An Efficient Basket Trial Design
Supplementary Materials

Kristen M. Cunanan, Alexia Iasonos, Ronglai Shen, Colin B. Begg, Mithat Gönen

1 Optimization

The results presented in Section 3.2 of the manuscript optimize our design for 4 of the 8 design parameters: $\gamma, \alpha_S, \alpha_C$, and n_{2k} (stage 2 sample size per basket for the heterogeneous design track). Selection of the 4 fixed design parameters: N_1, N_2, r_S, r_C is discussed in Section 2.4 of the main document. Optimization is accomplished through a dynamic grid search. The purpose of the grid search is to first identify candidate designs that are calibrated with the reference design with respect to family wise error rate (α) and power ($1-\beta$) and then to make an "optimal" selection from these candidates on the basis of a utility function. Since the marginal power depends on the number of active baskets (denoted A), we perform the calibration for a suitably selected alternative scenario. In our primary comparison, where there are $K = 5$ baskets, we chose to calibrate the optimal design to achieve $1-\beta$ power when $A = 2$ baskets are truly active. Since we calibrate to achieve $1-\beta$ power for the setting of $A = 2$ active baskets, when there is only one active basket ($A = 1$) the marginal power is less than $1-\beta$. To restrict the loss of power in this configuration we use the concept of *minimum acceptable (marginal) power*: $(1-\beta)_{min}$ and restrict candidate designs to those for which the marginal power is $\geq (1-\beta)_{min}$, for the case when the drug only works in a single basket ($A = 1$). We then select the best combination of (n_{2k}, $\gamma, \alpha_S, \alpha_C$) with a desirable trade-off between minimizing expected trial size over all scenarios while maximizing power when the drug truly works in all (or most) baskets.

The optimization is accomplished as follows. For each n_{2k} and all possible combinations of $(\gamma, \alpha_S, \alpha_C)$, we evaluate the family wise error rate using simulations when $A = 0$. Then, limiting attention to combinations of $(\gamma, \alpha_S, \alpha_C)$ with a FWER within an ϵ_α margin of α, we evaluate the marginal power using simulations when $A = 1$. Similarly, using the truncated combinations of $(\gamma, \alpha_S, \alpha_C)$ with marginal power within an ϵ_β margin of $(1-\beta)_{min}$, we evaluate the marginal power using simulations when $A = 2$. This process is designed to calibrate the design to the reference design with respect to FWER and power when $A = 2$, and additionally restrict options to those designs with power of $(1-\beta)_{min}$ when $A = 1$. To accomplish this for each n_{2k}, we determine the $(\gamma, \alpha_S, \alpha_C)$ combinations that satisfy the selection function S, where

$$S = |FWER(A = 0) - \alpha| + |P_1(A = 1) - (1-\beta)_{min}| + |P_1(A = 2) - (1-\beta)|$$

We then declare the optimal design to be the one with the $(n_{2k},\, \gamma, \alpha_S, \alpha_C)$ combination that maximizes our utility function U, where

$$U = \sum_{j=3}^{5} P_1(A = j) * 100\% - \sum_{j=0}^{5} EN(A = j).$$

Since power increases as the number of active baskets increases, we want to define a utility function that selects a design maximizing the average power over the scenarios when the drug works in all or most baskets. However, since the expected trial size also increases as the number of active baskets increases, we want our utility function to select a design that considers a trade-off between maximizing power and minimizing the expected sample size across all scenarios (average). Therefore, we consider the difference of the two averages as our utility function and declare the combination that maximizes this difference as optimal, since it maximizes power using the smallest average trial size.

The search space of the heterogeneity parameter γ is defined on [0.1,0.9], explored in increments of 0.01. The significance level $\alpha_S/K^\star$ for the separate analyses given the heterogeneous design path is defined on [0.01,0.1] and explored in increments of 0.01. The significance level for the combined analyses (α_C) given the homogeneous design path is defined on [0.005,0.05] and explored in increments on 0.005. We explored a grid of all possible combinations $(\gamma, \alpha_S, \alpha_C)$, using parallel processing to dramatically reduce computational time. We ran a simulation study assuming the null scenario ($A = 0$) for each $n_{2k} \in [15, 22]$ and identify the $(n_{2k}, \gamma, \alpha_S, \alpha_C)$ combinations satisfying |FWER - α| $\leq \epsilon_\alpha$. Using the truncated grid space of $(n_{2k}, \gamma, \alpha_S, \alpha_C)$ combinations, we ran a simulation study assuming $A = 1$ of the K baskets are truly active and identify the new $(n_{2k}, \gamma, \alpha_S, \alpha_C)$ combinations satisfying $|P_1 - (1 - \beta)_{min}| \leq \epsilon_\beta$. In our simulations with $K = 5$, we found using a difference of 10% between the desired power and minimum acceptable power provided the best trade-off in maximizing power at reduced sample sizes when $A > 2$. Therefore, we restrict all candidate designs to those with around $(1 - \beta)_{min}$ power when $A = 1$. Using the new truncated grid space of $(n_{2k}, \gamma, \alpha_S, \alpha_C)$ combinations, we then ran a simulation study assuming $A = 2$ of the K baskets are truly active and identify the $(n_{2k}, \gamma, \alpha_S, \alpha_C)$ combinations satisfying $|P_1 - (1 - \beta)| \leq \epsilon_\beta$. We then select the admissible designs using S and finally the optimal design using U, as previously described.

2 Sensitivity Analysis

To evaluate the impact of the 4 fixed parameters on our optimal design specifications and operating characteristics, we change each fixed parameter one at a time. These were originally set to be $n_1 = 7$, $N_2 = 20$, $r_S = 1$, and $r_C = 5$. For our sensitivity analysis, we consider the following changes, one at a time: (1) a reduced stage 1 sample size, from $n_1 = 7$/basket to $n_1 = 5$/basket, (2) an increased stage 2 sample size for the homogeneous design track, from $N_2 = 20$ to $N_2 = 30$ patients overall, (3) an increased stage 1 sample size, from $n_1 = 7$/basket to $n_1 = 9$/basket, and (4) increased stage 1 sample size with altered stopping rules, $n_1 = 9$ with $r_S = 2$ (increased from $r_S = 1$), $r_C = 10$ (increased from $r_S = 5$). Recall, this is for the setting of $K = 5$ baskets with null response rate $\theta_0 = 0.15$ and

alternative response rate $\theta_a = 0.45$.

For (1) when we reduce the stage 1 sample size from 7/basket to 5/basket, the optimal design parameters change to $n_{2k} = 20, \gamma = 0.57, \alpha_S = 0.07, \alpha_C = 0.05$ (originally, $n_{2k} = 16, \gamma = 0.46, \alpha_S = 0.07, \alpha_C = 0.05$). The corresponding operating characteristics are provided in the second stratum of SM Table 1. Here, we find very similar results with a slight decrease in the expected sample size and, under the alternative scenarios ($A \geq 1$), a slight increase in the expected trial duration and marginal rejection rates compared to the original results with $n_1 = 7$/basket (first stratum of SM Table 1). With the smaller sample sizes, the baskets are more likely to appear homogeneous at the interim analysis. Thus, the optimized value of γ is larger for the assessment of heterogeneity. The optimal significance levels α_S and α_C are the same as before.

When we reduce the second stage (homogeneous track) sample size to $N_2 = 30$, the optimal design parameters change to $n_{2k} = 14, \gamma = 0.45, \alpha_S = 0.07, \alpha_C = 0.05$ (originally, $n_{2k} = 16, \gamma = 0.46, \alpha_S = 0.07, \alpha_C = 0.05$). From the corresponding operating characteristics (SM Table 1 third stratum) we see very similar results as in the first stratum of SM Table 1 with a small increase in the expected trial duration. The optimal significance levels α_S and α_C are the same as before.

When we increase the stage 2 sample size from 7/basket to 9/basket, the optimal design parameters change to $n_{2k} = 12, \gamma = 0.46, \alpha_S = 0.07, \alpha_C = 0.05$ (originally, $n_{2k} = 16, \gamma = 0.46, \alpha_S = 0.07, \alpha_C = 0.05$). The corresponding operating characteristics (SM Table 1 fourth stratum) are very similar to those in the first stratum of SM Table 1. However, we do see a modest increase in the expected sample size and trial duration. With the larger stage 1, the optimal design sets a smaller value of n_{2k}. The optimal significance levels α_S and α_C are the same as before.

Finally, when we increase the stage 2 sample size to 9/basket and alter the interim stopping rules, the optimal design parameters change to $n_{2k} = 13, \gamma = 0.41, \alpha_S = 0.07, \alpha_C = 0.025$ (originally, $n_{2k} = 16, \gamma = 0.46, \alpha_S = 0.07, \alpha_C = 0.05$). Here, we increase the number of responders required in our futility rules for both tracks. The expected sample sizes are consistently smaller than those reported for the previous configuration and for the first stratum of SM Table 1. The optimal design decreases γ and α_C from their original values, to achieve the desired operating characteristics.

From these sensitivity analyses, we see that our results are robust to changes in our fixed parameters. The results are robust because for each altered fixed parameter we are able to find the new optimal design values for $n_{2k}, \gamma, \alpha_S, \alpha_C$ satisfying our FWER and power constraints but with modest changes to the other operating characteristics. These results suggest that there exist a variety of design options each of which provides similar gains in

efficiency relative to the reference design.

SM Table 1: **Power and Expected Sample Size: Sensitivity Analysis**

Altered Parameter(s)	Fixed Parameters	Optimal Parameters	Scenario (A)	FWER	Marginal Power*					EN	ET
					P_1	P_2	P_3	P_4	P_5		
Original Design	$n_{1k} = 7$	$n_{2k} = 15$	0 Active	5	2	2	2	2	2	58	7.0
	$(N_1 = 35)$	$\gamma = 0.52$	1 Active		70	7	7	7	7	74	9.5
	$N_2 = 20$	$\alpha_S = 0.07$	2 Active		80	80	11	11	11	83	10.4
	$r_S = 1$	$\alpha_C = 0.05$	3 Active		84	85	85	17	17	86	10.5
	$r_C = 5$		4 Active		86	85	86	86	23	88	10.2
			5 Active		88	90	88	88	88	78	8.3
$n_{1k} = 5$	$N_2 = 20$	$n_{2k} = 20$	0 Active	5	1	2	2	2	2	49	6.7
$(N_1 = 25)$	$r_S = 1$	$\gamma = 0.57$	1 Active		71	7	6	6	6	67	9.8
	$r_C = 5$	$\alpha_S = 0.07$	2 Active		79	80	14	13	14	78	10.8
		$\alpha_C = 0.05$	3 Active		86	83	84	18	18	86	11.1
			4 Active		88	88	87	86	30	84	10.3
			5 Active		89	88	89	88	87	79	8.9
$n_{1k} = 9$	$N_2 = 20$	$n_{2k} = 12$	0 Active	5	2	2	2	2	2	67	8.5
$(N_1 = 45)$	$r_S = 1$	$\gamma = 0.46$	1 Active		69	7	7	6	6	80	10.6
	$r_C = 5$	$\alpha_S = 0.07$	2 Active		80	81	11	11	11	88	11.4
		$\alpha_C = 0.05$	3 Active		84	84	84	16	16	90	11.4
			4 Active		85	85	87	87	24	89	11.0
			5 Active		89	90	90	89	90	78	9.0
$n_{1k} = 9$	$N_2 = 20$	$n_{2k} = 13$	0 Active	5	2	1	2	2	1	56	7.6
$(N_1 = 45)$		$\gamma = 0.41$	1 Active		71	5	4	5	4	68	10.3
$r_S = 2$		$\alpha_S = 0.07$	2 Active		78	81	11	11	10	77	11.3
$r_C = 10$		$\alpha_C = 0.025$	3 Active		85	83	82	15	14	83	11.4
			4 Active		89	88	89	87	28	85	11.0
			5 Active		90	91	91	91	91	76	8.8
$N_2 = 30$	$n_{1k} = 7$	$n_{2k} = 14$	0 Active	5	2	2	2	2	2	55	7.5
	$(N_1 = 35)$	$\gamma = 0.45$	1 Active		70	9	9	8	9	70	9.5
	$r_S = 1$	$\alpha_S = 0.07$	2 Active		80	79	16	16	16	78	10.2
	$r_C = 5$	$\alpha_C = 0.05$	3 Active		86	85	83	21	21	82	10.5
			4 Active		87	85	84	85	32	81	10.1
			5 Active		89	90	90	92	89	73	8.9

Panel 1 assumes the original design presented in the manuscript. In panels 2-5, the first column displays the altered parameter(s) in order to investigate the sensitivity of the original specification defined in panel 1. The second column displays the other fixed parameters. The third column displays the 4 optimal parameters, as described in SM Section 1. The remaining columns display similar results as those presented in the manuscript, where *marginal power displays the marginal error rates for inactive baskets.

3 Implementation: Trial Example

In this section we describe the process of designing a basket trial using our method. We deliberately chose different design parameters for this example than the ones we used in the manuscript. Suppose a clinical investigator wants to design a basket trial with $K = 7$ baskets under evaluation, assuming a $\theta_0 = 10\%$ response rate is ineffective and a $\theta_a = 30\%$ response rate is effective. We would like to control the family wise error rate at 10%, while achieving 80%

marginal power to detect active baskets when the drug truly works in 2 of the 7 baskets and at least 70% marginal power to detect the active basket when the drug truly works in only 1 basket. This is the complete specification of the problem that is required from the user.

We begin the design process by first assuming $r_S = 1$ for the minimum number of responses needed in a single basket to continue to stage 2 for the heterogeneous design path. This is a natural choice since most investigators view any response in stage 1 as positive and want to evaluate further in stage 2. Similarly, we assume $r_C = K = 7$ for the minimum number of responses needed across all baskets to continue to stage 2 for the homogeneous design path. Lastly, we assume $N_2 = 4 * K = 28$ for the homogeneous design path, since 4 additional patients per basket is the smallest number we think is acceptable for evaluating basket-specific efficacy in a secondary analyses. Note that we can vary these values to investigate the sensitivity of the operating characteristics to these choices.

We assume there are no specifications or restrictions on stage 1 sample size. To get an idea of a reasonable starting point for n_{1k}, we look at the reference (Simon optimal) design with $\alpha = 0.1/7 = 0.014$ and $\beta = 0.2$, which requires $n_1 = 14$ patients/basket in the first stage and an additional $n_2 = 32$ patients/basket if 3 or more responders are observed in stage 1. These values are obtained using the *ph2simon* function in the R library *clinfun* [1]. In our simulation studies presented in the manuscript we found the optimal n_{1k} to be slightly smaller than that of the reference design; based on this, we set $n_{1k} = 12$.

After downloading the R code in the manuscript, we can run the script with the above specifications: $K = 7$, $\theta_0 = 0.1$, $\theta_a = 0.3$, $\epsilon = 0.1$, $1 - \beta = 0.8$, $(1 - \beta)_{min} = 0.7$, $n_{1k} = 12$, $N_2 = 28$, $r_S = 1$, and $r_C = 7$. Below is a condensed table of the software output results.

SM Table 2: **Power and Expected Sample Size:** $n_{1k} = 12$

N_2	Scenario (A)	FWER	Marginal Power*							EN	ET
			P_1	P_2	P_3	P_4	P_5	P_6	P_7		
4*7 = 28	0 Active	9	2	2	2	2	2	2	2	156	13.6
	1 Active		73	3	3	3	3	3	3	180	15.8
	2 Active		80	79	4	4	4	4	4	194	16.6
	3 Active		80	80	78	4	5	5	5	202	16.9
	4 Active		82	79	80	80	7	7	7	209	17.0
	5 Active		81	80	78	79	83	9	10	210	16.8
	6 Active		81	82	79	80	81	80	15	209	16.2
	7 Active		84	84	82	84	84	84	86	189	13.9

* Marginal error rates for inactive baskets

The optimal design subject to the provided input values and calibrated to control the FWER at 10% when the drug has no effect in any of the baskets while achieving 80% power when the drug truly works in 2 of 7 baskets

has $n_{1k} = 12, n_{2k} = 22, \gamma = 0.7, \alpha_S = 0.1, \alpha_C = 0.03$. This design produces power in the region of 84% when all baskets are active and declines to 73% when only one basket is active. The expected sample size ranges from 156 to 210, depending on the number of active baskets. This compares with a range of 130 to 264 for the expected sample size if the reference design (parallel Simon) was used. We would recommend exploring optimal design options with different "fixed" values of N_1, N_2, r_C, and r_S before selecting the most suitable option.

References

[1] Seshan VE, Seshan MVE. R package "clinfun". *CRAN* 2015; :1–23.