


Variance prior specification for a basket trial design using Bayesian hierarchical modeling

Clinical Trials
2019, Vol. 16(2) 142–153
© The Author(s) 2018
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1740774518812779
journals.sagepub.com/home/ctj


Kristen M Cunanan, Alexia Iasonos, Ronglai Shen and Mithat Gönen

Abstract

Background: In the era of targeted therapies, clinical trials in oncology are rapidly evolving, wherein patients from multiple diseases are now enrolled and treated according to their genomic mutation(s). In such trials, known as basket trials, the different disease cohorts form the different baskets for inference. Several approaches have been proposed in the literature to efficiently use information from all baskets while simultaneously screening to find individual baskets where the drug works. Most proposed methods are developed in a Bayesian paradigm that requires specifying a prior distribution for a variance parameter, which controls the degree to which information is shared across baskets.

Methods: A common approach used to capture the correlated binary endpoints across baskets is Bayesian hierarchical modeling. We evaluate a Bayesian adaptive design in the context of a non-randomized basket trial and investigate three popular prior specifications: an inverse-gamma prior on the basket-level variance, a uniform prior and half-t prior on the basket-level standard deviation.

Results: From our simulation study, we can see that the inverse-gamma prior is highly sensitive to the input hyperparameters. When the prior mean value of the variance parameter is set to be near zero (≤ 0.5), this can lead to unacceptably high false-positive rates ($\geq 40\%$) in some scenarios. Thus, use of this prior requires a fully comprehensive sensitivity analysis before implementation. Alternatively, we see that a prior that places sufficient mass in the tail, such as the uniform or half-t prior, displays desirable and robust operating characteristics over a wide range of prior specifications, with the caveat that the upper bound of the uniform prior and the scale parameter of the half-t prior must be larger than 1.

Conclusion: Based on the simulation results, we recommend that those involved in designing basket trials that implement hierarchical modeling avoid using a prior distribution that places a majority of the density mass near zero for the variance parameter. Priors with this property force the model to share information regardless of the true efficacy configuration of the baskets. Many commonly used inverse-gamma prior specifications have this undesirable property. We recommend to instead consider the more robust uniform prior or half-t prior on the standard deviation.

Keywords

Basket trial, phase II, Bayesian method, adaptive design, variance prior

Background

Conventional phase II clinical trials evaluate a single drug in a single disease patient population. Increasingly, investigators are implementing master protocols that consider different subpopulations to investigate multiple drugs and/or multiple diseases and possibly multiple targets. Such trials have been called basket or umbrella trials. Basket trials evaluate a single drug targeting a single mutation in multiple disease cohorts, while umbrella trials evaluate multiple drugs (often targeting different mutations) in a single disease population. There has been overlap on labeling basket versus umbrella for describing the same clinical trial;

however, a key characteristic among all of the designs is multiple molecularly defined cohorts with a common element between cohorts, a common drug, mutation or disease group. We observed that the number of baskets in a study is a function of, among other things, how the

Department of Epidemiology & Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY, USA

Corresponding author:

Kristen M Cunanan, Department of Epidemiology & Biostatistics, Memorial Sloan Kettering Cancer Center, 2nd Floor, 485 Lexington Avenue, New York, NY 10017, USA.
Email: kristenmay206@gmail.com

baskets are defined (i.e. by disease, mutation, drugs or a combination of these). For basket trials investigating a single drug, typically three to nine disease-specific or disease mutation-specific baskets are defined at the beginning of the trial with perhaps an exploratory catch-all basket (e.g. clinicaltrials.gov identifiers NCT01524926 (six baskets), NCT03339843 (five baskets), NCT02568267 (nine baskets), NCT02576431 (eight baskets), NCT02811497 (three baskets) and¹⁻⁴ with five, two, nine and three baskets, respectively), and increasingly, investigators amend the baskets as the trial progresses possibly due to higher-than-expected accrual or new discoveries.^{1,4}

Current basket trials are often designed as a series of independent trials implemented in parallel for each of the different diseases or indications. This approach is simple and the overall false-positive rate can be controlled with strict decision rules in each basket; however, this approach fails to take into account the anticipated correlated responses. That is, if results are favorable in one basket, there is higher chance that they will also be favorable in other baskets. This poses a need for more creative designs that are simple to implement. One major design challenge is to capitalize on the correlated responses in baskets where the drug truly works while simultaneously screening and dropping baskets where the drug is truly futile. Empirical results from published trials suggest that we can expect heterogeneity in efficacy across baskets.^{1,4-6}

Several phase II designs applicable to basket trials have been proposed⁷⁻¹² to account for the anticipated correlated efficacy in a simple framework. Recently, more complex methods have been developed for more complicated settings, such as multiple covariates, adaptive randomization or a continuous outcome.¹³⁻¹⁶ In this article, we are interested in first understanding design implications in the simple basket trial setting evaluating a single drug in multiple pre-specified baskets. Most novel approaches for this setting use Bayesian methods that require a prior specification of at least one parameter that permits sharing of information across baskets. As compared to independent parallel designs, using methods such as Bayesian hierarchical modeling in an adaptive design can reduce the trial size and duration and improve power to identify individual baskets where the drug works. In our investigation of such methods, we have found the prior specification of the sharing parameter to be very influential. This motivated the research reported in this article, where we endeavor to provide recommendations on selecting a prior for the variance parameter in an adaptive basket trial design using hierarchical modeling.

Through a simulation study, we investigate three commonly used priors: an inverse-gamma prior on the basket-level variance, a uniform and half-t prior on the basket-level standard deviation. Inverse-gamma is by far the most popular prior of choice in Bayesian

hierarchical models. Most analysts are familiar with inverse-gamma as a prior because of its conditional conjugacy properties in simple models and they simply continue to use it in more complicated models. In most applications of hierarchical modeling, variances are nuisance parameters and their prior specification takes a back seat to the specification of the prior for the means. In our simulation study, we implement a Bayesian for the analysis and evaluate the overall performance of our design, model and prior selections based on conventional frequentist operating characteristics, such as power and false-positive rates. We believe that a thorough evaluation of such metrics is imperative for understanding the properties of any proposed basket trial design. Consequently, we calibrate the implemented designs and base our recommendations using these traditional metrics.

The remainder of this article proceeds as follows: section “Methods” presents the Bayesian adaptive design, hierarchical model, prior specifications used and presents the simulation study; section “Results” presents the results; and section “Conclusion” concludes with recommendations and a brief discussion.

Methods

In section “Bayesian adaptive design,” we present a Bayesian adaptive design following the design originally proposed by Berry et al.⁷ with pragmatic modifications in implementation to minimize the logistical burden of multiple interim analyses across diseases in different service departments or possibly centers. The original Berry design performed interim analyses in each basket when 10 patients have been observed in a given basket (and every 5 patients in a basket, thereafter); however, it did not take into account different accrual rates which could potentially require performing interim analyses in different baskets after every few patients depending on enrollment rates. We have modified the design to perform interim analyses based on enrollment for the entire trial rather than each basket and have added an eligibility rule for analysis.

Let y_{ik} be a binary indicator of response for patient i in basket k , for $i = 1, \dots, n_k$ and p_k be the probability of response in basket k for $k = 1, \dots, K$. Define p_a to be the target response rate indicating the drug displays promising activity and p_0 to be the null response rate indicating absence of activity for the drug. These quantities could be basket-specific, but in this article, we assume common target and null response rates.

Bayesian adaptive design

1. Treat the first $10K$ patients. Perform the first interim analysis and apply early stopping rules to baskets with at least n_{\min} patients, where n_{\min} is the

minimum number of patients required in a basket to evaluate efficacy (otherwise continue to next interim analysis). Stop an individual basket for futility if

$$\Pr(p_k > p_{mid} | data) < 0.05$$

where p_{mid} is the midpoint between p_0 and p_a . Stop an individual basket for efficacy if

$$\Pr(p_k > p_{mid} | data) > 0.90$$

2. Perform additional interim analyses in baskets with at least n_{\min} patients after every $5K^{\star}$ patients, where K^{\star} is the number of remaining evaluable baskets.
3. Perform the final analysis after the maximum sample size (n_{\max}) is enrolled and apply the final decision rule to remaining baskets with at least n_{\min} patients

$$\Pr(p_k > p_0 | data) > \gamma$$

We note that interim analyses use all available data, including baskets with less than n_{\min} patients. For some motivating studies, the early-stopping rule for efficacy may not be appropriate and can easily be dropped from the adaptive design. The final analysis occurs after the numbers of patients enrolled in the remaining baskets have reached the maximum sample size per basket. Bayesian inference is based on Markov chain Monte Carlo (MCMC) sampling from the posterior distribution using the Gibbs sampler.

Bayesian hierarchical model

We assume that $\sum_i y_{ik}$ follows a binomial distribution of size n_k with probability p_k . Similar to Berry et al.,⁷ to obtain the decision probabilities described in section “Bayesian adaptive design,” we apply a *logit* transformation to the basket-specific probabilities of a response to facilitate a Bayesian model, as follows

$$\theta_k = \text{logit}(p_k) - \text{logit}(p_0)$$

While the original Berry design modeled the change in log-odds from the target response rate, we instead model the change in log-odds from the null response rate since we believe this quantity can be better pre-specified by investigators. We define a hierarchical model for the basket-specific model parameters as follows

$$\begin{aligned}\theta_k &\sim \text{Normal}(\mu, \sigma^2) \\ \mu &\sim \text{Normal}(m_\mu, v_\mu) \\ \sigma^2 &\sim g(\cdot)\end{aligned}$$

where m_μ and v_μ are pre-specified mean and variance hyperparameters and $g(\cdot)$ is an appropriate distribution for the variance (i.e. sharing) parameter σ^2 , such as the inverse-gamma (α, β) . A more interpretable re-parameterization of the inverse-gamma distribution specifies a prior mean of σ^2 (define as m_{σ^2}) and prior effective sample size, that is, weight (define as w_{σ^2}),¹⁷ where $\alpha = w_{\sigma^2}/2$ and $\beta = m_{\sigma^2}^2 w_{\sigma^2}/2$.

Other functional forms for $g(\cdot)$ are proposed in the literature. Gelman¹⁸ has investigated numerous prior distributions for the variance parameter in conventional hierarchical linear models, including inverse-gamma and uniform on σ^2 , half-t and uniform on σ and uniform on $\log(\sigma)$. In his discussion of the latter, he notes “in a hierarchical model the data can never rule out a group-level variance of zero, and so the prior distribution cannot put an infinite mass in this area.” In his recommendations, he suggests starting with a uniform density on σ , but mentions that the uniform $(0, b)$ prior on σ can lead to overestimation of σ and less than optimal sharing of information across groups when the number of groups is small (say, below five). In his recommendations, he suggests working with the half-t family of priors when more prior information is desired (i.e. to restrict σ away from large values); he mentions this prior is more flexible with better behavior near 0, compared to the inverse-gamma prior. Finally, he does not recommend using the inverse-gamma prior as it is sensitive to input values when small values of σ are possible in the data, and we note that this is likely to occur in our basket setting (due to small K and also in homogeneous scenarios, where the drug works in all baskets or none). The conservative artifact of overestimation of σ has the cost of efficiency (in less than optimal sharing of information across baskets), but this could be a desirable alternative to underestimation of σ at or near zero which can lead to an unacceptably high overall false-positive rate. Gelman’s results were derived in the traditional framework of hierarchical models with large K . To study whether his findings apply to the basket trial setting (adaptive trial with small K), we investigate three main family of priors: inverse-gamma on σ^2 , uniform (a, b) and half-t (s, df) on σ . For the half-t prior on σ , we note that assuming 1 degree of freedom is equivalent to assuming a half-Cauchy prior; furthermore, as the degrees of freedom approach infinity, this prior converges to a half-Normal prior. Finally, we note that specifying $a > 0$ results in a technical violation of the prior, since $a = 0$ is a possible sampling value and should be included with positive probability in the prior parameter space; however, we specify $a > 0$ in our simulation study to investigate the impact of this parameter on the overall operating characteristics.

Simulation study

We performed a simulation study motivated by our experience¹⁹ in these trials to compare the operating

characteristics of these three priors. We focus on the setting of $K = 5$ baskets and evaluate $K + 1$ configurations (i.e. scenarios) of the baskets' true effectiveness. That is, $A = 0$ baskets are active, $A = 1$ basket is active (assume basket 1 is active), $A = 2$ baskets are active (baskets 1 and 2 are active), and so forth to $A = K = 5$ baskets are active. We assume that in each basket, the true response rate p_k for $k = 1, \dots, K$ is either at a null response rate of $p_0 = 0.15$ or at a target effective response rate of $p_a = 0.45$. Consequently, the true basket-level standard deviation of the model parameters, that is, the log-odds of response $\sigma = 0$ when $A = 0$ or 5 active; $\sigma = 0.68$ when $A = 1$ or 4 active; and $\sigma = 0.84$ when $A = 2$ or 3 active. We assume that a maximum of $n_{\max} = 20$ patients per basket and at least $n_{\min} = 10$ patients within a basket are needed for interim and final analyses. We also assume equal accrual rates of two patients per month per basket.

Operating characteristics from 1000 simulated trials are presented in section "Results." For the inverse-gamma prior, we consider 35 combinations of $m_{\sigma^2} = \{0.1, 0.5, 1, 2, 10\}$ and $w_{\sigma^2} = \{0.01, 0.1, 0.5, 1, 2, 5, 10\}$; for the uniform prior, we consider 36 combinations of $a = \{0, 0.01, 0.05, 0.3, 0.5, 0.71\}$ and $b = \{1, 2, 3, 10, 100, 10000\}$; and for the half-t prior, we consider 35 combinations of scale parameters, $s = \{0.5, 1, 2.5, 10, 25, 100, 500\}$ and $df = \{1, 2, 5, 10, 100\}$ with each combination corresponding to a different design. The final decision rule (γ , see section "Bayesian adaptive design") for each prior specification is calibrated to achieve a family-wise error rate (FWER) of 10% ($\pm 2\%$ margin due to

simulation error) when the drug does not work in any of the baskets, where the FWER is the proportion of simulated trials in which at least one inactive basket is incorrectly declared active (or the probability of at least one type 1 error). All simulations were completed in R version 3.4.0 and Gibbs sampling was completed in JAGS as called from R using *rjags*.²⁰ Within each simulated trial, 10,000 MCMC iterations were kept for inference with 2000 MCMC iterations for burn-in. We set the hyperparameters for the shared mean μ to be $m_\mu = 0$ and $v_\mu = 10$, to reflect uncertainty that there is no treatment effect.

For each scenario, we consider the following *operating characteristics*: marginal probabilities of declaring the drug active in each basket (i.e. marginal power in active baskets and marginal false-positive rate in inactive baskets), FWER and trial size (N). In Tables 1–3 for select prior combinations, we present the average posterior mean estimate of the basket-level standard deviation ($\hat{\sigma}$). We display the performance of each design using the operating characteristics' average and range over all $K + 1$ scenarios, that is, $A = 0, 1, \dots, K = 5$.

Results

Table 1 displays three strata with the operating characteristics for three prior specifications using an inverse-gamma prior with different prior means and weights (see first column ($m_{\sigma^2}, w_{\sigma^2}$)). The second column

Table 1. Operating characteristics: inverse-gamma prior.

$(m_{\sigma^2}, w_{\sigma^2})$	Scenario	FWER	Marginal rejection probability (%)					N	$\hat{\sigma}$
			Basket 1	Basket 2	Basket 3	Basket 4	Basket 5		
(0.1, 10)	0 Active	8	4	4	4	4	4	66	0.11
	1 Active	48	52	38	38	38	38	85	0.11
	2 Active	84	85	85	80	80	80	102	0.11
	3 Active	98	99	99	99	98	98	99	0.11
	4 Active	100	100	100	100	100	100	81	0.11
	5 Active	—	100	100	100	100	100	67	0.11
(10, 10)	0 Active	9	2	2	2	2	2	85	9.05
	1 Active	8	86	2	2	2	2	85	8.99
	2 Active	6	87	87	2	2	2	85	8.94
	3 Active	4	87	87	87	2	2	87	8.87
	4 Active	2	87	87	87	87	2	86	8.81
	5 Active	—	86	86	86	86	86	86	8.74
(1, 2)	0 Active	10	2	2	2	2	2	81	1.00
	1 Active	13	88	3	3	3	3	89	1.19
	2 Active	14	90	90	5	5	5	92	1.24
	3 Active	15	95	95	95	8	8	93	1.18
	4 Active	11	96	96	96	96	11	90	1.04
	5 Active	—	97	97	97	97	97	83	0.85

m_{σ^2} is the prior mean value of σ^2 and w_{σ^2} is the prior weight value for m_{σ^2} ; scenario displays the number of baskets in which the drug truly works; FWER is the family-wise error rate; marginal rejection probabilities for declaring the drug works in active (power) and inactive baskets (false positive); N is the expected trial size; $\hat{\sigma}$ is the average posterior estimate of the standard deviation.

Table 2. Operating characteristics: uniform prior.

(a, b)	Scenario	FWER	Marginal rejection probability (%)					N	$\hat{\sigma}$
			Basket 1	Basket 2	Basket 3	Basket 4	Basket 5		
(0, 1)	0 Active	10	4	3	3	3	3	75	0.44
	1 Active	22	86	8	8	9	8	90	0.63
	2 Active	32	94	94	14	15	14	98	0.68
	3 Active	36	98	98	98	24	21	98	0.67
	4 Active	39	99	99	99	99	39	91	0.58
(0, 100)	5 Active	—	100	100	100	99	100	76	0.39
	0 Active	10	2	3	2	2	3	79	1.40
	1 Active	20	87	6	6	5	7	89	2.08
	2 Active	19	93	94	8	7	7	92	2.15
	3 Active	20	95	96	96	11	12	94	1.77
(0.3, 10)	4 Active	21	96	98	96	97	21	90	1.25
	5 Active	—	99	98	99	98	98	78	0.61
	0 Active	12	3	3	3	3	3	82	1.38
	1 Active	14	88	4	4	4	4	88	1.88
	2 Active	15	93	93	6	6	6	92	1.93
	3 Active	14	95	95	95	8	8	92	1.74
	4 Active	13	96	96	96	96	13	89	1.37
	5 Active	—	98	98	98	98	98	82	0.82

a and *b* are the lower and upper bounds, respectively; scenario displays the number of baskets in which the drug truly works; FWER is the family-wise error rate; marginal rejection probabilities for declaring the drug works in active (power) and inactive baskets (false positive); N is the expected trial size; $\hat{\sigma}$ is the average posterior estimate of the standard deviation.

Table 3. Operating characteristics: half-t prior.

(s, df)	Scenario	FWER	Marginal rejection probability (%)					N	$\hat{\sigma}$
			Basket 1	Basket 2	Basket 3	Basket 4	Basket 5		
(0.5, 100)	0 Active	8	3	3	3	3	2	72	0.35
	1 Active	28	81	11	11	12	12	89	0.57
	2 Active	38	94	95	19	17	20	99	0.67
	3 Active	41	98	97	98	27	29	99	0.64
	4 Active	49	99	99	99	99	49	90	0.52
(500, 100)	5 Active	—	100	100	100	100	100	73	0.31
	0 Active	9	2	2	3	2	2	79	1.48
	1 Active	18	86	5	5	6	5	88	2.15
	2 Active	18	93	94	8	7	7	93	2.20
	3 Active	18	96	95	95	11	10	93	1.82
(10, 1)	4 Active	17	96	97	97	97	17	89	1.29
	5 Active	—	99	99	98	98	98	78	0.60
	0 Active	10	2	3	3	2	3	79	1.26
	1 Active	16	88	5	5	5	4	88	1.87
	2 Active	20	94	94	8	9	8	93	2.04
	3 Active	17	95	96	96	10	11	93	1.70
	4 Active	20	96	97	98	98	20	89	1.28
	5 Active	—	99	98	99	99	98	78	0.58

s and *df* are the scale and degrees of freedom, respectively; scenario displays the number of baskets in which the drug truly works; FWER is the family-wise error rate; marginal rejection probabilities for declaring the drug works in active (power) and inactive baskets (false positive); N is the expected trial size; $\hat{\sigma}$ is the average posterior estimate of the standard deviation.

displays the number of active baskets. Next, we display the FWER, marginal rejection probabilities (of the null hypothesis of no treatment effect) and expected trial size. The final column presents the average posterior mean estimate of the basket-level standard deviation.

The first and second strata have the same weight ($w_{\sigma^2} = 10$) but the prior mean is increased away from zero (where a value of zero indicates homogeneity across all baskets). Both specifications represent fairly extreme prior specifications: for $m_{\sigma^2} = 0.1$, the prior strongly suggests little heterogeneity between baskets.

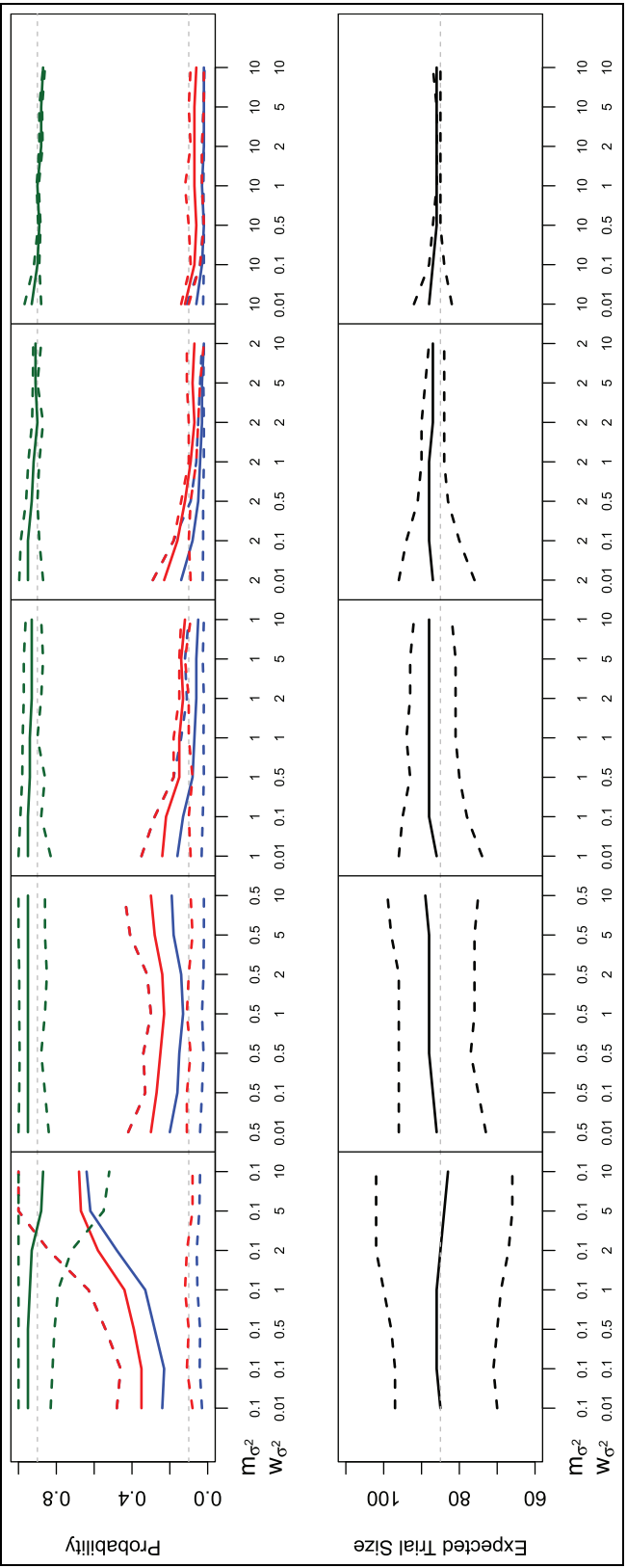


Figure 1. The x-axis displays the assumed prior mean and weight hyperparameters for the inverse-gamma prior distribution on σ^2 . (Top) The average and range of the power (green), marginal false-positive rate (blue) and family-wise error rate (red) across all scenarios ($A = 0, 1, \dots, K = 5$ active baskets). (Bottom) The average and range of the expected trial size across all scenarios.

Subsequently, the FWER rapidly increases from the calibrated 8% when the drug is not efficacious in any baskets (0 Active) to 48% when the drug works in only one basket (1 Active) to as high as 100% when the drug works in all but one basket (4 Active). For $w_{\sigma^2} = 10$, the prior strongly suggests heterogeneity; the FWER decreases as the number of truly active baskets increases, from the calibrated 9% (in the null scenario) to 2% (4 Active). Note that these results show that the design works essentially like a set of independent trials, with no information sharing. In the third stratum when $m_{\sigma^2} = 1$ and $w_{\sigma^2} = 2$, we place a modest prior belief that the treatment effect varies between baskets. This specification displays desirable results over all scenarios. The FWER ranges from the calibrated 10% in the null scenario to 15% in the more heterogeneous scenario (3 Active). When the drug works in one basket, the design has 88% power to identify this basket with a 3% false-positive rate in each of the four inactive baskets. Since this prior pushes more mass away from zero compared to assuming a smaller prior mean (i.e. first stratum), there is a loss in efficiency for the expected trial size in homogeneous scenarios such as 0 or 5 active, but there is a gain in efficiency in heterogeneous scenarios such as 2 or 3 active baskets.

Figure 1 displays summaries of the full simulation results assuming the inverse-gamma prior. In Figure 1, the top plot displays the average (solid line) and range (dashed lines) over all scenarios ($A = 0 - 5$) of the marginal power (green lines), marginal false-positive rates (blue lines) and FWER (red lines); the bottom plot displays the average (solid line) and range (dashed lines) of the trial size. In both plots, the x -axis displays the 35 combinations of the mean, m_{σ^2} (top x -axis value) and weight, w_{σ^2} (bottom x -axis value) hyperparameter inputs for the inverse-gamma prior, ordered by the mean values. For example, the design in the first stratum of Table 1 is represented in the first panel of Figure 1 on the seventh tick of the x -axis for ($m_{\sigma^2} = 0.1, w_{\sigma^2} = 10$); the FWER of this design ranges from 8% (0 Active) to 100% (4 Active) and this range is represented in Figure 1 with the bottom and top red dashed lines, respectively. In the first top panel of Figure 1 for a small prior value of σ^2 , as the prior weight increases, the prior distribution more strongly supports a small estimate of σ^2 by putting more mass near zero, which results in an increase in the average FWER and decrease in the average power but the range of both metrics increases (i.e. worse performance in heterogeneous scenarios but better performance in homogeneous scenarios). As the prior mean value increases (from 0.1 to 10), the average and range of our operating characteristics become more desirable. However, for a fixed prior mean value >1 , the range of our operating characteristics dramatically narrows and the design displays properties similar to implementing independent designs.

In short, when the prior distribution places too much mass near zero, we see a large range in operating characteristics; that is, the design can be overpowered when the drug works in all or most baskets but can have high false-positive rates when the drug works in only some baskets. We observe that this is the case for many seemingly reasonable prior specifications of the inverse-gamma. However, when we increase the prior mean value, the range of our operating characteristics narrows. This is because there is a small number of baskets to estimate σ^2 , and so as we push more prior mass away from zero, the model encourages less sharing across baskets and results in a loss of efficiency and decrease in power. Based on these results, the inverse-gamma prior in an adaptive basket trial is highly sensitive to input values, which is consistent with the findings of Gelman.¹⁸

Similar to Table 1, Table 2 displays three strata with the operating characteristics for three prior specifications using a uniform prior with different lower and upper bounds. Here, the first column displays the assumed lower and upper bounds (a, b). The first and second strata of Table 2 have the same lower bound ($a = 0$) but the upper bound, that is, domain of σ is increased. When $a = 0$ and $b = 1$ (Table 2, Stratum 1), the FWER increases from the calibrated 10% under the null scenario to 22% when the drug only works in basket 1 to 39% when the drug works in all but one basket. The narrow domain of the uniform prior from the small upper bound $b = 1$ imposes little heterogeneity between baskets and forces the model to share a certain level of information regardless of the truth which results in the large error rates when the drug works in some baskets but not all or none. Assuming a larger upper bound of $b = 100$ with $a = 0$ (Table 2, Stratum 2) results in a more desirable range of FWERs (10%–21%) while observing similar power across all scenarios. However, increasing b results in a more efficient trial should the true configuration of the baskets be heterogeneous (see 2 Active rows) at the cost of efficiency should the baskets be homogeneous (see 0 Active rows). Finally, $a = 0.3$ and $b = 10$ (Table 2, Stratum 3) result in better operating characteristics than in the first two strata of Table 2. Here, the largest FWER observed is 15% in the most heterogeneous scenario ($A = 2$) while observing similar power to the other two designs across all scenarios. This design observes similar efficiency to the design with the larger upper bound $b = 100$ when the baskets are truly heterogeneous but the cost is a larger expected trial size in the homogeneous scenarios.

Figure 2 displays summaries of the full simulation results from our investigation of the uniform prior. Similar to Figure 1, the top plot of Figure 2 displays the average and range of the marginal rejection probabilities and FWERs, and the bottom plot displays the average and range of the trial size. In both plots, the x -axis displays the 36 combinations of a and b

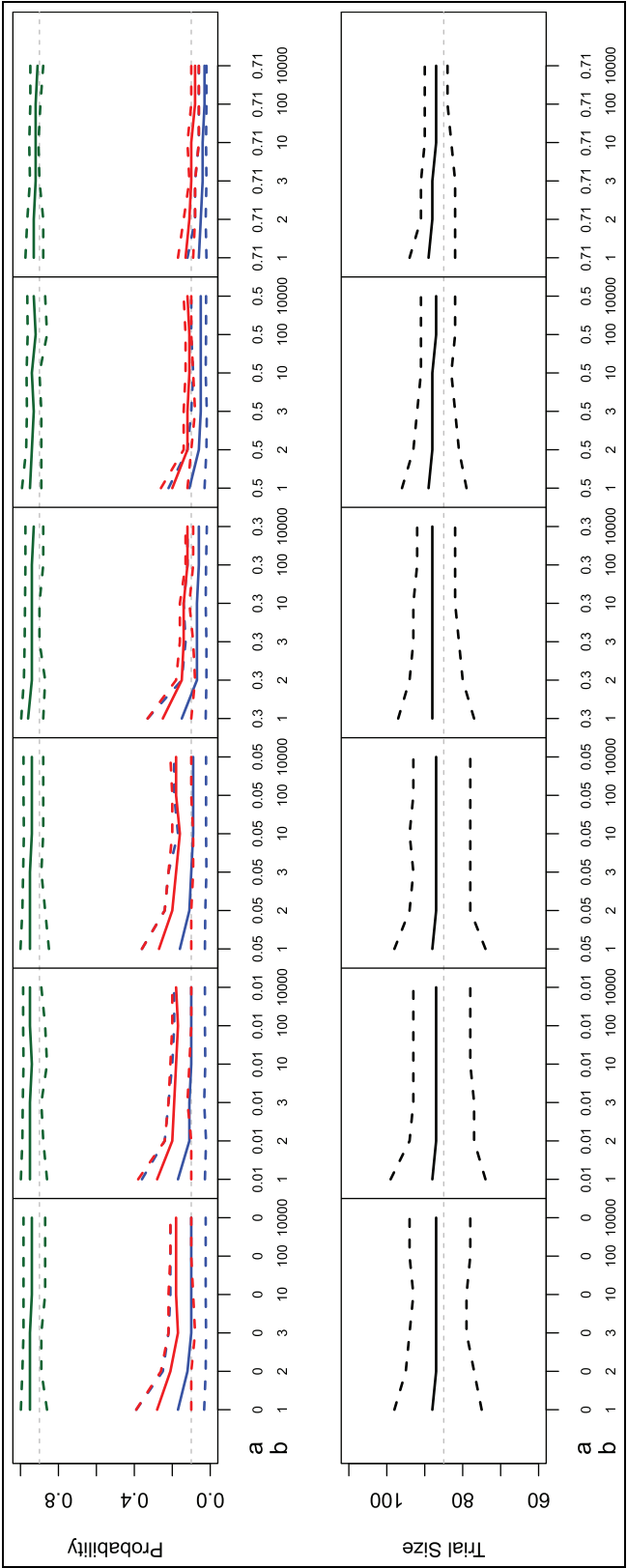


Figure 2. The x-axis displays the assumed prior lower and upper bound hyperparameters for the Uniform (a, b) prior distribution on σ . (Top) The average and range of the power (green), marginal false-positive rate (blue) and family-wise error rate (red) across all scenarios ($A = 0, 1, \dots, K = 5$ active baskets). (Bottom) The average and range of the expected trial size across all scenarios.

hyperparameter inputs for the uniform (a, b) prior, ordered by a . Looking across panels in Figure 2, we can see the range for the power, error rates and sample size decrease as we increase the lower bound of the uniform prior on σ away from zero. Looking within each panel, we can see fairly robust results when the upper bound b is greater than 1. In this simulation study, for large enough a , the performance of the design is similar to that of independent designs. This is because we are artificially imposing at least a^2 amount of variability into the model which encourages less sharing of information across baskets as a increases. While some specifications of $a > 0$ displayed desirable performance in our simulation study, we cannot recommend these prior specifications as a robust choice in a practical trial without supporting prior information for $a > 0$. Based on these results, the uniform prior is fairly robust assuming a lower bound of 0 and upper bound greater than 1.

Table 3 displays three strata with the operating characteristics for three prior specifications using a half-t prior with different scales and degrees of freedom. Here, the first column displays the assumed scale (s) and degree of freedom (df). The first and second strata of Table 3 assume a large degree of freedom ($df = 100$) which approximates the half-Normal prior on σ , with increasing scale, s . For small scale $s = 0.5$, the FWER increases from the calibrated 8% under the null scenario to 28% under the one active scenario to 49% under the four active scenarios. For larger $s = 500$, the FWER is controlled at around 18% across all heterogeneous scenarios (calibrated at 8% for 0 active). In the last two strata, we can see the half-Normal prior with large scale and half-Cauchy prior with modest scale display very similar and desirable results.

Figure 3 displays summaries of the full simulation results from our investigation of the half-t prior. The top plot of Figure 3 displays the average and range of the marginal rejection probabilities and FWERs, and the bottom plot displays the average and range of the trial sample size. In both plots, the x-axis displays the 35 combinations of s and df hyperparameter inputs for the half-t(s, df) prior, ordered by df . Looking within a panel in the top plot of Figure 3, we can see that the average and range for the power and error rates decrease as we increase the scale of our half-t prior on σ ; in the bottom plot, the range of the expected sample sizes decreases as we increase the scale parameter (s). Looking across panels, we observe very similar performance for all degrees of freedom considered ($df = 1 - 100$).

In short, similar to the uniform prior, decreasing the scale parameter s results in efficient trial sizes and slightly higher power in homogeneous scenarios, at the cost of slightly higher error rates in heterogeneous scenarios. This is because as the scale parameter decreases, more mass is placed near zero and less in the tail of the

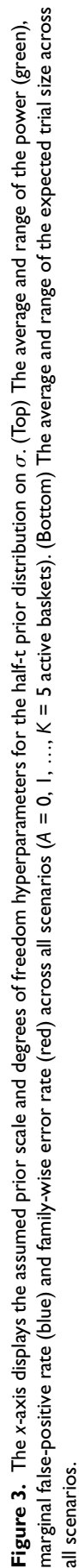
prior distribution. Based on these results, the half-t prior in an adaptive basket trial is very robust assuming a scale parameter greater than 1. We note that for the half-Cauchy prior ($df = 1$), as the scale parameter (s) approaches infinity, the prior distribution is equivalent to assuming a uniform prior on σ .¹⁸

In the foregoing analyses, the combination of hyperparameters (m_{σ^2} and w_{σ^2} for inverse-gamma; a and b for uniform; s and df for half-t) was not selected to achieve comparable prior distributions and subsequently are not equally calibrated in regard to prior information incorporated into the model; instead, hyperparameters were selected to capture both commonly used input values and extreme prior specifications. To relate the two prior specifications, Supplementary Table 1 displays quantiles of each prior distribution considered. For example, assuming $(m_{\sigma^2} = 0.1, w_{\sigma^2} = 0.01)$ ⁷ places 99% of the prior distribution below $4e-06$. This prior is akin to assuming a uniform prior with an extremely small domain, say $(0, 4e-06)$, which would never be considered a credible prior in practice. This points to another advantage of the uniform prior: the parameters are readily interpretable and weaknesses of a particular prior choice would be immediately evident. Similarly, half-t priors are easy to use since performance is robust to the df and any weaknesses can be attributed to s , the scale parameter.

Conclusion

When designing an adaptive basket trial using Bayesian hierarchical modeling, we suggest investigators avoid a prior distribution that is densely concentrated near zero and instead consider a prior distribution with sufficient mass in the tail, such as a uniform ($a = 0, b > 1$) or half-t ($s > 1, df$) prior on the basket-level standard deviation. In our investigation, we found the inverse-gamma prior to be very sensitive to input values depending on the true configuration of the baskets. On the other hand, in our simulation study, we found the half-t prior to display desirable and robust operating characteristics over a wide range of prior distributions considered. In our investigation of the uniform prior, specifying a lower bound $a > 0$ is technically incorrect. Notwithstanding this violation, we found that setting a larger a resembles a more conservative approach in protecting against oversharing in heterogeneous scenarios although it comes with a loss of efficiency and decrease in power in homogeneous scenarios. Nevertheless, in our simulation study, we found that the uniform prior with a lower bound set at 0 and an upper bound greater than 1 produced desirable and robust performance as well. Furthermore, our conclusions remain consistent when we vary accrual rates (see Supplementary Figures 1–4).

Bayesian hierarchical models are widely studied and used for larger experimental or observational studies. It



is important to note that our findings are limited to the cases where the number of groups (to share information across) is small and there is limited information in the data about σ^2 . This is a challenge that is particular to basket trials; most other applications of hierarchical models will have several (in some cases hundreds of) random effects.¹⁸ There is also the issue that the variance parameter in hierarchical models is central to the questions posed by a basket trial, whereas in many applications, it is considered a nuisance parameter. Finally, as we argued before, the choice of inverse-gamma is more habitual than carefully considered in many cases and specifying such a prior in a conventional Bayesian manner can have severe implications in erroneously declaring the drug works in futile baskets.

The ability to estimate the basket-level variability is gravely limited if the number of baskets is small (say 4 or 5), which is often the case in the setting of basket trials, and it is clear that the prior distribution strongly influences the final posterior of σ^2 . Therefore, it is our conclusion that it is impossible for a prior distribution to be non-informative in this basket trial setting, and thus it is essential to use a prior distribution with more robust and conservative properties such as the uniform or half-t distributions.

The heterogeneous scenarios where the drug works in some baskets, but not all or none, have been empirically shown to be likely, based on previously published basket trials, and in such cases, the particular exchangeability we assumed in this Bayesian hierarchical model is violated. Alternate approaches that do not require such an assumption can be pursued; however, other prior specifications can be just as cumbersome and less easily understood. More complex modeling approaches to remedy the lack of exchangeability across all baskets, such as Bayesian hierarchical mixture modeling, have been proposed in the literature¹¹ to design a basket trial. We believe that these approaches have the potential to be beneficial in many basket trial settings and have found in preliminary work the results in our simulation study are applicable to these other complex models (such as mixture models) that use a shared variance parameter in an adaptive basket trial but more work is needed to verify.

In our investigation, we examined the average and range of operating characteristics across all scenarios to evaluate the performance of the various prior specifications. In practice, a more formal utility function could be developed to help guide prior selection, taking into account the desired trade-off between efficiency, power and false-positive rates across all possible configurations. Furthermore, we chose to calibrate the decision rule for each design to weakly control the FWER at 10% when $A = 0$ active baskets; we acknowledge that other calibration schemes may be optimal and should

be investigated, such as calibrating under different scenarios and/or for FWER and power. Other preliminary results (not shown) reveal that model shrinkage behavior is consistent for other calibration approaches considered. The purpose of this simulation study is to evaluate three commonly used variance priors in a basket trial with recommendations, and we recommend that investigators consider a uniform prior with a modest domain or a half-t prior with a modest scale parameter on the standard deviation.

R code for the simulation study presented in section “Simulation study” is provided at <https://github.com/kristenmay206/BTcode>.

Acknowledgements

The authors are very thankful to Dr Colin Begg of the Department of Epidemiology and Biostatistics at Memorial Sloan Kettering Cancer Center for his helpful comments and suggestions.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

This research was funded in part through the NCI awards CA008748 and CA163251.

Supplemental material

Supplemental material for this article is available online.

References

1. Hyman DM, Puzanov I, Subbiah V, et al. Vemurafenib in multiple nonmelanoma cancers with BRAF V600 mutations. *N Engl J Med* 2015; 373(8): 726–736.
2. Hyman DM, Smyth LM, Donoghue MT, et al. AKT inhibition in solid tumors with AKT1 mutations. *J Clin Oncol* 2017; 35(20): 2251–2259.
3. Hyman DM, Piha-Paul SA, Won H, et al. HER kinase inhibition in patients with HER2- and HER3-mutant cancers. *Nature* 2018; 554(7691): 189.
4. Li BT, Shen R, Buonocore D, et al. Ado-trastuzumab emtansine for patients with HER2-mutant lung cancers: results from a phase II basket trial. *J Clin Oncol* 2018; 36(24): 2532–2537.
5. Hainsworth JD, Meric-Bernstam F, Swanton C, et al. Targeted therapy for advanced solid tumors based on molecular profiles: early results from MyPathway, an open-label, phase IIa umbrella basket study. *Am Soc Clin Oncol*.
6. Chenard-Poirier M, Kaiser M, Boyd K, et al. Results from the biomarker-driven basket trial of RO5126766 (CH5127566), a potent RAF/MEK inhibitor, in RAS- or RAF-mutated malignancies including multiple myeloma. *Am Soc Clin Oncol*.

7. Berry SM, Broglio KR, Groshen S, et al. Bayesian hierarchical modeling of patient subpopulations: efficient designs of phase II oncology clinical trials. *Clin Trials* 2013; 10(5): 720–734.
8. Neuenschwander B, Wandel S, Roychoudhury S, et al. Robust exchangeability designs for early phase clinical trials with multiple strata. *Pharm Stat* 2015; 15(2): 123–134.
9. Simon R, Geyer S, Subramanian J, et al. The Bayesian basket design for genomic variant driven phase II trials. *Semin Oncol* 2016; 43(1): 13–18.
10. Cunanan KM, Iasonos A, Shen R, et al. An efficient basket trial design. *Stat Med* 2017; 36(10): 1568–1579.
11. Liu R, Liu Z, Ghadessi M, et al. Increasing the efficiency of oncology basket trials using a Bayesian approach. *Contemp Clin Trials* 2017; 63: 67–72.
12. Zhou W, Yuan A, Thieu T, et al. Phase II basket group sequential clinical trial with binary responses. *Austin Biom Biostat* 2017; 4(1): 1033.
13. Xu Y, Mueller P, Mitra R, et al. A nonparametric Bayesian basket trial design, 2016, <https://arxiv.org/abs/1612.02705>
14. Trippa L and Alexander BM. Bayesian baskets: a novel design for biomarker-based clinical trials. *J Clin Oncol* 2017; 35: 681–687.
15. Guo W, Ji Y and Catenacci DVT. A subgroup cluster-based Bayesian adaptive design for precision medicine. *Biometrics* 2017; 73(2): 367–377.
16. Ventz S, Barry WT, Parmigiani G, et al. Bayesian response-adaptive designs for basket trials. *Biometrics* 2017; 73(3): 905–915.
17. Browne WJ and Draper D. A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Anal* 2006; 1(3): 473–514.
18. Gelman A. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Anal* 2006; 1(3): 515–534.
19. Cunanan KM, Gonen M, Shen R, et al. Basket trials in oncology: a trade-off between complexity and efficiency. *J Clin Oncol* 2017; 35(3): 271.
20. Plummer M. rjags: Bayesian graphical models using MCMC (R package version 3–10), 2011, <http://CRAN.R-project.org/package=rjags>