Contemporary Clinical Trials 85 (2019) 105807

Contents lists available at ScienceDirect

Contemporary Clinical Trials

journal homepage: www.elsevier.com/locate/conclintrial

Optimal two-stage designs for exploratory basket trials

Heng Zhou[a,*], Fang Liu[a], Cai Wu[a], Eric H. Rubin[b], Vincent L. Giranda[b], Cong Chen[a]

[a] Biostatistics and Research Decision Sciences, Merck & Co., Inc, Kenilworth, NJ 07033, USA
[b] Oncology Early development, Merck & Co., Inc, Kenilworth, NJ 07033, USA

ARTICLE INFO

Keywords:
Efficacy screening
Immunotherapy
Master protocol
Trial optimization

ABSTRACT

The primary goal of an exploratory oncology clinical trial is to identify an effective drug for further development. To account for tumor indication selection error, multiple tumor indications are often selected for simultaneous testing in a basket trial. In this article, we propose optimal and minimax two-stage basket trial designs for exploratory clinical trials. Inactive tumor indications are pruned in stage 1 and the active tumor indications are pooled at end of stage 2 to assess overall effectiveness of the test drug. The proposed designs explicitly control the type I and type II error rates with closed-form sample size formula. They can be viewed as a natural extension of Simon's optimal and minimax two-stage designs for single arm trials to multi-arm basket trials. A simulation study shows that the proposed design method has desirable operating characteristics as compared to other commonly used design methods for exploratory basket trials.

1. Introduction

In the exploratory phase of oncology drug development, it is often unknown how many tumor indications are likely active and what they are. Under resource constraint, exploring fewer tumor indications means greater power for each, but at the expense of an opportunity cost of not looking at additional tumor indications [1]. To account for the opportunity cost, multiple tumor indications are often selected for simultaneous testing in a basket trial so that a positive signal in any of the tumor indications may trigger subsequent development. A high level cost-effective strategy for determining the number of tumor indications in a basket can be found in Chen et al. [2]. Conventionally, the tumor indications in a basket trial are designed and analyzed separately by applying a Simon's two-stage design [3] or some other designs. Oftentimes, the overall type I error is not controlled, which may result in many inactive tumor indications being carried forward to costly late development. On the other hand, in order to properly control the overall type I error, each indication has to be tested at a stringent level after multiplicity adjustment, rendering large sample size. With the explosion of new compounds and new trials crowding the pipeline of oncology drug development, the conventional approach is unsustainable [4].

In order to gain efficiency under limited resources, the basket trials are routinely used in the exploratory phase of oncology drug development, and some of them have also been used for registration purpose [5]. In spite of its popularity, designing and analyzing an efficient basket trial remains challenging. Previous basket trials either analyze individual tumor response separately [6] or simply pool the data across tumor indications to declare an overall drug effect [7]. Independent evaluation is straightforward but lacks efficiency in detecting a treatment effect when the sample size is small in an exploratory trial. On the other hand, pooled evaluation that combines data from all tumor indications together is also straightforward but is at risk of treatment effect dilution when majority of the tumor indications are inactive. For example, vemurafenib is effective in BRAF- V600 mutant NSCLC and melanoma but not in colorectal cancer [8], thus pooling them together could have resulted in an overall negative trial especially when sample size is small. To tackle such dilemma, various statistical methods have been proposed to borrow information across tumor indications in basket trials [9–15]. These approaches are usually required to test for homogeneity across all indications to determine the degree of information borrowing.

The pruning and pooling approach, which was originally proposed in the confirmatory two-stage basket trial setting [16,17], is also intuitive under the exploratory setting especially when the practical question is “Does the drug work?” instead of “Which indication is most active?”. In this article, we will apply it to the design of exploratory basket trials by extending Simon's optimal design and minimax two-stage design from single arm trials to multi-arm trials. The overall power of the study is calculated under a non-informative prior assumption on the number of active tumor indications, a common situation in practice when there is no strong evidence to suggest

* Corresponding author at: MAILSTOP UG-1CD44, 351 North Sumneytown Pike, North Wales, PA 19454, USA.
E-mail address: heng.zhou@merck.com (H. Zhou).

https://doi.org/10.1016/j.cct.2019.06.021
Received 9 January 2019; Received in revised form 28 May 2019; Accepted 28 June 2019
Available online 29 June 2019

otherwise. Sample sizes and other design parameters are calculated from closed-form formula. The designs can be easily modified accordingly when an informative prior is incorporated into the calculation.

2. Optimal two-stage basket designs

Consider a phase II basket trial with tumor response rate as the primary endpoint. Suppose there are K tumor indications of interest, and we define the global null hypothesis as no treatment effect on any of the K tumor indications. Sharing the spirit of the classical Simon's two-stage designs for single arm trials, a design that minimizes the expected sample size under null (*optimal*) and a design that minimizes the maximum sample size (*minimax*), we propose two corresponding two-stage designs for basket trials. Inactive tumor indications are pruned in stage 1. Additional patients are enrolled in the active tumor indications, and data are pooled across the active ones at end of stage 2 to assess overall effectiveness of the test drug.

For each of the K tumor indications, we define homogenous null and alternative hypotheses on the true response rate p_k, $k = 1, \ldots, K$, such that $H_{0k}: p_k = p_0$ and $H_{ak}: p_k = p_a$, where $p_a > p_0$. The pruning decision will be made after n_1 subjects are enrolled in stage 1 for each indication. The indications with at least r_1 responders will continue to stage 2, where additional n_2 subjects will be enrolled to each continuing tumor indication. Let X_{k1} and X_{k2} denote the number of responses observed in stage 1 and stage 2 for the k^{th} indication respectively, such that $X_{ki} \sim Binomial(n_i, p_k)$, $i = 1, 2$. Note that here we allow a minimum pruning criterion $r_1 = 0$ in stage 1, and when it is met, the design will be simplified to a one-stage design, i.e., the analysis will be conducted based on the data of all K tumor indications without any pruning. The final analyses will be performed by pooling the indications completing stage 2. Suppose there are $M \geq 1$ indications pooled in stage 2, we will claim that the test drug is effective in at least one indication (i.e., rejecting the global null hypothesis) if there are at least R_M responders. Without losing generality, we assume that indications $k = 1, \ldots, M$, are the pooled indications, and indications $k = M + 1, \ldots, K$ are the pruned indications. Then the probability of rejecting the global null hypothesis is

$$F(r_1, n_1, n_2, R_M, p_k \mid K, M) = \prod_{(k=M+1)}^{K} \times B(r_1 - 1; n_1, p_k) \times \sum_{x_{11}=r_1}^{n_1} \cdots \sum_{x_{M1}=r_1}^{n_1} \left\{ \Pr(X_{k1} = x_{k1}, k = 1,\ldots,M) \times \Pr\left(\sum_{k=1}^{M} X_{k2} \geq R_M - \sum_{k=1}^{M} X_{k1} \right) \right\}, \quad (1)$$

where $B(\cdot; n, p)$ is the cumulative distribution function (CDF) of a binomial variable. Under the global null hypothesis that $p_k = p_0$, $k = 1, \ldots, K$, we control the significance level at $\alpha^\star$ for the pooled analysis to claim an effective drug for the pooled indications, i.e., $1 - B(R_M - 1; M(n_1 + n_2), p_0) \leq \alpha^\star$. Then given $M \geq 1$ indications pooled in stage 2, the overall probability that the test statistic of pooling analysis in stage 2 being significant under the global null hypothesis is

$$F_0(r_1, n_1, n_2, \alpha^\star, p_0 \mid K, M) = \{B(r_1 - 1; n_1, p_0)\}^{K-M} \times \sum_{x_{11}=r_1}^{n_1} \cdots \sum_{x_{M1}=r_1}^{n_1} \{\Pr(X_{k1} = x_{k1}, k = 1,\ldots,M) \times \Pr\left(\sum_{k=1}^{M} X_{k2} > Q_{1-\alpha^\star}(M(n_1 + n_2), p_0) - \sum_{k=1}^{M} X_{k1} \right)\}, \quad (2)$$

where $Q_{1-\alpha^\star}(n, p)$ is the $100(1 - \alpha^\star)\%$ quantile of $Binomial(n, p)$ and $\sum_{k=1}^{M} X_{k2} \sim Binomial(Mn_2, p_0)$. Hence, to control the global type I error rate at the α level, we can solve $\alpha^\star$ from the following equation

Table 1
Optimal basket trial parameters with type I and II error rates controlled at (0.05, 0.20). The optimal design minimizes expected sample size under global null, and the minimax design minimizes the maximum sample size.

p_0	p_a	Optimal				Minimax			
		r_1	n_1	$\alpha^\star$	n	r_1	n_1	$\alpha^\star$	n
		$K = 2$							
0.01	0.1	1	14	0.035	28	1	18	0.030	24
0.01	0.15	1	8	0.041	23	1	13	0.033	14
0.05	0.15	2	22	0.032	48	3	39	0.044	43
0.05	0.20	1	7	0.048	33	2	19	0.024	22
0.10	0.25	3	17	0.037	37	3	22	0.048	34
0.10	0.30	2	9	0.041	25	2	16	0.043	20
		$K = 4$							
0.01	0.1	1	9	0.026	25	0	NA	0.048	20
0.01	0.15	1	6	0.041	15	0	NA	0.051	9
0.05	0.15	2	19	0.024	37	2	31	0.014	32
0.05	0.20	1	7	0.022	22	3	17	0.013	18
0.10	0.25	2	10	0.024	34	3	24	0.015	27
0.10	0.30	2	7	0.046	21	3	13	0.022	17
		$K = 6$							
0.01	0.1	1	9	0.023	24	1	15	0.010	16
0.01	0.15	1	4	0.023	12	1	6	0.010	8
0.05	0.15	2	15	0.011	35	2	22	0.009	30
0.05	0.20	1	4	0.025	23	1	9	0.014	15
0.10	0.25	2	8	0.023	32	2	18	0.006	22
0.10	0.30	1	4	0.017	17	2	9	0.008	14
		$K = 8$							
0.01	0.1	1	7	0.022	23	1	11	0.008	14
0.01	0.15	0	NA	0.017	10	0	NA	0.004	7
0.05	0.15	2	13	0.007	34	2	20	0.005	26
0.05	0.20	1	5	0.013	17	1	8	0.006	13
0.10	0.25	2	8	0.013	27	3	16	0.005	20
0.10	0.30	1	3	0.016	17	2	9	0.005	12
		$K = 10$							
0.01	0.1	1	6	0.017	20	1	10	0.005	12
0.01	0.15	1	4	0.013	9	1	6	0.008	7
0.05	0.15	2	11	0.006	33	2	20	0.001	22
0.05	0.20	1	4	0.009	16	1	10	0.004	11
0.10	0.25	1	4	0.014	23	2	12	0.003	18
0.10	0.30	1	2	0.017	19	2	8	0.003	11

$$\sum_{M=1}^{K} \binom{K}{M} F_0(r_1, n_1, n_2, \alpha^\star, p_0 \mid K, \mathrm{M}) = \alpha \quad (3)$$

Suppose G out of K tumor indications are truly active with the response rate p_a, while the other $K - G$ indications have response rate p_0. Given that $M \geq 1$ indications are pooled in stage 2 and J out of M indications are truly active, the probability that the test statistic of the pooled analysis in stage 2 being significant is

$$F_1(r_1, n_1, n_2, \alpha^\star, p_0, p_a \mid K, M, G, J) = \{B(r_1 - 1; n_1, p_0)\}^{K-M-G+J} \{B(r_1 - 1; n_1, p_a)\}^{G-J} \times \sum_{x_{11}=r_1}^{n_1} \cdots \sum_{x_{M1}=r_1}^{n_1} \{\Pr(X_{k1} = \mathrm{x}_{k1}, k = 1,\ldots,M) \times \Pr\left(\sum_{k=1}^{M} X_{k2} > Q_{1-\alpha^\star}(M(n_1 + n_2), p_0) - \sum_{k=1}^{M} X_{k1} \right)\}, \quad (4)$$

where $\sum_{k=1}^{M} X_{i2}$ follows the Poisson Binomial distribution as the summation of J Binomial variables $Binomial(n_2, p_a)$ and $M - J$ Binomial variables $Binomial(n_2, p_0)$. Thus, the basket design is powered at

$$1 - \beta(G \mid K) = \sum_{M=1}^{K} \sum_{J=\max(0,M+G-K)}^{\min(M,G)} \binom{G}{J} (K - GM - J) F_1(r_1, n_1, n_2, \alpha^\star, p_0, p_a \mid K, M, G, J), \quad (5)$$

where $\beta(G|K)$ is the type II error rate when there are G truly active tumor indications. Notice that the power varies across the values of $G = 1, \ldots, K$. In our proposed design, we assume a non-informative uniform distribution on the number of truly active tumor indications with response rate p_a, thus the expected overall type II error is $\beta = \frac{1}{K}\sum_{G=1}^{K}\beta(G \mid K)$ and the expected overall power is $1 - \beta$. For simplicity, the global type I error rate and expected overall type II error rate are also loosely referred to be type I and II error rates respectively in the following sections.

Our goal in this article is to find the design parameters r_1, n_1, n_2, $\alpha^\star$ with the type I/II error rates at the target levels. Similar to Simon's *optimal* two-stage design, our proposed basket trial design is optimal in the sense of minimizing the expected sample size under the global null hypothesis (i.e., minimize patients' exposure to an ineffective treatment). The expected sample size is defined as:

$$EN(r_1, n_1, n_2, K, p_0) = n_1 * K + \sum_{M=1}^{K} n_2 * M * \binom{K}{M}\{B(r_1 - 1; n_1, p_0)\}^{K-M}\{1 - B(r_1 - 1; n_1, p_0)\}^{M}. \quad (6)$$

An alternative design can be derived by minimizing the maximum sample size $n = n_1 + n_2$ for each tumor indication, similar to Simon's minimax two-stage design. In practice, this design may be favorable when enrollment for the tumor indications is challenging and the budget is capped. When $K = 1$, our proposed designs degenerate to the respective Simon's two-stage designs.

Table 1 shows some examples of the optimal design parameters in different cases with type I and II error rates controlled at (0.05, 0.20), where we present the maximum planned sample size $n = n_1 + n_2$ instead of n_2 for each tumor indication. As an illustration of the optimal design under $K = 4$, $p_0 = 0.10$, $p_a = 0.25$, we enroll 10 subjects for each indication in stage 1. If there are at least 2 responders in an indication, this indication will be included in the pooled analysis in stage 2. Otherwise, the indication will be excluded. We enroll additional 24 subjects for each of the pooled indications in stage 2 and conduct the pooled analysis in the end at significance level 0.024. Depending on the number of tumor indications in the pooled analysis, the sample size of the pooled population may range from 34 (in case of one tumor indication) to 136 (in case of four tumor indications). Regardless of the sample size, the drug is considered effective as long as the p-value for testing the null hypothesis ($p_0 = 0.10$) is <0.024. The number of responses required for the basket trial to be considered positive depends on the actual sample size in the pooled population. For example, when 34 patients (one tumor indication) are included, 8 or more responses ($\geq$24% response rate) are required. When 68 patients (two tumor indications) are included, 13 or more responses ($\geq$19% response rate) are required.

In Table 1, we controlled the type I and type II error rates strictly at 0.05 and 0.20 respectively. Due to the discreteness of the binomial distribution, in rare cases, the overall type I error rate can be controlled under 0.05 even when $\alpha^\star$ is slightly above 0.05. Given the exploratory nature of the trials, the control of type I/II error rates may be relaxed to potentially reduce the sample size in practice. For example, when $K = 4$, under the null hypothesis of $p_0 = 0.01$ and alternative hypothesis of $p_a = 0.1$, the sample size under the minimax design is $n = 20$ per tumor indication. An alternative design has parameters $r_1 = 1$, $n_1 = 14$, $\alpha^\star = 0.014$ and $n = 18$, which has type I/II error rates of 0.0522/0.199. The new design may be preferred in practice if small inflation of type I error is not a concern. As also shown in Table 1, n_1 and n can be very close in some cases (e.g., the minimax design for $K = 2$ with $p_0 = 0.01$, $p_a = 0.15$). Since the operating characteristics is not expected to drastically change with a few patients, the two-stage design may be reduced to a one-stage design for simplicity by forcing $n_1 = n$ and the design parameters can be calculated accordingly. The two-stage designs yield $r_1 = 0$ for several cases with $p_0 = 0.01$, which effectively reduce to a one-stage design (i.e., n_1 is not applicable). Due to the computational complexity to calculate the exact type I/II error rates, we have used the empirical type I/II error rates from simulations for $K > 2$ as a starting point. The R code for computing both exact and empirical type I/II error rates can be found in Appendix A and it can be shown that the empirical error rates accurately approximate the exact ones.

Table 2
True response rate in the simulation setup and expected sample size using the optimal two-stage basket design.

Scenario	True response rate of each tumor indication						Expected sample size using optimal basket design
	1	2	3	4	5	6	
1	0.05	0.05	0.05	0.05	0.05	0.05	8
2	0.05	0.05	0.05	0.05	0.05	0.20	9
3	0.05	0.05	0.05	0.05	0.05	0.40	10
4	0.05	0.05	0.05	0.05	0.20	0.20	10
5	0.05	0.05	0.05	0.05	0.20	0.30	11
6	0.05	0.05	0.05	0.10	0.20	0.30	11
7	0.05	0.05	0.20	0.20	0.20	0.20	13
8	0.20	0.20	0.20	0.20	0.20	0.20	15

The planned sample size is greater under the optimal design than under the minimax design, just like Simon's two-stage design counterparts. In general, $\alpha^\star$ decreases with K as higher K means more cherry-picking, which requires greater penalty for multiplicity adjustment. As one might expect, under same p_0 and p_a, the planned sample size (n) for each tumor indication decreases as number of tumor indications increases while the overall sample size (Kn) also increases. It is tempting to wonder if this implies that we should include smaller number of tumor indications in a basket trial to reduce the overall sample size. Apparently, opportunity cost will need to be incorporated to properly address this question, and this topic is beyond the scope of this article.

3. Simulation comparison

The Simon's optimal two-stage design is more commonly applied in practice than the minimax design. We conducted a simulation study to evaluate the performance of the proposed optimal two-stage basket trial design with three additional design methods: a model-based Simon's Bayesian basket design [13] and two straightforward non-model-based methods. The purpose of the simulation study is to illustrate the relative performance of the proposed design in select scenarios of practical interest. It is not meant to draw any general conclusion, and extensive comparison between our proposed design and Simon's Bayesian basket design, or for this matter, any other basket design method in the literature is not the primary focus of this article.

Freidlin and Korn [18] have claimed that Bayesian hierarchical modeling approaches do not provide much information sharing for the sample sizes typically used in phase II exploratory trials. Unlike

Table 3
Operating characteristics of the optimal basket design and three other design methods.

Scenario	Independent evaluation	Pooled evaluation	Simon's Bayesian basket design	Optimal basket design
	Probability of claiming the drugs work (%)			
1	3.4	3.2	5.0	3.6
2	29.2	10.3	29.3	39.3
3	62.0	48.3	78.4	86.1
4	23.1	39.3	54.3	64.5
5	52.5	67.3	77.6	81.9
6	53.3	74.4	80.4	83.8
7	69.0	91.9	91.2	89.1
8	92.6	99.7	99.2	97.0
	Expected number of true positives			
1 (0 active)	0.00	0.00	0.00	0.00
2 (1 active)	0.26	0.10	0.27	0.38
3 (1 active)	0.62	0.48	0.78	0.86
4 (2 active)	0.24	0.79	0.67	0.97
5 (2 active)	0.59	1.35	1.05	1.26
6 (2 active)	0.59	1.49	1.08	1.27
7 (4 active)	1.01	3.67	2.11	2.26
8 (6 active)	2.11	5.98	4.50	3.51
	Expected number of false positives			
1 (6 inactive)	0.03	0.19	0.07	0.08
2 (5 inactive)	0.04	0.51	0.14	0.40
3 (5 inactive)	0.01	2.42	0.14	0.80
4 (4 inactive)	0.00	1.57	0.14	0.50
5 (4 inactive)	0.01	2.70	0.12	0.62
6 (4 inactive)	0.02	2.98	0.24	0.77
7 (2 inactive)	0.01	1.84	0.15	0.33
8 (0 inactive)	0.00	0.00	0.00	0.00

Bayesian hierarchical modeling approaches, Simon's Bayesian basket design assumed either homogeneity across all tumor indications (S_0) or independence of all indications (S_1). The posterior probability that tumor indication k is active was expressed as a weighted average under S_0 and S_1, as shown below:

$$Pr[p_k = p_a \mid data] = Pr[p_k = p_a \mid data, S_0]\, Pr[S_0 \mid data] + Pr[p_k = p_a \mid data, S_1]\, Pr[S_1 \mid data]$$

Tumor indication k is claimed as active when $\Pr(p_k = p_a | data) > \gamma$, where γ is the threshold that can be adjusted to reach the target type I error rate α or target power. The analysis method of Simon's Bayesian basket design is included in the simulation for comparison purpose. Two straightforward methods with type I error controlled at α are also included: 1) the independent evaluation method that tests at level $\alpha^* = 1 - (1-\alpha)^{\frac{1}{K}}$ for each individual tumor indication and the test drug is considered effective if it works in any of them, and 2) the pooled evaluation method that tests at level α after pooling all tumor indications without any pruning.

3.1. Simulations setup

We considered a basket trial with $K = 6$ indications and planned for targeting $H_{0k}: p_k = 0.05$ and $H_{ak}: p_k = 0.20$ ($k = 1, \ldots, K$). In order to control the type I error rate (α) at 0.05 and the type II error rate at 0.2, we found $r_1 = 1, n_1 = 4, \alpha^\star = 0.025$ and $n = 23$ from Table 1 under the optimal basket design. While a common alternative hypothesis $p_a = 0.20$ is used for trial planning purpose, the true response rates are unknown in practice and it is of interest in the investigation of the operating characteristics under various alternatives. Eight scenarios were considered in the simulation, as presented in Table 2. Scenario 1 was used to show the global type I error rate as the true response rates for the six indications are all 0.05. Other scenarios consist of one (Scenarios 2–3), two (Scenarios 4–6), four (Scenario 7) or six (Scenario 8) active indications. The response rates for the active indications under Scenarios 2–3 are 0.2 and 0.4, respectively, reflecting the possibility of an exceptionally active indication in practice. The impact of heterogeneity of responses rates is only explored when there are two active indications, which is arguably more realistic to expect than four or six in the exploratory phase of a new drug.

The three design methods for comparison are all one-stage trial designs. To make a fair comparison, the expected sample size calculated from the optimal two-stage basket design under each scenario was used as the sample size of the one-stage trial designs so that the expected sample sizes are the same under all designs. As seen in Table 2, the average sample size increases with the average response rate in a basket. The decision rules for the three methods were chosen so that the global type I error rate is controlled at 0.05, which requires fine-tuning of γ in Simon's Bayesian basket design. We considered three metrics for evaluating the operating characteristics of these four methods:

a. Probability of claiming the drug works, which is defined as the percentage of the simulated trials in which the drug was claimed as effective in at least one indication.
b. The expected number of true positives, which is defined as the average number of active indications correctly identified as active in the simulated trials.
c. The expected number of false positives, which is defined as the average number of inactive indications incorrectly identified as active in the simulated trials.

3.2. Simulations results

Results based on 1,000,000 simulations are presented in Table 3. Due to the discrete nature of the binomial distribution, the global type I error rate (probability of claiming the drug works under Scenario 1) cannot be controlled exactly at 0.05 for independent evaluation, pooled evaluation or the optimal basket design, while it can be controlled for Simon's Bayesian basket design at the exact level with proper choice of the threshold γ.

In general, the independent evaluation method yields the lowest probability in claiming the drug works. This is not surprising as the power for an individual tumor indication is low after the multiplicity adjustment. The existence of an exceptionally active indication in Scenario 3 helped a bit, but not enough to surpass our proposed optimal basket design or Simon's Bayesian basket design. As expected, when majority of the indications are active (Scenarios 7–8), the pooled evaluation method performed the best. However, both the optimal design and Simon's Bayesian design had decent power as well in the two scenarios. Overall, the two basket designs performed better than the two straightforward methods, and the optimal basket design was generally more robust than Simon's Bayesian basket design. The latter finding wasn't surprising because the proposed design is optimized to have the highest power when the number of active indications is uncertain, and the associated analysis method doesn't involve any homogeneity assessment. The optimal basket design tends to yield more true positive

indications than Simon's Bayesian basket design, but it also tends to yield more false positive indications. However, it is arguably more important to identify true positives than to reduce false positives in early stage drug development because the risk of carrying an inactive tumor indication forward can be mitigated later.

4. Discussion

Historically, oncology drugs were independently tested on one tumor indication at a time. To expedite drug development, master protocols for clinical trials are being actively considered. A basket trial is one example of master protocols that simultaneously tests one study drug on multiple tumor indications (another example is to simultaneously test multiple drugs on one tumor indication, or an umbrella trial). It signifies a paradigm change to the molecular view of cancer research. Although a traditional basket trial is stratified by tumor histological sites in patients with a common biomarker of molecular aberration, FDA's most recent draft guideline (www.fda.gov) on master protocols [19] has broadened its definition, including but not limited to stratification by disease stage, number of prior therapies, or demographic characteristics, in patients with or without a common biomarker. Regardless of how the patients are stratified, the primary objective of such basket-like trials in the exploratory phase is to identify an effective drug. In this article, for ease of illustration, we have followed the conventional definition of a basket stratified by tumor indications to present optimal and minimax two-stage designs. It is understood that our proposed designs equally apply to the broader definition of a basket.

In this article we have proposed optimal and minimax two-stage basket trial designs to test if a drug is effective in at least one of the tumor indications, which can be viewed as a natural extension of the two classical Simon's two-stage designs for single arm trials to multi-arm basket trials. The proposed designs can control the global type I error and the expected overall type II error rates with respect to a non-informative prior assumption about drug activity at the pre-specified levels. Design parameters are solved from closed-form formula. The simulation study has demonstrated the desired properties of the optimal basket design, and the properties of the minimax basket design are expected to be in line with expectation. When an informative prior assumption can be made (e.g., higher response rate of few active indications) or when the indications can be ordered by response rate, the design parameters can be calculated similarly after modifying the equation for the expected type II error rate. Note that, the proposed designs for exploratory basket trials use the same pruning and pooling method as in the two-stage designs for confirmatory basket trials [17]. In this regard, the confirmatory trial designs can be optimized similarly, and in this case normal approximation may be applied to facilitate the computation.

Though same response rates are assumed across the tumor indications under the null and alternative hypotheses, the proposed design methods can also be adapted to the trials with different null and alternative hypotheses across indications in practice. Under such circumstances, we can set the pruning bar in stage 1 using a common significance level for observed response rate instead of using a common number of responses as in this article. This is a topic of an ongoing research. Another practical issue is that the accrual speed in different indications may differ a lot in a basket trial. To address this, we can adjust the boundaries for both stages based on the optimal design parameters according to the accrual situation. Simulation is needed to evaluate the operating characteristics (OC) of adjusted boundaries and some efforts may be needed to tune the adjusted boundaries to achieve desirable OC. Also, additional interim analyses beyond two stages can be added in practice, though the derivation of optimal design parameters would be more complicated.

In this article, we illustrated the performance of the proposed optimal design with independent evaluation method, pooled evaluation method and Simon's Bayesian basket design in the simulation study. As Freidlin and Korn [18] pointed out, the Bayesian hierarchical model (BHM) does not work properly and may lead to inflated type I error rates for basket trials with 10 or fewer subgroups because there is usually not sufficient information to be borrowed in the observed data. Chu and Yuan [11] has shown that for the scenarios in which the treatment effects are heterogenous across the subgroups, the BHM method would yield biased estimates and its calibrated version would yield unbiased estimates similar to the independent approach. After all, our proposed optimal design with pruning and pooling technique is aimed to find a balance between the risk of undesirably including the inactive indications by simply pooling the data and the lack of efficiency in independently evaluating the treatment effect in each indication, just like most of the existing designs. It is not our intention to draw any general conclusion between different methods and more comprehensive comparison between our proposed design and others is beyond the scope of this article. Considering the recent advancement and the attractive flexibility of Bayesian designs in phase II exploratory clinical trials [20–23], we will work on extending our optimal design to Bayesian settings in the future.

As we have seen in the simulation study, inactive (or less active) tumor indications may be inadvertently included in the pooled analysis. Therefore, the fact that the test drug is demonstrated to work doesn't mean all the tumor indications in the pool are truly active. We can apply certain criteria to guide go/no-go decisions on cohort expansion for each pooled indication given the observed data. Alternatively, additional patients may be enrolled to those active tumor indications to confirm the positive findings. With or without confirmation, caution must be exerted before planning a Phase 3 confirmatory trial in any of the active tumor indications. Risk-mitigation design strategies in late development such as the 2-in-1 design [24] should be actively considered whenever possible.

Acknowledgments

We thank two anonymous referees for their helpful comments on this article.

Appendix A. R code to calculate power in Eq. (5)

```
### load packages ###
library(gtools)
library(poibin)

### Exact power calculated as shown in equation (5). When g=0, it is the type I error ###
PowerCalculation <- function(K,g,r1,n1,alphastar,n2,p0,pa)
{

 prob <- function(vec,n1,m,j,r,n2,p0,pa)
 {
  prod(dbinom(vec,n1,c(rep(pa,j),rep(p0,m-j))))*(1-ppoibin(kk=r-sum(vec),pp=c(p0,pa),wts=c(n2*(m-
j),n2*j)))
 }

 PrAccDrug <- function(K,g,r1,n1,m,j,r,n2,p0,pa)
 {
  options(expressions=1e5)
  dp <- permutations(n1-r1+1,m,r1:n1,repeats.allowed = TRUE)
  r = qbinom(1-alphastar,m*(n1+n2),p0)

  return(choose(g,j) * choose(K-g,m-j) * (pbinom(r1-1,n1,p0))^(K-m-g+j) * (pbinom(r1-1,n1,pa))^(g-j) *
         sum(apply(dp,1,prob,n1,m,j,r,n2,p0,pa)))
 }

 sum.inner <- function(K,g,r1,n1,m,r,n2,p0,pa)
 {
  temp = 0
  for (j in max(0,m+g-K):min(m,g)) temp = temp + PrAccDrug(K,g,r1,n1,m,j,r,n2,p0,pa)
  return(temp)
 }

 sum.outer <- 0
 for (k in 1:K) sum.outer = sum.outer + sum.inner(K,g,r1,n1,m=k,r,n2,p0,pa)

 return(sum.outer)

}


### Empirical power used in optimization. When g=0, it is the empirical type I error ###
PowerEmpirical <- function(K,g,r1,n1,alphastar,n2,p0,pa,nsim=10000)
{
 accept = rej = 0
 set.seed(10)
 for (sim in 1:nsim)

 {
  probs = c(rep(pa,g),rep(p0,K-g))
  resp1 = rbinom(K,n1,probs)
  if(all(resp1 < r1)) rej = rej + 1
  else {
   index = which(resp1 >= r1)
   resp2 = rbinom(length(index),n2,probs[index])
   respall = resp1[index] + resp2
   r <- qbinom(1-alphastar, length(index)*(n1+n2), p0)
   if(sum(respall) <= r) rej = rej + 1
   else accept = accept + 1
  }
 }
 return(accept/nsim)
}
```

References

[1] R.A. Beckman, J. Clark, C. Chen, Integrating predictive biomarkers and classifiers into oncology clinical development programmes, Nat. Rev. Drug Discov. 10 (2011) 735–748.

[2] C. Chen, Q. Deng, L. He, D.V. Mehrotra, E.H. Rubin, R.A. Beckman, How many tumor indications should be initially studied in clinical development of next generation immunotherapies? Contemp. Clin. Trials 59 (2017) 113–117.

[3] R. Simon, Optimal two-stage designs for phase II clinical trials, Contemp. Clin. Trials 10 (1) (1989) 1–10.

[4] J. Tang, A. Shalabi, V.M. Hubbard-Lucey, Comprehensive analysis of the clinical immuno-oncology landscape, Ann. Oncol. 29 (1) (2017) 84–91.

[5] T.Y. Seiwert, B. Burtness, R. Mehra, J. Weiss, et al., Safety and clinical activity of pembrolizumab for treatment of recurrent or metastatic squamous cell carcinoma of the head and neck (KEYNOTE-012): an open-label, multicentre, phase 1b trial, Lancet Oncol. 17 (7) (2016) 956–965.

[6] M.C. Heinrich, H. Joensuu, G.D. Demetri, C.L. Corless, J. Apperley, J.A. Fletcher, ... A. McKinley, Phase II, open-label study evaluating the activity of imatinib in treating life-threatening malignancies known to be associated with imatinib-sensitive tyrosine kinases, Clin. Cancer Res. 14 (9) (2008) 2717–2725.

[7] S. Lemery, P. Keegan, R. Pazdur, First FDA approval agnostic of cancer site-when a biomarker defines the indication, N. Engl. J. Med. 377 (15) (2017) 1409–1412.

[8] D.M. Hyman, I. Puzanov, V. Subbiah, J.E. Faris, I. Chau, J.Y. Blay, ... R. Gervais, Vemurafenib in multiple nonmelanoma cancers with BRAF V600 mutations, N. Engl. J. Med. 373 (8) (2015) 726–736.

[9] P.F. Thall, J.K. Wathen, B.N. Bekele, R.E. Champlin, L.H. Baker, R.S. Benjamin, Hierarchical Bayesian approaches to phase II trials in diseases with multiple subtypes, Stat. Med. 22 (5) (2003) 763–780.

[10] S.M. Berry, K.R. Broglio, S. Groshen, D.A. Berry, Bayesian hierarchical modeling of patient subpopulations: efficient designs of phase II oncology clinical trials, Clin. Trials 10 (5) (2013) 720–734.

[11] Y. Chu, Y. Yuan, A Bayesian basket trial design using a calibrated Bayesian hierarchical model, Clin. Trials 15 (2) (2018) 149–158.

[12] B. Neuenschwander, S. Wandel, S. Roychoudhury, S. Bailey, Robust exchangeability designs for early phase clinical trials with multiple strata, Pharm. Stat. 15 (2) (2016) 123–134.

[13] Simon, R., Geyer, S., Subramanian, J., & Roychowdhury, S. (2016, February). The Bayesian basket design for genomic variant-driven phase II trials. In Seminars in Oncology (vol. 43, No. 1, pp. 13–18). WB Saunders.

[14] K.M. Cunanan, A. Iasonos, R. Shen, C.B. Begg, M. Gönen, An efficient basket trial design, Stat. Med. 36 (10) (2017) 1568–1579.

[15] R. Simon, New designs for basket clinical trials in oncology, J. Biopharm. Stat. 28 (2) (2018) 245–255.

[16] R.A. Beckman, Z. Antonijevic, R. Kalamegham, C. Chen, Adaptive design for a confirmatory basket trial in multiple tumor indications based on a putative predictive biomarker, Clin. Pharmacol. Ther. 100 (6) (2016) 617–625.

[17] C. Chen, X. Li, S. Yuan, Z. Antonijevic, R. Kalamegham, R.A. Beckman, Statistical design and considerations of a phase 3 basket trial for simultaneous investigation of multiple tumor indications in one study, Stat. Biopharm. Res. 8 (3) (2016) 248–257.

[18] B. Freidlin, E.L. Korn, Borrowing information across subgroups in phase II trials: is it useful, Clin. Cancer Res. 19 (2013) 1326–1334.

[19] Master protocols, Efficient Clinical Trial Design Strategies to Expedite Development of Oncology Drugs and Biologits, FDA Guidance for Industry, 2019, www.fda.gov.

[20] P.F. Thall, R. Simon, Practical Bayesian guidelines for phase IIB clinical trials, Biometrics (1994) 337–349.

[21] P.F. Thall, R.M. Simon, E.H. Estey, Bayesian sequential monitoring designs for single-arm clinical trials with multiple outcomes, Stat. Med. 14 (4) (1995) 357–379.

[22] J.J. Lee, D.D. Liu, A predictive probability design for phase II cancer clinical trials, Clin. Trials 5 (2) (2008) 93–106.

[23] H. Zhou, J.J. Lee, Y. Yuan, BOP2: Bayesian optimal design for phase II clinical trials with simple and complex endpoints, Stat. Med. 36 (21) (2017) 3302–3314.

[24] C. Chen, K. Anderson, D.V. Mehrotra, E.H. Rubin, A. Tse, A 2-in-1 adaptive phase 2/3 design for expedited oncology drug development, Contemp. Clin. Trials 64 (2018) 238–242.