*Memorial Sloan-Kettering Cancer Center*
Memorial Sloan-Kettering Cancer Center, Dept. of Epidemiology & Biostatistics Working Paper Series

*Year* 2017 *Paper* 35

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

Kristen Cunanan* Alexia Iasonos†
Ronglai Shen‡ Mithat Gonen**

*Memorial Sloan Kettering Cancer Center, cunanank@mskcc.org
†Memorial Sloan Kettering Cancer Center, iasonosa@mskcc.org
‡Memorial Sloan-Kettering Cancer Center, shenr@mskcc.org
**Memorial Sloan-Kettering Cancer Center, gonenm@mskcc.org



http://biostats.bepress.com/mskccbiostat/paper35

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

Kristen Cunanan, Alexia Iasonos, Ronglai Shen, and Mithat Gonen

**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 method used to capture the correlated endpoints across baskets is Bayesian hierarchical modeling. We evaluate a Bayesian adaptive design in the context of a basket trial and investigate two popular prior specifications: an inverse-gamma prior on the basket-level variance and a uniform prior on the basket-level standard deviation.

**Results:** From our simulation study, we see 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 ($>40\%$) in some scenarios. Thus, use of this prior requires a fully comprehensive sensitivity analysis before implementation. Alternatively, we see that a prior that moves the mass of the variance parameter away from zero, such as the uniform 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 must be larger than 1.

**Conclusion:** Based on our results, we recommend that those involved in designing basket trials that implement hierarchical modeling avoid using a prior distribution that places a large 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 on the standard deviation.

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

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

Department of Epidemiology and Biostatistics

Memorial Sloan Kettering Cancer Center

**Running Head**: Variance priors in basket trials

**Word Count**: 3988 (Abstract: 340)

**Corresponding Author**:

Kristen M. Cunanan, PhD

Research Scholar

Department of Biostatistics and Epidemiology

Memorial Sloan-Kettering Cancer Center

485 Lexington Avenue 2nd Floor, New York, NY 10017

Email: kristenmay206@gmail.com

Tel.: 646-888-8306

## 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 method used to capture the correlated endpoints across baskets is Bayesian hierarchical modeling. We evaluate a Bayesian adaptive design in the context of a basket trial and investigate two popular prior specifications: an inverse-gamma prior on the basket-level variance and a uniform prior on the basket-level standard deviation.

**Results:** From our simulation study, we see 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 ($\leq 0.5$), 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 moves the mass of the variance parameter away from zero, such as the uniform 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 must be larger than 1.

**Conclusion:** Based on our results, we recommend that those involved in designing basket trials that implement hierarchical modeling avoid using a prior distribution that places a large 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 on the standard deviation.



**Keywords:** basket trial; phase II; Bayesian method; adaptive design; variance prior

# 1 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, either a common drug or a common disease group.

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 more 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, 2, 3, 4].

Several phase II designs applicable to basket trials have been proposed [5, 6, 7, 8, 9, 10] to account for the anticipated correlated endpoints 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 [11, 12, 13, 14]. 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 two commonly used priors: an inverse-gamma prior on the basket-level variance and a uniform 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 design and model for the analysis but 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 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 manuscript proceeds as follows: Section 2 presents the Bayesian adaptive design, hierarchical model, prior specifications used, and presents the simulation study, in Section 3 the results, and Section 4 concludes with recommendations and a brief discussion.

## 2 Methods

In Section 2.1, we present a Bayesian adaptive design following the design originally proposed by Berry *et al.* [5] 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. 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 $i = 1, \ldots, n_k$ and $p_k$ be the probability of response in basket $k$ for $k = 1, \ldots, K$. 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.

## 2.1 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 (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:

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

where $n_{min}$ is the minimum number of patients required in a basket to evaluate efficacy.

The final analysis occurs after the numbers of patients enrolled in the remaining basket(s) have treated 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.

## 2.2 Bayesian Hierarchical Model

We assume $\sum_i y_{ik}$ follows a binomial distribution of size $n_k$ with probability $p_k$. Similar to Berry *et al.* [5], to obtain the decision probabilities described in Section 2.1 above we apply a *logit* transformation to the basket-specific probabilities of a response to facilitate a Bayesian model, as follows:

$$\theta_k = logit(p_k) - logit(p_0),$$

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,

$$\theta_k \sim Normal(\mu, \sigma^2)$$

$$\mu \sim Normal(m_\mu, v_\mu)$$
$$\sigma^2 \sim g(.)$$

where $m_\mu$ and $v_\mu$ are pre-specified mean and variance hyperparameters and $g(.)$ is an appropriate distribution for the variance (i.e. sharing) parameter $\sigma^2$, such as the inverse-gamma$(\alpha, \beta)$. A more interpretable re-parameterization of the inverse-gamma distribution specifies a prior mean of $\sigma^2$ (define as $m_{\sigma^2}$) and prior effective sample size, i.e. weight (define as $w_{\sigma^2}$) [15], where $\alpha = w_{\sigma^2}/2$ and $\beta = m_{\sigma^2}^2 w_{\sigma^2}/2$.

Other functional forms for $g(.)$ are proposed in the literature. Gelman has investigated numerous prior distributions for the variance parameter in conventional hierarchical linear models [16], including: inverse-gamma and uniform on $\sigma^2$, half-Cauchy and uniform on $\sigma$, and uniform on $\log(\sigma^2)$. He notes that a " prior distribution cannot put an infinite mass near zero, since the data can never rule out a group-level variance of zero in a hierarchical linear model", and goes on to recommend using a uniform density on $\sigma$, but mentions the uniform$(0, b)$ prior on $\sigma$ can lead to overestimation of $\sigma$ and less than optimal sharing of information across groups when the number of groups is small. Also, he mentions the inverse-gamma prior is sensitive to input values when small values of $\sigma$ are possible in the data, and we note 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 $\sigma$ has the cost of efficiency (in less than optimal sharing of information across baskets) but this could be a desirable alternative to underestimation of $\sigma$ 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 (small $K$), we investigate a uniform$(a, b)$ prior on $\sigma$ and compare the results with the preceding inverse-gamma prior.

## 2.3 Simulation Study

We performed a simulation study motivated by our experience [17] in these trials to compare the operating characteristics of these two 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 $k = 1, \ldots, K$ 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, i.e. 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 a maximum of $n_{max} = 20$ patients per basket and at least $n_{min} = 10$ patients within a basket are needed for inferences. We also assume equal accrual rates of 2 patients per month.

Operating characteristics from 1000 simulated trials are presented in Section 3. For the inverse-gamma prior, we consider 35 combinations of $m_{\sigma^2} = \{0.1, 0.5, 1, 2, 10\}$ and $w_{\sigma^2} = \{0.01, 0.1, 0.5, 1, 2, 5, 10\}$; and for the uniform prior, we consider 36 combinations of $a = \{0, 0.01, 0.05, 0.3, 0.5, 0.71\}$ and $b = \{1, 2, 3, 10, 100, 10000\}$, with each combination corresponding to a different design. The final decision rule ($\gamma$, see Section 2.1) for each prior specification is calibrated to achieve a family-wise error rate of 10% ($\pm$ 2% margin due to simulation error) when the drug does not work in any of the baskets, where the *family-wise error rate* (FWER) is the proportion of simulated trials in which at least one inactive basket(s) is incorrectly declared active. All simulations were completed in R version 3.4.0 and Gibbs sampling was completed in JAGS as called from R using *rjags* [18]. Within each simulated trial, 10000 MCMC iterations were kept for inference with 2000 MCMC iterations for burn-in. We set the hyperparameters for the shared mean $\mu$ 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), family-wise error rate (FWER), and trial size (N). In Tables 1 and 2 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, i.e. $A = 0, 1, \ldots, K = 5$.

# 3 Results

Table 1 displays 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 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). The second column displays the number of active baskets. Next, we display the family-wise error rate, 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.



When $m_{\sigma^2} = 0.1$ and $w_{\sigma^2} = 10$ (Table 1 Stratum 1), the prior strongly suggests little heterogeneity

between baskets. Subsequently, the family-wise error rate 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) with only 52% power to identify the active basket. The family-wise error rate is as high as 100% when the drug works in all but one basket (4 Active), while achieving 100% power in the active baskets. Clearly, this prior specification encourages the model to share information readily regardless of the baskets' true configuration, as reflected in the consistent estimate of $\hat{\sigma} = 0.11$ across all scenarios. When $m_{\sigma^2}$ is instead set to 10 with $w_{\sigma^2} = 10$ (Table 1 Stratum 2), which strongly suggests heterogeneity, the family-wise error rate decreases from the calibrated 9% (in the null scenario) as the number of truly active baskets increases. Regardless of the number of truly active baskets this design achieves 86-87% power to identify active baskets with a 2% false positive rate for inactive baskets. Observe that these results show that the design works essentially like a set of independent trials, with no information sharing. When $m_{\sigma^2} = 1$ and $w_{\sigma^2} = 2$ (Table 1 Stratum 3), we place a modest prior belief the treatment effect varies between baskets and we see this specification displays desirable results over all scenarios. The family-wise error rate 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 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 family-wise error rate (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, $m_{\sigma^2}$ (top x-axis value) and weight, $w_{\sigma^2}$ (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 family-wise error rate 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 $\sigma^2$, as the prior weight increases, the prior distribution more strongly supports a small estimate of $\sigma^2$ by putting more mass near zero, which results in an increase in the average family-wise error rate 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 over-powered 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 little data to estimate $\sigma^2$, 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 [16].

Similar to Table 1, Table 2 displays 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 have the same lower bound ($a = 0.05$) but the upper bound, i.e. domain of $\sigma$, is increased. When $a = 0.05$ and $b = 1$ (Table 2 Stratum 1), the family-wise error rate increases from the calibrated 9% under the null scenario to 25% when the drug only works in Basket 1 with 87% power to identify this basket. As the number of truly active baskets increases, the family-wise error rate continues to increase to 39% when the drug works in all but one basket, with 99% power to identify the four active baskets. 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 $a = 0.05$ (Table 2 Stratum 2), results in a more desirable range of family-wise error rates (10-19%, with the largest rate in the most heterogeneous scenarios) 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), results in better operating characteristics than in the first two strata of Table 2. Here, the largest family-wise error rate 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 family-wise error rates; 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 $b$ hyperparameter inputs for the uniform$(a, b)$ prior, ordered by $a$. For example, the design Table 2 Stratum 1 is represented in Figure 2 with the first tick in the third panel; here, the family-wise error rate ranges from 9% to 39% and is displayed with the bottom and top red dashed lines in Figure 2, while the average family-wise error rate across all scenarios (not displayed in Table 1) is displayed with the red solid line. Looking across panels in the top plot of Figure 2, we see the average and range for the power and error rates decrease as we increase the lower bound of our uniform prior on $\sigma$ away from zero; in the bottom plot, the range of the expected sample sizes decreases as we increase $a$ but the average remains fairly constant.

In short, decreasing the lower bound $a$ results in efficient trial sizes and slightly higher power in homogeneous scenarios, at the cost of slightly higher error rates in heterogeneous scenarios. Similar to the inverse-gamma's hyperpameter $m_{\sigma^2}$, as we increase the lower bound $a$ of the uniform prior away from zero, we see the range of our operating characteristics narrows. 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, and results in a loss of efficiency and decrease in power in more homogeneous scenarios. Based on these results, the uniform prior in an adaptive basket trial is fairly robust assuming an upper bound greater than 1.

In the foregoing analyses the combination of hyperparameters ($m_\sigma^2$ and $w_\sigma^2$ for inverse-gamma; $a$ and $b$ for uniform) were not selected to achieve comparable prior distributions and subsequently are not equally calibrated in regards to prior information incorporated into the model; instead hyperparameters were selected to capture both commonly used input values and extreme model behavior. To relate the two prior specifications, Supplementary Materials Table 1 displays quantiles of each prior distribution considered. For example, assuming ($m_{\sigma^2} = 0.1$, $w_{\sigma^2} = 0.01$) [5] places 99% of the prior distribution below 4e-06. This prior is akin to assuming a uniform prior with a 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.

# 4 Conclusion

When designing an adaptive basket trial using Bayesian hierarchical modeling, we recommend using a prior distribution that places a large density mass away from zero, such as a uniform$(a, b)$ prior on the basket-level standard deviation with the upper bound $b$ set to a value greater than 1. 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 uniform prior to display desirable and robust operating characteristics over a wide range of prior distributions considered. 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 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 $\sigma^2$. 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 [16]. 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 the prior distribution strongly influences the final posterior of $\sigma^2$. 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 distribution.

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 literature [9] to design a basket trial. We believe 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 each design to weakly control the FWER at 10% when $A = 0$ active baskets; we acknowledge other calibration schemes may be optimal and should be investigated. Other preliminary results (not shown), reveal model shrinkage behavior is consistent for other calibration approaches considered. The purpose of this simulation study is to evaluate two commonly used variance priors in a basket trial with recommendations; and we recommend that investigators avoid using the inverse-gamma prior and instead consider a uniform prior with a modest domain on the standard deviation.

R code for the simulation study presented in Section 2.3 is provided at: `https://gist.github.com/kristenmay206/461384bb6c082c49bf855447db5c66cd`

## Acknowledgment

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. The authors gratefully acknowledge this research was funded in part through the NCI awards CA008748 and CA163251.

## References

[1] Hyman DM, Puzanov I, Subbiah V, et al. Vemurafenib in multiple nonmelanoma cancers with BRAF V600 mutations. *New England Journal of Medicine*, 373(8):726–736, 2015.

[2] 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. American Society of Clinical Oncology, 2016.

[3] Li BT, Shen R, Buonocore D, et al. Ado-trastuzumab emtansine in patients with HER2 mutant lung cancers: Results from a phase II basket trial. American Society of Clinical Oncology, 2017.

[4] 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. American Society of Clinical Oncology, 2017.

[5] Berry SM, Broglio KR, Groshen S, et al. Bayesian hierarchical modeling of patient subpopulations: Efficient designs of phase II oncology clinical trials. *Clinical Trials*, 10:720–734, 2013.

[6] Neuenschwander B, Wandel S, Roychoudhury S, et al. Robust exchangeability designs for early phase clinical trials with multiple strata. *Pharmaceutical Statistics*, 15(2):123–134, 2015.

[7] Simon R, Geyer S, Subramanian J, et al. The Bayesian basket design for genomic variant driven phase II trials. *Seminars in Oncology*, pages 1–6, 2016.

[8] Cunanan KM, Iasonos A, Shen R, et al. An efficient basket trial design. *Statistics in Medicine*, 36(10):1568–1579, 2017.

[9] Liu R, Liu Z, Ghadessi M, et al. Increasing the efficiency of oncology basket trials using a Bayesian approach. *Contemporary Clinical Trials*, 2017.

[10] Zhou W, Yuan A, Thieu T, et al. Phase II basket group sequential clinical trial with binary responses. *Austin Biometrics and Biostatistics*, 4(1):1033, 2017.

[11] Xu Y, Mueller P, Mitra R, et al. A nonparametric Bayesian basket trial design. *(Online pre-print)*, 2016.

[12] Trippa L and Alexander BM. Bayesian baskets: A novel design for biomarker-based clinical trials. *Journal of Clinical Oncology*, 2016.

[13] Guo W, Ji Y, and Catenacci DV. A subgroup cluster-based Bayesian adaptive design for precision medicine. *Biometrics*, 73(2):367–377, 2017.

[14] Ventz S, Barry WT, Parmigiani G, et al. Bayesian response-adaptive designs for basket trials. *Biometrics*, 2017.

[15] Browne WJ, Draper D, et al. A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis*, 1(3):473–514, 2006.

[16] Gelman A et al. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, 1(3):515–534, 2006.

[17] Cunanan KM, Gonen M, Shen R, et al. Basket trials in oncology: A trade-off between complexity and efficiency. *Journal of Clinical Oncology*, 35(3):271–273, 2016.

[18] Plummer M. *rjags: Bayesian graphical models using MCMC*, 2011. R package version 3-10.

Table 1: **Operating Characteristics: Inverse-Gamma Prior**

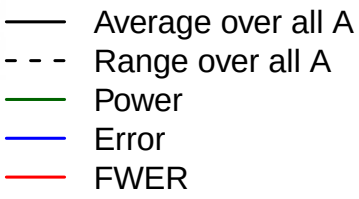
| $(m_{\sigma^2}, w_{\sigma^2})$ | Scenario | FWER | Marginal Rejection Probability (%) | | | | | N | $\hat{\sigma}$ |
|---|---|---|---|---|---|---|---|---|---|
| | | | Basket 1 | Basket 2 | Basket 3 | Basket 4 | Basket 5 | | |
| (0.1,10) | 0 Active | 8 | 4 | 4 | 4 | 4 | 4 | 66 | 0.11 |
| | 1 Active | 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 |

$m_{\sigma^2}$ is the prior mean value of $\sigma^2$ and $w_{\sigma^2}$ is the prior weight value for $m_{\sigma^2}$; 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 | | |
| (0.05, 1) | 0 Active | 9 | 3 | 3 | 3 | 3 | 3 | 74 | 0.47 |
| | 1 Active | 25 | 87 | 8 | 8 | 8 | 8 | 90 | 0.64 |
| | 2 Active | 32 | 94 | 94 | 15 | 15 | 15 | 99 | 0.68 |
| | 3 Active | 39 | 98 | 98 | 98 | 25 | 25 | 98 | 0.67 |
| | 4 Active | 39 | 99 | 99 | 99 | 99 | 39 | 90 | 0.59 |
| | 5 Active | - | 100 | 100 | 100 | 100 | 100 | 76 | 0.42 |
| (0.05, 100) | 0 Active | 10 | 2 | 2 | 2 | 2 | 2 | 79 | 1.49 |
| | 1 Active | 19 | 89 | 6 | 6 | 6 | 6 | 89 | 2.10 |
| | 2 Active | 19 | 93 | 93 | 7 | 7 | 7 | 92 | 2.11 |
| | 3 Active | 19 | 96 | 96 | 96 | 11 | 11 | 93 | 1.88 |
| | 4 Active | 18 | 97 | 97 | 97 | 97 | 18 | 89 | 1.32 |
| | 5 Active | - | 98 | 98 | 98 | 98 | 98 | 79 | 0.63 |
| (0.3, 10) | 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.



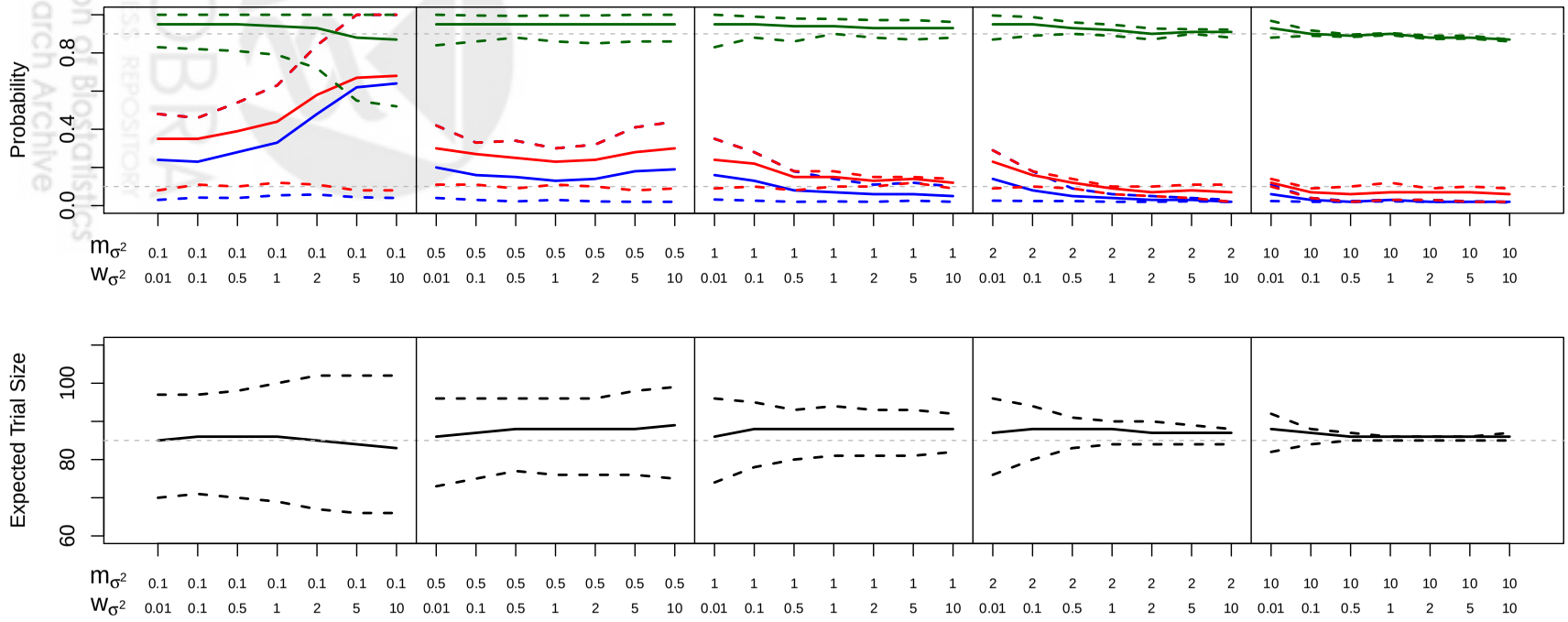



Figure 1: The x-axis displays the assumed prior mean and weight hyperparameters for the Inverse Gamma prior distribution on $\sigma^2$. (Top) Presented are the average and range of the power (green), marginal false positive rate (blue), and family wise error rate (red) across all scenarios (A = 0, 1,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

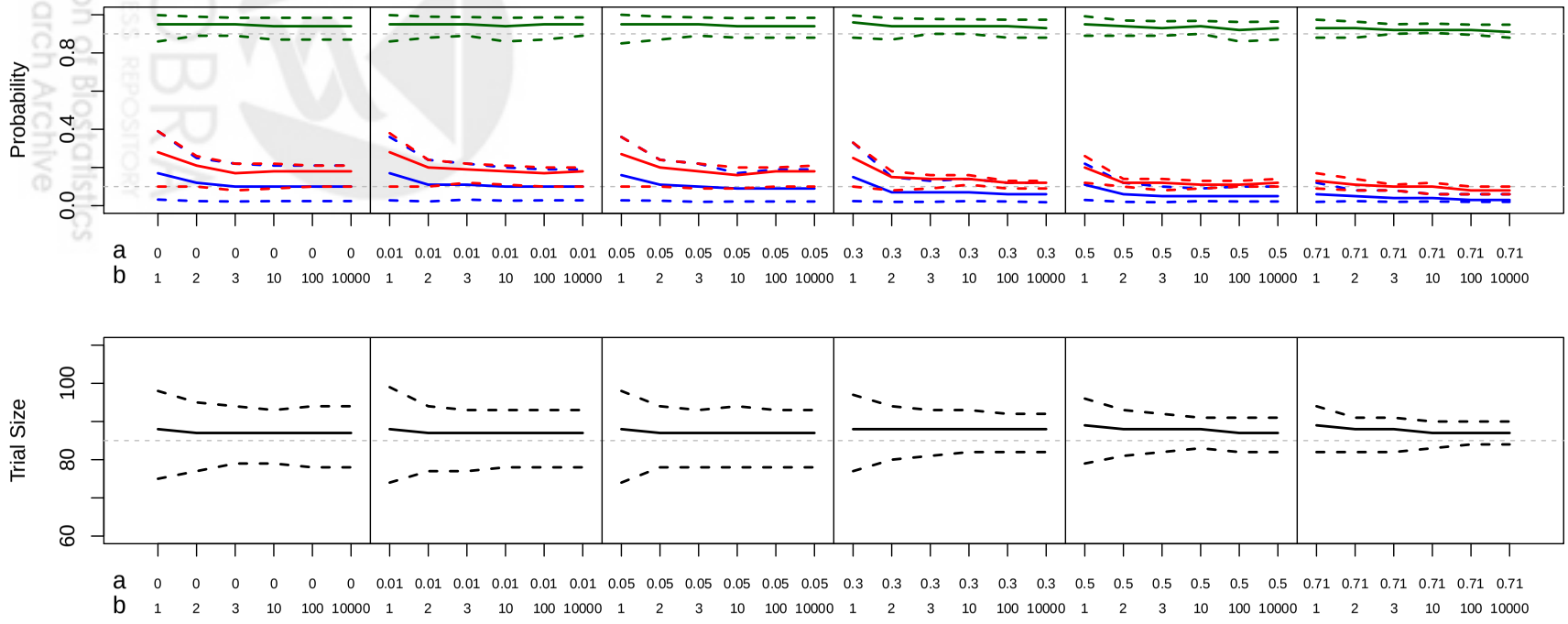



Figure 2: The x-axis displays the assumed prior lower and upper bound hyperparameters for the Uniform(a,b) prior distribution on $\sigma$. (Top) Presented are the average and range of the power (green), marginal false positive rate (blue), and family wise error rate (red) across all scenarios (A = 0,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