Efficient Basket Trial Designs

Kee-Young Shin, Jun Lu

Abstract

There has been an increased impetus for basket trials in which treatments are tested across a group of patients. These designs are especially catered towards cases wherein mutational targets occur only in a small proportion of tumors, but are also present in many types of tumors. In this study, we sought to determine whether a new trial design characterized by an interim heterogeneity assessment of baskets would lead to greater efficiency in terms of expected sample size and marginal power compared to independent Simon Two-Stage designs and Bayesian hierarchical adaptive design. Through simulated trials, the Heterogeneity Assessment design was shown to perform better than independent Simon Two-Stage design in scenarios in which more baskets are active.

Introduction

Basket Trial

The current oncology landscape is changing dramatically due to the advancements in tumor sequencing [1], which enable us to differentiate cancers by their genetic mutations rather than their body locations. The success in targeting genetic mutations has not only led to a better understanding of the mechanism of cancer but has also fueled many innovative cancer treatments. Different from traditional cancer therapies, innovative therapies termed targeted therapies are designed to specifically target and treat cancer on the basis of their genetic mutations. However, the U.S. Food and Drug Administration (FDA) usually approves drugs for

use in specific disease sites while the prescription for other types of cancer is still considered "off-label". In addition, it is unrealistic to investigate cancer on different sites in one trial by conventional trial designs.

In order to address these problems, basket trials have become an increasingly popular approach to evaluate treatment effects of targeted therapies across different types of cancer, in which the centering principle is molecular alteration, irrespective of histology. In basket studies, the drug is tested simultaneously in different baskets (i.e. sub-groups of different tumor types). Basket trials are commonly used in early or mid-stage trials to discover potential indications of drug efficacy, in which good candidates will be investigated further in lager trials in terms of location of cancer. However, in some cases, a basket trial has been considered as adequate evidence for approval. For example, imatinib mesylate[2] was approved for 5 different types of cancer (KIT mutation) on the basis of a single phase II trial. Additionally, it has been suggested that borrowing information across baskets or pooling baskets collectively can make basket trials more efficient, but these approaches are still being questioned and under the research.

Existing methods for basket trials

One traditional way to conduct basket trials is to evaluate drug effects independently in each basket using the Simon two-stage design [3]. Kristen et al.[4] modified this approach by adding a heterogeneity assessment in stage 1. In their approach, a combined basket evaluation was performed for the homogeneous path.

Apart from this traditional way, the use of a Bayesian framework has also become common in proposed trial designs. Thall et al.[5] proposed a hierarchical Bayesian model for multiple subtypes, wherein the effects are exchangeable and correlated under a presupposed a

priori. Berry et al.[6] proved that using a hierarchical Bayesian model can increase power and provide a reduction in sample size using frequent interim analyses. Neuenschwander et al.[7] improved the hierarchical Bayesian model by proposing an exchangeability-nonexchangeability approach for more heterogeneous populations.

In this paper, we focus on illustrating the method proposed by Kristen et al. and compare its performance to an independent Simon Two-stage design and a Bayesian hierarchical adaptive design. Implement code can be found on https://github.com/Lujun1995/Efficient-Basket-Trial-Designs.

Methods

Design parameters and decision rule

We assume there are **K** baskets. In each basket, the true response rate is $\square_k$ for the $\mathbf{k^{th}}$ basket (for k = 1, 2, …, K). The null value and effective value for each basket are $\square_0$ and $\square_a$, respectively. We define A as the number of active baskets, or baskets in which the drug truly works.

The total sample size for all baskets is denoted as **N** ($\mathbf{N_1}$ for the first stage; $\mathbf{N_2}$ for the second stage) while the sample size for each basket is denoted as $\mathbf{n_k}$ ($\mathbf{n_{1k}}$ for the first stage; $\mathbf{n_{2k}}$ for the second stage). The number of responses for each basket is denoted as $\mathbf{r_k}$ ($\mathbf{r_{1k}}$ for the first stage; $\mathbf{r_{2k}}$ for the second stage). The workflow is demonstrated in Figure 1.

The first decision (node (a))

We conduct an exact test of a K×2 contingency table at decision node (a) to determine whether a homogeneous or a heterogeneous path will be taken. Here, we defined a design parameter □ used to select the most appropriate path.

$$H_0: \square_1 = \square_2 = \ldots = \square_K$$

H_a: At least one □k is different from others

Pr_{exact} = Pr(Data or more extreme situations for the first stage | H0: $\square_1 = \square_2 = \ldots = \square_K$) If **Pr1** > □, fail to reject H_0 and go to the homogenous path

If **Pr1** ≤ □, reject H_0 and go to the heterogeneous path

The heterogeneous path

In this path, we evaluate treatment effects in each basket independently following the Simon Two-stage design. At stage 1 (node (b)), we have the minimum number of responses (in a single basket) needed $\mathbf{r_S}$. For the $\mathbf{k^{th}}$ basket:

If $\mathbf{r_{1k} < r_S}$, stop the trial and conclude futility.

If $\mathbf{r_{1k} \geq r_S}$, continue the trial to stage 2.

Let us define **K*** as the subset of baskets continuing to stage 2. For baskets that display encouraging response rates, we enroll and treat $\mathbf{n_{2k}}$ patients in stage 2, for all $\mathbf{k \in K^*}$. For these baskets, we evaluate the response rate separately using one-sided Binomial exact tests with a correction for multiple comparisons as represented by decision node (d). Our design parameter at this node is the significance level for each individual test, defined as $\boldsymbol{\alpha_S/K^*}$.

The homogenous path

If the homogeneous track is taken, we first assess the futility of all baskets collectively at node (c). We create a new design parameter $\mathbf{r_C}$ (i.e., the minimum number of responses needed across all baskets combined in stage 1 to warrant enrolling additional patients to stage 2 for all baskets). For all baskets:

$$\text{If } \sum \mathbf{r_{1k}} \geq \mathbf{r_C}, \text{ continue to stage 2}$$

$$\text{if } \sum \mathbf{r_{1k}} < \mathbf{r_C}, \text{ stop the trial and conclude futility for all baskets}$$

If we determine that stage 1 results appear encouraging, we enroll and treat an additional $\mathbf{N_2}$ patients sampled from all baskets. At node (e), we evaluate the overall response rate using a one-sided Binomial exact test and all the available data. Our design parameter at this node is the significance level for the combined basket $\boldsymbol{\alpha_C}$.

Pooling baskets in the Heterogeneous Track

We extend the Kristen et al. Heterogeneity Assessment design to create a trial design wherein baskets are pooled even on the heterogeneous track (Design 2). For this model, we allow for baskets to be pooled for final analysis similar to the homogeneous track in situations in which baskets are found to be heterogeneous and three or more baskets are continued to the second stage. The pooled baskets are used to evaluate the overall response rate using a one-sided Binomial exact test at significance level $\boldsymbol{\alpha_C}$. We sought to see if this "middle ground" where only some of the baskets were pooled would lead to better performance metrics.

Performance Metrics

With regard to basket trials, the conventional type 1 and type 2 error rates are no longer appropriate. Thus, the designs are evaluated using the following metrics: Family Wise Error Rate (**FWER**), marginal power ($\mathbf{P_k}$), and expected trial sample size (**EN**). FWER is defined as the probability of incorrectly declaring activity in one or more baskets when in fact the drug does not work in any basket. The marginal power (**Pk**) for basket k is defined as the probability of correctly declaring activity in basket k when in fact the drug works in basket k. Expected sample size is calculated through the following expression:

$$EN = \sum_{k=1}^{K} n_{1k} + \sum_{k \in K^*}^{K} Pr(r_{sk} \geq 1 \mid \textit{heterogeneous design path})\, Pr(\textit{heterogeneous design path})$$

$$+ N_2 Pr(r_C \geq 5 \mid \textit{homogeneous design path}) Pr(\textit{homogeneous design path})$$

The expected sample size for Design 2 is calculated in the same manner since only the way in which the baskets were analyzed differed.

Calibrations

In this design, there are 8 unknown design parameters: $\mathbf{N_1}$, $\mathbf{n_{2k}}$, $\mathbf{N_2}$, □, $\mathbf{r_S}$, $\mathbf{r_C}$, $\boldsymbol{\alpha_S}$, $\boldsymbol{\alpha_C}$ that are designed to optimize the utility function based on marginal power and expected sample size.

Since 8 unknown design parameters are too complicated for computation, we fix four of the design parameters: $\mathbf{N_1}$, $\mathbf{N_2}$, $\mathbf{r_S}$, $\mathbf{r_C}$ based on preliminary simulations and common practices.

1. Choose a modest value for $\mathbf{N_1}$(the total number for the first stage) based on common practice.
2. Choose $\mathbf{N_2}$, the total number of stage 2 patients for the homogeneous design track, to be smaller than $\sum \mathbf{n_{2k}}$ ($k \in K^*$) in the heterogeneous path.

3. Set $\mathbf{r_S} = 1$ and $\mathbf{r_C} = \mathbf{K}$ (which equivalently requires around 1 responder per basket in order to continue all baskets to stage 2)

After fixing the above four parameters, we determine the remaining design parameters that optimize the power and sample size, and declare the corresponding design as optimal.

The optimization is accomplished through a dynamic grid search. The purpose of dynamic grid search is to identify candidate designs that satisfy the restrictions below after which we select an optimal design within those candidates.

Restrictions for Designs ($\mathbf{K} = 5$)

1. Achieve the same **FWER** as the reference design (Simon two-stage) when the drug is not active in any baskets ($\mathbf{A} = 0$).
2. Achieve the same power when the drug is active specifically in two baskets ($\mathbf{A} = 2$), while ensuring that the power achieves a minimum target level when the drug is active in only a single basket ($\mathbf{A} = 1$).

For each $\mathbf{n_{2k}}$, we selected all possible combinations (□, $\boldsymbol{\alpha_S}$, $\boldsymbol{\alpha_C}$) which satisfy the restriction. Then the optimal design is the one ($\mathbf{n_{2k}}$, □, $\boldsymbol{\alpha_S}$, $\boldsymbol{\alpha_C}$) which maximizes the utility function of U, where

$$U = \sum_{j=3}^{5} P_{marginalPower}(A = j) - \sum_{j=0}^{5} EN(A = j)$$

Reference Designs

There are two reference designs in our study: independent Simon two-stage design and the Bayesian hierarchical adaptive design. For the Bayesian hierarchical adaptive design, we model the log-odds of response, including an adjustment for the targeted rates θ_a.

$$y_k = log(\frac{\theta_k}{1-\theta_k}) - log(\frac{\theta_a}{1-\theta_a})$$

Then we model $\mathbf{y_k}$ with a normal distribution with unknown mean μ and variance σ^2.

$$\mathbf{y_k} \sim N(\mu, \sigma^2)$$

The μ and σ^2 follow prior distribution below

$$\mu \sim N(-1.34, 10^2), \sigma^2 \sim \text{Inverse-Gamma}(0.0005, 0.000005)$$

The Final evaluation criteria is

If $Pr(\theta_k > \theta_0 \mid Data) > \mathbf{P_c}$ ($\mathbf{P_c}$ a calibrated cutoff), the basket is active.

The interim evaluation criteria is

$$\theta_{mid} = (\theta_0 + \theta_a)/2$$

If $Pr(\theta_k > \theta_{mid} \mid Data) < 5\%$, stops early

Simulated Studies

In the simulated studies, we set K = 5, $\theta_0 = 15\%$, $\theta_a = 45\%$. The true response is either the null value θ_0 or the effective value θ_a. We tested the designs of interest under 6 different scenarios (**A** = 0, 1, 2, 3, 4, 5).

For the reference Simon two-stage design, type I error in each basket is controlled at 1% while the the type II error is controlled at 20%. The sample size for the stage 1 is 9 while the

minimum number of responses for passing is 3. We require a least 9 responders over all 27 patients in each basket to declare the drug works

For the reference Bayesian hierarchical adaptive design, we enroll patients in 4 cohorts (5, 5, 5, 7). The calibrated efficacy cutoff based on null scenario (**A** = 0) is 93%.

For the Heterogeneity Assessment Design, the assessment of heterogeneity tuning parameter is 0.52. The total sample size for the first stage $\mathbf{N_1}$ is 35. In the heterogeneity path, the minimum number of responders needed $\mathbf{r_S}$ is 1, the sample size required in the second stage for each passed basket $\mathbf{n_{2k}}$ is 7, and the significance level for each individual test $\boldsymbol{\alpha_S}$ is 0.7. In the homogenous path, the minimum number of responses needed $\mathbf{r_C}$ is 5, the total sample size for the second stage $\mathbf{N_2}$ is 20, and the significance level for combined basket test $\boldsymbol{\alpha_C}$ is 0.05. We compare the operating characteristics of the proposed design with the reference designs based on this setting through 1000 simulations.

Results

Heterogeneity Assessment Design

First, the Heterogeneity Assessment design is compared to the independent Simon Two-Stage design. For one active basket, the marginal power is 79% for the reference design (Table 1) while the proposed design is 70.1% (Table 2). For two active baskets, the marginal powers increased for the proposed design at 79.5% and 80.1%, but are still lower than the reference design at 81% and 82%. The expected sample sizes are also lower for the reference design compared to the proposed for these two scenarios with 69 and 83 vs. 77 and 85, respectively.

The benefits of the proposed design is seen when the number of active baskets increase or all baskets are inactive. In the null scenario (no active baskets) the type I error rates are

controlled at 1% for all baskets in the reference design, while the proposed design experience similar rates at around 2-3%, though the reference design has a lower expected sample size. However, the proposed design begins performing better with increased active baskets. When four baskets are active, the marginal power for the proposed design is 3-6% higher (except for basket 2 in which they are the same) while the expected sample size is 18% smaller compared to the reference design. The difference is even more pronounced when all baskets are active, with the proposed design experiencing 5-8% increase in marginal power across all baskets and a 33% decrease in expected sample size. In terms of the Family Wise Error Rate, the proposed design show increased error rates as active baskets increase, unlike the Simon Two-Stage design.

When □ is lowered to 0.20

The smaller □ is more likely to pursue the homogeneous design path and the expected sample will become same smaller. When lowering the □ to 0.20, we saw inflations in the FWER for the proposed design (Table 3) in addition to smaller sample sizes (i.e. 63.66 vs. 80.49 for A=5). Additionally, the marginal power increases when $A \geq 3$ and the marginal power decreases when $A < 3$.

Pooling heterogeneous baskets

Design 2 extends the proposed design to allow for pooling of baskets when they are found to be heterogeneous. Expected sample sizes stayed the same since the recruitment of stage 1 and stage 2 samples followed the same logic as in the original Heterogeneity Assessment design. Marginal power is increased compared to the original proposed design for active baskets greater than or equal to 2 (Table 4). For example, the marginal powers under scenario 6 (all

baskets active) for Design 2 are 97%, 96%, 96%, 97%, and 97% compared to 88%, 87%, 88%, 88%, and 87% for the original proposed design. However, the FWER and the type I error rates are inflated in all scenarios, with the FWER reaching as high as 66.4% (4 active baskets).

Due to computational limitations, we only run 100 simulations for the Bayesian hierarchical adaptive design. The result is presented in the Table. 5. Compared with the calibrated Heterogeneity Assessment design (Table. 2), in the setting we chose there is a more serious issue of inflated FWER in the Bayesian hierarchical adaptive design. However, the marginal powers are also larger in all scenarios. The calibrated Heterogeneity Assessment design achieves a better balance between the FWER and the marginal power. Additionally, the calibrated Heterogeneity Assessment design requires less computation.

Discussion

In this study, we sought to determine whether or not a basket trial design, wherein an interim heterogeneity assessment that could lead to the pooling of baskets, would lead to greater efficiency compared to performing several independent Simon Two-Stage designs. After running 1000 simulations, the results showed that the proposed design performed better in scenarios in which there were more active baskets. For example, in the case that all baskets were active, this Heterogeneity Assessment design would pool the baskets together and treat them like one sample leading to smaller expected sample sizes, while independent Simon Two-Stages would run separate trials for each of the baskets to completion.

However, when there were only one or two baskets active, the independent Simon Two-Stage designs led to higher marginal powers and lower expected sample sizes. In these situations wherein the majority of baskets display homogenous futile results, a pooled sample analysis may

provide less favorable results since ineffective baskets may be aggregated with effective baskets. The Heterogeneity Assessment design would not be recommended in these cases.

In Design 2, some of the baskets were allowed to be pooled for final analysis of the treatment's overall efficacy. This trial design led to higher marginal powers for baskets in scenarios that had two or greater active baskets, in return for higher type I error rates for the inactive baskets. Since now baskets are pooled in both heterogeneous and homogeneous tracks, the effect of aggregating inactive baskets with active baskets is more pronounced, leading to an increase in the number of false positives. The marginal gain in power does not seem to justify the large increases in type I error rate and FWER. Similarly, when □ the homogeneous path becomes more favorable leading to increased marginal power and decreased sample size while type I error rates are inflated. Again, it seems hard to justify the large increase false positives with the slightly improved power and sample size.

These trade-offs are also present in the Hierarchical Bayesian design. When there are more information borrowing across different baskets and the drug is active in most baskets, the FWER will be inflated and the sample size will decrease or the marginal power will increase.

All in all, the Heterogeneity Assessment design was shown to be a more suitable framework for basket trials in situations in which baskets are uniformly inactive or active in all or most of the baskets. By determining whether or not a drug/treatment's effect is homogenous across baskets, we are able to proceed to subsequent analysis to determine its effectiveness with greater confidence while utilizing a smaller sample size. In scenarios wherein the majority of baskets are homogenous futile, the independent Simon Two-Stage is recommended since this prevents the possibility of aggregating ineffective baskets with effective baskets in analysis. However, it is often the case that investigators are more focused on not missing active baskets,

especially in situations where the drugs being assessed are for the treatment of rare and damaging conditions. Therefore, this Heterogeneity Assessment design represents an attractive framework for basket trial studies.

There were several shortcomings present in our review of the proposed design. For one, we did not optimize across all design parameters due to computational limitations. Additionally, futility decision rules were chosen based on common practice rather than any researched logic. With this in mind, it is possible that there exists a greater optimized set of parameters that lead to better results. These issues could potentially be addressed in future studies. Additionally, it may be interesting to incorporate Bayesian principles into the Heterogeneity Assessment design to see how greater efficiencies are gained in terms of the performance metrics.

Tables and Figures

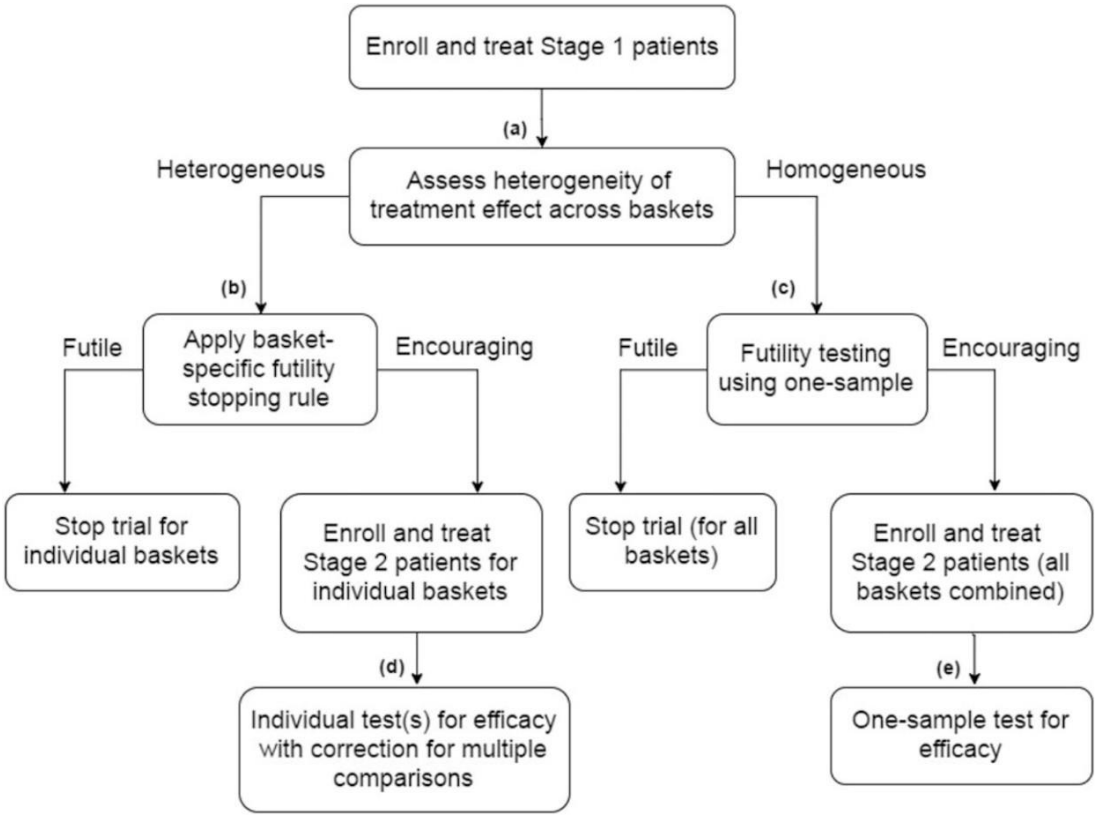

Figure 1. Flow chart of proposed design.

Scenario	FWER	Expected Sample Size	Probability of rejecting H0				
			P1	P2	P3	P4	P5
0 Active (0,0,0,0,0)	5%	58	1%	1%	1%	1%	1%
1 Active (1,0,0,0,0)		69	79%	1%	2%	1%	2%
2 Active (1,1,0,0,0)		83	81%	82%	1%	1%	1%
3 Active (1,1,1,0,0)		96	80%	82%	81%	1%	1%
4 Active (1,1,1,1,0)		108	82%	84%	80%	80%	1%
5 Active (1,1,1,1,1)		121	82%	81%	80%	80%	82%

Table 1. Power and Expected Sample Size for Optimal Simon Two-stage Design

Scenario	FWER	Expected Sample Size	Probability of rejecting H0				
			P1	P2	P3	P4	P5
0 Active (0,0,0,0,0)	5.7%	64.09	2.8%	2.8%	2.7%	3%	2.8%
1 Active (1,0,0,0,0)	7%	77.57	70.1%	5.4%	5.2%	4.9%	5.3%
2 Active (1,1,0,0,0)	11.9%	85.42	79.5%	80.1%	9.9%	9.7%	10.1%
3 Active (1,1,1,0,0)	15.7%	89.07	82.8%	80.5%	81.8%	15.1%	15.1%
4 Active (1,1,1,1,0)	26.9%	88.54	86.9%	83.6%	83.2%	85.8%	26.9%
5 Active (1,1,1,1,1)	NA	80.49	88.4%	87.3%	88.1%	87.9%	87.1%

Table 2. Power and Expected Sample Size for Proposed Design

Scenario	FWER	Expected Sample Size	Probability of rejecting H0				
			P1	P2	P3	P4	P5
0 Active (0,0,0,0,0)	7.6%	75.51	4.9%	4.3%	4.1%	4.3%	7.6%
1 Active (1,0,0,0,0)	15.8%	67.15	57.9%	14.6%	14.5%	14.7%	15.8%
2 Active (1,1,0,0,0)	29.7%	73.07	74.4%	75.5%	29.1%	28.6%	28.6%
3 Active (1,1,1,0,0)	42.4%	76.19	86.3%	85.9%	86.1%	42%	42.1%
4 Active (1,1,1,1,0)	54.9%	74.18	89.3%	89.5%	90.8%	89%	54.9%
5 Active (1,1,1,1,1)	NA	63.66	96.2%	95.3%	94.3%	95.4%	95.4%

Table 3. Power and Expected Sample Size for Proposed Design with Gamma=0.20

Scenario	FWER	Expected Sample Size	Probability of rejecting H0				
			P1	P2	P3	P4	P5
0 Active (0,0,0,0,0)	5.4%	64.09	3.4%	2.9%	3.2%	3.3%	3.6%
1 Active (1,0,0,0,0)	29.7%	77.57	62.3%	24.2%	24.7%	23%	23%
2 Active (1,1,0,0,0)	62.2%	85.42	85.6%	85.7%	50.6%	50.5%	48.3%
3 Active (1,1,1,0,0)	79.3%	89.07	91.3%	91.6%	91.8%	63.3%	60.2%
4 Active (1,1,1,1,0)	66.4%	88.54	96.8%	95.9%	95.4%	96.9%	66.4%
5 Active (1,1,1,1,1)	NA	80.49	96.4%	96.3%	96%	97.2%	96.5%

Table 4. Marginal Power and Expected Sample Size for Proposed Design with Pooling of Heterogeneous Baskets (Design 2)

Scenario	FWER	Expected Sample Size	Probability of rejecting H0				
			P1	P2	P3	P4	P5
0 Active (0,0,0,0,0)	6%	50.14	2%	3%	3%	3%	5%
1 Active (1,0,0,0,0)	20%	75.60	73%	8%	7%	11%	11%
2 Active (1,1,0,0,0)	37%	95.70	94%	92%	18%	13%	17%
3 Active (1,1,1,0,0)	39%	104.70	98%	98%	95%	24%	24%
4 Active (1,1,1,1,0)	47%	109.38	96%	100%	99%	99%	47%
5 Active (1,1,1,1,1)	NA	110.00	100%	100%	100%	100%	100%

Table 5. Marginal Power and Expected Sample Size for the Bayesian Hierarchical Adaptive Design

Code For Simulation

https://github.com/Lujun1995/Efficient-Basket-Trial-Designs.git

Reference

[1] Tao, Jessica J., Alison M. Schram, and David M. Hyman. "Basket studies: redefining clinical trials in the era of genome-driven oncology." *Annual review of medicine* 69 (2018): 319-331.

[2] Iqbal, Nida, and Naveed Iqbal. "Imatinib: a breakthrough of targeted therapy in cancer." *Chemotherapy research and practice* 2014 (2014).

[3] Simon, Richard. "Optimal two-stage designs for phase II clinical trials." *Controlled clinical trials* 10.1 (1989): 1-10.

[4] Cunanan, Kristen M., et al. "An efficient basket trial design." *Statistics in medicine* 36.10 (2017): 1568-1579.

[5] Thall, Peter F., et al. "Hierarchical Bayesian approaches to phase II trials in diseases with multiple subtypes." *Statistics in medicine* 22.5 (2003): 763-780.

[6] Berry, Scott M., et al. "Bayesian hierarchical modeling of patient subpopulations: efficient designs of phase II oncology clinical trials." *Clinical Trials* 10.5 (2013): 720-734.

[7] Neuenschwander, Beat, et al. "Robust exchangeability designs for early phase clinical trials with multiple strata." *Pharmaceutical statistics* 15.2 (2016): 123-134.