Bayesian Model Averaging Continual Reassessment Method in Phase I Clinical Trials

Guosheng YIN and Ying YUAN

The continual reassessment method (CRM) is a popular dose-finding design for phase I clinical trials. This method requires that practitioners prespecify the toxicity probability at each dose. Such prespecification can be arbitrary, and different specifications of toxicity probabilities may lead to very different design properties. To overcome the arbitrariness and further enhance the robustness of the design, we propose using multiple parallel CRM models, each with a different set of prespecified toxicity probabilities. In the Bayesian paradigm, we assign a discrete probability mass to each CRM model as the prior model probability. The posterior probabilities of toxicity can be estimated by the Bayesian model averaging (BMA) approach. Dose escalation or deescalation is determined by comparing the target toxicity rate and the BMA estimates of the dose toxicity probabilities. We examine the properties of the BMA-CRM approach through extensive simulation studies, and also compare this new method and its variants with the original CRM. The results demonstrate that our BMA-CRM is competitive and robust, and eliminates the arbitrariness of the prespecification of toxicity probabilities.

KEY WORDS: Adaptive design; Bayesian inference; Maximum tolerated dose; Model selection; Posterior model probability; Robustness; Toxicity probability.

1. INTRODUCTION

The primary goal of a phase I clinical trial is to identify the maximum tolerated dose (MTD) of a new drug. The MTD is typically defined as the dose with the toxicity probability closest to the target toxicity rate. A phase I clinical trial is critically important, because it determines the MTD that will be further investigated in the subsequent phase II or III trials. Misidentification of the MTD could result in an inconclusive trial, thereby wasting enormous resources, or a trial in which a substantial number of patients are treated at excessively toxic doses. In addition, inappropriate selection of a dose with low toxicity and negligible efficacy as the MTD might cause researchers to overlook a promising drug.

Many statistical methods have been developed for phase I dose-finding studies. The standard $3 + 3$ design is an algorithm-based procedure that typically defines the MTD as the highest dose with a toxicity probability $< 33\%$ (Storer 1989). In practice, many clinical trials are carried out using the standard $3 + 3$ design, which is easy to understand and implement. But the operating characteristics of this design are not satisfactory, and the estimates of the toxicity probabilities are not reliable. Moreover, the $3 + 3$ design is "memoryless" with no convergence property and is suitable only for targeting a toxicity probability $<33\%$ (O'Quigley and Chevret 1991; O'Quigley and Shen 1996). A well-known alternative to the conventional $3+3$ design is the continual reassessment method (CRM) of O'Quigley, Pepe, and Fisher (1990). The CRM is a model-based dose-finding approach that uses a single unknown parameter to link the true toxicity probabilities with the prespecified toxicity probabilities. During the trial, the unknown parameter is continuously updated using the accrued information to identify the dose with a given target toxicity level. In related work, Whitehead and Brunier (1995) introduced a decision-theoretic approach based on Bayesian decision theory; Durham, Flournoy, and Rosenberger (1997) described a family of random-walk rules that is a nonparametric method with a completely workable distribution theory; Babb, Rogatko, and Zacks (1998) proposed a dose escalation method with overdose control that directly controls the probability of overdosing; Gasparini and Eisele (2000) developed a curve-free method that can be reformulated in the CRM framework under a particular prior distribution (O'Quigley 2002); and Stylianou and Flournoy (2002) proposed the biased-coin design with an isotonic regression estimator. Comprehensive coverage of dose-finding methods and up-to-date developments has been provided by Chevret (2006) and Ting (2006).

Of the aforementioned dose-finding methods, here we focus on the CRM, which has been widely used in phase I clinical trials. Many revisions to the CRM aimed at improving its properties have been proposed. Faries (1994) introduced several conservative modifications of the CRM. Goodman, Zahurak, and Piantadosi (1995) developed practical improvements to the CRM. They suggested assigning more than one subject at a time to each dose level and limiting each dose escalation by a single dose level. Møller (1995) extended the CRM using a preliminary up-and-down design to reach the neighborhood of the target dose during a successive escalation. Piantadosi, Fisher, and Grossman (1998) proposed a practical implementation of a modified CRM. They used a simple dose–toxicity model to guide data interpolation and grouped three patients into a cohort to minimize calculations and stabilize estimates. Heyd and Carlin (1999) further refined the CRM by allowing the trial to stop earlier when the width of the posterior 95% probability interval for the MTD becomes sufficiently narrow. Ishizuka and Ohashi (2001) proposed monitoring a posterior density function of toxicity to reduce the number of patients treated at doses exceeding the MTD. Leung and Wang (2002) extended the CRM using decision theory to optimize the number of patients allocated to the highest dose with toxicity not exceeding the tolerable level. Braun (2002) extended the CRM to model bivariate competing outcomes. Yuan, Chappell, and Bailey (2007) developed a quasi-likelihood approach to accommodate multiple

Guosheng Yin is Associate Professor (E-mail: *gsyin@mdanderson.org*) and Ying Yuan is Assistant Professor, Department of Biostatistics, M. D. Anderson Cancer Center, University of Texas, Houston, TX 77230. The authors thank the editor, the associate editor, and two anonymous referees for their insightful and constructive comments that substantially improved the article.

© 2009 American Statistical Association
Journal of the American Statistical Association
September 2009, Vol. 104, No. 487, Applications and Case Studies
DOI: 10.1198/jasa.2009.ap08425

toxicity grades. For a comprehensive introduction and information on the practical use of the CRM in phase I clinical trials, see the tutorial by Garrett-Mayer (2006).

But despite the immense success of the CRM and its modified versions, a major issue associated with it is the required prespecification of toxicity probabilities for the doses to be considered in the trial. Because the toxicity profile of a new drug often is unknown, this prespecification can be arbitrary and very subjective. The set of values of the mean probability of toxicity at each dose to be considered in the trial is known as the "skeleton" of the CRM. The use of different skeletons may lead to quite different design properties. Shen and O'Quigley (1996) showed that in large samples, the CRM is robust to the misspecification of the skeleton, and the recommended dose level in general converges to the target level. But a typical phase I trial has a very small sample size, often as low as 20–40 patients. The CRM's asymptotic behavior is not very relevant, and its performance may be compromised if the elicited toxicity probabilities in the skeleton do not fit the assumed dose–toxicity model. As we see in our simulation studies (described in Sec. 3), the selection probability of the target dose can be 40% lower under one skeleton than that under another skeleton. Unfortunately, in practical situations, practitioners have no information to determine whether or not a specific skeleton is reasonable, because the underlying true toxicity probabilities are unknown. In clinical practice, we also have found that the requirement to specify a skeleton is one of the major obstacles to physicians' acceptance and application of the CRM. The lack of previous knowledge of a new drug often leads to uncertainty regarding the specification of the skeleton. Because physicians may have several different guesses about a drug's toxicity profile a priori, they may be reluctant to choose one as the best skeleton to use to carry out the CRM design.

To overcome the arbitrariness in this prespecification of toxicity probabilities, we propose conducting the CRM design using multiple skeletons in parallel. Each skeleton represents a prior guess of the toxicity profile of the drug, which may be close to or far from the true toxicity profile. We view the CRM under different sets of prespecified toxicity probabilities as separate models. We take a Bayesian model averaging (BMA) approach to obtain the posterior estimates for the true toxicity probabilities by weighing the estimates from each model with the corresponding posterior model probability. The decision for dose escalation or deescalation is then made by comparing the BMA estimates of the toxicity probabilities and the target toxicity level. In other words, instead of using a single CRM for the trial, we carry out a set of parallel CRMs and rely on a BMA estimator for decision making. During the trial, the BMA method automatically and adaptively assigns a larger weight to a model with a better fit. Thus the BMA procedure ensures that the estimates of the toxicity probabilities are always close to the best estimates given a set of models. In the proposed procedure, we no longer consider the prior guesses of the toxicity probabilities as fixed; instead, we associate them with model uncertainties. After a new cohort of patients enters the trial, in light of the most recent observations, we update our knowledge on the estimated probabilities of toxicity using the BMA approach. As a result, we are able to find the dose with the desired toxicity level in a more reliable way and to treat the patients in the trial at more appropriate doses.

The remainder of the article is organized as follows. In Section 2 we briefly review the original CRM and BMA methodologies and propose the Bayesian model averaging continual reassessment method (BMA-CRM) and its alternative versions. In Section 3 we present simulation studies to compare the operating characteristics of the new methods with those of the original CRM. In Section 4 we conduct extensive sensitivity analysis to further investigate the properties of the BMA-CRM. In Section 5 we illustrate the proposed designs with two phase I clinical trials, and we conclude with a brief discussion in Section 6.

2. METHODS

2.1 Continual Reassessment Method

We assume that the dose-limiting toxicity (DLT) is recorded as a binary outcome and that the true dose toxicity monotonically increases with respect to the dose level. The CRM assumes a prior dose–toxicity curve, then continuously updates this curve given the observed cumulating toxicity outcomes from patients in the trial. Based on the updated dose–toxicity curve, a new cohort of patients is assigned to the dose with an estimated toxicity probability closest to the prespecified target. Let $(d_1, \ldots, d_J)$ denote a set of J prespecified doses for the drug under investigation, and let $(p_1, \ldots, p_J)$ be the prespecified toxicity probabilities (skeleton) at those doses, $p_1 < \cdots < p_J$. Let ϕ be the target toxicity rate specified by physicians. The first cohort of patients receives the lowest dose, d_1. For the CRM, we assume a working dose–toxicity model, such as

$$\text{pr}(\text{toxicity at } d_j) = \pi_j(\alpha) = p_j^{\exp(\alpha)} \tag{1}$$

for $j = 1, \ldots, J$, where α is an unknown parameter and the p_j's can be viewed as "imputed" values for the toxicity probabilities.

Suppose that among n_j patients treated at dose level j, y_j patients have experienced DLT. Let D denote the observed data, $D = \{(n_j, y_j), j = 1, \ldots, J\}$. Based on the binomial distribution for the toxicity outcome, the likelihood function is given by

$$L(D|\alpha) = \prod_{j=1}^{J} \{p_j^{\exp(\alpha)}\}^{y_j} \{1 - p_j^{\exp(\alpha)}\}^{n_j - y_j}.$$

We estimate the toxicity probabilities using the corresponding posterior means of $\pi_j(\alpha)$. Using the Bayes theorem, we can compute the posterior means of the dose toxicity probabilities given D by

$$\hat{\pi}_j = \int p_j^{\exp(\alpha)} \frac{L(D|\alpha) f(\alpha)}{\int L(D|\alpha) f(\alpha)\, d\alpha}\, d\alpha,$$

where $f(\alpha)$ is a prior distribution for the parameter α. We take a normal prior distribution $\text{N}(0, \sigma^2)$ for α,

$$f(\alpha) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\alpha^2}{2\sigma^2}\right).$$

After updating the posterior estimates of the toxicity probabilities at all of the doses considered, the recommended dose level for the next cohort of patients is the one that has a toxicity probability closest to the target ϕ. Thus a new cohort of patients is assigned to dose level j^*, such that

$$j^* = \underset{j \in (1, \ldots, J)}{\text{argmin}} \, |\hat{\pi}_j - \phi|.$$

The trial continues until the total sample size is exhausted, after which the dose with a posterior toxicity probability closest to ϕ is selected as the MTD.

2.2 Bayesian Model Averaging Continual Reassessment Method

A major issue associated with the CRM is that prespecification of the toxicity probabilities $(p_1, \ldots, p_J)$ is arbitrary. Because of the lack of toxicity information on a new drug, physicians may have quite different opinions on the toxicity probabilities. If the p_j's deviate far from the true dose–toxicity curve (i.e., the true toxicity probabilities cannot be recovered even after being adjusted by α), this may lead to poor operating characteristics and a high probability of selecting the wrong dose as the MTD. To avoid subjectivity in specifying the skeleton, we propose prespecifying multiple skeletons, each representing a set of prior estimates of the toxicity probabilities. We view each skeleton as corresponding to a CRM model of the form (1) with a different set of p_j's. During the trial, conditional on the observed data, these different models usually yield different estimates of the toxicity probabilities $(\hat{\pi}_1, \ldots, \hat{\pi}_J)$. Some of these estimates may be close to the true values, whereas others may not, depending on how well the models fit the cumulating data. To accommodate the uncertainty in the specification of these skeletons, we take a BMA approach to average $\hat{\pi}_j$ across the CRM models to obtain the BMA estimate of the toxicity probability for dose level j. BMA is known to provide a better predictive performance than any single model (Raftery, Madigan, and Hoeting 1997; Hoeting et al. 1999). In other words, we incorporate the uncertainty in the prespecification of the toxicity probabilities into the estimation procedure, such that the potential estimation bias caused by a misspecification of the p_j's can be averaged out.

Let $(M_1, \ldots, M_K)$ be the models corresponding to each set of prior guesses of the toxicity probabilities $\{(p_{11}, \ldots, p_{1J}), \ldots, (p_{K1}, \ldots, p_{KJ})\}$. Model M_k $(k = 1, \ldots, K)$ in the CRM is given by

$$\pi_{kj}(\alpha_k) = p_{kj}^{\exp(\alpha_k)}, \qquad j = 1, \ldots, J,$$

which is based on the kth skeleton $(p_{k1}, \ldots, p_{kJ})$. Let $\text{pr}(M_k)$ be the prior probability that model M_k is the true model; that is, the probability that the kth skeleton $(p_{k1}, \ldots, p_{kJ})$ matches the true dose–toxicity curve. If there is no preference a priori for any single model in the CRM case, then we can assign equal weights to the different skeletons by simply setting $\text{pr}(M_k) = 1/K$. When there is prior information on the importance of each set of the prespecified toxicity probabilities, we can incorporate such information into $\text{pr}(M_k)$. For example, if a certain set of the prespecification is more likely to be true, then we can assign it a higher prior model probability. At a certain stage of the trial, based on the observed data $D = \{(n_j, y_j), j = 1, \ldots, J\}$, the likelihood function under model M_k is

$$L(D|\alpha_k, M_k) = \prod_{j=1}^{J} \{p_{kj}^{\exp(\alpha_k)}\}^{y_j} \{1 - p_{kj}^{\exp(\alpha_k)}\}^{n_j - y_j}.$$

The posterior model probability for M_k is given by

$$\text{pr}(M_k|D) = \frac{L(D|M_k)\,\text{pr}(M_k)}{\sum_{i=1}^{K} L(D|M_i)\,\text{pr}(M_i)},$$

where $L(D|M_k)$ is the marginal likelihood of model M_k,

$$L(D|M_k) = \int L(D|\alpha_k, M_k) f(\alpha_k|M_k)\, d\alpha_k,$$

α_k is the power parameter in the CRM associated with model M_k, and $f(\alpha_k|M_k)$ is the prior distribution of α_k under model M_k.

The posterior model probability can be naturally linked to the Bayes factor. The Bayes factor, B_{10}, for a model M_1 against another model M_0 given data D is defined as the ratio of posterior to prior odds,

$$B_{10} = \frac{\text{pr}(D|M_1)}{\text{pr}(D|M_0)},$$

which is the ratio of the marginal likelihoods, i.e., $\text{pr}(D|M_k) = L(D|M_k)$. We can construct such Bayes factors for each of the models $(M_1, \ldots, M_K)$ against model M_0, denoted by $(B_{10}, \ldots, B_{K0})$. Then the posterior model probability of M_k is

$$\text{pr}(M_k|D) = \frac{\eta_k B_{k0}}{\sum_{i=1}^{K} \eta_i B_{i0}},$$

where $\eta_k = \text{pr}(M_k)/\text{pr}(M_0)$ is the prior odds for M_k against M_0, $k = 1, \ldots, K$.

The BMA estimate for the toxicity probability at each dose level is given by

$$\bar{\pi}_j = \sum_{k=1}^{K} \hat{\pi}_{kj}\,\text{pr}(M_k|D), \qquad j = 1, \ldots, J, \tag{2}$$

where $\hat{\pi}_{kj}$ is the posterior mean of the toxicity probability of dose level j under model M_k, that is,

$$\hat{\pi}_{kj} = \int p_{kj}^{\exp(\alpha_k)} \frac{L(D|\alpha_k, M_k) f(\alpha_k|M_k)}{\int L(D|\alpha_k, M_k) f(\alpha_k|M_k)\, d\alpha_k}\, d\alpha_k.$$

By assigning $\hat{\pi}_{kj}$, a weight of $\text{pr}(M_k|D)$, the BMA method automatically identifies and favors the best-fitting model; thus $\bar{\pi}_j$ is always close to the best estimate. Therefore, the decision of dose escalation or deescalation in the trial is based on $\bar{\pi}_j$ as opposed to $\hat{\pi}_{kj}$.

The original CRM is based on only one set of prespecified toxicity probabilities, $(p_1, \ldots, p_J)$. But in our approach, we consider multiple sets of the prespecified toxicity probabilities. We not only estimate α_k for each set of p_{kj}'s, but also update the posterior model probabilities for all sets of p_{kj}'s during the trial. BMA provides a coherent mechanism to account for the model uncertainty associated with each skeleton. Madigan and Raftery (1994) noted that by averaging over all of the considered models, we can provide a better average predictive ability than can be attained using any single model based on a logarithm scoring rule.

2.3 Dose-Finding Algorithm

Let ϕ be the physician-specified toxicity target. Patients are treated in cohorts, for example, with a cohort size of three. To be conservative, we restrict dose escalation or deescalation by one dose level of change at a time. The dose-finding algorithm in our BMA-CRM method is as follows:

1. Patients in the first cohort are treated at the lowest dose, d_1, or the physician-specified dose.

2. At the current dose level j^{curr}, we obtain the BMA estimates for the toxicity probabilities, $\bar{\pi}_j$ $(j = 1, \ldots, J)$, based on the cumulated data. We then find dose level j^* that has a toxicity probability closest to ϕ, that is,

$$j^* = \underset{j\in(1,\ldots,J)}{\text{argmin}} |\bar{\pi}_j - \phi|.$$

   If $j^{\text{curr}} > j^*$, then we deescalate the dose level to $j^{\text{curr}} - 1$, and if $j^{\text{curr}} < j^*$, then we escalate the dose level to $j^{\text{curr}} + 1$; otherwise, the dose stays at the same level as j^{curr} for the next cohort of patients.
3. Once the maximum sample size is reached, we choose the dose with toxicity probability closest to ϕ as the MTD.

In addition, we add a stopping rule in our algorithm: If pr(toxicity rate at $d_1 > \phi) > 0.9$, then the trial is terminated for safety. In the BMA-CRM, we require early termination of a trial if the lowest dose is too toxic, as noted by

$$\sum_{k=1}^{K} \text{pr}\{\pi_{k1}(\alpha_k) > \phi | M_k, D\} \,\text{pr}(M_k|D) > 90\%.$$

The BMA method automatically assigns a higher weight to a better-fitting model. When averaging across all of the models considered, it would effectively downweight the impact of the poorly fitting models. However, if the fit of a model were far worse than that of the best-fitting model, then excluding that model from the model-averaging set would be reasonable. This procedure can be carried out using Occam's window criterion. More specifically, model M_k will be included in the model-averaging set only if it satisfies

$$\frac{\text{pr}(M_k|D)}{\max_{i\in(1,\ldots,K)} \text{pr}(M_i|D)} > \delta.$$

We can calibrate δ to obtain desirable operating characteristics for simulated trials, such as yielding a high MTD selection percentage. In a clinical trial, patients are sequentially accrued and assigned to a dose in a cohort; thus as the trial proceeds, the BMA estimates of the dose toxicity probabilities using Occam's window may have a different set of models from which the model averaging is taken. A further refinement based on Occam's razor can exclude more complex models that are not as well supported by the data as the simpler models. In our case, all of the CRM models have the same structure or complexity but different sets of underlying prespecifications of the toxicity probabilities; thus Occam's window serves as an adequate criterion for the purpose of BMA-CRM dose finding.

In contrast to model averaging, model selection takes a different perspective in regression models. Among a set of competing models, we can simply select the best-fitting model according to a suitable model selection criterion. In the CRM case, we consider that each skeleton corresponds to a CRM model. As more data are collected in the trial, we can select the most appropriate skeleton each time that a decision on dose assignment is made. Thus the skeleton in the CRM is not fixed, but can be updated and adaptively chosen from a set of skeletons. A natural candidate for the model selection criterion is based on the posterior model probability, as described previously. After observing the outcomes of each cohort of patients, we select the skeleton or the CRM that yields the highest posterior model probability with probability 1.

3. SIMULATION STUDIES

We investigated the operating characteristics of the proposed BMA-CRM design through simulation studies under nine different toxicity scenarios. We considered eight dose levels and assumed that toxicity increased monotonically with respect to the dose. We prepared four sets of initial guesses of the toxicity probabilities:

$$(p_1, p_2, p_3, p_4, p_5, p_6, p_7, p_8) = \begin{cases} (0.02, 0.06, 0.08, 0.12, 0.20, 0.30, 0.40, 0.50), \\ \quad \text{skeleton 1} \\ (0.01, 0.05, 0.09, 0.14, 0.18, 0.22, 0.26, 0.30), \\ \quad \text{skeleton 2} \\ (0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80), \\ \quad \text{skeleton 3} \\ (0.20, 0.30, 0.40, 0.50, 0.60, 0.65, 0.70, 0.75), \\ \quad \text{skeleton 4.} \end{cases}$$

The first skeleton is for the case in which toxicity increases slowly at the low doses but increases quickly at the high doses. The second skeleton is more concentrated at the low toxicity levels; the highest dose has a toxicity probability of 0.3. The toxicity probabilities in the third skeleton are spread evenly over a range of 0.1–0.8. The fourth skeleton starts at a relatively high toxicity probability of 0.2 and increases quickly at the low doses, before leveling off at the high doses. Thus these four sets of skeletons represent four different prior opinions on the true dose–toxicity curve. We focus on different features in the four skeletons that are expected to capture the true dose–toxicity curve more effectively when combined together using the BMA. We refer to the individual CRMs using each of these four skeletons as CRM 1, CRM 2, CRM 3, and CRM 4.

In Table 1, under each scenario we list the true toxicity probabilities in the first row, the dose selection probability and the average number of patients treated at each dose separately for the CRM using each of the four skeletons in rows 2–9, the results obtained using the proposed BMA-CRM and the BMA-CRM based on Occam's window with $\delta = 0.6$ (denoted by BMAO-CRM) in rows 10–13, and the results obtained by the CRM using the Bayesian model selection procedure (referred to as BMS-CRM) in the last two rows. We also report the percentage of inconclusive trials (denoted by "none"), the average number of patients experiencing toxicity, and the total number of patients in the trial. For each single CRM (1–4), we used the modified version (Goodman, Zahurak, and Piantadosi 1995; Møller 1995; Chevret 2006), with a stopping rule: If pr(toxicity rate at $d_1 > \phi$) was >0.9, then the trial was terminated for safety. The target toxic probability was $\phi = 30\%$. We took the prior distribution of α as a normal distribution with mean 0 and standard deviation $\sigma = 2$. With no preference for any specific skeleton, we assigned the prior model probability of $1/4$ to each CRM model, that is, $\text{pr}(M_k) = 1/4$ for $k = 1, \ldots, 4$. We took the cohort size 3 and treated the first cohort of patients at the lowest dose level. The maximum sample size was 30, and for each scenario we carried out 10,000 simulated trials.

In scenario 1, the seventh dose was the MTD, and the four individual CRMs using different skeletons selected the MTD with very different probabilities. In particular, CRM 2 (corresponding to the design using skeleton 2) had the lowest selection percentage of 30.8% for the MTD, but selected the eighth dose

Table 1. Simulation study comparing the CRM, BMA-CRM, BMA-CRM with Occam's window (BMAO-CRM), and BMS-CRM with a toxicity target $\phi = 30\%$

	Recommendation percentage at dose level									Average	Average
Design	1	2	3	4	5	6	7	8	None	toxicity	# patients
Scenario 1	2	3	4	6	8	10	30	50			
CRM 1	0	0	0	0.1	1.4	16.0	**52.6**	29.9	0	4.7	30
# patients	3.2	3.0	3.0	3.1	3.6	4.7	5.7	3.6			
CRM 2	0	0	0	0.1	1.0	11.2	**30.8**	56.9	0	5.6	30
# patients	3.2	3.0	3.1	3.1	3.2	3.5	4.3	6.6			
CRM 3	0	0	0	0.8	4.6	22.1	**59.3**	13.1	0	3.9	30
# patients	3.2	3.0	3.2	3.5	4.1	5.2	6.4	1.4			
CRM 4	0	0	0	0.6	3.6	18.0	**44.8**	33.0	0	4.7	30
# patients	3.2	3.0	3.1	3.5	3.8	4.2	5.2	3.8			
BMA-CRM	0	0	0	0.2	1.5	16.2	**51.5**	30.6	0	4.7	30
# patients	3.2	3.0	3.1	3.2	3.5	4.4	6.3	3.2			
BMAO-CRM	0	0	0	0.2	1.3	15.4	**54.5**	28.6	0	4.7	30
# patients	3.2	3.0	3.1	3.2	3.4	4.5	6.1	3.4			
BMS-CRM	0	0	0	0.1	1.5	19.2	**50.5**	28.6	0	4.8	30
# patients	3.2	3.0	3.1	3.2	3.6	4.5	5.4	4.0			
Scenario 2	2	6	8	12	20	30	40	50			
CRM 1	0	0	0	2.9	23.9	**43.6**	22.7	6.9	0	5.9	30
# patients	3.2	3.1	3.2	3.6	6.3	6.6	3.0	0.8			
CRM 2	0	0	0.3	4.3	17.1	**28.4**	25.5	24.4	0	6.5	30
# patients	3.2	3.1	3.4	3.8	4.6	4.8	3.8	3.2			
CRM 3	0	0	0.6	6.3	32.6	**40.8**	18.1	1.6	0	5.2	30
# patients	3.2	3.1	3.6	4.8	6.9	5.7	2.4	0.2			
CRM 4	0	0	0.4	7.5	27.8	**35.3**	20.7	8.2	0	5.5	30
# patients	3.2	3.1	3.6	4.9	6.4	4.8	2.9	1.0			
BMA-CRM	0	0	0.3	4.3	23.9	**41.6**	22.7	7.3	0	5.7	30
# patients	3.2	3.1	3.4	4.3	5.9	5.8	3.3	0.8			
BMAO-CRM	0	0	0.2	4.3	23.2	**40.0**	24.3	7.9	0	5.7	30
# patients	3.2	3.1	3.4	4.4	5.8	6.2	3.0	0.9			
BMS-CRM	0	0	0.2	3.8	26.1	**38.4**	21.2	10.3	0	5.8	30
# patients	3.2	3.1	3.4	4.1	6.3	5.4	3.1	1.3			
Scenario 3	6	15	30	55	60	65	68	70			
CRM 1	0.9	27.8	**48.5**	21.0	1.5	0.2	0	0	0	9.1	30
# patients	4.3	7.4	9.7	6.5	1.9	0.2	0	0			
CRM 2	0.2	22.6	**60.8**	15.1	1.0	0.2	0	0	0	8.8	30
# patients	3.9	7.5	11.7	5.1	1.5	0.3	0	0			
CRM 3	0.3	19.6	**65.1**	14.4	0.6	0	0	0	0	8.4	30
# patients	4.1	7.2	13.0	5.0	0.6	0	0	0			
CRM 4	0.4	19.3	**65.6**	14.2	0.5	0	0	0	0	8.5	30
# patients	4.1	7.2	12.7	5.2	0.7	0.1	0	0			
BMA-CRM	0.3	20.6	**62.0**	16.1	0.9	0	0	0	0	8.6	30
# patients	4.1	7.2	12.2	5.6	0.8	0.1	0	0			
BMAO-CRM	0.3	19.9	**63.0**	15.9	0.8	0	0	0	0	8.6	30
# patients	4.1	7.2	12.3	5.4	0.8	0.1	0	0			
BMS-CRM	0.2	20.0	**64.9**	13.7	1.0	0.1	0	0	0	8.6	30
# patients	4.1	7.2	12.4	5.2	1.0	0.1	0	0			
Scenario 4	20	30	40	50	60	65	70	75			
CRM 1	24.8	**42.6**	20.1	8.1	0.8	0	0	0	3.6	9.0	29.3
# patients	11.3	9.0	5.4	2.8	0.7	0.1	0	0			
CRM 2	23.0	**46.7**	21.9	4.6	0.4	0	0	0	3.4	8.9	29.3
# patients	10.7	10.4	5.7	1.9	0.6	0.1	0	0			
CRM 3	22.9	**46.1**	22.4	3.9	0.2	0	0	0	4.5	8.6	29.1
# patients	11.2	10.0	6.0	1.6	0.2	0	0	0			

Table 1. (Continued)

Design	Recommendation percentage at dose level									Average toxicity	Average # patients
	1	2	3	4	5	6	7	8	None		
Scenario 4 (Continued)											
CRM 4	23.5	**46.0**	22.3	3.7	0.2	0	0	0	4.3	8.6	29.2
# patients	11.3	9.9	6.0	1.7	0.2	0	0	0			
BMA-CRM	22.3	**46.7**	21.6	4.8	0.4	0	0	0	4.2	8.7	29.2
# patients	11.2	9.9	5.8	1.9	0.3	0	0	0			
BMAO-CRM	22.4	**46.5**	21.9	4.5	0.4	0	0	0	4.2	8.7	29.2
# patients	11.1	10.0	5.8	1.9	0.3	0	0	0			
BMS-CRM	21.6	**46.9**	22.3	4.0	0.5	0.1	0	0	4.6	8.7	29.1
# patients	11.1	10.0	5.8	1.8	0.4	0	0	0			
Scenario 5	10	20	30	40	50	60	70	80			
CRM 1	2.3	23.2	**33.2**	31.5	8.8	0.8	0	0	0.2	8.6	30.0
# patients	5.5	7.1	7.4	6.5	3.1	0.4	0	0			
CRM 2	1.6	25.8	**41.9**	23.4	5.9	1.2	0.1	0	0.2	8.4	30.0
# patients	5.0	8.0	8.8	5.1	2.2	0.6	0.1	0			
CRM 3	1.6	25.4	**45.7**	22.8	4.0	0.3	0	0	0.2	8.0	30.0
# patients	5.3	8.0	9.8	5.2	1.4	0.2	0	0			
CRM 4	2.0	23.9	**45.1**	24.6	3.9	0.2	0	0	0.2	7.9	30.0
# patients	5.5	7.9	9.6	5.4	1.4	0.2	0	0			
BMA-CRM	1.7	24.4	**42.1**	25.8	5.2	0.5	0	0	0.2	8.1	29.9
# patients	5.3	7.9	9.1	5.7	1.6	0.3	0	0			
BMAO-CRM	1.6	23.8	**43.8**	24.6	5.3	0.6	0	0	0.2	8.1	30.0
# patients	5.3	7.8	9.2	5.6	1.6	0.3	0	0			
BMS-CRM	1.8	23.9	**43.6**	23.6	6.4	0.4	0.1	0	0.2	8.2	30.0
# patients	5.3	7.8	9.1	5.3	2.0	0.3	0	0			
Scenario 6	2	3	5	7	30	50	70	80			
CRM 1	0	0	0	7.3	**60.3**	30.6	1.7	0	0	7.6	30
# patients	3.2	3.0	3.1	3.8	8.6	6.9	1.3	0.1			
CRM 2	0	0	0.8	20.7	**49.7**	24.4	4.0	0.4	0	8.1	30
# patients	3.2	3.0	3.1	4.4	6.9	5.6	3.0	0.6			
CRM 3	0	0	0.1	9.9	**64.2**	24.9	0.9	0	0	6.7	30
# patients	3.2	3.0	3.2	4.9	9.6	5.3	0.7	0			
CRM 4	0	0	0.1	12.5	**62.3**	23.6	1.5	0	0	6.9	30
# patients	3.2	3.0	3.2	5.0	9.4	4.8	1.3	0.1			
BMA-CRM	0	0	0.1	10.4	**60.2**	27.9	1.5	0	0	7.2	30
# patients	3.2	3.0	3.1	4.5	8.6	6.2	1.3	0			
BMAO-CRM	0	0	0	10.4	**61.3**	26.7	1.6	0	0	7.2	30
# patients	3.2	3.0	3.1	4.5	8.4	6.7	0.9	0			
BMS-CRM	0	0	0	10.1	**62.7**	25.7	1.5	0	0	7.1	30
# patients	3.2	3.0	3.1	4.4	9.7	5.2	1.2	0.1			
Scenario 7	3	7	10	15	20	30	50	70			
CRM 1	0	0	0.4	5.2	27.3	**49.3**	16.5	1.2	0	6.1	30
# patients	3.4	3.2	3.3	4.0	6.7	6.6	2.5	0.4			
CRM 2	0	0	0.9	7.0	24.0	**38.9**	23.3	5.7	0	7.0	30
# patients	3.3	3.2	3.7	4.1	4.8	5.2	3.8	1.8			
CRM 3	0	0	1.7	10.9	33.7	**41.9**	11.6	0.1	0	5.4	30
# patients	3.4	3.2	4.1	5.5	6.7	5.4	1.7	0.1			
CRM 4	0	0	1.5	12.2	31.1	**39.2**	14.7	1.2	0	5.8	30
# patients	3.3	3.2	4.0	5.5	6.3	4.7	2.5	0.4			
BMA-CRM	0	0	1.0	7.4	27.3	**46.6**	16.2	1.3	0	6.0	30
# patients	3.4	3.2	3.7	4.9	6.0	5.7	2.8	0.3			
BMAO-CRM	0	0	1.1	7.4	26.6	**45.5**	17.6	1.9	0	6.0	30
# patients	3.4	3.2	3.7	4.9	5.7	6.1	2.6	0.4			
BMS-CRM	0	0.1	1.1	6.6	29.9	**44.4**	15.7	2.3	0	6.1	30
# patients	3.4	3.2	3.7	4.5	6.3	5.5	2.7	0.7			

Table 1. (Continued)

Design	Recommendation percentage at dose level 1	2	3	4	5	6	7	8	None	Average toxicity	Average # patients
Scenario 8	2	3	5	6	7	9	10	30			
CRM 1	0	0	0	0.2	1.2	6.9	22.1	**69.7**	0	3.2	30
# patients	3.2	3.0	3.0	3.1	3.6	4.1	4.3	5.6			
CRM 2	0	0	0	0.1	0.4	1.3	7.6	**90.6**	0	3.7	30
# patients	3.2	3.0	3.1	3.1	3.1	3.1	3.2	8.0			
CRM 3	0	0	0	0.8	4.6	12.9	34.4	**47.3**	0	2.7	30
# patients	3.2	3.0	3.2	3.5	4.0	4.4	5.2	3.3			
CRM 4	0	0	0	0.7	3.1	5.7	18.2	**72.1**	0	3.3	30
# patients	3.2	3.0	3.2	3.6	3.7	3.7	3.8	5.9			
BMA-CRM	0	0	0	0.3	1.3	4.5	19.1	**74.8**	0	3.3	30
# patients	3.2	3.0	3.1	3.3	3.5	3.7	4.1	6.0			
BMAO-CRM	0	0	0	0.3	1.2	3.6	19.0	**76.0**	0	3.3	30
# patients	3.2	3.0	3.1	3.3	3.4	3.7	4.1	6.2			
BMS-CRM	0	0	0	0.1	1.2	4.4	18.6	**75.7**	0	3.4	30
# patients	3.2	3.0	3.1	3.2	3.5	3.6	3.8	6.6			
Scenario 9	40	50	60	70	80	90	95	99			
CRM 1	37.8	4.1	0.3	0	0	0	0	0	57.8	8.6	20.3
# patients	16.6	2.8	0.7	0.2	0	0	0	0			
CRM 2	41.0	5.2	0.2	0	0	0	0	0	53.6	9.0	21.1
# patients	17.0	3.3	0.8	0.1	0	0	0	0			
CRM 3	36.3	4.3	0.2	0	0	0	0	0	59.1	8.4	20.0
# patients	16.3	2.9	0.7	0.1	0	0	0	0			
CRM 4	34.8	4.3	0.2	0	0	0	0	0	60.7	8.3	19.7
# patients	16.2	2.8	0.7	0.1	0	0	0	0			
BMA-CRM	36.7	4.4	0.2	0	0	0	0	0	58.7	8.5	20.1
# patients	16.4	2.9	0.7	0.1	0	0	0	0			
BMAO-CRM	36.1	4.5	0.2	0	0	0	0	0	59.2	8.4	19.9
# patients	16.2	2.9	0.7	0.1	0	0	0	0			
BMS-CRM	34.5	4.3	0.3	0	0	0	0	0	60.9	8.3	19.7
# patients	16.0	2.8	0.7	0.1	0	0	0	0			

with a percentage of 56.9%. In contrast, the proposed BMA-CRM had an MTD selection percentage of 51.5%, and, using Occam's window, the BMAO-CRM yielded a slightly better selection percentage than the BMA-CRM. The BMS-CRM selected the MTD 50.5% of the time. The number of patients treated at each dose was similar across all of the seven designs, except that CRM 2 treated almost twice the number of patients at dose level 8 as the other designs. Therefore, if skeleton 2 had been recommended by physicians to carry out the CRM trial design, then the eighth dose likely would have been selected as the MTD. But the eighth dose was overly toxic, with a toxicity probability of 0.5. Scenario 2 had the MTD at the sixth dose level, and the MTD selection percentage using the BMA-CRM was the second best among the seven designs. The worst skeleton corresponded to CRM 2, which yielded an MTD selection percentage of $<30\%$, whereas the proposed designs recommended the MTD approximately 40% of the time. In scenario 3, the MTD was at the third dose level. CRM 1 behaved the worst in this scenario, with an MTD selection percentage of $<50\%$, compared with MTD selection percentages of $>60\%$ for the other three single CRMs. The BMA-CRM and BMAO-CRM performed well, with MTD selection probabilities of 62.0% and 63.0%, and the BMS-CRM produced a slightly better MTD selection probability. In scenario 4, all of the selection percentages using different designs were quite close. In scenario 5, the MTD was the third dose. CRM 1 performed the worst in this scenario, with an MTD selection percentage almost 10% lower than that of the others. In scenarios 6 and 7, again the proposed BMA-CRM was very robust, with an MTD selection percentage always close to that of the best-conducted CRM. Evaluating relative model performances showed that typically one or two CRMs did not perform well. In particular, in scenario 7, CRM 1 performed the best, with an MTD selection percentage 10% greater than that of the other three individual CRMs. That good-performing CRM using skeleton 1 was the highlight of the BMA-CRM, and it lifted the MTD selection percentage of the BMA-CRM to a level close to the best. Scenario 8 is an interesting case, because the MTD is the last dose, and the performance of the four individual CRMs differed dramatically in this scenario. CRM 2 performed the best, with an MTD selection probability $> 90\%$, compared with $<50\%$ for CRM 3.

These findings demonstrate that the skeleton indeed plays a critical role in the CRM design. There was a difference of $>40\%$ in the MTD selection probability when using different skeletons in scenario 8. However, our BMA-CRM, BMAO-CRM, and BMS-CRM performed similarly and second best, with MTD selection probabilities of around 75%. In scenario 9, in which all of the doses were overly toxic, all of the designs were able to terminate the trial early because of the safety rule that we implemented.

Based on these simulations, we conclude that the proposed BMA-CRM, BMAO-CRM, and BMS-CRM methods are quite robust in terms of dose selection probabilities. These methods typically cannot perform as well as the best single CRM in the BMA set, but their performance is always quite close to that of the best single CRM and can be much better than that of the worst single CRM. Our proposed methods carry the essence of the BMA by adaptively balancing among competing models, and thus offer more reliable and robust estimates for the toxicity probabilities. Occam's window in the BMAO-CRM may help to completely eliminate a model that makes a substantially bad prediction based on the cumulating data, whereas the BMA-CRM is also able to downweight poorly fitted models. A key difference between the BMAO-CRM and other BMA methods is that under the BMAO-CRM, the model space or the number of skeletons keeps changing as the trial proceeds. At each dose assignment, Occam's window criterion may select a different set of models over which to average, based on the cumulating data. We found similar performance from our applications of the BMA-CRM, BMAO-CRM, and BMS-CRM.

More interestingly, in scenarios 2, 4, and 5 we intentionally set the true toxicity probabilities at the eight doses to be the same as those in one of the prespecified skeletons. As expected, in scenario 2, CRM 1 (with the true skeleton) gave the highest MTD selection percentage, and the MTD selection percentage of the BMA-CRM was only 2% lower. In scenario 4, CRM 4 was based on the true skeleton, and the CRM was quite robust with respect to the other three specified skeletons. The true skeleton in scenario 5 corresponded to CRM 3, which indeed yielded the best selection probability of the MTD. The proposed designs with multiple skeletons also performed well under scenario 5.

4. SENSITIVITY ANALYSIS

In the scenarios considered, CRMs with certain skeletons may outperform the others; for example, in scenario 1, CRM 1 and CRM 3 substantially outperformed the other two CRMs. For scenarios 1, 2, 3, and 5, we examined the relationship between the performance of each individual CRM and the corresponding posterior model probabilities using the BMA approach. We took 30 cohorts of size 3 and simulated 5,000 trials. For each trial, we computed the posterior model probabilities for each CRM after every cohort was sequentially accrued. Figure 1 presents the average of these posterior model probabilities over 5,000 simulations versus the accumulating number of cohorts. In scenarios 1–3, the posterior model probabilities of the four CRM models began to separate after approximately four to eight cohorts and eventually approached the stabilized val-

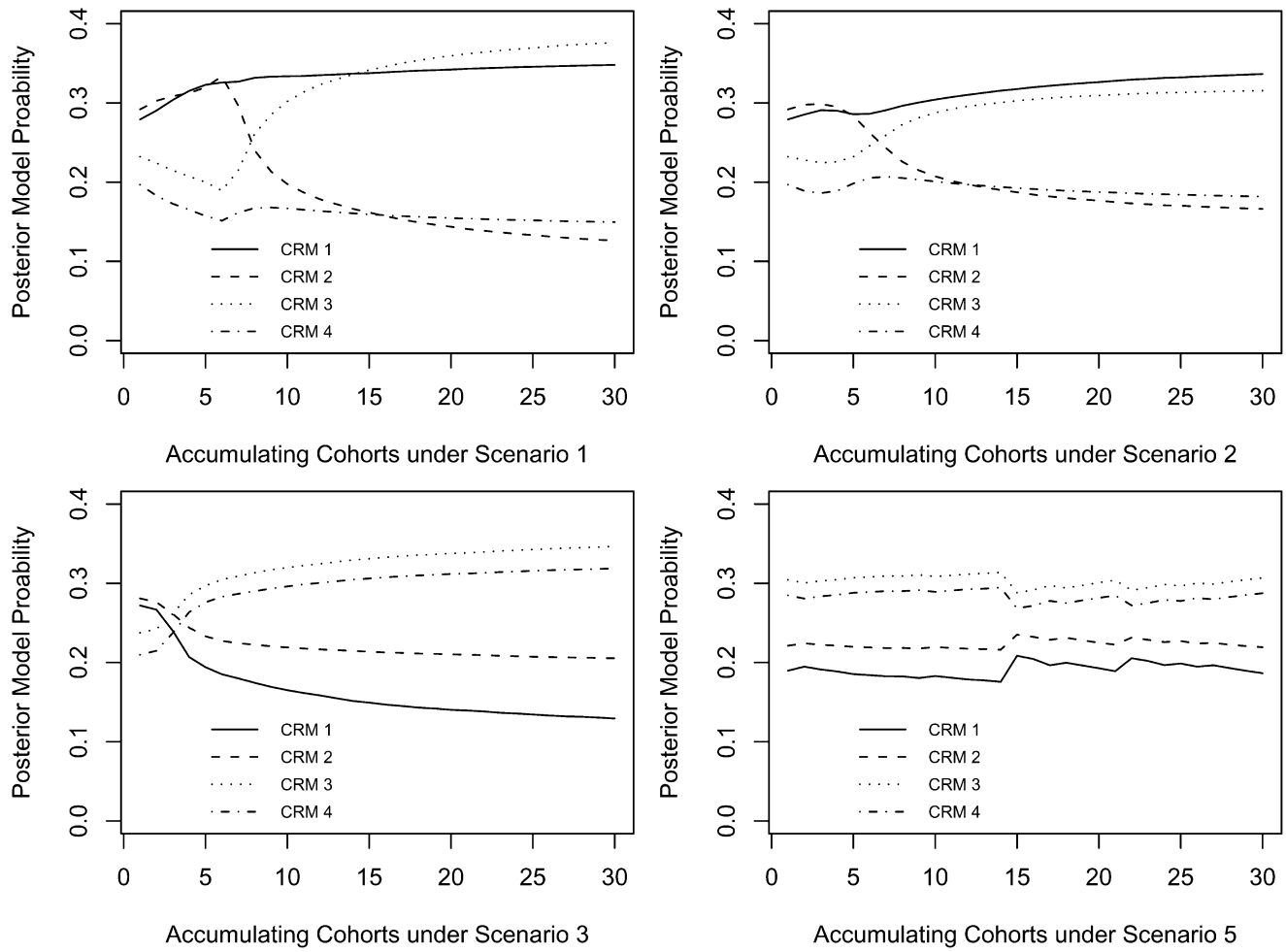

Figure 1. Posterior model probabilities of four CRMs versus the accumulating number of cohorts under scenarios 1, 2, 3, and 5.

ues in an order matching the performances of each individual CRM. This indicates that the BMA approach can indeed distinguish the model fitting as more data are collected in the trial, and thus select the better-performing CRM with a higher posterior model probability. For example, in scenario 1 the BMA exactly selected CRM 3 as the best-fitting model and CRM 2 as the worst-fitting model after approximately 15 cohorts were accrued. For scenarios 1 and 2, the true toxicity probabilities at the beginning of the dose range were extremely low and toxicity outcomes were quite rare, and thus more patients or data were needed to distinguish the model fit. In scenario 5 the order of the posterior model probabilities of the four CRMs matched their individual performance from the initiation of the trial. The BMA with Occam's window would begin to be more effective to remove the underperforming models as the posterior model probabilities stabilized. Figure 2 shows the frequency of each CRM model included in the BMA set as the trial proceeds.

In the BMA with Occam's window, we need to specify the threshold δ to ensure that only the models with an adequate fit are included. Under exactly the same setup as in scenario 1, we took $\delta = 0.5$ and 0.7 to examine its influence on the design properties. In addition, we experimented with more vague normal prior distributions for α by taking the corresponding variance as 25 and 100. Table 2 presents the simulation results, demonstrating little impact on the performance of the proposed designs in terms of both the dose selection percentage and the number of patients treated at each dose.

We also evaluated the performance of the proposed designs using different numbers of skeletons. Under scenario 5, we increased the number of skeletons from one up to six by successively adding one skeleton at a time in the original order. Table 1 presents the simulation results for the cases with one skeleton and four skeletons. Table 3 presents the selection percentage and number of patients treated at each dose when using two, three, five, and six skeletons. The fifth skeleton is (0.08, 0.15, 0.21, 0.29, 0.37, 0.44, 0.51, 0.58), and the sixth is (0.05, 0.10, 0.20, 0.25, 0.30, 0.40, 0.47, 0.55). Recall that with only the first skeleton in scenario 5, CRM 1 yielded the lowest MTD selection percentage. Adding the second skeleton slightly improved the design performance, and the proposed design with three to six skeletons produced similar results; they all increased the MTD selection percentage by approximately 10%. In practice, we recommend using three to five skeletons in the trial design, depending on the number of doses under consideration.

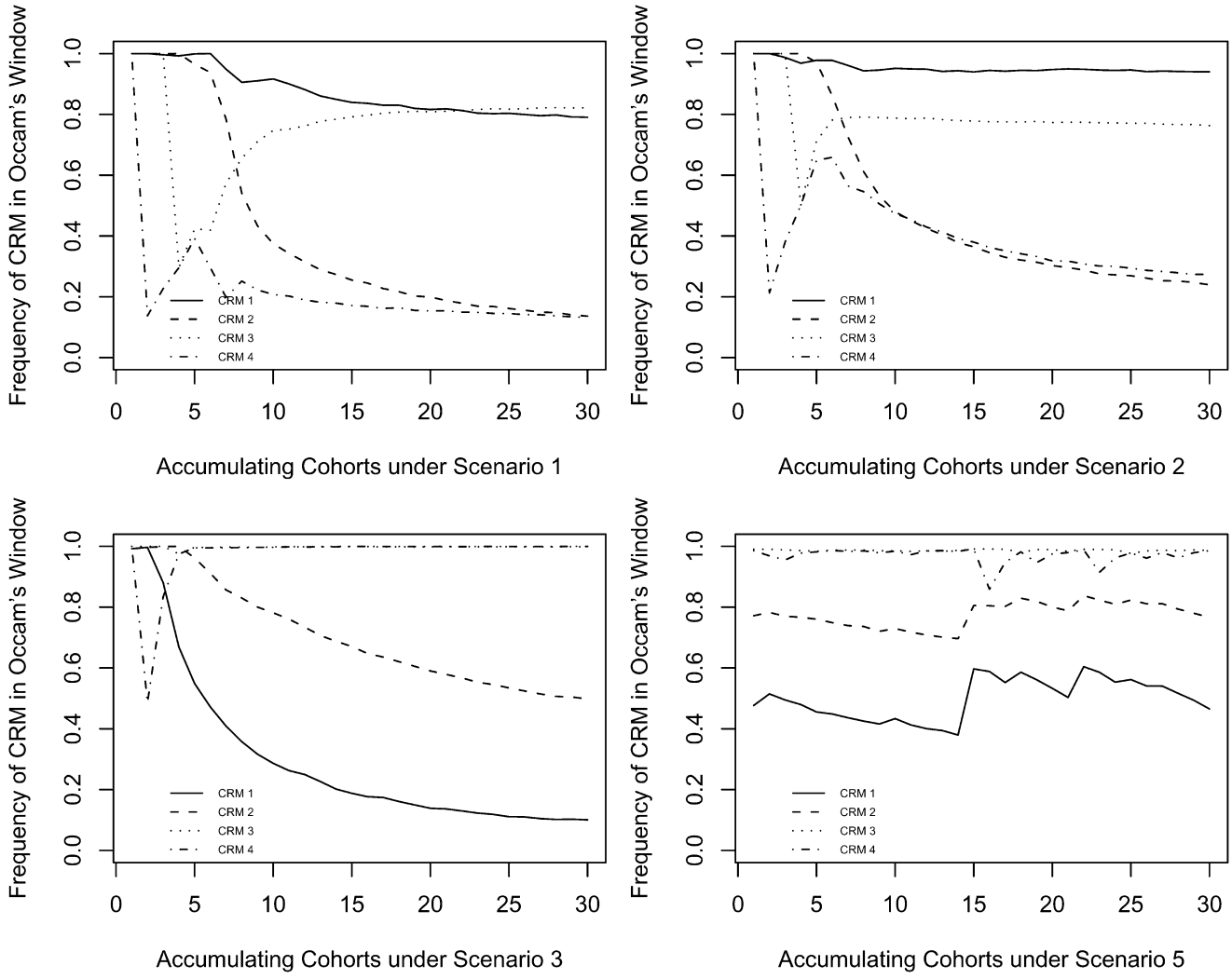

Figure 2. The frequency of each CRM model included in the BMA set using Occam's window with $\delta = 0.6$ versus the accumulating number of cohorts under scenarios 1, 2, 3, and 5.

Table 2. Sensitivity analysis of the BMA-CRM, BMAO-CRM, and BMS-CRM with different values of δ and σ under scenario 1

Design	Recommendation percentage at dose level									Average toxicity	Average # patients
	1	2	3	4	5	6	7	8	None		
Scenario 1	2	3	4	6	8	10	30	50			
	$\sigma = 2$ and $\delta = 0.5$										
BMA-CRM	0.0	0.0	0.0	0.2	1.5	16.2	**51.5**	30.6	0.0	4.7	30.0
# patients	3.2	3.0	3.1	3.3	3.5	4.4	6.3	3.2			
BMAO-CRM	0.0	0.0	0.0	0.2	1.4	15.7	**54.1**	28.6	0.0	4.7	30.0
# patients	3.2	3.0	3.1	3.3	3.5	4.5	6.1	3.3			
BMS-CRM	0.0	0.0	0.0	0.1	1.5	19.2	**50.5**	28.6	0.0	4.8	30.0
# patients	3.2	3.0	3.1	3.2	3.6	4.5	5.4	4.0			
	$\sigma = 2$ and $\delta = 0.7$										
BMA-CRM	0.0	0.0	0.0	0.2	1.5	16.2	**51.5**	30.6	0.0	4.7	30.0
# patients	3.2	3.0	3.1	3.3	3.5	4.4	6.3	3.2			
BMAO-CRM	0.0	0.0	0.0	0.2	1.4	16.7	**53.7**	28.0	0.0	4.7	30.0
# patients	3.2	3.0	3.1	3.2	3.5	4.6	5.9	3.5			
BMS-CRM	0.0	0.0	0.0	0.1	1.5	19.2	**50.5**	28.6	0.0	4.8	30.0
# patients	3.2	3.0	3.1	3.2	3.6	4.5	5.4	4.0			
	$\sigma = 5$ and $\delta = 0.6$										
BMA-CRM	0.0	0.0	0.0	0.2	1.6	14.5	**51.1**	32.5	0.0	4.8	30.0
# patients	3.2	3.0	3.1	3.2	3.5	4.2	6.1	3.6			
BMAO-CRM	0.0	0.0	0.0	0.3	1.3	14.7	**55.2**	28.5	0.0	4.7	30.0
# patients	3.2	3.0	3.1	3.3	3.4	4.2	6.4	3.3			
BMS-CRM	0.0	0.0	0.0	0.1	1.5	16.9	**54.2**	27.4	0.0	4.7	30.0
# patients	3.2	3.0	3.1	3.2	3.6	4.5	5.9	3.5			
	$\sigma = 10$ and $\delta = 0.6$										
BMA-CRM	0.0	0.0	0.0	0.2	1.5	14.2	**51.4**	32.6	0.0	4.8	30.0
# patients	3.2	3.0	3.1	3.2	3.5	4.2	6.1	3.6			
BMAO-CRM	0.0	0.0	0.0	0.3	1.4	15.0	**54.1**	29.2	0.0	4.8	30.0
# patients	3.2	3.0	3.1	3.3	3.5	4.2	6.4	3.3			
BMS-CRM	0.0	0.0	0.0	0.1	1.5	16.9	**55.4**	26.0	0.0	4.8	30.0
# patients	3.2	3.0	3.1	3.2	3.6	4.5	5.8	3.6			

To further examine the robustness of the proposed designs, we simulated one true toxicity scenario with six dose levels. The target toxicity rate was $\phi = 40\%$, and the true toxicity probabilities were generated from the inverse cumulative distribution function of a normal distribution. We took three different skeletons for each case, while keeping the true dose–toxicity curve the same. We used different sets of skeletons to reflect the changes in the model averaging set. As shown in Figure 3, we conducted four simulations, each with three different skeletons. Our goal was to examine whether different sets of skeletons would have a substantial impact on the performance of the proposed designs. From the simulation results summarized in Table 4, we can see that for the first and fourth sets of skeletons, the proposed BMA-, BMAO-, and BMS-CRM designs performed quite similarly, with each demonstrating an MTD selection percentage of around 70%. For the second and third sets of skeletons, the proposed designs yielded MTD selection percentages of around 63%. There was not much difference in the performance of the four trial designs, and they all correctly selected the MTD with the highest percentages, demonstrating the robustness of the proposed designs.

Of the three proposed methods, we recommend the BMA-CRM for practical use, because it is simple and coherent in the Bayesian framework. The BMAO-CRM requires specification of the δ value for Occam's window, which may be subjective. As for the BMS-CRM, model selection based on small samples may not be reliable, especially at the beginning of the trial when very few patients have been accrued. Although different skeletons may lead to quite different results, in general the original CRM of O'Quigley, Pepe, and Fisher (1990) is quite robust, and most skeletons will perform reasonably well under that method. If the local fit and prediction of the toxicity probabilities are reasonable, the CRM is usually able to identify the MTD accurately.

5. APPLICATION

We illustrate the proposed designs using a pediatric phase I clinical trial (Jakacki et al. 2008) that aimed to determine the MTD of erlotinib in children with refractory solid tumors. Erlotinib is an oral inhibitor of the epidermal growth factor receptor signal pathway that has been approved by the Food and Drug Administration for adults with recurrent non–small cell lung cancer and advanced pancreatic cancer. This clinical trial studied five dose levels of erlotinib: 35, 50, 65, 85, and

Table 3. Simulation study comparing the BMA-CRM, BMAO-CRM, and BMS-CRM with two, three, five, and six skeletons under scenario 5

Design	Recommendation percentage at dose level									Average toxicity	Average # patients
	1	2	3	4	5	6	7	8	None		
Scenario 5	10	20	30	40	50	60	70	80			
					Two skeletons						
BMA-CRM	1.7	24.7	**37.6**	26.5	8.3	1.0	0	0	0.2	8.5	30.0
# patients	5.2	7.6	8.4	5.5	2.7	0.6	0.1	0			
BMAO-CRM	1.7	25.0	**37.5**	26.3	8.5	0.8	0.1	0	0.2	8.5	30.0
# patients	5.1	7.6	8.4	5.4	2.7	0.6	0.1	0			
BMS-CRM	1.3	25.3	**37.6**	26.6	8.2	0.9	0	0	0.2	8.5	30.0
# patients	5.0	7.7	8.4	5.5	2.7	0.6	0.1	0			
					Three skeletons						
BMA-CRM	1.6	23.4	**42.0**	25.9	6.1	0.8	0	0	0.2	8.2	30.0
# patients	5.4	7.7	8.8	5.6	1.9	0.4	0	0			
BMAO-CRM	1.6	24.2	**41.1**	25.9	6.6	0.6	0.1	0	0.2	8.2	30.0
# patients	5.3	7.7	9.0	5.6	1.8	0.4	0	0			
BMS-CRM	1.6	25.4	**42.6**	23.6	6.1	0.6	0.1	0	0.2	8.2	30.0
# patients	5.3	8.0	9.0	5.2	2.0	0.3	0	0			
					Five skeletons						
BMA-CRM	1.8	23.2	**42.9**	25.6	5.7	0.7	0	0	0.2	8.1	30.0
# patients	5.4	7.7	9.0	5.6	1.8	0.3	0	0			
BMAO-CRM	1.7	23.8	**42.3**	25.6	5.8	0.5	0	0	0.2	8.2	30.0
# patients	5.3	7.7	9.2	5.6	1.8	0.3	0	0			
BMS-CRM	1.5	25.0	**43.1**	23.7	6.0	0.5	0	0	0.2	8.2	30.0
# patients	5.3	8.0	9.1	5.3	1.9	0.3	0	0			
					Six skeletons						
BMA-CRM	1.6	24.4	**41.8**	25.3	6.0	0.6	0	0	0.2	8.2	30.0
# patients	1.6	24.4	41.8	25.3	6.0	0.6	0	0			
BMAO-CRM	1.5	24.2	**42.3**	25.3	5.9	0.7	0	0	0.1	8.2	30.0
# patients	5.3	7.8	9.2	5.5	1.7	0.4	0	0			
BMS-CRM	1.6	24.2	**42.9**	23.9	6.6	0.6	0	0	0.2	8.2	30.0
# patients	5.4	7.8	8.9	5.1	2.3	0.4	0	0			

110 mg/m^2/day. A total of 19 assessable patients were used for dose escalation. DLT determination included any grade 3 or 4 thrombocytopenia or grade 4 neutropenia, or any grade 3 or 4 nonhematologic toxicity. We took the MTD as the dose with a DLT rate of 20% and elicited three different skeletons in the CRM,

$$(p_1, p_2, p_3, p_4, p_5) = \begin{cases} (0.20, 0.40, 0.60, 0.70, 0.80), & \text{skeleton 1} \\ (0.05, 0.10, 0.20, 0.30, 0.40), & \text{skeleton 2} \\ (0.01, 0.05, 0.10, 0.15, 0.20), & \text{skeleton 3.} \end{cases}$$

These three skeletons represent different prior opinions on the dose-response curve, from a toxicity increasing the most aggressively to that increasing the least aggressively with dose. We applied the BMA-CRM, BMAO-CRM, and BMS-CRM designs to the trial conduct and for comparison also conducted the CRM design under each of the three skeletons separately, designated CRM 1, CRM 2, and CRM 3.

Table 5 shows the path of dose escalation, the posterior mean of the power parameter α, and the selected MTD. The dose assignment followed exactly the same scheme under each of the six designs. The trial started with treating the first cohort of three patients at the lowest dose, 35 mg/m^2/day. Because no DLT was observed, the dose was escalated to 50 mg/m^2/day for the second cohort. Again, no DLT was observed, so the dose was escalated to 65 mg/m^2/day; still no DLT occurred. The dose was then escalated to 85 mg/m^2/day, at which point one of six patients experienced DLT. The trial ended by treating the last cohort of patients at a dose of 110 mg/m^2/day; this dose, two of four patients experienced DLT.

Although the paths of dose escalation were the same for all of these designs, the MTD selection differed slightly. CRM 1 selected dose 65 as the MTD, whereas both CRM 2 and CRM 3 selected dose 85 as the MTD. The inconsistency of MTD identification demonstrates the sensitivity of the CRM to the specification of the skeleton and its limitation when only a single skeleton is used in the trial. In contrast, all proposed designs BMA-CRM, BMAO-CRM, and BMS-CRM selected dose 85 as the MTD, which is consistent with the MTD identified by Jakacki et al. (2008) based on the "3 + 3" design. Figure 4 shows the posterior probabilities of CRMs 1–3 under the BMA-CRM and BMAO-CRM during the trial. For the BMA-CRM, after no DLT was observed in the first cohort, our dose-finding procedure correctly recognized that CRM 1 with the most ag-

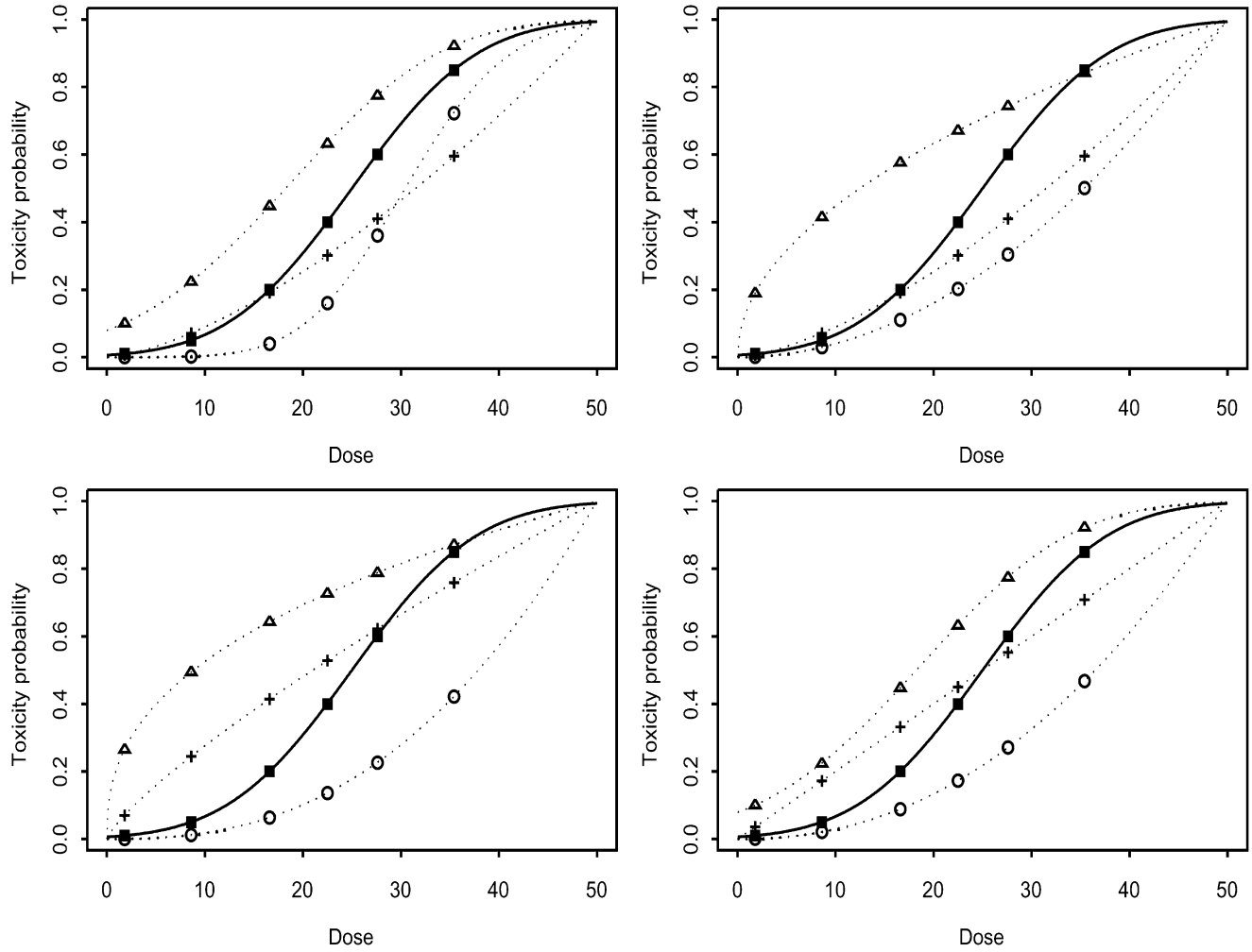

Figure 3. Robust analysis with six dose levels and four sets of three skeletons. The solid line is the true dose–toxicity curve, and the dotted lines are the dose–toxicity curves for the prespecified toxicity probabilities. The line with triangles represents skeleton 1; the line with crosses, skeleton 2; and the line with circles, skeleton 3.

gressive skeleton (skeleton 1) was the least supported by the observed data, as reflected by the smallest posterior model probability. Because no DLT was observed in cohorts 2 and 3, the posterior model probability of CRM model 1 kept decreasing. At this stage, both CRM models 2 and 3 were reasonably supported by the data, and their posterior model probabilities were comparable. After one DLT occurred in cohort 4 and two of four patients experienced DLTs in cohort 5, the data began to show more support to CRM 1, and its posterior model probability substantially increased. Similar patterns were observed in the BMAO-CRM design, except that the model with a negligible posterior probability was dropped from the Bayesian model averaging set by Occam's window. For example, model 1 was dropped after no DLTs were observed in cohort 2, but it was later brought back into the model-averaging set until cohort 5, when more DLTs were observed. The behavior of the BMS-CRM was consistent with that of the BMA-CRM as well. For the first three cohorts, CRM 3 fit the data the best. After one DLT was observed in cohort 4, CRM 2, with a slightly more aggressive skeleton (skeleton 2), became the best model; this was also true for cohort 5.

During the trial, the posterior estimate of the power parameter α was continuously updated to reflect the accumulating data. Because each individual CRM (1, 2, and 3) used rather different skeletons, $\hat{\alpha}$ differed dramatically across these methods, particularly at the end of the trial when all of the data were available. In contrast, the $\hat{\alpha}$'s using the BMA-based designs were much more stable.

In another illustrative example, we applied the proposed designs to a phase I prostate cancer clinical trial conducted at M. D. Anderson Cancer Center (Mathew et al. 2004). The goal was to find the MTD of docetaxel used in combination with daily 600 mg imatinib with a targeted DLT rate of 30%. Patients were treated in cohorts of six, and as many as eight cohorts were possible. Six potential doses of docetaxel were investigated: 20, 25, 30, 35, 40, and 45 mg/m^2 weekly for 4 weeks every 6 weeks. Three skeletons were elicited,

$$(p_1, p_2, p_3, p_4, p_5, p_6) = \begin{cases} (0.30, 0.40, 0.50, 0.60, 0.70, 0.80), & \text{skeleton 1} \\ (0.07, 0.16, 0.30, 0.40, 0.46, 0.53), & \text{skeleton 2} \\ (0.01, 0.05, 0.10, 0.15, 0.20, 0.30), & \text{skeleton 3,} \end{cases}$$

representing different prior opinions on the location of the MTD. Skeleton 2 is the skeleton used in the original CRM.

Table 4. Robust analysis based on four different sets, each with three skeletons, comparing the BMA-CRM, BMAO-CRM, and BMS-CRM with a toxicity target of $\phi = 40\%$

Design	Recommendation percentage at dose level							Average toxicity	Average # patients
	1	2	3	4	5	6	None		
Toxicity prob.	1	5	20	40	60	85			
BMA-CRM	0	0	11.3	**71.5**	17.2	0	0	9.5	30
# patients	3.0	3.1	6.0	13.1	4.7	0			
BMAO-CRM	0	0	11.4	**70.8**	17.7	0	0	9.6	30
# patients	3.0	3.1	6.0	12.9	4.8	0.1			
BMS-CRM	0	0	12.4	**68.5**	19.0	0.1	0	9.6	30
# patients	3.0	3.1	6.6	11.6	5.5	0.2			
BMA-CRM	0	0	13.0	**63.7**	23.2	0.1	0	10.1	30
# patients	3.1	3.1	5.9	10.9	6.6	0.4			
BMAO-CRM	0	0	13.4	**62.8**	23.5	0.2	0	10.1	30
# patients	3.1	3.1	5.9	10.8	6.6	0.5			
BMS-CRM	0	0	12.7	**63.4**	23.6	0.2	0	10.2	30
# patients	3.1	3.2	5.5	11.1	6.5	0.7			
BMA-CRM	0	0	13.5	**64.7**	21.7	0.1	0	10.0	30
# patients	3.1	3.1	6.0	10.9	6.4	0.4			
BMAO-CRM	0	0	13.7	**64.5**	21.6	0.2	0	10.0	30
# patients	3.1	3.1	6.0	10.8	6.4	0.5			
BMS-CRM	0	0	13.9	**64.6**	21.2	0.2	0	10.0	30
# patients	3.1	3.1	6.2	10.8	6.1	0.7			
BMA-CRM	0	0	11.6	**69.1**	19.3	0	0	9.6	30
# patients	3.0	3.1	6.1	12.6	5.0	0.2			
BMAO-CRM	0	0	12.0	**70.4**	17.6	0.1	0	9.5	30
# patients	3.1	3.1	6.2	12.8	4.6	0.2			
BMS-CRM	0	0	12.6	**70.6**	16.6	0.1	0	9.5	30
# patients	3.0	3.1	6.6	12.3	4.5	0.4			

Table 5. Application of the proposed designs to the phase I pediatric solid tumor trial with erlotinib

Method		Cohort sequence					Selected MTD
		1	2	3	4	5	
CRM 1	Dose ($mg/m^2/day$)	35	50	65	85	110	65
	# of tox/# of pts	0/3	0/3	0/3	1/6	2/4	
	$\hat{\alpha}$	0.69	1.07	1.36	1.34	1.06	
CRM 2	Dose ($mg/m^2/day$)	35	50	65	85	110	85
	# of tox/# of pts	0/3	0/3	0/3	1/6	2/4	
	$\hat{\alpha}$	0.44	0.69	0.89	0.56	0.31	
CRM 3	Dose ($mg/m^2/day$)	35	50	65	85	110	85
	# of tox/# of pts	0/3	0/3	0/3	1/6	2/4	
	$\hat{\alpha}$	0.28	0.54	0.72	0.17	−0.13	
BMA-CRM	Dose ($mg/m^2/day$)	35	50	65	85	110	85
	# of tox/# of pts	0/3	0/3	0/3	1/6	2/4	
	$\hat{\alpha}$	0.44	0.70	0.87	0.60	0.59	
BMAO-CRM	Dose ($mg/m^2/day$)	35	50	65	85	110	85
	# of tox/# of pts	0/3	0/3	0/3	1/6	2/4	
	$\hat{\alpha}$	0.44	0.61	0.79	0.37	0.59	
BMS-CRM	Dose ($mg/m^2/day$)	35	50	65	85	110	85
	# of tox/# of pts	0/3	0/3	0/3	1/6	2/4	
	$\hat{\alpha}$	0.28	0.54	0.72	0.56	0.31	

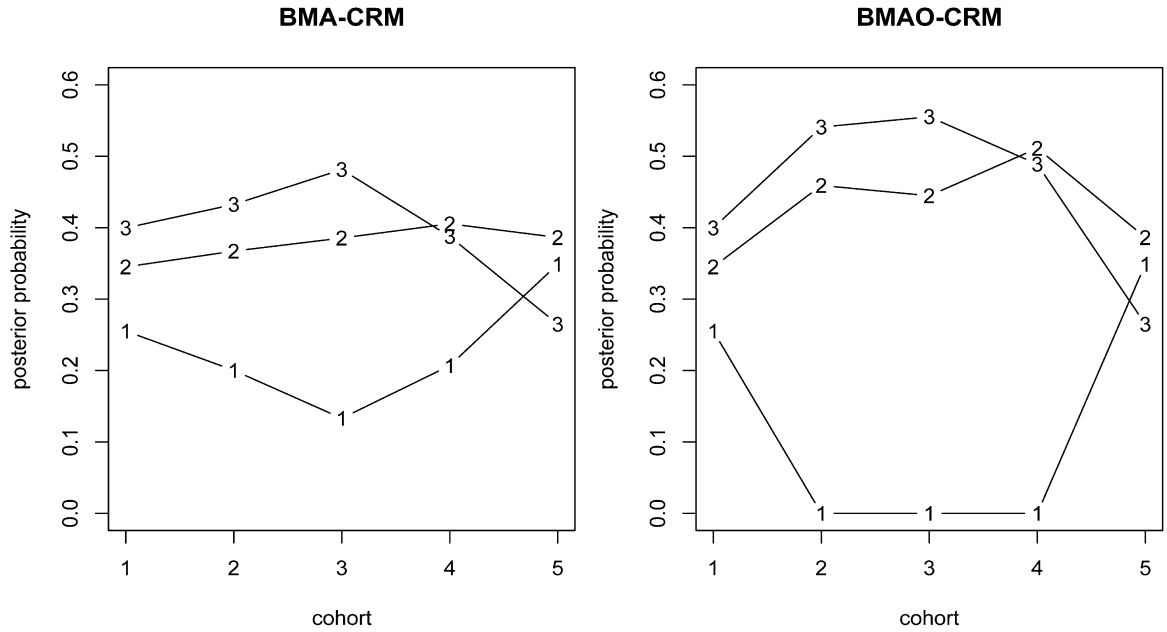

Figure 4. Posterior model probabilities under BMA-CRM and BMAO-CRM for the pediatric phase I clinical trial with erlotinib. Symbols 1–3 denote CRMs 1–3 in the Bayesian model-averaging set.

Based on these skeletons, we applied the BMA-CRM, BMAO-CRM, and BMS-CRM to locate the MTD.

Table 6 presents the dose-finding results of the trial under the proposed BMA-CRM and the CRM used by Mathew et al. (2004). Using the BMAO-CRM and BMS-CRM led to the same results as those obtained from the BMA-CRM. The main difference between the BMA-CRM and CRM was that after no DLT was observed in the first cohort at dose 30 mg/m^2, the CRM took an aggressive dose escalation to 45 by skipping two intermediate dose levels for the second cohort, at which three of four patients experienced DLTs, whereas the BMA-CRM escalated the dose to a more appropriate level of 35 mg/m^2. Although both designs eventually deescalated to a dose of 30 mg/m^2 and subsequently selected it as the MTD, the BMA-CRM avoided exposing patients to the overly toxic dose of 45 mg/m^2 and required fewer cohorts. These two illustrative trials demonstrate that the CRM may be sensitive to the specified skeleton in practice and that the proposed BMA-based designs are more robust and reliable.

6. CONCLUSION

We have proposed a new dose-finding algorithm based on Bayesian model averaging and the original CRM, using multiple sets of prespecified toxicity probabilities. The performance of the proposed designs can be substantially improved over that of the original CRM if the skeleton in the CRM happens to be very far from the true model. The BMA-CRM method is straightforward to implement and very easy to compute based on the Gaussian quadrature approximation or the Markov chain Monte Carlo procedure. This method requires specifying multiple skeletons to cover different potential scenarios for the underlying dose–toxicity curve. It provides a nice compromise for the initial guesses of toxicity probabilities from different physicians. If one skeleton corresponds to the true toxicity probabilities, then the BMA-CRM would perform very well, because it often performs similarly to the best-performing CRM. In practice, as long as one skeleton in the BMA set leads to a well-behaved CRM, then the performance of the BMA-CRM will be close to that of the CRM. This Bayesian model-averaging procedure dramatically improves the robustness of the CRM. As shown in the simulations, a certain skeleton often yields underperforming results; however, simultaneously specifying multiple skeletons reduces the likelihood of all sets of toxicity probabilities leading to a poorly performing CRM design. The arbitrariness in the specification of the skeleton is eliminated by incorporating the uncertainties associated with each skeleton into the Bayesian model-averaging procedure.

Table 6. Application of the proposed designs to the phase I prostate cancer trial

Method		Cohort sequence 1	2	3	4	Selected **MTD**
CRM	Dose (mg/m^2)	30	45	35	30	30
	# of tox/# of pts	0/6	3/4	5/6	3/6	
BMA-CRM	Dose (mg/m^2)	30	35	30		30
	# of tox/# of pts	0/6	5/6	3/6		

In our numerical studies we focused on the power model of the CRM because of its simplicity; other model structures can be used as well, such as the one-parameter logistic model or the parabolic function. As a referee mentioned, the power model $p_j^{\exp(\alpha)}$ is exactly equivalent to $(p_j^w)^{\exp(\alpha)}$ for $w > 0$. Therefore, we need to take precautions to propose "reasonable" skeletons in the BMA-CRM, because the skeleton with $p_j^{w_1}$ and that with $p_j^{w_2}$ $(w_1 \neq w_2)$ are redundant in the BMA set. Moreover, the spacing between the adjacent p_j's is more critical than the values of the p_j's themselves. In our simulations we used a cohort size of three; however, cohort sizes of one or two also could be used. Our setup is based on the improved versions of the

CRM to optimize its practical performance. As an extension of the CRM, the BMA-CRM makes this trial design more widely applicable and reliable for phase I clinical trials.

[Received August 2008. Revised February 2009.]

REFERENCES

Babb, J., Rogatko, A., and Zacks, S. (1998), "Cancer Phase I Clinical Trials: Efficient Dose Escalation With Overdose Control," *Statistics in Medicine*, 17, 1103–1120.

Braun, T. M. (2002), "The Bivariate Continual Reassessment Method: Extending the CRM to Phase I Trials of Two Competing Outcomes," *Controlled Clinical Trials*, 23, 240–256.

Chevret, S. (2006), *Statistical Methods for Dose-Finding Experiments*, New York: Wiley.

Durham, S. D., Flournoy, N., and Rosenberger, W. F. (1997), "A Random Walk Rule for Phase I Clinical Trials," *Biometrics*, 53, 745–760.

Faries, D. (1994), "Practical Modification of the Continual Reassessment Methods for Phase I Cancer Clinical Trials," *Journal of Biophamaceutical Statistics*, 4, 147–164.

Garrett-Mayer, E. (2006), "The Continual Reassessment Method for Dose-Finding Studies: A Tutorial," *Clinical Trials*, 3, 57–71.

Gasparini, M., and Eisele, J. (2000), "A Curve-Free Method for Phase I Clinical Trials," *Biometrics*, 56, 609–615.

Goodman, S. N., Zahurak, M. L., and Piantadosi, S. (1995), "Some Practical Improvements in the Continual Reassessment Method for Phase I Studies," *Statistics in Medicine*, 14, 1149–1161.

Heyd, J. M., and Carlin, P. B. (1999), "Adaptive Design Improvements in the Continual Reassessment Method for Phase I Studies," *Statistics in Medicine*, 18, 1307–1321.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999), "Bayesian Model Averaging: A Tutorial," *Statistical Science*, 14, 382–401.

Ishizuka, N., and Ohashi, Y. (2001), "The Continual Reassessment Method and Its Applications: A Bayesian Methodology for Phase I Cancer Clinical Trials," *Statistics in Medicine*, 20, 2661–2681.

Jakacki, R. I., Hamilton, M., Gilbertson, J. R., Blaney, S. M., Tersak, J., Krailo, M. D., Ingle, A. M., Voss, S. D., Dancey, J. E., and Adamson, P. C. (2008), "Pediatric Phase I and Pharmacokinetic Study of Erlotinib Followed by the Combination of Erlotinib and Temozolomide: A Children's Oncology Group Phase I Consortium Study," *Journal of Clinical Oncology*, 26, 4921–4927.

Leung, D. H.-Y., and Wang, Y.-G. (2002), "An Extension of the Continual Reassessment Method Using Decision Theory," *Statistics in Medicine*, 21, 51–63.

Madigan, D., and Raftery, A. E. (1994), "Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window," *Journal of the American Statistical Association*, 89, 1535–1546.

Mathew, P., Thall, P., Jones, D., Perez, C., Bucana, C., Troncoso, P., Kim, S., Fidler, I. J., and Logothetis, C. (2004), "Platelet-Derived Growth Factor Receptor Inhibitor Imatinib Mesylate and Docetaxel: A Modular Phase I Trial in Androgen-Independent Prostate Cancer," *Journal of Clinical Oncology*, 22, 3323–3329.

Møller, S. (1995), "An Extension of the Continual Reassessment Methods Using a Preliminary Up-and-Down Design in a Dose Finding Study in Cancer Patients, in Order to Investigate a Greater Range of Doses," *Statistics in Medicine*, 14, 911–922.

O'Quigley, J. (2002), Comment on "A Curve-Free Method for Phase I Clinical Trials," by M. Gasparini and J. Eisele, *Biometrics*, 58, 245–249.

O'Quigley, J., and Chevret, S. (1991), "Methods for Dose Finding Studies in Cancer Clinical Trials: A Review and Results of a Monte Carlo Study," *Statistics in Medicine*, 10, 1647–1664.

O'Quigley, J., and Shen, L. Z. (1996), "Continual Reassessment Method: A Likelihood Approach," *Biometrics*, 52, 673–684.

O'Quigley, J., Pepe, M., and Fisher, L. (1990), "Continual Reassessment Method: A Practical Design for Phase 1 Clinical Trials in Cancer," *Biometrics*, 46, 33–48.

Piantadosi, S., Fisher, J., and Grossman, S. (1998), "Practical Implementation of a Modified Continual Reassessment Method for Dose Finding Trials," *Cancer Chemotherapy and Pharmacology*, 41, 429–436.

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997), "Bayesian Model Averaging for Linear Regression Models," *Journal of the American Statistical Association*, 92, 179–191.

Shen, L., and O'Quigley, J. (1996), "Consistency of Continual Reassessment Method Under Model Misspecification," *Biometrika*, 83, 395–405.

Storer, B. E. (1989), "Design and Analysis of Phase I Clinical Trials," *Biometrics*, 45, 925–937.

Stylianou, M., and Flournoy, N. (2002), "Dose Finding Using the Biased Coin Up-and-Down Design and Isotonic Regression," *Biometrics*, 58, 171–177.

Ting, N. (2006), *Dose Finding in Drug Development*, Cambridge, MA: Springer.

Whitehead, J., and Brunier, H. (1995), "Bayesian Decision Procedures for Dose Determining Experiments," *Statistics in Medicine*, 14, 885–893.

Yuan, Z., Chappell, R., and Bailey, H. (2007), "The Continual Reassessment Method for Multiple Toxicity Grades: A Bayesian Quasi-Likelihood Approach," *Biometrics*, 63, 173–179.