

A Bayesian basket trial design using a calibrated Bayesian hierarchical model

Clinical Trials
2018, Vol. 15(2) 149–158
© The Author(s) 2018
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/1740774518755122
journals.sagepub.com/home/ctj



Yiyi Chu¹  and Ying Yuan²

Abstract

Background: The basket trial evaluates the treatment effect of a targeted therapy in patients with the same genetic or molecular aberration, regardless of their cancer types. Bayesian hierarchical modeling has been proposed to adaptively borrow information across cancer types to improve the statistical power of basket trials. Although conceptually attractive, research has shown that Bayesian hierarchical models cannot appropriately determine the degree of information borrowing and may lead to substantially inflated type I error rates.

Methods: We propose a novel calibrated Bayesian hierarchical model approach to evaluate the treatment effect in basket trials. In our approach, the shrinkage parameter that controls information borrowing is not regarded as an unknown parameter. Instead, it is defined as a function of a similarity measure of the treatment effect across tumor subgroups. The key is that the function is calibrated using simulation such that information is strongly borrowed across subgroups if their treatment effects are similar and barely borrowed if the treatment effects are heterogeneous.

Results: The simulation study shows that our method has substantially better controlled type I error rates than the Bayesian hierarchical model. In some scenarios, for example, when the true response rate is between the null and alternative, the type I error rate of the proposed method can be inflated from 10% up to 20%, but is still better than that of the Bayesian hierarchical model.

Limitation: The proposed design assumes a binary endpoint. Extension of the proposed design to ordinal and time-to-event endpoints is worthy of further investigation.

Conclusion: The calibrated Bayesian hierarchical model provides a practical approach to design basket trials with more flexibility and better controlled type I error rates than the Bayesian hierarchical model. The software for implementing the proposed design is available at http://odin.mdacc.tmc.edu/~yyuan/index_code.html

Keywords

Basket trials, Bayesian hierarchical model, adaptive borrowing, borrow information, Bayesian adaptive trial design

Introduction

Traditional phase II oncology clinical trials have been designed to evaluate a single treatment in patients of a particular cancer type. With tremendous advances in cancer biology and genomic medicine, the forefront of cancer research has shifted from conventional chemotherapy to targeted therapy that treats cancer by targeting a specific genetic or molecular aberration.¹ The basket trial is a trial design that accommodates such a paradigm shift.^{2,3} As illustrated in Figure 1, under the basket trial, patients with the same genetic or molecular aberration, regardless of their cancer types, are enrolled in the trial for evaluating the effect of a targeted agent. The basket trial allows for the incorporation of precision medicine into clinical trials that also evaluate molecular aberrations that are too rare to

study solely within a tumor-specific context.⁴ In addition, the basket trial often requires fewer patients and a shorter duration to identify a favorable response to the targeted therapy.^{2,5,6}

Despite increasing recognition of trial designs based on genetic or molecular aberrations as opposed to cancer types, evaluating targeted therapies in basket trials

¹Department of Biostatistics and Data Science, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX, USA

²Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

Corresponding author:

Ying Yuan, Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA.
Email: yyuan@mdanderson.org

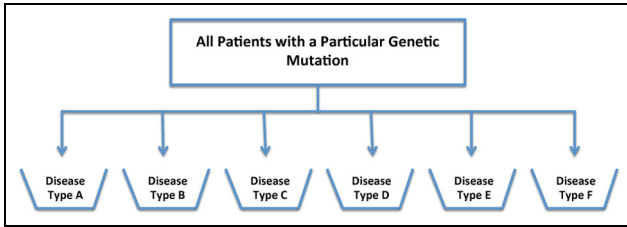


Figure 1. An illustration of basket trials.

is challenging. Although the patients enrolled in a basket trial have the same genetic or molecular aberration, that does not necessarily mean that they will respond homogeneously to a targeted agent regardless of the primary tumor site. Tumor type often has profound effects on the treatment effect, and it is not uncommon for a targeted agent to be effective for some tumor types, but not others. As a result, when evaluating the treatment effect in basket trials, an analysis that simply pools the data across tumor types is often problematic. It leads to large biases and inflated type I error rates if the treatment effect actually is heterogeneous across different tumor types. The independent approach, which evaluates the treatment effect in each tumor type independently, avoids these issues, but is less efficient and often lacks power to detect the treatment effect due to the limited sample size in each tumor type or tumor subgroup. For convenience, we use tumor type and tumor subgroup to mean the same thing.

To overcome the drawbacks of pooled and independent approaches, Thall et al.⁷ first proposed using a Bayesian hierarchical model (BHM) to adaptively borrow information across different tumor subgroups, and Berry et al.⁸ applied it to basket trials. In that approach, the shrinkage parameter, which controls the strength of information borrowing, is treated as an unknown parameter following a noninformative prior and let the data determine how much information should be borrowed across tumor subgroups. Although conceptually attractive, Freidlin and Korn⁹ showed that such a fully Bayesian approach does not work appropriately and leads to substantially inflated type I error rates for basket trials with 10 or fewer cancer subgroups, because there is insufficient information in the observed data to determine whether borrowing across subgroups is appropriate (i.e. to accurately estimate the shrinkage parameter). Our numerical study also confirms that result.

We propose a Bayesian phase II basket trial design based on a novel calibrated Bayesian hierarchical model (CBHM). The treatment effect in cancer subgroups is modeled using a hierarchical model. However, unlike Berry et al.,⁸ in our approach, the shrinkage parameter is not regarded as an unknown parameter. Instead, it is defined as a function of a similarity measure of the treatment effect across tumor subgroups. The key is

that the function is calibrated using simulations such that information is strongly borrowed across subgroups if their treatment effects are similar and barely borrowed if the treatment effects are heterogeneous. The simulation study shows that our method has substantially better controlled type I error rates compared to those of the BHM.

Our research is motivated by a phase II basket trial at MD Anderson Cancer Center for patients with Neurotrophic tropomyosin receptor kinase (NTRK) aberration advanced solid tumors. The trial investigated a novel tropomyosin receptor kinase (TRK) inhibitor that targets the NTRK fusion, which has been found in various types of advanced tumors and contributes to tumorigenesis by driving activation of intracellular signaling molecules.¹⁰ NTRK aberration appears in non small cell lung cancer, thyroid cancer, sarcoma, and colorectal cancer. The goal of the trial is to evaluate the efficacy of the TRK inhibitor in patients with cancers that harbor a NTRK gene fusion. The trial includes patients with the four cancer types listed above and will enroll up to 30 patients for each cancer type. The treatment efficacy will be scored using the Response Evaluation Criteria in Solid Tumors, version 1.1, and coded as “response” if the patient achieves complete or partial remission (CR/PR), otherwise “no response.” The targeted agent will be regarded as unpromising if the response rate is lower than 20% and promising if the response rate is higher than 35%.

The remainder of this article is organized as follows. In section “Methods,” we propose the CBHM approach for designing basket trials. In section “Simulation studies,” we present the simulation studies to evaluate the operating characteristics of the proposed design. We conclude with a brief discussion in section “Discussion.”

Methods

BHM

Consider a phase II basket trial that evaluates the efficacy of a new targeted agent in J different tumor subgroups that share the same genetic or molecular aberrations. Let p_j and N_j respectively denote the response rate and maximum sample size for the j tumor subgroups. The objective of the trial is to test whether the targeted agent is effective in each of the tumor subgroups

$$H_0 : p_j \leq q_0 \quad \text{versus} \quad H_a : p_j \geq q_1 \quad \text{for } j = 1, \dots, J$$

where q_0 is the response rate cutoff under which the drug is deemed futile, and q_1 is the target response rate under which the drug is regarded as promising.

Suppose at an interim go/no-go decision time, n_j patients from tumor subgroup j have been enrolled, among which y_j patients responded favorably to the

treatment. We assume that y_j follows a hierarchical model

$$\begin{aligned} y_j|p_j &\sim \text{Bin}(p_j) \\ \theta_j &= \log\left(\frac{p_j}{1-p_j}\right) \\ \theta_j|\theta, \sigma^2 &\sim N(\theta, \sigma^2) \\ \theta &\sim N(\alpha_0, \omega_0^2) \end{aligned} \tag{1}$$

where $\text{Bin}(\cdot)$ denotes a binomial distribution, α_0 and ω_0^2 are hyperparameters. This hierarchical model shrinks the subgroup-specific treatment effect θ_j toward the common mean θ , thereby borrowing information across different tumor subgroups. The degree of shrinkage or information borrowing is controlled by the shrinkage parameter σ^2 . A small value of σ^2 induces strong information borrowing across tumor subgroups; whereas a large value of σ^2 induces little information borrowing. As the extreme cases, $\sigma^2 = 0$ is equivalent to the pooled approach, which assumes that the drug is equivalently effective across all different tumor subgroups, and $\sigma^2 = \infty$ is equivalent to the independent approach, where the data in different tumor subgroups are analyzed independently, which is appropriate when the drug is effective for some subgroups, but not for other subgroups.

The fully Bayesian approach, such as that of Berry et al.,⁸ assigns a prior distribution to σ^2 and estimates it jointly with other parameters. Ideally, given the interim data, we would like the model to automatically sense the similarity or homogeneity in the treatment effect across tumor subgroups, based on which we could determine the appropriate degree of information borrowing. That is, if the treatment effect is homogeneous across tumor subgroups (i.e. the drug is effective for all subgroups), the BHM strongly shrinks θ_j toward θ to borrow information and improve the statistical power for detecting the treatment effect, and when the treatment effect is heterogeneous (i.e. the drug is effective for some tumor subgroups but not for other subgroups), the BHM does not induce shrinkage to maintain a proper type I error rate. Unfortunately, this is not the case for typical basket trials that have a small or moderate number (e.g. 3–10) of tumor subgroups. Freidlin and Korn⁹ showed that when there are 10 or fewer tumor subgroups, the BHM approach cannot provide accurate estimation of the shrinkage parameter σ^2 and thus fails to make appropriate information borrowing. Our simulation study (described in section “Simulation studies”) shows that the BHM approach can inflate the type I error rate from the nominal value of 10% to over 50%.

The problem stems from the fact that the shrinkage parameter σ^2 represents the variance between tumor subgroups and the observation unit used to estimate σ^2 is the tumor subgroups, not patients. Therefore, even

when there are a large number of patients in each tumor subgroup, but limited number of tumor subgroups, the data cannot provide adequate information to estimate σ^2 reliably. This is analogous to a random-effects model-based meta-analysis, for which it is difficult to obtain a reasonably precise estimate for the between-trial variability if only a few trials are available.¹¹

CBHM

To address the aforementioned issues, we propose a CBHM approach to adaptively borrow information across tumor subgroups for phase II basket trials. Unlike the BHM approach, which assigns a prior to σ^2 and estimates it from the data, our approach defines σ^2 as a function of the measure of homogeneity among the tumor subgroups. The key is that the function is prespecified and calibrated in a way such that when the treatment effects in the tumor subgroups are homogeneous, strong information borrowing occurs and thus improves power, and when the treatment effects in the tumor subgroups are heterogeneous, little or no borrowing across groups occur, thereby controlling the type I error rate. In what follows, we first describe a homogeneity measure and then describe a procedure to determine and calibrate the function that links the homogeneity measure and shrinkage parameter σ^2 .

A natural measure of homogeneity is the chi-squared test statistic of homogeneity, given by

$$T = \sum_{j=1}^J \frac{(O_{0j} - E_{0j})^2}{E_{0j}} + \sum_{j=1}^J \frac{(O_{1j} - E_{1j})^2}{E_{1j}}$$

where O_{0j} and O_{1j} denote the observed counts of failures and responses for subgroup j (i.e. $n_j - y_j$ and y_j), and E_{0j} and E_{1j} are the expected counts of failures and responses, given by

$$E_{0j} = n_j \frac{\sum_j n_j - \sum_j y_j}{\sum_j n_j} \quad \text{and} \quad E_{1j} = n_j \frac{\sum_j y_j}{\sum_j n_j}$$

A smaller value of T indicates higher homogeneity in the treatment effect across subgroups. Note that the chi-squared test statistic T is used here for measuring the strength of homogeneity, not conducting hypothesis testing. Therefore, when some cell counts (i.e. O_{0j} and O_{1j}) are small, it does not cause any issue because our procedure does not rely on the large-sample distribution of T .

We link the shrinkage parameter σ^2 with T through

$$\sigma^2 = g(T) \tag{2}$$

where $g(\cdot)$ is a monotonically increasing function. Although different choices of $g(\cdot)$ are certainly possible, our numerical studies show that the following two-parameter exponential model yields good and robust operating characteristics

$$\sigma^2 = g(T) = \exp\{a + b \times \log(T)\} \tag{3}$$

where a and b are tuning parameters that characterize the relationship between σ^2 and T . We require $b > 0$ such that greater homogeneity (i.e. a small value of T) leads to stronger shrinkage (i.e. a small value of σ^2).

The key to our approach is that the values of a and b are calibrated such that strong shrinkage occurs when the treatment effect is homogeneous across the tumor subgroups, and no or little shrinkage occurs when the treatment effect is heterogeneous. This can be done using the following three-step simulation-based procedure:

1. Simulate the case in which the treatment is effective for all tumor subgroups; thus, we should borrow information across tumor subgroups. Specifically, we generate R replicates of data by simulating $\mathbf{y} = (y_1, \dots, y_J)$ from $\text{Bin}(\mathbf{N}, \mathbf{q}_1)$, where $\mathbf{N} = (N_1, \dots, N_J)$ and $\mathbf{q}_1 = (q_1, \dots, q_1)$ and then calculate T for each simulated dataset. Let H_B denote the median of T from R simulated datasets.
2. Simulate the cases in which the treatment effect is heterogeneous across tumor subgroups; thus, we should not borrow information across tumor subgroups. Let $\mathbf{q}(j) = (q_1, \dots, q_1, q_0, \dots, q_0)$ denote the scenario in which the treatment is effective for the first j subgroups with the response rate of q_1 , but not effective for subgroups $j + 1$ to J with the response rate of q_0 . Given a value of j , we generate R replicates of data by simulating \mathbf{y} from $\text{Bin}(\mathbf{N}, \mathbf{q}(j))$, calculate T for each simulated dataset and then obtain its median H_{Bj} . We repeat this for $j = 1, \dots, J - 1$ and define

$$H_B = \min_j (H_{Bj}) \tag{4}$$

3. Let σ_B^2 denote a prespecified small value (e.g. 1) for shrinkage parameter σ^2 under which strong shrinkage or information borrowing occurs under the hierarchical model (equation (1)), and let σ_B^2 denote a prespecified large value (e.g. 80) of shrinkage parameter σ^2 , under which little shrinkage or information borrowing occurs. Solve a and b in equation (3) based on the following two equations

$$\sigma_B^2 = g(H_B; a, b) \tag{5}$$

$$\sigma_B^2 = g(H_B; a, b) \tag{6}$$

where the first equation enforces strong shrinkage (i.e. information borrowing) when the treatment is effective for all subgroups, and the second equation enforces little shrinkage (i.e. weak information borrowing) when the treatment effect is heterogeneous across all subgroups. The solution of the equations is given by

$$a = \log(\sigma_B^2) - \frac{\log(\sigma_B^2) - \log(\sigma_B^2)}{\log(H_B) - \log(H_B)} \log(H_B) \tag{7}$$

$$b = \frac{\log(\sigma_B^2) - \log(\sigma_B^2)}{\log(H_B) - \log(H_B)} \tag{8}$$

Remarks. In step 2, we define H_B as the minimum value of the $\{H_{Bj}\}$, that is, equation (4), and impose that little shrinkage occurs when $T = H_B$, as dictated by equation (6) in step 3. This is to reflect that when the treatment is effective for some subgroups, but not effective for the other subgroup(s), the treatment effect is regarded as heterogeneous and no information should be borrowed across subgroups. For example, in a basket trial with four tumor subgroups, if the treatment is effective for three subgroups, but not effective for one subgroup, the treatment effect is regarded as heterogeneous and no information should be borrowed across subgroups. Such “strong” definition of heterogeneity and “strong” control of borrowing is necessary for controlling type I error, and any information borrowing will inflate the type I error of the ineffective subgroup due to the shrinkage of that subgroup’s treatment effect toward the effective subgroups, that is, overestimating the treatment effect for the ineffective subgroup. Because the shrinkage parameter σ^2 is a monotonically increasing function of T , as long as we control that little shrinkage occurs when $T = H_B$, we automatically ensure that little shrinkage occurs for cases with larger values of H_{Bj} (i.e. higher levels of heterogeneity), for example, when the treatment is effective for two subgroups but not effective in the other two subgroups. In some situations, for example, when the majority of subgroups are responsive and only one subgroup is not responsive, it may be debatable whether controlling type I error for each of subgroups is the best strategy. We might be willing to tolerate a certain type I error inflation in one subgroup in exchange for power gain in the majority of subgroups. This can be conveniently done by relaxing the “strong” definition of heterogeneity. For example, rather than defining $H_B = \min_j (H_{Bj})$, we can define H_B as the minimal value of H_{Bj} when the treatment is not effective in at least two subgroups. That is, if the treatment is not effective for only one subgroup, we do not treat the treatment effect as heterogeneous. Consequently, information can be borrowed across subgroups, but at the expense of some inflated type I error for the ineffective subgroup. Actually, this is one of important advantages of the proposed CBHM over the standard BHM. The CBHM provides us abundant flexibility to control the degree of borrowing in an intuitive and straightforward way.

Another advantage of the proposed CBHM is that the proposed calibration procedure relies only on the null response rate q_0 , alternative response rate q_1 , and

sample size of tumor groups N_j , which are known before the trial is conducted. This is an important and very desirable property because it allows the investigator to determine the values of a and b and to include them in the study protocol before the onset of the study. This avoids the common concern about the method for borrowing information; that is, the method could be abused by choosing the degree of borrowing to favor a certain result, for example, statistical significance. When the true response rates of some subgroups are between q_0 and q_1 , the CBHM induces partial information borrowing, depending on the actual value of homogeneity measure T . In addition, the resulting CBHM has the following desirable large-sample property. The proof is provided in Appendix 1.

Theorem 1. When the sample size in each subgroup is large, the CBHM achieves full information borrowing when the treatment effect is homogeneous across subgroups, and no information borrowing when the treatment effect is heterogeneous across subgroups.

In contrast, to achieve the above asymptotic property, the standard BHM requires an additional assumption that the number of subgroups is large because the shrinkage parameter σ^2 represents the inter-subgroup variance. To precisely estimate σ^2 and ensure appropriate shrinkage behavior, we must increase the number of subgroups. This extra requirement, unfortunately, is restrictive and often unrealistic in practice because the number of tumor subgroups with a certain genetic or molecular aberration is often fixed and cannot be manipulated within the trial design.

Trial design

The proposed phase II basket trial design has a total of K interim looks, with the k th interim observation occurring when the sample size of the j th subgroup reaches n_{jk} . Let $\mathcal{D}_k = \{(n_{jk}, y_{jk}), j = 1, \dots, J, k = 1, \dots, K\}$ denote the data from the k th interim look, where y_{jk} is the number of responses from n_{jk} patients. The proposed phase II basket trial design with K interim looks is described as follows:

1. Enroll n_{j1} patients in the j th subgroup, $j = 1, \dots, J$.
2. Given the data \mathcal{D}_k from the k th interim look
 - (a) (*Futility stopping*) if $\Pr(p_j > (q_0 + q_1)/2 | \mathcal{D}_k) < C_f$, suspend the accrual for the j th subgroup, where $(q_0 + q_1)/2$ denotes the rate halfway between the null and target response rate and C_f is a probability cutoff for futility stopping.
 - (b) otherwise, continue to enroll patients until reaching the next interim analysis.
3. Once the maximum sample size is reached or the treatment of all subgroups is stopped early due to

futility, evaluate the efficacy for each subgroup based on all the observed data. If $\Pr(p_j > q_0 | \mathcal{D}) > C$, then the treatment for the j th group will be declared effective; otherwise, the treatment for that group will be declared ineffective, where C is a probability cutoff.

In step 2a, to facilitate the simulation comparison of the proposed design with the BHM design,⁸ we use $(q_0 + q_1)/2$ as the boundary for assessing futility, as the latter design. Another natural boundary is q_0 , that is, if there is a high posterior probability that p_j is less than q_0 , we stop the accrual for the j th subgroup for futility. To ensure good operating characteristics, the probability cutoffs C_f and C should be calibrated through simulations to achieve a desired type I error rate and early stopping rate for each subgroup. This simulation-based calibration procedure is widely used in Bayesian clinical trial designs.^{12,13}

Simulation studies

We investigated the operating characteristics of the proposed CBHM design through simulation studies. Following the setting of NTRK basket trial, we considered four subgroups with null $q_0 = 0.2$ and alternative $q_1 = 0.35$. The maximum sample size for each subgroup was 30, with two interim analyses conducted when the sample size in each subgroup reached 10 and 20. We compared the proposed CBHM design to the independent approach, where each subgroup is analyzed independently without information borrowing, and to the BHM approach. In the proposed CBHM approach, we calibrated the values of a and b in equation (3) using the procedure described in section “CBHM,” with $\sigma_B^2 = 1$ and $\sigma_B^2 = 80$, resulting in $a = -5.98$ and $b = 6.83$. We assigned θ a vague normal prior $N(-1.39, 100)$, where the prior mean -1.39 is obtained as the average of $\theta_j, j = 1, \dots, 4$, under the null hypothesis. For the BHM approach, following Berry et al.,⁸ we used a noninformative inverse gamma prior $IG(0.0005, 0.000005)$ for σ^2 . We also considered an alternative half-normal prior $HN(0, 0.5)$ for σ^2 , that is, a normal distribution $N(0, 0.5)$ left truncated at 0. We denote the resulting design as BHM-HN. To ensure fair comparison between the different approaches, we set $C_f = 0.05$ for all the considered scenarios and calibrated C for each design such that when all tumor subgroups are not responsive to the drug (i.e. scenario 1 in Table 1), the type I error rate is 10% in each subgroup. We considered a total of 33 scenarios with various response rates for the subgroups. Under each scenario, we carried out 5000 simulated trials.

Table 1 shows the simulation results of 14 scenarios. The results of the other scenarios can be founded in Table A1 of the Supplementary Files. As described

Table 1. Rejection rate of the null hypothesis under independent, BHM, BHM-HN, and CBHM approaches.

Scenario	Method	Response rate of subgroup				Sample size
		1	2	3	4	
1	Independent	0.2	0.2	0.2	0.2	106.1
	BHM	0.099	0.101	0.101	0.101	92.1
	BHM-HN	0.100	0.098	0.098	0.098	96.4
	CBHM	0.098	0.098	0.099	0.098	95.5
2	Independent	0.35	0.35	0.35	0.35	118.6
	BHM	0.716	0.702	0.715	0.734	119.5
	BHM-HN	0.971	0.974	0.972	0.969	119.6
	CBHM	0.949	0.951	0.946	0.945	118.5
3	Independent	0.803	0.801	0.804	0.816	113.0
	BHM	0.45	0.2	0.2	0.45	117.6
	BHM-HN	0.954	0.101	0.097	0.952	117.8
	CBHM	0.978	0.394	0.401	0.977	112.9
4	Independent	0.978	0.333	0.333	0.977	116.4
	BHM	0.978	0.333	0.333	0.977	119.6
	BHM-HN	0.954	0.109	0.109	0.953	119.7
	CBHM	0.954	0.109	0.109	0.953	116.4
5	Independent	0.2	0.45	0.45	0.45	109.4
	BHM	0.104	0.952	0.953	0.953	107.2
	BHM-HN	0.553	0.991	0.993	0.990	108.5
	CBHM	0.491	0.992	0.993	0.989	105.9
6	Independent	0.124	0.956	0.957	0.957	109.4
	BHM	0.2	0.2	0.2	0.35	107.2
	BHM-HN	0.106	0.100	0.095	0.721	108.5
	CBHM	0.258	0.258	0.265	0.573	105.9
7	Independent	0.206	0.199	0.200	0.653	108.8
	BHM	0.128	0.127	0.123	0.685	103.1
	BHM-HN	0.128	0.127	0.123	0.685	105.4
	CBHM	0.128	0.127	0.123	0.685	103.3
8	Independent	0.3	0.2	0.2	0.2	108.8
	BHM	0.507	0.096	0.098	0.103	103.1
	BHM-HN	0.383	0.212	0.213	0.216	105.4
	CBHM	0.460	0.170	0.169	0.162	103.3
9	Independent	0.476	0.114	0.122	0.123	112.6
	BHM	0.2	0.2	0.35	0.45	116.7
	BHM-HN	0.101	0.095	0.714	0.951	116.7
	CBHM	0.420	0.421	0.844	0.952	112.1
10	Independent	0.325	0.322	0.838	0.965	112.1
	BHM	0.127	0.123	0.727	0.953	109.4
	BHM-HN	0.45	0.2	0.2	0.2	111.9
	CBHM	0.952	0.096	0.096	0.099	112.2
11	Independent	0.864	0.285	0.293	0.289	108.3
	BHM	0.922	0.225	0.227	0.221	100.9
	BHM-HN	0.937	0.225	0.227	0.221	92.2
	CBHM	0.937	0.109	0.113	0.112	94.4
12	Independent	0.15	0.15	0.15	0.35	95.1
	BHM	0.019	0.019	0.019	0.721	100.9
	BHM-HN	0.039	0.043	0.040	0.409	92.2
	CBHM	0.030	0.035	0.020	0.532	94.4
13	Independent	0.018	0.020	0.022	0.674	95.1
	BHM	0.15	0.15	0.35	0.35	106.7
	BHM-HN	0.021	0.016	0.714	0.719	109.7
	CBHM	0.144	0.147	0.676	0.683	109.9
14	Independent	0.087	0.082	0.712	0.719	109.9
	BHM	0.031	0.026	0.702	0.712	105.0
	BHM-HN	0.15	0.2	0.2	0.2	107.8
	CBHM	0.271	0.097	0.098	0.101	98.1
15	Independent	0.227	0.157	0.161	0.154	98.1
	BHM	0.227	0.157	0.161	0.154	101.3
	BHM-HN	0.258	0.135	0.136	0.128	101.3

(continued)

Table 1. Continued

Scenario	Method	Response rate of subgroup				Sample size
		1	2	3	4	
12	CBHM	0.261	0.108	0.112	0.112	99.7
	Independent	0.3	0.3	0.3	0.2	114
	BHM	0.498	0.494	0.495	0.102	115.2
	BHM-HN	0.725	0.725	0.731	0.523	115.7
	CBHM	0.687	0.692	0.685	0.362	112.3
13	Independent	0.2	0.25	0.25	0.35	112.8
	BHM	0.103	0.266	0.268	0.726	113.1
	BHM-HN	0.431	0.541	0.542	0.732	114
	BHM-HN	0.301	0.47	0.463	0.759	114
	CBHM	0.178	0.321	0.321	0.725	110.8
14	Independent	0.2	0.2	0.25	0.35	111
	BHM	0.103	0.098	0.268	0.72	110.7
	BHM	0.339	0.337	0.448	0.657	111.5
	BHM-HN	0.251	0.25	0.406	0.713	108.5
	CBHM	0.142	0.143	0.295	0.704	

BHM: Bayesian hierarchical model; CBHM: calibrated Bayesian hierarchical model; BHM-HN: Bayesian hierarchical model with a half-normal prior for the shrinkage parameter.

above, scenario 1 is used to calibrate the three designs such that they have the same type I error rate of 10% when the treatment is not effective for all subgroups. In scenario 2, the treatment is effective for all subgroups. The proposed CBHM had higher power than the independent approach. The power of the CBHM was around 81% for the subgroups, and that of the independent design was about 72%. The BHM and BHM-HN yielded higher power than the CBHM and independent designs; however, both failed to control the type I error rate when the subgroups were heterogeneous. For example, in scenario 3, subgroups 1 and 4 are responsive and subgroups 2 and 3 are not responsive. The type I error rate from the BHM for subgroups 2 and 3 was inflated to 39.4% and 40.1%, respectively. This result is consistent with previous findings⁹ that the BHM cannot accurately determine whether and how much information borrowing is appropriate across subgroups. Figure 2 shows the posterior distribution of shrinkage parameter σ^2 under scenarios 2 and 3. We can see that these two posterior distributions are essentially identical, suggesting that the BHM failed to distinguish the case when shrinkage is needed (i.e. scenario 2) from the case when shrinkage is not needed (i.e. scenario 3). We found that the BHM and BHM-HN tended to strongly shrink or borrow information across subgroups no matter whether the subgroups were homogeneous or heterogeneous. In contrast, the CBHM correctly detected that in scenario 3, the subgroups were dissimilar and no information should be borrowed. The type I error rate of the CBHM was close to the nominal value of 10% for subgroups 2 and 3 (10.9%). In addition, the CBHM design had smaller

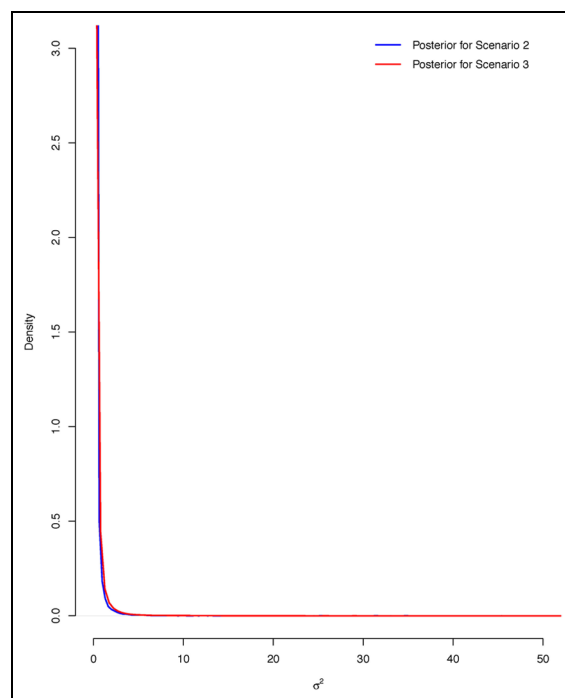


Figure 2. The posterior distributions of σ^2 under the BHM approach in scenario 2, in which strong shrinkage or information borrowing is appropriate, and scenario 3, in which little shrinkage or information borrowing is appropriate.

sample sizes than the BHM and BHM-HN design (112.9 patients vs 117.6 and 117.8). This is because the CBHM is more likely to recognize that subgroups 2 and 3 are not responsive and early terminate these two arms. The power of the CBHM was comparable to

that of the independent design for the two responsive subgroups 1 and 4. In scenario 4, only subgroup 1 is non-responsive to the drug of interest. The CBHM yielded a reasonable type I error rate (i.e. 12.4%); whereas the type I error rate of the BHM and BHM-HN were inflated to 55.3% and 49.1%, respectively. Again, the sample size of the CBHM was smaller than that of the BHM and BHM-HN. In scenario 5, subgroups 1, 2, and 3 are not responsive to the treatment. The type I error rate for the CBHM design was inflated slightly to 12%. In contrast, the type I error rate for the BHM design was inflated to 26%. Moreover, compared to the BHM design, the CBHM design had higher power to detect the responsive subgroup (i.e. 68.5% vs 57.3% for subgroup 4). In scenario 6, subgroup 1 is responsive to the treatment, and its true response rate lies in between the null and alternative rates. Our CBHM design was capable of recognizing the heterogeneous treatment effect across subgroups and maintained the type I error rate around 12%, whereas the BHM and BHM-HN designs had the inflated type I error rates of 22% and 17%, respectively. In addition, the CBHM design yielded higher power than the BHM design (i.e. 47.6% vs 38.3%). Similar results were observed in scenarios 7 to 11. In scenario 12, subgroups 1, 2, and 3 have true response rate of 0.3, between the null (i.e. 0.2) and the alternative (i.e. 0.35), while subgroup 4 is not responsive to the treatment. The type I error rate for subgroup 4 is 20.3% under the CBHM and 52.3% under the BHM. The reason that the CBHM led to an inflated type I error rate is that the CBHM is calibrated based on the null and alternative response rates to ensure that little information is borrowed when the subgroups are heterogeneous with some subgroups having the response rate of 0.35 and some subgroups having the response rate of 0.2. In scenario 12, however, the response rate of the responsive subgroups (i.e. subgroups 1, 2, and 3) is 0.3 (i.e. between null and alternative). The heterogeneity among subgroups is not large enough to forbid information borrowing, thereby leading to the inflated type I error. Nevertheless, the type I error rate of the CBHM is less than one half of that of the BHM. Similar results were observed in scenarios 13 and 14, where different numbers of subgroups have the true response rate between the null and the alternative. The type I error rate of CBHM is 17.8% and 14.2% in scenarios 13 and 14, respectively, substantially lower than that of the BHM (i.e. 43.1% and 33.9%).

One may note that although the CBHM design has better type I error control than the BHM approach, its performance seems comparable to that of the independent approach in most scenarios (i.e. scenarios 3–14). This is simply because the treatment effect is heterogeneous in most scenarios, under which no information should be borrowed across subgroups. In other words, we prefer the design performing as the independent

design in scenarios 3–14. In the case that the treatment effect is homogeneous (e.g. scenario 2) and borrowing is preferable, we can see that the CBHM provides substantial gain in power over the independent design. That is, the CBHM design is able to adaptively borrow information according to the homogeneity of the treatment effect in subgroups.

To gain insight on the different operating characteristics of CBHM and BHM designs and how much pooling is done during the trial, Table 2 shows the posterior estimate of shrinkage parameter σ^2 at the interims under different designs. Ideally, the estimate of σ^2 should be small in scenarios 1 and 2, where the treatment effect is homogeneous, to pool information across subgroups, and be large in scenarios 3 and 4, where the treatment effect is heterogeneous, so that no or little pooling occurs. The BHM and BHM-HN designs failed to differentiate the scenarios: the estimates of σ^2 are always small (i.e. $\hat{\sigma}^2 < 1$) in all four scenarios, inducing strong shrinkage (or pooling) throughout the trial no matter whether the treatment effect is homogeneous or heterogeneous. That is the reason why the BHM and BHM-HN yielded substantially inflated type I errors. In contrast, the CBHM design is responsive to the scenarios. When the treatment effect is homogeneous (i.e. scenarios 1 and 2), it generated relatively small values of σ^2 (< 3.75) to facilitate information borrowing across subgroups, and when the treatment is heterogeneous (i.e. scenarios 3 and 4), it generated large values of σ^2 to depress pooling and maintain appropriate type I errors. For example, the value of σ^2 is about 23,150 and 4335 at the end of the trial in scenarios 3 and 4, respectively.

Table 2. The posterior estimate of σ^2 at interim analyses under the BHM, BHM-HN, and CBHM designs.

Design	Interim		
	1	2	3
Scenario 1			
BHM	0.36	0.16	0.09
BHM-HN	0.22	0.17	0.14
CBHM	1.37	3.2	3.75
Scenario 2			
BHM	0.22	0.11	0.07
BHM-HN	0.18	0.13	0.10
CBHM	1.32	1.35	1.49
Scenario 3			
BHM	0.67	0.77	1.01
BHM-HN	0.30	0.38	0.45
CBHM	130.8	3520.9	23,150.5
Scenario 4			
BHM	0.44	0.40	0.49
BHM-HN	0.26	0.28	0.31
CBHM	43.2	835.6	4335.0

BHM: Bayesian hierarchical model; CBHM: calibrated Bayesian hierarchical model; BHM-HN: Bayesian hierarchical model with half-normal prior.

Table 3. Sensitivity analysis of the CBHM design based on different values of σ_B^2 and σ_B^2 .

Scenario	σ_B^2	σ_B^2	Response rate of subgroup				Sample size
			1	2	3	4	
1	0.5	80	0.2	0.2	0.2	0.2	93.1
	1.5	80	0.102	0.098	0.098	0.1	96.0
	1.0	50	0.1	0.101	0.101	0.101	95.5
2	0.5	80	0.099	0.101	0.1	0.099	95.5
	1.5	80	0.35	0.35	0.35	0.35	118.4
	1.0	50	0.812	0.8	0.818	0.828	118.5
3	0.5	80	0.799	0.799	0.802	0.815	118.5
	1.5	80	0.806	0.808	0.808	0.817	118.5
	1.0	50	0.45	0.2	0.2	0.45	112.8
4	0.5	80	0.960	0.106	0.109	0.955	112.8
	1.5	80	0.958	0.116	0.115	0.955	112.9
	1.0	50	0.962	0.124	0.124	0.956	112.9
5	0.5	80	0.2	0.45	0.45	0.45	116.4
	1.5	80	0.137	0.952	0.955	0.963	116.4
	1.0	50	0.125	0.958	0.96	0.961	116.4
6	0.5	80	0.133	0.963	0.965	0.962	116.4
	1.5	80	0.2	0.2	0.2	0.35	104.6
	1.0	50	0.139	0.130	0.130	0.688	104.6
7	0.5	80	0.131	0.128	0.131	0.692	106.1
	1.5	80	0.135	0.135	0.132	0.688	105.9
	1.0	50	0.3	0.2	0.2	0.2	101.6
8	0.5	80	0.487	0.120	0.126	0.132	101.6
	1.5	80	0.480	0.117	0.121	0.123	103.6
	1.0	50	0.492	0.125	0.126	0.120	103.4

We also examined the estimates of response rates at the end of the trial under different designs. To focus on the relative performance of the designs and untangle the effect of early stopping on the estimate, we used the estimates of independent approach as the benchmark for comparison. In addition, because the independent approach treated subgroups and estimated their treatment effects independently as conventional phase II trials, such comparison also provides a natural comparison of the novel designs with the conventional design. As shown in Figures A1 and A2 in Supplementary Files, when the treatment effect is homogeneous (e.g. scenarios 1 and 2), the posterior mean estimates were similar under different designs. The estimate of the CBHM was more efficient than the independent approach with a smaller variance (e.g. scenario 2). The BHM had the smallest variance in three designs, as it induced the strongest shrinkage. However, when the treatment effect is heterogeneous, as shown in scenarios 3–6, the BHM yielded biased estimates. For example, in scenario 3, the posterior mean estimates under the BHM for non-responsive subgroup 2 and 3 were overestimated by about 5%, and those for responsive subgroups 1 and 4 were underestimated by 5%. These biases stem from the fact that the BHM tends to strongly shrink across subgroups even when subgroups are heterogeneous. In contrast, as described previously,

the CBHM sensed the heterogeneity in these scenarios and induced little shrinkages, and thus yielded unbiased estimates similar to the independent approach.

Finally, we studied the sensitivity of the CBHM method with respect to the values of σ_B^2 and σ_B^2 used in the calibration procedure. We considered three alternative values: $(\sigma_B^2, \sigma_B^2) = (0.5, 80), (1.5, 80)$ and $(1, 50)$. The results (see Table 3) are generally similar to those reported in Table 1, suggesting that as long as σ_B^2 is adequately small such that little shrinkage occurs and σ_B^2 is adequately large such that strong shrinkage occurs, the choice of σ_B^2 and σ_B^2 has little impact on the performance of the proposed design.

Discussion

We have proposed a CBHM approach to evaluate the treatment effect for basket trials. By linking the shrinkage parameter with a measure of homogeneity among subgroups through an appropriately calibrated link function, the CBHM allows information borrowing when the treatment effect is homogeneous across subgroups and yields a much better controlled type I error rate than the BHM when the treatment effect is heterogeneous across subgroups.

Similar to the BHM, one limitation of the proposed CBHM is that it assumes that subgroups are

exchangeable. As a result, in order to control type I error for each of subgroups, the heterogeneity (in treatment effect) has to be defined in a “strong” sense that if the treatment is effective in some subgroups, but not effective in the other subgroup(s), the treatment effect is regarded as heterogeneous and no information should be borrowed across the subgroups. This is the reason why the CBHM performs similarly to the independent approach in most simulation scenarios (e.g. scenarios 3–11) where the treatment effect is heterogeneous. For the same reason, the CBHM does not fully achieve “adaptive information borrowing,” which implies that the methodology can use the observed data to accurately distinguish which baskets have similar response rates and pool information accordingly. Recently, Chu and Yuan¹⁴ proposed a Bayesian latent subgroup trial (BLAST) design that allows such adaptive information borrowing for basket trials. The BLAST design employs a latent subgroup/class model to group together the baskets that have similar response rates and then pool information accordingly within each homogeneous latent class. As a result, the BLAST design yields high power to detect the treatment effect for sensitive cancer types that are responsive to the treatment, while maintaining a reasonable type I error rate for insensitive cancer types that are not responsive to the treatment.

We described our method using cancer basket trials. The proposed methodology and design can be used for other diseases as well. Although we focused on a binary outcome, the CBHM can be easily extended to continuous, ordinal, and survival outcomes. The key is that given a hierarchical model for these outcomes, we do not directly estimate the shrinkage parameter from the data, but prespecify it as a function of a homogeneity measure. The same link function and similar calibration procedure described previously can be used to determine the form of the function.


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

Funding Sources: National Cancer Institute (grant/award number: “P30 CA016672,” “P50 CA098258,” “R01 CA154591”).

ORCID iD

Yiyi Chu  <https://orcid.org/0000-0001-9067-1366>

References

1. Simon R and Roychowdhury S. Implementing personalized cancer genomics in clinical trials. *Nat Rev Drug Discov* 2013; 12: 358–369.

2. Redig AJ and Jänne PA. Basket trials and the evolution of clinical trial design in an era of genomic medicine. *J Clin Oncol* 2015; 33: 975–977.
3. Ornes S. Core concept: basket trial approach capitalizes on the molecular mechanisms of tumors. *Proc Natl Acad Sci* 2016; 113: 7007–7008.
4. Billingham L, Malottki K and Steven N. Research methods to change clinical practice for patients with rare cancers. *Lancet Oncol* 2016; 17: e70–e80.
5. Berry DA. The brave new world of clinical cancer research: adaptive biomarker-driven trials integrating clinical practice with clinical research. *Mol Oncol* 2015; 9: 951–959.
6. Renfro LA and Sargent DJ. Statistical controversies in clinical research: basket trials, umbrella trials, and other master protocols: a review and examples. *Ann Oncol* 2017; 28: 34–43.
7. Thall PF, Wathen JK, Bekele BN, et al. Hierarchical Bayesian approaches to phase II trials in diseases with multiple subtypes. *Stat Med* 2003; 22: 763–780.
8. 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: 720–734.
9. Freidlin B and Korn EL. Borrowing information across subgroups in phase II trials: is it useful? *Clin Cancer Res* 2013; 19: 1326–1334.
10. Greco A, Miranda C and Pierotti M. Rearrangements of NTRK1 gene in papillary thyroid carcinoma. *Mol Cell Endocrinol* 2010; 321: 44–49.
11. Higgins J, Thompson SG and Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. *J R Stat Soc Ser A Stat Soc* 2009; 172: 137–159.
12. Thall PF, Simon RM and Estey EH. Bayesian sequential monitoring designs for single-arm clinical trials with multiple outcomes. *Stat Med* 1995; 14: 357–379.
13. Yuan Y, Nguyen HQ and Thall PF. *Bayesian designs for phase I-II clinical trials*. Boca Raton, FL: CRC Press, 2016.
14. Chu Y and Yuan Y. Blast: Bayesian latent subgroup design for basket trials. *J R Stat Soc* 2018; doi:10.1111/rssc.12255.

Appendix I

Proof of Theorem 1

When the treatment effect is homogeneous across subgroups, $T \rightarrow 0$ when $N_j \rightarrow \infty$ because the chi-squared test statistic of homogeneity is consistent. As $b > 0$, it follows that $b \times \log(T) \rightarrow -\infty$, and thus $\sigma^2 = \exp\{a + b \times \log(T)\} \rightarrow 0$. In other words, the CBHM achieves full information borrowing, as with the pooled analysis. Similarly, when the treatment effect is heterogeneous across subgroups, $T \rightarrow \infty$ when $N_j \rightarrow \infty$. Thus, $\sigma^2 = \exp\{a + b \times \log(T)\} \rightarrow \infty$, which means that no information will be borrowed across subgroups, as with the independent analysis.