Biostatistics (2021) **22**, 1, *pp.* 19–34
doi:10.1093/biostatistics/kxz014
Advance Access publication on May 20, 2019

Bayesian adaptive basket trial design using model averaging

MATTHEW A. PSIODA*, JIAWEI XU

Department of Biostatistics, University of North Carolina, McGavran-Greenberg Hall, CB#7420,
Chapel Hill, North Carolina 27599, USA
matt_psioda@unc.edu

QI JIANG

Seattle Genetics, 21717–30th Drive S.E., Building 3, Bothell, WA 98021, USA

CHUNLEI KE

Biogen, 300 Binney St, Cambridge, MA 02142, USA

ZHAO YANG

Amgen Inc., One Amgen Center Drive, 24-1-B, Thousand Oaks, CA 91320, USA

JOSEPH G. IBRAHIM

Department of Biostatistics, University of North Carolina, McGavran-Greenberg Hall, CB#7420,
Chapel Hill, North Carolina 27599, USA

SUMMARY

In this article, we develop a Bayesian adaptive design methodology for oncology basket trials with binary endpoints using a Bayesian model averaging framework. Most existing methods seek to borrow information based on the degree of homogeneity of estimated response rates across all baskets. In reality, an investigational product may only demonstrate activity for a subset of baskets, and the degree of activity may vary across the subset. A key benefit of our Bayesian model averaging approach is that it explicitly accounts for the possibility that any subset of baskets may have similar activity and that some may not. Our proposed approach performs inference on the basket-specific response rates by averaging over the complete model space for the response rates, which can include thousands of models. We present results that demonstrate that this computationally feasible Bayesian approach performs favorably compared to existing state-of-the-art approaches, even when held to stringent requirements regarding false positive rates.

Keywords: Adaptive design; Basket trials; Bayesian model averaging; Clinical trial design.

*To whom correspondence should be addressed.

1. Introduction

Oncology drug development has traditionally focused on the histology of cancer as this was the key known determinant for whether a tumor would respond to a given treatment. With the advent of genomic technologies that allow for the characterization of specific genomic alterations within the tumor (e.g., mutations), the focus has broadened to include developing targeted therapies for specific alterations (Redig and Jänne, 2015). For an investigational drug which targets a specific genomic alteration (e.g., BRAF V600E mutation), it may be expected that the drug will show activity for multiple tumor histologies as long as the alteration is present. As such, standard enrichment designs (Simon and Maitournam, 2004; Maitournam and Simon, 2005; Mandrekar and Sargent, 2011) which focus on a particular histology and screen for the genomic alteration are not ideally suited for this scenario. Increasingly, so-called *basket trials* are being considered for this purpose. Basket trials are usually nonrandomized, focus on multiple cancer histologies, and include patients with a specified genomic alteration who then receive a regimen to which their tumors are expected to be responsive based on prior information (Simon, 2017). Often these trials use binary Response Evaluation Criteria in Solid Tumors (RECIST) endpoints, reflecting a compromise between clinical relevance and the need to have timely information on treatment activity.

In the simplest case, a basket trial can be used to evaluate the activity of a single investigational drug on multiple tumor histologies that possess a common genomic alteration (i.e., baskets). One of the key goals of basket trials is to determine the subset of baskets for which the investigational product has activity (i.e., patients treated with the product have a desirable probability of response). For example, Hyman *and others* (2015) conducted a basket trial for Vemurafenib, a selective oral inhibitor of the BRAF V600 kinase that included six pre-specified non-melanoma BRAF V600E mutation-positive cancers: nonâŁ“small-cell lung cancer (NSCLC), ovarian cancer, colorectal cancer, cholangiocarcinoma, breast cancer, and multiple myeloma. Vemurafenib had previously received Food and Drug Administration (FDA) approval for the treatment of BRAF V600E mutation-positive metastatic melanoma and so melanomas were not included in the trial. Based on a RECIST endpoint, the basket trial demonstrated that Vemurafenib was active in NSCLC and several other histologies, but not in colorectal cancer.

In this article, we propose a flexible Bayesian adaptive basket trial design methodology that accommodates early stopping of individual baskets due to inactivity (futility) or activity (efficacy) using the Bayesian model averaging (BMA) framework (Madigan and Raftery, 1994; Raftery, 1995; Draper, 1995). For a thorough tutorial on BMA, see Hoeting *and others* (1999). Bayesian model averaging is naturally suited for the design of basket trials wherein one expects many or all of the baskets to respond similarly to treatment based on the common genomic alteration targeted by the investigational product. The key benefit of the BMA approach over many existing methods is that it naturally allows for information borrowing over any subset of baskets that have similar activity.

Basket trial design is an active topic for statistical methods research. Much of the innovative work to date was recently summarized in a review article by Simon (2017). We briefly discuss the relevant literature here. One of the most popular approaches for phase II oncology trials has been to use Simon's Two-Stage Design (Simon, 1989). When applied in the basket trial setting, one essentially performs separate two-stage trials for each basket. This type of approach will be inefficient when the drug is inactive in all baskets or has similar activity in a subset of baskets. Cunanan *and others* (2017) proposed a frequentist approach that improves efficiency of parallel two-stage evaluation of each histology by assessing the homogeneity of the response rates across baskets at an interim analysis using a calibrated Fisher's exact test based on the contingency table of response and non-response counts for the baskets. Their method allows for early stopping for inactivity at the end of stage one. The second stage of the design either pools all the data or analyzes the data for each basket separately with a Bonferroni-type multiplicity adjustment. Simon *and others* (2016) proposed a Bayesian approach that performs inference for each basket by averaging results from two competing models: (1) a model that assumes that all baskets have different response rates and

(2) a model that assumes all baskets have the same response rate. Whereas the method by Simon *and others* averages inference from models (1) and (2), the method by Cunanan *and others* uses interim data to choose between them. In either case, the explicit focus on these two models is likely a concession made to achieve a more feasible design method to implement in practice. A core tenet guiding the development of the proposed method is that for most basket trials the underlying truth will fall somewhere between the two extremes defined by (1) and (2) and therefore it is desirable to have a method that explicitly accounts for the myriad of possibilities which may occur.

Several basket trial design methods have been proposed that make use of Bayesian hierarchical models (BHM). For example, Thall *and others* (2003), Berry *and others* (2013), Liu *and others* (2017), and Chu and Yuan (2018) all propose design methods based on variations of the BHM. The approach taken by Thall *and others* and Berry *and others* is a traditional Bayesian adaptive design with multiple analyses at which it is possible to stop enrollment for individual baskets due to their having demonstrated activity or a high likelihood of inactivity. Berry *and others* propose using a weakly informative hierarchical prior to let the data determine the degree of borrowing as much as possible. Chu and Yuan (2018) instead proposed a supervised or *calibrated* BHM (CBHM), where the variance parameter in the hyperprior is not specified *a priori* as in a fully Bayesian approach but instead is constructed as a log-linear function of a Pearson chi-square test statistic based on the observed data at the time of the analysis. Their approach allows for early stoppage of enrollment in individual baskets for inactivity but not for demonstrated activity.

The rest of this article is organized as follows. In Section 2, we provide a motivating discussion for the BMA approach. In Section 3, we develop the BMA design methodology in detail. In Section 4, we present simulation studies that compare our BMA design approach with several recently proposed design methods for basket trials, focusing primarily on two-stage designs. We close the article with some discussion in Section 5.

2. Motivation for the Bayesian model averaging approach

The primary purpose of the proposed trial design methodology is to allow practitioners to evaluate whether an investigational treatment improves binary response rates in each of K distinct baskets using an inference procedure that permits borrowing information across subsets of baskets to the extent that such borrowing is reasonable based on the observed data. We assume throughout the article that all patients within a basket are independent and that they share a common probability of response when administered the investigational treatment. Thus, the data for each basket will result in a binomial likelihood indexed by a basket-specific probability of response. Henceforth, we will use the term *response rate* as a synonym for *probability of response*.

For illustration, consider the case where there are $K = 3$ baskets. In this setting, the simplest model (i.e., the most parsimonious model) for the basket-specific response rates would constrain them all to be equal. The least parsimonious model would allow them all to differ. Let $\pi_{(j,p)}$ be the pth *distinct* response rate for the jth of J possible models for the basket-specific response rates. We denote the jth model by M_j. The term *distinct* implies that $\pi_{(j,h)} \neq \pi_{(j,l)}$ for $h \neq l$. Necessarily, the number of distinct response rates is bounded above by K.

Each of the J models constrain different subsets of the basket-specific response rates to be equal (with the exception of the least parsimonious model which imposes no constraints). Table 1 presents all possible models for the response rates for the case where $K = 3$. If one were certain that baskets one and two had equal response rates which differed from the response rate for basket three (i.e., model M_2), the prudent analysis would be to pool the data for the two baskets that share a common response rate. Of course, in reality one cannot know definitively which of the five possible models corresponds to the *true* model, but one can use the data to help decide which of the competing models are plausible and act accordingly.

Table 1. *Possible response rate models for $K = 3$ baskets*

	Basket		Distinct	
	response rate		response	
$k = 1$	$k = 2$	$k = 3$	rates	Model
$\pi_{(1,1)}$	$\pi_{(1,1)}$	$\pi_{(1,1)}$	1	M_1
$\pi_{(2,1)}$	$\pi_{(2,1)}$	$\pi_{(2,2)}$	2	M_2
$\pi_{(3,1)}$	$\pi_{(3,2)}$	$\pi_{(3,1)}$	2	M_3
$\pi_{(4,1)}$	$\pi_{(4,2)}$	$\pi_{(4,2)}$	2	M_4
$\pi_{(5,1)}$	$\pi_{(5,2)}$	$\pi_{(5,3)}$	3	M_5

As noted above, the methods proposed by Cunanan *and others* (2017) and Simon *and others* (2016) do not take the intermediate models $M_2 - M_4$ explicitly into account. Our simulations illustrate that such approaches yield relatively high false positive rates (FPRs) when these intermediate models best describe the underlying generative processes for the data. Methods based on shrinkage using the BHM (Thall *and others*, 2003; Berry *and others*, 2013) assume model M_5 but borrow information based on the similarity of observed basket-specific response rates across *all* baskets. When intermediate models such as $M_2 - M_4$ hold, this technique can lead to undesirable false positive and/or false negative rates due to the BHM shrinking all estimates towards a common average. Adaptations such as the CBHM (Chu and Yuan, 2018) have recently been developed to correct this deficiency.

The approach that we develop in this article is guided by the perspective that it is unlikely that scientific knowledge will exist such that many (or any) of the possible competing models can be justifiably discarded. It would be ideal to let the observed data identify the set of plausible models from the complete model space and then average inference results from the competing models in accordance with their posterior probabilities given the data that have been observed.

3. Methods

3.1. *Overview of the Bayesian adaptive design*

We assume that K baskets will be evaluated for activity in the trial and propose an adaptive design framework that accommodates multiple analyses and allows for individual baskets to be permanently closed for enrollment at any of them. Baskets may be closed to enrollment as a result of having demonstrated activity or a sufficiently high likelihood of inactivity.

Let $n_{i,k}$ and $y_{i,k}$ denote the number of patients and number of responders, respectively, for basket k at the time of analysis i and let $\mathbf{D}_i = \left\{y_{i,k}, n_{i,k} : k = 1, ..., K\right\}$ represent the complete dataset at that time. Further, let π_0 be a user-specified response rate associated with inactivity and π_{A} be a hypothesized plausible response rate associated with activity. For each basket open to enrollment at the time of analysis i, one computes $P\left(\pi_k > \frac{\pi_{\mathrm{A}}+\pi_0}{2}|\mathbf{D}_i\right)$ to determine a futility action and $P\left(\pi_k > \pi_0|\mathbf{D}_i\right)$ to evaluate activity. If $P(\pi_k > \pi_0|\mathbf{D}_i) > \phi_1$ for a pre-specified evidence threshold ϕ_1, then the basket may be closed to enrollment due to having demonstrated sufficient probability of activity so as to warrant further study in a confirmatory setting. Conversely, if $P\left(\pi_k > \frac{\pi_{\mathrm{A}}+\pi_0}{2}|\mathbf{D}_i\right) \leq \phi_0$ for a pre-specified evidence threshold ϕ_0, then the basket may be closed to enrollment due to having a low likelihood of demonstrating the hypothesized level of activity in the trial.

The trial may terminate for one of several reasons. First, the trial will terminate when the number of baskets open to enrollment, denoted by R, reaches zero. Otherwise the trial will terminate when the maximum number of analyses, denoted by I, has been reached. Between the $(i - 1)$th and ith interim

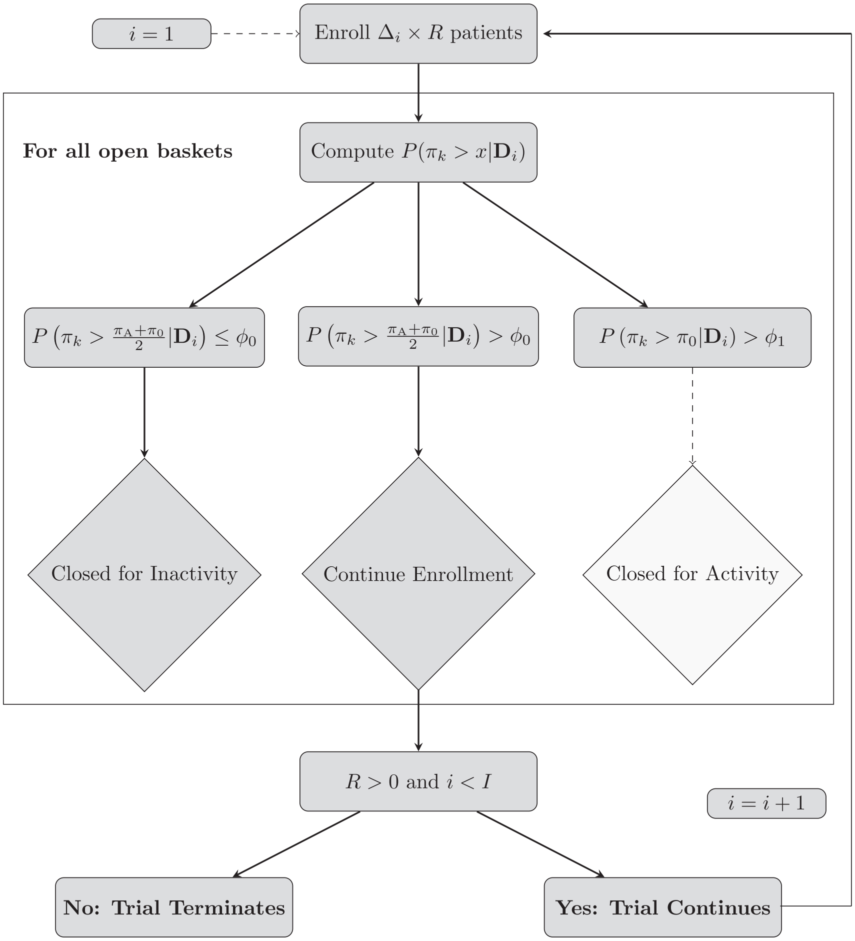
$i = 1$
Enroll $\Delta_i \times R$ patients
For all open baskets
Compute $P(\pi_k > x|\mathbf{D}_i)$
$P\left(\pi_k > \frac{\pi_A+\pi_0}{2}|\mathbf{D}_i\right) \leq \phi_0$
$P\left(\pi_k > \frac{\pi_A+\pi_0}{2}|\mathbf{D}_i\right) > \phi_0$
$P\left(\pi_k > \pi_0|\mathbf{D}_i\right) > \phi_1$
Closed for Inactivity
Continue Enrollment
Closed for Activity
$R > 0$ and $i < I$
$i = i + 1$
No: Trial Terminates
Yes: Trial Continues

Fig. 1. Overview of Basket Trial Study Design.

analysis, *at least* $\Delta_i \times R$ patients are enrolled to the baskets open for enrollment. For the sake of pragmatism, we do not require that *exactly* Δ_i patients enroll in each open basket between consecutive analyses. The quantity Δ_i simply serves as a target for enrollment in each basket that would be reached in an ideal setting. We do require that at least $\Delta_{i,\min} \leq \Delta_i$ patients enroll in each open basket but place no cap on the maximum number of patients. Thus, the actual enrollment between the $(i-1)$th and ith analysis will be no less than $\Delta_i \times R$. The proposed Bayesian adaptive design framework is illustrated in Figure 1. Note that some trials may only evaluate activity (e.g., whether $P(\pi_k > \pi_0|\mathbf{D}_i) > \phi_1$) at the end of the trial due to ethical concerns about closing a basket that has demonstrated activity (i.e., when no other proven treatment exists) while other baskets are still open for enrollment.

3.2. *Inference through model averaging*

3.2.1. *Likelihood formulation and the model space.* For ease of exposition, in this section, we omit the index i. We assume that the patients enrolled in basket k constitute a random sample with each having the same probability π_k of responding to treatment. Thus, $y_k \sim \text{Binomial}(\pi_k, n_k)$ and the likelihood for the observed data can be written as follows:

$$\mathcal{L}\left(\boldsymbol{\pi}\middle|\mathbf{D}\right) \propto \prod_{k=1}^{K}\binom{n_k}{y_k}\pi_k^{y_k}\left(1-\pi_k\right)^{n_k-y_k},$$

where $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_K)$ is the vector of response rates. Denote the number of distinct response rates for model M_j by P_j. Further, let $\Omega_{j,p}$ be the set of basket labels corresponding to baskets having the pth distinct response rate for model M_j $(p = 1,\ldots,P_j)$. For example, from Table 1 when $j = 2$, $\Omega_{2,1} = \{1,2\}$ and $\Omega_{2,2} = \{3\}$. Conditional on model M_j being the true model, the likelihood for the observed data can be rewritten as follows:

$$\mathcal{L}\left(\boldsymbol{\pi}_{(j)}\middle|\mathbf{D},M_j\right) \propto \prod_{p=1}^{P_j}\left\{\prod_{k\in\Omega_{j,p}}\binom{n_k}{y_k}\pi_{(j,p)}^{y_k}\left(1-\pi_{(j,p)}\right)^{n_k-y_k}\right\},$$

where $\boldsymbol{\pi}_{(j)} = \left(\pi_{(j,1)},\ldots,\pi_{(j,P_j)}\right)$ and $\pi_{(j,p)}$ is the pth distinct response rate for model M_j.

Let $\mathcal{M}_{K,P}$ denote the model space over the K baskets under the assumption that there are no more than P distinct response rates (necessarily $P \leq K$). Each model in $\mathcal{M}_{K,P}$ corresponds to a distinct classification of the K baskets into sets, within which response rates are equivalent and across which the response rates differ. In the *complete* model space, denoted as $\mathcal{M}_{K,K}$, the two most extreme models are the fully constrained model where $\pi_h = \pi_l$ for all h and l (a common response rate for all baskets) and the unconstrained model where all basket-specific response rates are allowed to differ (i.e., $\pi_h \neq \pi_l$ for $h \neq l$).

The number of models in $\mathcal{M}_{K,P}$, denoted by J, is given by the following formula:

$$J = \sum_{p}^{P}\left[\frac{1}{p!}\sum_{j=0}^{p}(-1)^{p-j}j^K\binom{p}{j}\right].$$

Note that when K is moderately large and P is unrestricted, the number of models is substantial (e.g., $J = 115\,975$ for $K = P = 10$). By exploiting closed-forms for posterior quantities used to make inference, the proposed BMA design approach remains computationally feasible even when averaging over complete model spaces for $K \geq 10$. Details on the computational feasibility of our BMA approach are provided in Section 4.3.

3.2.2. *Prior elicitation.*

We envision the proposed design methodology being applied in settings where it is unlikely that substantial prior information will exist beyond the basic belief that the investigational treatment will have similar activity levels across baskets and that baskets will likely have some degree of activity. To elicit a prior in a scenario with multiple competing models, one must elicit a prior probability for each model M_j and a prior for the distinct response rates $\boldsymbol{\pi}_{(j)}$ under model M_j.

We propose the following default prior over the model space:

$$p(M_j) \propto P_j^{\,\alpha},$$

where $\alpha \geq 0$ is a tuning parameter for the design. For $\alpha = 0$, the prior model probabilities are uniform. For $\alpha > 0$, the prior model probabilities are elicited such that models with more parameters (and therefore less borrowing) are given greater weight *a priori*. We investigate the influence of the prior model probability tuning parameter in Section 4.1.

We propose taking $\pi_{(j,p)}\middle|M_j \sim \text{Beta}\,(a_0, b_0)$ for each $j = 1, ..., J$ and $p = 1, ..., P_j$. This prior is consistent with the *a priori* belief that all baskets will have a similar level of activity. As a default choice, we propose choosing a_0 and b_0 such that

$$\frac{a_0}{a_0 + b_0} = \pi_{\text{A}}$$

and $a_0 + b_0 = 1.0$ to obtain a weakly informative prior with mean equal to the hypothesized response rate associated with activity π_{A}.

Given the above formulation of the prior, it follows that

$$\pi_{(j,p)}\middle|\mathbf{D}, M_j \sim \text{Beta}\left(a_{(jp)}, b_{(jp)}\right),$$

where $a_{(jp)} = a_0 + \sum_{k\in\Omega_{j,p}} y_k$ and $b_{(jp)} = b_0 + \sum_{k\in\Omega_{j,p}} (n_k - y_k)$ and that

$$P\left(\pi_k > x\middle|M_j, \mathbf{D}\right) = \prod_{p=1}^{P_j} \left[1 - F\left(x\middle|a_{(jp)}, b_{(jp)}\right)\right]^{1^{\left[k\in\Omega_{j,p}\right]}}, \tag{3.1}$$

where $F\left(\cdot\middle|a_{(jp)}, b_{(jp)}\right)$ is the cumulative distribution function of a beta random variable with parameters $a_{(jp)}$ and $b_{(jp)}$. The marginal likelihood for the data conditional on model M_j being the true model, denoted by $p\left(\mathbf{D}\middle|M_j\right)$, is

$$p\left(\mathbf{D}\middle|M_j\right) = \prod_{k=1}^{K} \binom{n_k}{y_k} \times \prod_{p=1}^{P_j} \frac{\mathcal{B}\left(a_{(jp)}, b_{(jp)}\right)}{\mathcal{B}\left(a_0, b_0\right)}, \tag{3.2}$$

where $\mathcal{B}\,(\cdot,\cdot)$ is the complete beta function. The posterior probability for model M_j given the data has the following general representation

$$p\left(M_j\middle|\mathbf{D}\right) = \frac{p\left(\mathbf{D}\middle|M_j\right)p(M_j)}{\sum_{j'} p\left(\mathbf{D}\middle|M_{j'}\right)p(M_{j'})},$$

and can be computed easily in closed-form based on (3.2) and the prior model probabilities. Inference for basket k is based on the posterior probability $P(\pi_k > x\middle|\mathbf{D})$ which can be expressed as a function of the model-specific posterior probabilities (3.1) and posterior model probabilities (3.2) as follows:

$$P\left(\pi_k > x\middle|\mathbf{D}\right) = \sum_j P\left(\pi_k > x\middle|M_j, \mathbf{D}\right) p\left(M_j\middle|\mathbf{D}\right).$$

Thus, having fit each model separately, $P\left(\pi_k > x\middle|\mathbf{D}\right)$ is straightforward to compute for any x.

3.2.3. *Simulation-based tuning for the design.* Basket trial designs are typically calibrated using simulation to ensure a desired set of operating characteristics are obtained. At minimum, the sample sizes required for the design must be determined via simulation to ensure the trial will have a sufficiently high true positive rate (TPR) for each active basket given a specified level of activity (i.e., sufficiently high power). Often constraints are placed on the design so that it controls the FPR at some desired level for the subset of inactive baskets. For example, the methods proposed by Cunanan *and others* (2017) and

Chu and Yuan (2018) use such a strategy. For our BMA design approach, we employ a similar strategy. We consider I, K, π_0, and π_{A} to be fixed design parameters and the quantities $\{\{\Delta_i : i = 1, ..., I\}, \alpha, \phi_0, \phi_1\}$ as customizable.

The quantities α, ϕ_0, and ϕ_1 need not be determined blindly by grid search although some degree of grid search may be necessary to finely tune the design's performance if so desired. Specifically, the value of ϕ_1 should correspond to a level of evidence thought to be compelling to investigators or to ensure that the basket-specific FPR is controlled at some acceptable level. For example, if it is desired to have a basket-specific FPR of approximately δ, then taking $\phi_1 = 1 - \delta$ will achieve that approximately and also provide a family-wise FPR of approximately $K \times \delta$ when all baskets are inactive. Guided by our exploration of BMA designs through simulation, we propose taking $\alpha = 2$, which balances information borrowing and stability of the design under varying assumptions on patient accrual. This choice for the tuning parameter is discussed in more detail in Section 4.1. The choice of ϕ_0 (which defines the futility criteria $P\left(\pi_k > \frac{\pi_{\mathrm{A}}+\pi_0}{2} | \mathbf{D}_i\right) \leq \phi_0$) can be guided by how conservative one wishes to be regarding enrollment termination in baskets where there is uncertainty regarding whether the treatment has the hypothesized level of activity. A value of $\phi_0 = 0.5$ would result in closing enrollment in a basket if it is more likely than not to have an activity level less than $\frac{\pi_{\mathrm{A}}+\pi_0}{2}$ (i.e., half of the hypothesized effect). In contrast, a choice of ϕ_0 much less than 0.5 would be more reasonable when *any* activity above the threshold π_0 would be clinically meaningful. We have found that choices of $\phi_0 \in [0.2, 0.4]$ provide a reasonable balance regarding closing inactive baskets early and keeping active baskets open in cases where early data are not highly compelling.

4. Simulation studies

In this section, we compare our proposed BMA design with the recently proposed basket trial designs by Cunanan *and others* (CUN) and Chu and Yuan (CBH) as well as to a basket trial that implements Simon's optimal two-stage design (Simon, 1989) independently in each basket (SIM). Both the SIM and CUN designs were proposed as two-stage designs whereas the BMA and CBH designs can be implemented with any number of analyses. Two-stage designs are appealing for logistical purposes since it is known ahead of time that upon reaching the end of the second stage, future activities can begin in earnest. Because of this and to facilitate a fair comparison of the four designs, we consider application of the BMA and CBH designs as two-stage designs in our example applications.

All of the comparator designs allow for the early stoppage of enrollment in a basket due to having a low likelihood of proving activity in the trial. That is to say, they all allow early stoppage for futility but not for efficacy. To facilitate fair comparison across methods, we apply the same restriction to our BMA design. In cases where patients enrolled in the trial have no other treatment options, it would be unethical to stop enrolling patients in a basket that has demonstrated activity while other baskets are still being evaluated. We compare the performance of the BMA design with and without early stopping for demonstrated activity in Appendix A supplementary material available at *Biostatistics* online.

In the following design examples, we consider the case where there are $K = 5$ baskets and assume an inactive (null) response rate equal to $\pi_0 = 0.15$ and a target alternative response rate equal to $\pi_{\mathrm{A}} = 0.45$. Our focus is testing the hypothesis $H_0 : \pi_k \leq \pi_0$ versus $H_1 : \pi_k > \pi_0$ for each $k = 1, \ldots, K$. The hypotheses tested in our example and the chosen inactive (null) response rate mirrors that used for the Vemurafenib basket trial (Hyman *and others*, 2015).

The SIM design requires that each basket enroll a fixed number of patients. For example, for $\pi_0 = 0.15$, $\pi_{\mathrm{A}} = 0.45$, and to achieve a basket-specific FPR of 0.01, each basket must enroll 9 patients in stage 1 and an additional 18 patients in stage 2 (if proceeding to that stage). The requirement that each basket has a precise number of patients is not ideal (and perhaps not practical) for a basket trial. Thus, in our evaluation of the BMA, CUN, and CBH designs, we placed more realistic restrictions on the minimum number of patients required for each basket (i.e., $\Delta_{i,\min}$) to trigger an analysis. For the optimal BMA, CUN, and CBHM designs, the target number of patients to enroll in stage 1 for each basket was 7, 8, and 10, respectively. We imposed the requirement that baskets would continue to enroll until each basket had at least four patients. Reaching the per-basket minimum increment in all active baskets and the overall enrollment total would trigger the stage 1 analysis. The requirements were imposed on baskets that proceeded into stage 2 for the BMA and CBH designs. Such constraints are not necessary for the CUN design, because in stage 2 all baskets are either analyzed separately (and hence require a fixed number of patients for each basket) or they are pooled (and hence only the total enrollment would be fixed).

Subject to the minimum enrollment requirements in each basket, we evaluated the performance of the designs based on varied accrual assumptions including uniform accrual across all baskets, slow accrual in active baskets, and fast accrual in active baskets. For the uniform accrual scenario, we assumed accrual in each basket followed an independent homogeneous Poisson process (Ross, 1996) with rate parameter $\lambda = 2$. This implies that interarrival times for patients in each basket are independent and identically distributed according to an exponential distribution with mean 0.5 so that approximately two patients enroll in each basket each month. For the case where active baskets accrue patients slowly, we set $\lambda = 1$ for the active baskets and $\lambda = 2$ for the inactive ones. Thus, the active baskets enrolled half as fast as the inactive baskets in this scenario. The slow enrolling case provides a conservative viewpoint for evaluating design performance in the presence of unbalanced enrollment since all the lower enrolling baskets were the active ones. For the case where active baskets accrue patients more quickly, we reversed the rate parameters from those described above so that active baskets had rate parameter $\lambda = 2$ and inactive baskets had rate parameter $\lambda = 1$.

We required certain desirable operating characteristics to be met for each design method to facilitate comparisons of their performance. Define the *family-wise* FPR (i.e., the family-wise type I error rate) as the probability of declaring at least one basket is active among the set that are inactive and the *basket-specific* FPR as the probability of declaring a given basket is active when it is actually inactive. The basket-specific TPR (i.e., basket-specific power) is defined analogously. We required all designs to meet the following three criteria:

(1) For any accrual scenario, the family-wise FPR must be ≤ 0.05 when all baskets are inactive (i.e., $\pi_k = \pi_0$ for $k = 1, ..., K$).
(2) Under the uniform accrual scenario with exactly one active basket, its basket-specific TPR must be ≥ 0.78 (assuming response rate equal to π_A).
(3) Under the slow active accrual scenario with exactly one active basket, its basket-specific TPR must be ≥ 0.60 (assuming response rate equal to π_A).

Criterion (1) requires that the family-wise FPR is well-controlled when the investigational product is inactive for all baskets. Criterion (2) requires that when sample sizes in each basket are balanced in expectation, the designs provide high basket-specific TPR values even in the case where only one basket is active. For reference, the implemented SIM design has a basket-specific TPR approximately equal to 0.81. Thus, criterion (2) mandates that in the case where only one basket is active with activity level π_A, its TPR can drop by more than 0.03 relative to the SIM design as a result of borrowing information across baskets. Criterion (3) places an analogous bound for the extreme case of slow active accrual. Criteria (2) and (3) determine what one is willing to trade in terms of TPR degradation in order to potentially gain efficiency over non-borrowing designs when multiple baskets are active as expected. Our choice of threshold values 0.05, 0.78, and 0.60 used for criterion (1), (2), and (3), respectively, should not be viewed as the *correct* choices. They are simply reasonable choices in the judgment of the authors. Given the imposed constraints above, we can compare the four designs on their ability to control the family-wise FPR when a proper subset of baskets are inactive (i.e., less than K of them), as well as with regard to their basket-specific TPR values, and expected samples sizes and trial durations.

Table 2. *Estimated true and false positive rates under varying accrual assumptions*

	#	—— Family-wise FPR ——				—— Basket-specific TPR ——			
Accrual	**Act.**	**BMA**	**CBH**	**CUN**	**SIM**	**BMA**	**CBH**	**CUN**	**SIM**
Uniform	0	0.05	0.05	0.05	0.05	—	—	—	—
	1	0.05	0.05	0.06	0.04	0.78	0.78	0.78	0.82
	2	0.05	0.04	0.07	0.03	0.81	0.83	0.83	0.81
	3	0.04	0.03	0.10	0.02	0.83	0.84	0.85	0.81
	4	0.02	0.03	0.16	0.01	0.85	0.86	0.86	0.81
	5	—	—	—	—	0.87	0.89	0.89	0.81
Fast active	0	0.05	0.05	0.05	0.05	—	—	—	—
	1	0.06	0.05	0.08	0.04	0.91	0.91	0.84	0.81
	2	0.05	0.04	0.10	0.03	0.90	0.91	0.86	0.81
	3	0.05	0.04	0.14	0.02	0.90	0.90	0.87	0.81
	4	0.03	0.07	0.22	0.01	0.89	0.89	0.88	0.81
	5	—	—	—	—	0.87	0.89	0.89	0.81
Slow active	0	0.05	0.05	0.05	0.05	—	—	—	—
	1	0.05	0.05	0.05	0.04	0.65	0.62	0.74	0.81
	2	0.04	0.04	0.05	0.03	0.72	0.71	0.80	0.81
	3	0.03	0.03	0.06	0.02	0.78	0.77	0.83	0.81
	4	0.02	0.02	0.09	0.01	0.82	0.82	0.85	0.81
	5	—	—	—	—	0.87	0.89	0.89	0.81

In what follows, we present the operating characteristics of the BMA, CBH, CUN, and SIM designs after optimization through simulation (i.e., tuning) to achieve goals (1)–(3) and to provide the minimum expected sample size under the uniform accrual scenario when all baskets are inactive. Appendix B of the supplementary material available at *Biostatistics* online presents the optimal design inputs for each method (e.g., target sample size for each stage). The optimal BMA design was obtained by taking on $\Delta_1 = 7$, $\Delta_2 = 16$, $\phi_1 = 0.985$, $\phi_0 = 0.275$, and $\alpha = 2$. All operating characteristics presented were estimated based on $\geq$200 000 simulated basket trials. Analyses for the CBH design required Markov chain Monte Carlo (MCMC) methods and each used 20 000 posterior samples obtained from a Metropolis–Hastings sampler. Analyses for all other methods, including the BMA approach, are based on exact calculations (i.e., obtainable without approximation or Monte Carlo error).

Table 2 presents the estimated TPR and FPR for each design under the three accrual scenarios. For these simulations, all inactive baskets have response rates equal to $\pi_0 = 0.15$ and all active baskets have response rates equal to $\pi_A = 0.45$. Thus, the basket-specific TPR is the same for all active baskets in a given scenario.

Focusing first on the family-wise FPR results, the only striking detail is that the CUN method is unable to control the family-wise FPR at the targeted level (0.05) when the number of active baskets is near K. The performance is exceedingly poor under the fast active accrual scenario (over four times the nominal level) but is also poor for the uniform accrual scenario (over three times the nominal level). In particular, when only one basket is inactive, the family-wise FPR is 2–3 times the targeted level. The BMA and CBH designs control family-wise FPR quite well with the only inflation for the BMA and CBH designs coming in the fast active accrual scenario. In that case, the degree of inflation is quite small.

For the basket-specific TPR results, one can see that the information borrowing designs (i.e., BMA, CBH, and CUN) all provide increased basket-specific TPR values compared to the SIM design when

Table 3. *Estimated expected sample size and trial duration under varying accrual assumptions*

	#	—— Expected sample size ——				—— Expected trial duration ——			
Accrual	**Act.**	**BMA**	**CBH**	**CUN**	**SIM**	**BMA**	**CBH**	**CUN**	**SIM**
Uniform	0	59.7	61.6	63.5	57.7	9.9	8.4	9.7	10.4
	1	70.3	75.9	78.6	70.4	10.7	11.2	12.5	13.6
	2	80.9	87.4	89.6	83.3	11.1	11.8	13.6	14.8
	3	91.4	97.7	98.0	96.1	11.3	12.0	14.0	15.5
	4	100.7	107.6	102.8	108.7	11.4	12.0	13.7	16.0
	5	109.0	116.7	96.8	121.5	11.4	12.0	11.8	16.3
Fast active	0	59.7	61.6	63.5	57.7	19.7	16.7	19.5	20.8
	1	76.2	80.0	78.9	70.4	16.7	17.3	22.4	20.3
	2	89.8	93.1	88.5	83.2	15.1	15.7	21.3	19.5
	3	100.1	104.1	96.3	96.0	13.9	14.4	18.9	18.6
	4	106.3	112.6	100.5	108.7	12.6	13.3	15.5	17.5
	5	109.0	116.7	96.7	121.6	11.4	12.0	11.8	16.3
Slow active	0	59.8	61.5	63.5	57.7	9.9	8.4	9.7	10.4
	1	72.0	75.0	83.7	70.4	14.6	14.3	21.1	24.7
	2	82.8	86.0	97.4	83.2	17.5	17.7	25.6	28.8
	3	92.4	95.7	106.6	96.0	19.6	20.0	27.6	30.6
	4	101.1	105.6	111.3	108.8	21.2	21.9	28.2	31.8
	5	109.0	116.7	96.8	121.6	22.7	24.0	23.5	32.6

the number of active baskets is at least two under the uniform accrual scenario. The basket-specific TPR gains are even greater for the fast active accrual scenario with the gains in the BMA and CBH designs outpacing those of the CUN design. However, for the slow active accrual scenario the basket-specific TPR values for the information borrowing designs only exceed those of the SIM design when there are at least three (CUN) or four (BMA and CBH) baskets active. In this case, it is the CUN design that has the least reduction in TPR values relative to the SIM design but, as noted before, it has a notably inflated family-wise FPR in the four active basket case.

When evaluating basket-specific TPR values across the three accrual scenarios, two points are critical to keep in mind. First, the SIM design requires equal enrollment in all baskets (regardless of how slowly or quickly patients could be accrued in the baskets). The end result of this is that basket-specific TPR values are invariant to accrual rates but also that it can take substantially longer to complete study of all baskets as shown in Table 3. In the slow active accrual scenario where there is only one inactive basket, the expected trial duration associated with the SIM design is 50% longer than the BMA design ($31.8/21.2 \approx 1.50$) .

The second point is that the maximum basket-specific TPR value decrease is associated with the case where there is one active basket that enrolls half as fast as all others. This is an extreme scenario and although it is possible, it should not be construed as being the most likely scenario in general. The slow active accrual scenario provides a reasonable bound on how much the basket-specific TPR values can be degraded due to substantial deviations from uniform accrual.

Table 3 presents the expected trial durations for each design method as well as the expected sample size. As was the case regarding FPR and TPR, performance of the BMA and CBH designs are quite similar. The BMA design has modestly lower expected sample size than the CBH across all scenarios (3–6.5% lower). The BMA design comparison to the CUN design is less one-sided. For the uniform and slow active accrual scenarios the BMA design has comparatively smaller samples sizes for all scenarios

Table 4. *Design operating characteristics for varied prior model probability tuning parameter*

		$\alpha = 0$			$\alpha = 2$			$\alpha = 4$		
Accrual	# Active	FW-FPR	TPR	Sample size	FW-FPR	TPR	Sample size	FW-FPR	TPR	Sample size
Uniform	0	0.03	—	52.5	0.05	—	59.7	0.07	—	64.1
	1	0.05	0.70	67.4	0.05	0.79	70.3	0.05	0.82	72.1
	4	0.06	0.87	103.2	0.02	0.85	100.7	0.02	0.83	98.9
	5	—	0.92	104.3	—	0.87	108.9	—	0.83	108.0
Fast active	0	0.03	—	52.5	0.05	—	59.8	0.07	—	64.0
	1	0.06	0.86	75.6	0.06	0.90	76.2	0.06	0.92	76.7
	4	0.08	0.91	110.6	0.03	0.89	106.3	0.02	0.86	105.3
	5	—	0.92	104.4	—	0.87	109.1	—	0.83	108.1
Slow active	0	0.03	—	52.5	0.05	—	59.8	0.07	—	64.1
	1	0.04	0.52	67.4	0.05	0.65	71.9	0.05	0.71	74.8
	4	0.04	0.83	102.7	0.02	0.82	101.1	0.01	0.80	99.9
	5	—	0.92	104.4	—	0.87	109.0	—	0.83	108.0

except when all baskets are active. For the fast active accrual scenario, the CUN design provides smaller sample sizes unless less than or equal to one basket is active. Relative to the SIM design, the BMA design has smaller sample size except for the case where all baskets are inactive, and as much as a 10% reduction when all baskets are active. As noted previously, the information borrowing designs lead to shorter trials on average compared to the SIM design as a result of allowing variable enrollment in the baskets. Among the information borrowing designs, the BMA design generally leads to the shortest trial duration except in the case where all baskets are inactive where the CBH design performs best.

In Appendix C of the supplementary material available at *Biostatistics* online, we compare the performance of the four designs with respect to point estimate bias for the basket-specific response rates. In general, the degree of bias in point estimates of the basket-specific response rates is comparable across the four methods. In Appendix D of the supplementary material available at *Biostatistics* online, we discuss the performance of the designs when activity levels are heterogeneous. In particular, we evaluate the performance of the four designs when four out of five baskets are active but when three of those four have half the hypothesized level of benefit (i.e., response rate equal to 0.30 instead of 0.45). We also look at a case with extreme response rate heterogeneity with basket-specific response rates ranging from 0.05 up to 0.45. In both cases, the BMA design has TPR values greater than or equal to those for all other designs and provides control of the family-wise FPR at the nominal level.

4.1. *Understanding the prior model probability tuning parameter*

Table 4 presents estimated operating characteristics for the optimal BMA design described above (corresponding to a prior model probability tuning parameter value of $\alpha = 2$) as well as BMA designs with design parameters equal to those from the optimal design except that we modified α to be either 0 or 4 for comparison purposes. Note that taking $\alpha = 0$ corresponds to a uniform prior over the model space. Comparing the design with $\alpha = 0$ to the design with $\alpha = 2$ (i.e., the optimal design), one can see that the design based on uniform prior model probabilities tends to result in family-wise FPR values that increase with the number of active baskets, a behavior not exhibited when $\alpha = 2$. The opposite behavior occurs when $\alpha = 4$. For the case where $\alpha = 4$, the family-wise FPR is greatest when all baskets are inactive.

As one compares the basket-specific TPR values for the three values of α, it is apparent that as α increases there is more stability in the TPR values across accrual scenarios but sample sizes generally increase as well. This is consistent with the fact that less information is being borrowed across baskets and therefore a smaller likelihood that baskets will close early, leading to longer trials and greater enrollment. The value $\alpha = 2$ provides a degree of balance between more information borrowing/more variable TPR values ($\alpha = 0$) and less information borrowing/more stable TPR values ($\alpha = 4$).

4.2. *Estimable quantities specific to BMA approach*

Aside from identifying which baskets have a response rate above a specified threshold, it is also of interest to characterize which baskets have the same or similar response rates. BMA provides a unique and appealing framework for doing this. Specifically, BMA provides a natural metric to quantify the likelihood that two baskets have the same response rate:

$$P\left(\pi_k = \pi_l \middle| \mathbf{D}\right) = \sum_j \sum_{p=1}^{P_j} 1\left[k \in \Omega_{j,p}, l \in \Omega_{j,p}\right] p(M_j|\mathbf{D}),$$

where $1\left[k \in \Omega_{j,p}, l \in \Omega_{j,p}\right]$ is an indicator that baskets k and l have the same parameter for model j. In other words, to compute the posterior probability of response rate equivalence for two baskets, one only needs to sum the posterior model probabilities for the models where the two baskets share a common parameter.

By way of example, consider a dataset corresponding to $K = 5$ baskets with sample size vector $\mathbf{n} = (20, 20, 20, 20, 20)$ and corresponding response vector $\mathbf{y} = (3, 4, 9, 10, 10)$. Using the same BMA design inputs previously described (e.g., $\alpha = 2$, etc.), one obtains $P\left(\pi_1 = \pi_2 \middle| \mathbf{D}\right) = 0.31$, $P\left(\pi_3 = \pi_4 \middle| \mathbf{D}\right) = 0.27$, and $P\left(\pi_4 = \pi_5 \middle| \mathbf{D}\right) = 0.28$. All other posterior probabilities of response rate equivalence are less than or equal to 0.07. If instead one takes $\alpha = 0$, they obtain $P\left(\pi_1 = \pi_2 \middle| \mathbf{D}\right) = 0.69$, $P\left(\pi_3 = \pi_4 \middle| \mathbf{D}\right) = 0.62$, and $P\left(\pi_4 = \pi_5 \middle| \mathbf{D}\right) = 0.66$. In this case, all other posterior probabilities of response rate equivalence are less than or equal to 0.14. For either choice of α, in this hypothetical dataset, one can see that the data suggest a natural grouping of baskets 1 and 2 as well as baskets 4 and 5 into classes of similar responders.

4.3. *Computational efficiency of BMA*

One of the major limitations that BMA has faced is that it is computationally demanding to perform. In order to perform BMA without approximation, one needs to be able to compute the marginal likelihood for the observed data under competing models and such computations are generally expensive. Our proposed approach makes use of closed-form expressions for the marginal likelihood of the data and model-specific posterior quantities, greatly reducing the computational burden of the method. Moreover, many of the calculations (e.g., evaluations of the Gamma function) are highly repetitive in large scale simulations and our software uses pre-computation of expensive quantities and look-up tables to avoid unnecessary repeated computation in large scale simulations.

Nonetheless, since the proposed approach averages over *all possible models*, the sheer number of models does present a computational challenge as the number of baskets increases. For example, for $K = 10$ there are 115 975 models and for $K = 12$ there are 4 213 597 models. Figure S1 of the supplementary material available at *Biostatistics* online shows the average time required to perform 10 000 simulation studies to investigate the $K+1$ possibilities for the number of active baskets for each of the three accrual scenarios. Each estimated run time is the average of four identical runs of the set of simulations. Thus, the total number of simulation studies is $10\,000 \times (K+1) \times 3$ for each point in the figure. For $K = 5$ this corresponds 180 000 simulation studies and for $K = 10$ to 330 000 simulation studies.

Using a single compute core (e.g., a single computer without making use of multi-core processing), the R (R Core Team, 2016) code written by the authors can perform the necessary simulations for $K = 5$ baskets in less than 15 min. For $K = 10$ baskets the simulations can be completed in less than 1 day. We would add that the idea that users are restricted to a single computing core is antequated. The reality is that with pay-for-use high performance computing services (e.g., Google Cloud Platform) virtually anyone can reap the benefits of substantial computing power without significant long-term cost. By making use of multiple computing cores (from a single user's machine, an on-premises high performance computing cluster, or from a pay-for-use cloud computing platform), simulation studies can be divided into sets and performed in parallel fashion to reduce wall-clock run time. For example, by making use of 25 available computing cores on the Longleaf computing cluster at the University of North Carolina at Chapel Hill, the authors were able to reduce the time required for simulations to complete in the $K = 10$ case from 16.9 h to 1.1 h. Note that the total time reduction is not by a factor of 25 due to variable performance of the computing cores.

Lastly, for applications with K larger than what we consider here, it is possible to transition from a direct computation approach to a stochastic algorithm that use reversible jump MCMC to compute posterior quantities. We refer interested readers to the seminal work by Green (1995). In particular, Green considers a partition problem similar to that discussed in this article and proposes a simple MCMC algorithm for model fitting.

5. Discussion

Because of the tradeoffs inherent in adopting an information borrowing approach to designing basket trials, one should not expect one approach to be uniformly better than all others. Even when comparing the information borrowing designs to the SIM design, we see there are instances where the latter is the top performer. The motivation behind information borrowing designs for basket trials is the belief that many or most baskets will have similar activity and it is under that reality that such designs provide notable efficiency gains (e.g., increased TPRs over standard designs). The particular BMA design we evaluated in our simulation studies attempted to strike a balance in performance between scenarios where most baskets have activity and scenarios where most baskets do not. Others may prefer a design that is more or less aggressive regarding information borrowing. Being more or less aggressive regarding information borrowing is closely related to the choice of the prior model probability tuning parameter as illustrated by Table 4. There is no correct choice in that regard but it should be clear that a tradeoff is always being made.

Our results indicate that the performance of the BMA design using $\alpha = 2$ is quite similar to that of the CBH design. We would highlight two points that tilt the argument in favor of the BMA design. First, the BMA design has significantly decreased computational burden for trials with a moderate number of baskets as a result of not requiring MCMC for model fitting. Second, BMA has a very natural mechanism for classifying baskets into groups of similar responders using posterior model probabilities which is be helpful for planning future studies.

According to U.S. FDA guidance on adaptive designs (Food and Drug Administration, 2018), "for simulations intended to estimate Type I error probability, hypothetical clinical trials would be simulated under a series of assumptions compatible with the null hypothesis." The ability to perform large scale simulations that investigate variable accrual rates and activity levels in baskets is critical and the computational efficiency of the proposed BMA framework allows for this. Our proposal to examine the performance of basket trial designs under uniform patient accrual, slow active accrual, and fast active accrual and for every possible number of active baskets under each accrual scenario is an attempt to define a set of scenarios that adhere to the spirit of the FDA guidance. Evaluating the performance of the design over this broad set of scenarios provides a comprehensive picture of potential efficiency gains and losses associated with using BMA (or any information borrowing design).

In our design application, we evaluated activity levels in each basket against a common inactivity threshold of $\pi_0 = 0.15$ which was motivated by the Vemurafenib basket trial (Hyman *and others*, 2015). However, nothing about the BMA approach we have developed requires evaluation against such a common threshold. It is, however, true that the framework we have developed borrows information to the degree that basket-specific response rates are homogeneous across subsets of baskets. In contrast, the traditional BHM allows for information borrowing to be governed by the degree to which the response rates in each basket differ from hypothesized values which can differ across baskets. In future work, we plan to extend the BMA framework developed in this article to support this more general strategy for information borrowing.

In this article, we focused on the design of a basket trial evaluating a single treatment that targets a genomic alteration present in multiple tumor histologies. Given the favorable performance of BMA-based designs in this setting, we hope to extend our method to more general settings evaluating multiple treatments in response-adaptive randomized trials. Recent innovations in this more complex setting include the works of Steffen *and others* (2017) and Trippa and Alexander (2017). The authors are optimistic that an approach based on BMA in this more complicated setting could be beneficial if the computational challenges can be adequately solved.

Supplementary material

Supplementary material is available at http://biostatistics.oxfordjournals.org.

Acknowledgments

The authors wish to thank the associate editor and reviewers for helpful comments and suggestions, which have led to improvements of this article. *Conflict of Interest:* None declared.

Funding

National Institutes of Health (NIH) (#GM 70335 and P01CA142538), in part.

References

Berry, S. M., Broglio, K. R., Groshen, S. and Berry, D. A. (2013). Bayesian hierarchical modeling of patient subpopulations: efficient designs of Phase II oncology clinical trials. *Clinical Trials*, **10**, 720–734.

Chu, Y. and Yuan, Y. (2018). A Bayesian basket trial design using a calibrated Bayesian hierarchical model. *Clinical Trials*, **15**, 149–158.

Cunanan, K. M., Iasonos, A., Shen, R., Begg, C. B. and Gönen, M. (2017). An efficient basket trial design. *Statistics in Medicine*, **36**, 1568–1579.

Draper, D. (1995). Assessment and propagation of model uncertainty. *Source Journal of the Royal Statistical Society. Series B (Methodological)*, **57**, 45–97.

Food and Drug Administration (2018). Adaptive Designs for Clinical Trials of Drugs and Biologics Guidance for Industry. https://www.fda.gov/downloads/drugs/guidances/ucm201790.pdf [online; last accessed February 21, 2018].

GREEN, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

HOETING, J. A., MADIGAN, D., RAFTERY, A. E. AND VOLINSKY, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical Science*, **14**, 382–401.

HYMAN, D. M., PUZANOV, I., SUBBIAH, V., FARIS, J. E., CHAU, I., BLAY, J.-Y., WOLF, J., RAJE, N. S., DIAMOND, E. L., HOLLEBECQUE, A. *and others* (2015). Vemurafenib in multiple nonmelanoma cancers with BRAF V600 mutations. *New England Journal of Medicine*, **373**, 726–736.

LIU, R., LIU, Z., GHADESSI, M. AND VONK, R. (2017), Increasing the efficiency of oncology basket trials using a Bayesian approach. *Contemporary Clinical Trials*, **63**, 67–72.

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.

MAITOURNAM, A. AND SIMON, R. (2005). On the efficiency of targeted clinical trials. *Statistics in Medicine*, **24**, 329–339.

MANDREKAR, S. J. AND SARGENT, D. J. (2011). All-comers versus enrichment design strategy in phase II trials. *Journal of Thoracic Oncology: Official Publication of the International Association for the Study of Lung Cancer*, **6**, 658–660.

R CORE TEAM (2016). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

RAFTERY, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, **25**, 111–163.

REDIG, A. J. AND JÄNNE, P. A. (2015). Basket trials and the evolution of clinical trial design in an era of genomic medicine. *Journal of Clinical Oncology*, **33**, 975–977.

ROSS, S. M. (1996). *Stochastic Processes*, 2nd edition. New York: Wiley.

SIMON, R. (1989). Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials*, **10**, 1–10.

SIMON, R. (2017). Critical review of umbrella, basket, and platform designs for oncology clinical trials. *Clinical Pharmacology & Therapeutics*, **102**, 934–941.

SIMON, R., GEYER, S., SUBRAMANIAN, J. AND ROYCHOWDHURY, S. (2016). The Bayesian basket design for genomic variant-driven phase II trials. *Seminars in Oncology*, **43**, 13–18.

SIMON, R. AND MAITOURNAM, A. (2004). Evaluating the efficiency of targeted designs for randomized clinical trials. *Clinical Cancer Research*, **10**, 6759–6763.

STEFFEN, V., T., B. W., GIOVANNI, P. AND LORENZO, T. (2017). Bayesian response adaptive designs for basket trials. *Biometrics*, **73**, 905–915.

THALL, P. F., WATHEN, J. K., BEKELE, N. B., CHAMPLINE, R. E., BAKER, L. H. AND BENJAMIN, R. S. (2003). Hierarchical Bayesian approaches to phase II trials in diseases with multiple subtypes. *Statistics in Medicine*, **22**, 763–780.

TRIPPA, L. AND ALEXANDER, B. M. (2017). Bayesian baskets: a novel design for biomarker-based clinical trials. *Journal of Clinical Oncology*, **35**, 681–687.

[*Received September 13, 2018; revised March 4, 2019; accepted for publication March 24, 2019*]