2, 3, and 4 are acceptable and the respective desirabilities of the five doses based on their efficacy–toxicity trade-offs are 41, 57, 96, 79, and 66.
ESS: effective sample size.

Table 2. Figure 2 gives Prob(select dose level k) for $k = 3$, 4, and 5 and Prob(stop early) plotted as functions of ESS. Selection probabilities are not plotted for doses $k = 1$ or 2 since, as shown in Table 2, these probabilities are negligible in this scenario. Since priors with very small ESS are pathological with high incorrect early stopping probabilities, as shown in Table 2 for $ESS = 0.02$, we consider the range $0.02 \leq ESS \leq 10$ in Figure 2. Note that, from Table 2, the respective dose desirabilities under Scenario 2 are (41,57,96,79,66) for doses $k = (1,2,3,4,5)$, so dose $k = 3$ is best, and dose $k = 5$ is unacceptable since $\pi_{T,5}^{true} = 0.45$ is above the design's upper limit $\pi_T^* = 0.30$. Figure 2 shows that, as the prior ESS increases beyond about 1.4, the probability of selecting dose $k = 4$ begins to exceed the probability of selecting the best dose $k = 3$. At the lower end of the domain, for smaller values of ESS, the probability of incorrectly stopping the trial early begins to become large. Based on these results, and the behavior of similar plots under other scenarios, we recommend priors with $0.3 \leq ESS \leq 1.0$. However, in a given application, it may be worthwhile to perform a similar sensitivity analysis of design behavior as a function of ESS under one's chosen simulation scenarios.

Practical guidelines for trial design

The following steps should be carried out in collaboration with the physician or physicians planning the trial. All necessary computations may be done using the most recent version of the EffTox computer program.

STEP 1. Determine the disease and trial entry criteria, treatment, the K doses, and precise definitions of efficacy and toxicity.
STEP 2. Determine the maximum sample size N, which may be a range of feasible values, and one or more cohort sizes c, which should be at most 3. Larger K (more dose levels) requires a larger N, and the maximum number of cohorts, N/c, should be large enough to explore the K dose levels reliably. $N = 24$ is roughly the smallest trial that can be run feasibly using EffTox, although $N = 18$ may be feasible for $K = 3$. Larger N provides more reliable posterior estimates of all quantities. It should be kept in mind when choosing N that this design completes both phase I and phase II simultaneously.
STEP 3. Elicit prior mean probabilities of efficacy and toxicity at each dose, and probability pairs to determine an initial target contour $\mathcal{C}$. Because it is not intuitively obvious why a $\mathcal{C}$ that is not sufficiently steep will produce a design that may not escalate as desired, this should be explained carefully to the physician. The EffTox program interface gives a picture of $\mathcal{C}$ that can easily be modified interactively, and this tool should be used during the contour elicitation process.
STEP 4. Specify the ESS targets and derive the prior hyperparameters.
STEP 5. Specify a set of simulation scenarios in terms of $(\pi_{E,k},\pi_{T,k})^{true}$, $k = 1, \ldots, K$. Conduct a preliminary set of simulations to help choose and possibly calibrate the design parameters. This may be done using 100–400 replications per scenario to save time. Sensitivity analyses in N, c, and ESS also may be done at this stage. If appropriate, show these preliminary simulations to the physician to help establish N, c, π_T^*, π_E^*, $\mathcal{C}$, and the ESS targets.
STEP 6. Iterate some or all of Steps 1–5, if necessary, until an acceptable design is obtained.
STEP 7. Conduct a final set of simulations, with 1000–4000 replications per scenario, to establish the design's OCs. Include a table of these OCs in the trial protocol, as part of the design specification.

Discussion

We have provided an algorithm that exploits the idea of prior ESS to compute prior hyperparameters based on elicited outcome probabilities. Additionally, we have provided guidelines for specifying the EffTox design's target trade-off contour. Together, the algorithm and guidelines provide substantive improvements in the EffTox design's OCs. The algorithm is general, and may be applied in other settings. It provides a practical way to avoid priors that are either overly informative or overly disperse, both of which can lead to designs with pathological behavior. As our sensitivity analyses show, prior informativeness and consequently design performance vary substantially with the magnitudes of the hypervariances. Since the numerical values of these parameters have no intuitive meaning, ESS is a convenient and intuitive tool for quantifying prior informativeness.

A simple but important point is that prior informativeness cannot be a function of $\dim(\boldsymbol{\theta})$ alone, since all of our sensitivity analyses were conducted in a setting where $\dim(\boldsymbol{\theta})$ did not vary. How prior informativeness, quantified in terms of ESS or possibly some other index, may vary with $\dim(\boldsymbol{\theta})$ for given $\tilde{\boldsymbol{\sigma}}$ values, either under different parameterizations of a given model or possibly different dose–outcome models, is an interesting related issue.

The simple requirement that the trade-off contour should be sufficiently steep greatly improves the design's ability to escalate in cases where efficacy increases with dose while toxicity remains acceptable. A new version of the EffTox computer program for implementation has been provided. The methodology is labor intensive, however, requiring a large number of elicited probabilities and design parameters, as well as computer simulations to establish an acceptable

design. Together, these improvements greatly enhance the design's practical usefulness.

Declaration of conflicting interests

The authors declare that there is no conflict of interest.

Funding

This research was supported by National Cancer Institute (NCI) grant RO1 CA 083932.

References

1. Braun TM and Wang S. A hierarchical Bayesian design for phase I trials of novel combinations of cancer therapeutic agents. *Biometrics* 2010; 66: 805–812.
2. Cheung Y-K. *Dose finding by the continual reassessment method*. New York: Chapman & Hall/CRC, 2011.
3. Thall PF, Szabo A, Nguyen HQ, et al. Optimizing the concentration and bolus of a drug delivered by continuous infusion. *Biometrics* 2011; 67: 1638–1646.
4. Thall PF and Cook JD. Dose-finding based on efficacy-toxicity trade-offs. *Biometrics* 2004; 60: 684–693.
5. Thall PF, Cook JD and Estey EH. Adaptive dose selection using efficacy-toxicity trade-offs: illustrations and practical considerations. *J Biopharm Stat* 2006; 16: 623–638.
6. Morita S, Thall PF and Müller P. Determining the effective sample size of a parametric prior. *Biometrics* 2008; 64: 595–602.
7. Morita S, Thall PF and Müller P. Evaluating the impact of prior assumptions in Bayesian biostatistics. *Stat Biosci* 2010; 2: 1–17.
8. Nelder JA and Mead R. A simplex method for function minimization. *Comput J* 1965; 7: 308.
9. Tsutakawa RK and Lin HY. Bayesian estimation of item response curves. *Psychometrika* 1986; 51: 251–267.
10. Bedrick EJ, Christensen R and Johnson W. A new perspective on priors for generalized linear models. *J Am Stat Assoc* 1996; 91: 1450–1460.
11. Conaway MR and Petroni GR. Designs for phase II trials allowing for a trade-off between response and toxicity. *Biometrics* 1996; 52: 1375–1386.
12. Thall PF and Russell KT. A strategy for dose-finding and safety monitoring based on efficacy and adverse outcomes in phase I/II clinical trials. *Biometrics* 1998; 54: 251–264.
13. Thall PF, Wooten LH and Shpall EJ. A geometric approach to comparing treatments for rapidly fatal diseases. *Biometrics* 2006; 62: 193–201.
14. Gooley TA, Martin PJ, Fisher LD, et al. Simulation as a design tool for phase I/II clinical trials: an example from bone marrow transplantation. *Control Clin Trials* 1994; 15: 450–462.
15. Braun TM.The bivariate continual reassessment method: extending the CRM to phase I trials of two competing outcomes. *Control Clin Trials* 2002; 23: 240–256.
16. Yuan Y and Yin G. Bayesian dose-finding by jointly modeling toxicity and efficacy as time-to-event outcomes. *J Roy Stat Soc C: App* 2009; 58: 954–968.
17. Zhang W, Sargent DJ and Mandrekar S. An adaptive dose-finding design incorporating both efficacy and toxicity. *Stat Med* 2006; 25: 2365–2383.
18. Chevret S. *Statistical methods for dose-finding experiments*. New York: Wiley, 2006.
19. Zohar S and Chevret S. Recent developments in adaptive designs for phase I/II dose-finding studies. *J Biopharm Stat* 2007; 17: 1071–1083.
20. Jin I-H, Liu S, Thall PF, et al. Using data augmentation to facilitate conduct of phase I/II clinical trials with delayed outcomes. *J Am Stat Assoc* 2014; 109: 525–536.