OPTIM-ARTS An Adaptive Phase II Open Platform Trial Design

Anthony D'Amelio Jr., Associate Director, Biostatistics
Co-authors: Amy Racine-Poon, Alex Sverdlov, Tomas Haas
NJ and PT ASA 2023 Spring Symposium
June 23, 2023

Introduction to Master Protocols

- Increasing interest in performing innovative trials allowing for simultaneous evaluation of multiple treatments in one disease or one treatment in multiple diseases within the same overall trial structure.
- Such designs are referred to as **master protocols**

Table 1. Types of Master Protocols.

Type of Trial	Objective
Umbrella	To study multiple targeted therapies in the context of a single disease
Basket	To study a single targeted therapy in the context of multiple diseases or disease subtypes
Platform	To study multiple targeted therapies in the context of a single disease in a perpetual manner, with therapies allowed to enter or leave the platform on the basis of a decision algorithm

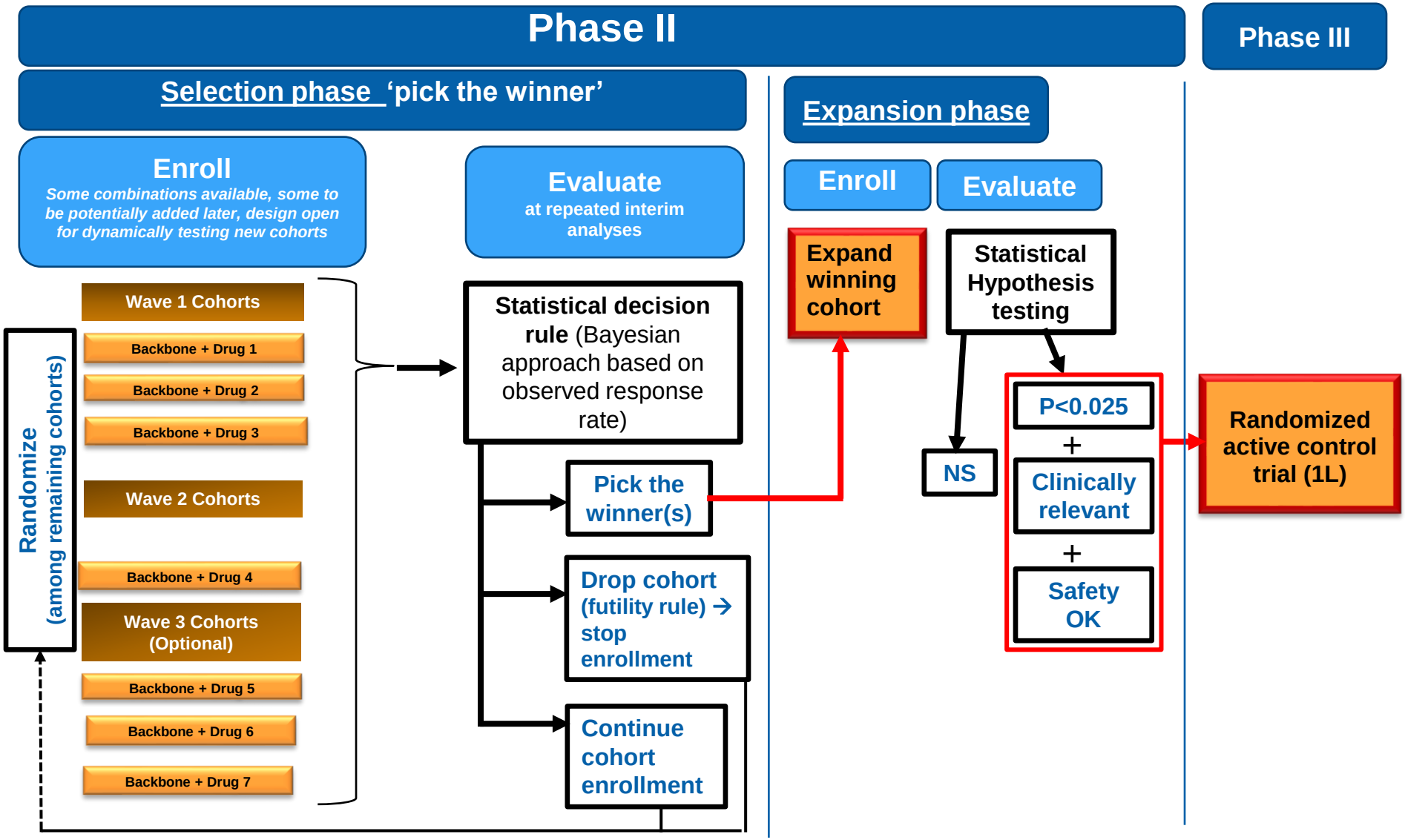
From Woodcock and LaVange (NEJM 2017)

FRACTION-GC
FRACTION-lung

Design Elements

Adaptive Platform Trial (Phase II)

Evaluation of multiple treatment combinations in a phase II setting



Open platform design - Pros and Cons

PROS	CONS
Multiple combinations with experimental treatment from different classes evaluated within 1 single trial	Time between enrollment & efficacy outcome can reduce benefits of a platform design.
Single master protocol and shared infrastructure across experimental treatments • Resource need for one single trial is lower than multiple trials. • Platform trial cost set up is substantially lower than multiple trials • No competing studies; only 1 trial. • Allows one IRB/IEC approval for one disease population • Less Time to add an additional cohort/wave vs starting a new study • Streamlined recruitment process • Increases data quality and trial efficiencies • Patients benefit through more opportunities to participate in investigational research • Building a network of specialized centers for future development	Potential for population shift as cohorts are enrolling in a staggered manner, potentially creating bias due to improvement in general care
Design adaptive/flexible: • Drop arms for futility, add new arms • Declare one or more winners	Complex trial logistics & operations (eligibility, drug supply, etc.)
In general, a platform design requires smaller sample size as compared to a standard design of separate Phase II studies	However, does require more start-up time upfront compared to a standard phase II trial

Statistical Methodology
High level overview

- **Randomized design** but **not a traditional comparative design;** instead each combination arm compared against predefined thresholds
- **Response rate** (ORR) used as a primary endpoint
  - Other endpoints such as DoR, PFS, OS, Safety, biomarker, etc.. can be considered as well, but are not part of the statistical decision making
- **Adaptive design features:**
  - allows for dropping arms for futility or selection of one or more winners at each interim analysis
  - allows for adding new arms
  - ‘dynamic’ sample size in selection phase and adaptive sample size in expansion
- **Bayesian methodology** used in the selection phase: probabilistic assessments of ORR in adaptive decisions making
- **Classical hypothesis testing** of ORR used at the end of expansion phase
- Extensive **simulations** performed
  - to fine-tune decisions rules and to explore operating characteristics
  - to determine sample size for both selection and expansion phases
  - to assess alpha (type I error)

Decision criteria in selection phase and interim analyses

- ***Clinical thresholds*** for ORR:
  - ORR<10% futile;
  - ORR≥25% worth further development
- Bayesian ***probabilistic decision rules***:
  - Prob(ORR<15%)>70% ➔ drop for futility
  - Prob(ORR≥20%)>70% ➔ declare winner

❖ Statistical decision rules selected based on simulations so that clinical thresholds 'achieved' with high probability

❖ Decision rules to be applied at each ***interim analysis (IA):***

- The 1st IA after ~10 subjects have been enrolled in each of the first 3 arms and have completed the 2nd post-baseline tumor assessment (or have discontinued prior to completing the 2nd post-baseline tumor assessment).

- Subsequent interim analyses in part 1 will be conducted thereafter until # of sufficient subjects is enrolled to assess efficacy/futility or until cap is reached

- Although the formal interim decisions rules and statistical thresholds are based on the observed confirmed ORR, all available efficacy, safety and biomarker data will be considered at each decision point.

Statistical Bayesian model for selection phase

- **Patients are assessed in batches of size n per arm**
  - **Size n determined by enrollment rate,**
  - **number of patients enrolled per month**
  - **number of months per efficacy assessment**
  - **Inclusion of any new arms during efficacy assessment**

- **Uninformative prior for true ORR rates p_1, p_2, p_3, ..., p_k in treatment arms 1, 2, ...k**

- **At completion of batch 1 (total N=n*T)**
  1. **p_i ~ beta($y1_i$+1, $n1_i$-$y1_i$ +1)**
  2. **Decisions**
     - Winner Prob($p_i \geq r_W$|data) $\geq P_W$
     - Futile Prob($p_i \leq r_F$|data) $\geq P_F$
     - Continue otherwise

n = batch size, T number of arms available.

r_W and r_F are ORR thresholds for success (winner) and futility, resp.

P_W and P_F are probability thresholds for success (winner) and futility, resp.

Red numbers are design parameters:

Sample Size – Expansion Phase (Part 2)

- **Sample size needed to reject** the null hypothesis value of ORR≤10% if the targeted ORR ≥ 25%
  - the standard approach to sample size and power calculations that would not condition on selection part result
  - However, a more meaningful approach would be to utilize the information obtained in selection part 1 and condition on the part 1 results to determine the sample size for part 2 with an adaptive approach.
  - ***'Adaptive' sample size*** in expansion phase based on **predictive power** conditioned on the selection phase results and based on **shrinkage estimator**

$$Predictive\ power = 1 - \sum_{j=0}^{m} \binom{n}{j} p^j (1-p)^{n-j}$$

where m is the maximum number of responses that can be observed in the expansion phase and still be unable to reject the null hypothesis,
n is sample size of part 2, and
p is the Markov Chain Monte Carlo (MCMC) sample of the shrinkage estimator adjusted objective response rate from part 1

- Based on the simulation results a sample size of 50 subjects for the selected efficacious arm for which the true ORR is 25% is sufficient to achieve the mean predictive power of at least 70%.

Predictive power modeling for expansion phase

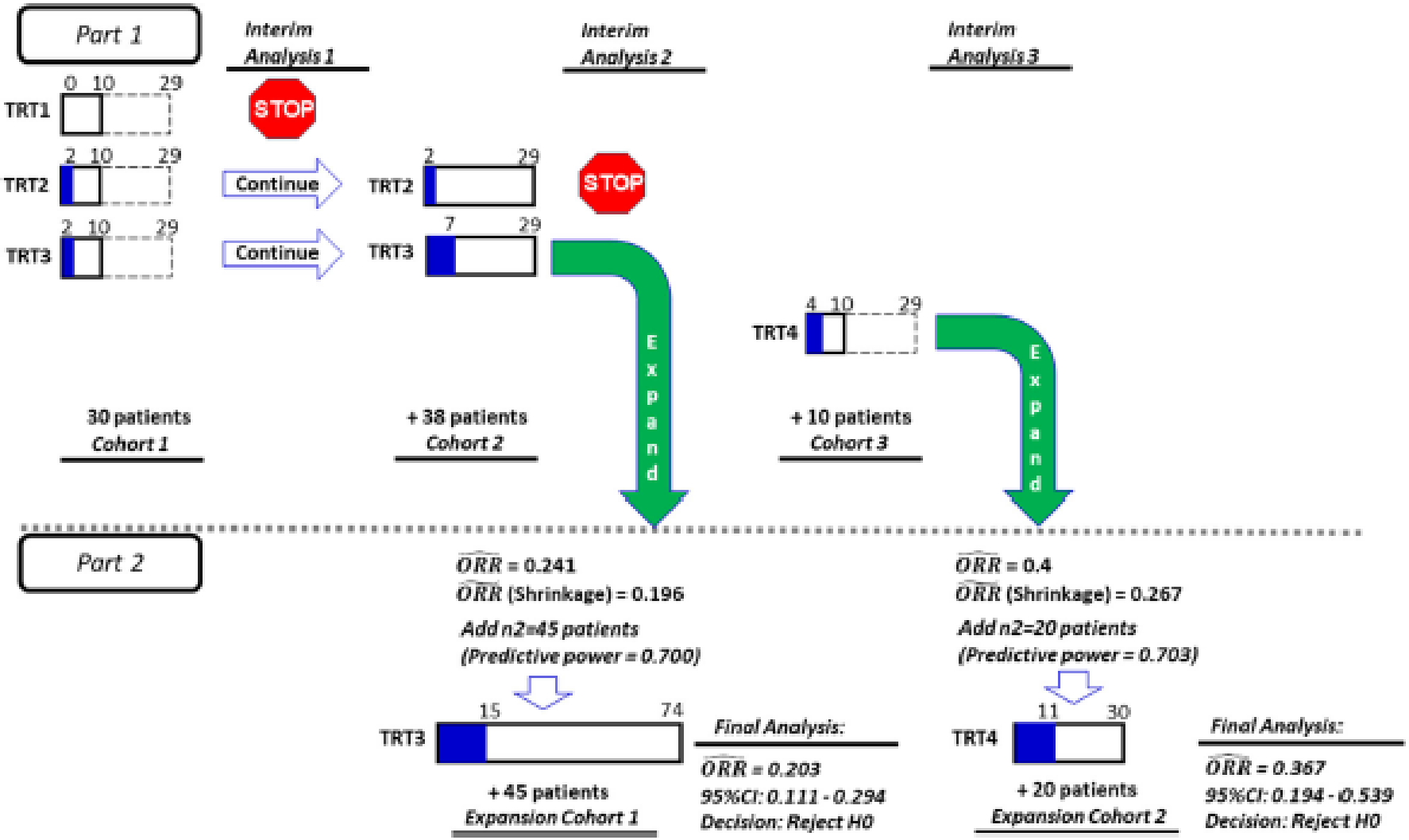
- Predictive power is defined as the probability of success, i.e.
  - probability of a statistically significant result at the end of the study given the results obtained at the selection of an efficacious arm.
- Shrinkage estimator overview:
  - Estimates of ORR observed in part 1 can potentially be biased due to random high or random low values
  - Thus the arm(s) to be expanded can potentially be selected based on a random-high ORR observed in part 1.
  - Consequently the estimated sample size and power can potentially be affected as well.
  - Therefore, a shrinkage estimator based on a hierarchical model that takes into account the ORR from all the combination arms will be used.
  - The shrinkage estimator for ORR will be calculated using the observed ORR from each available arm at the time when an efficacious combination arm can be determined in part 1 so that predictive power can be determined.

Predictive power assumptions

- Assume 1-sided alpha level of 0.025
- Assume observed ORR's and number of subjects enrolled from part 1 at the time of selecting a winner
- 1000 simulations for each ORR scenario
- 20000 iterations of the Markov chain Monte Carlo (MCMC) for the shrinkage estimator
- Predictive power of at least 70% is required where
  - Predictive power uses a Bayesian approach to determine probability of success where success is defined as statistically significant result at the end of the study; decision of statistical significance is based on exact 95% confidence interval (CI) for ORR
  - Conditioning is done using the shrinkage estimator of the ORR in the efficacious arm in part 1 and takes into account also part 1 ORR results at the time of selecting a winner. Shrinkage estimator reduces bias of the ORR estimate of the efficacious arm that arises from the selection process that can lead to random high or random low values.

Application to Simulation study

Illustration of a Single Trial



Interim decisions for simulation study

Interim analysis 1	n	x	$\widehat{ORR}$	Pr(Efficacy)[a]	Pr(Futility)[b]	Decision[c]
TRT1	10	0	0	0.086	**0.833**	Terminate
TRT2	10	2	0.2	0.617	0.221	Continue
TRT3	10	2	0.2	0.617	0.221	Continue
Interim analysis 2	n	x	$\widehat{ORR}$	Pr(Efficacy)	Pr(Futility)	Decision
TRT1	10	0	0	N/A	N/A	N/A
TRT2	29	2	0.069	0.044	**0.849**	Terminate
TRT3	29	7	0.241	**0.761**	0.070	Expand
Interim analysis 3	n	x	$\widehat{ORR}$	Pr(Efficacy)	Pr(Futility)	Decision
TRT1	10	0	0	N/A	N/A	N/A
TRT2	29	2	0.069	N/A	N/A	N/A
TRT3	29	7	0.241	N/A	N/A	N/A
TRT4	10	4	0.4	**0.950**	0.016	Expand

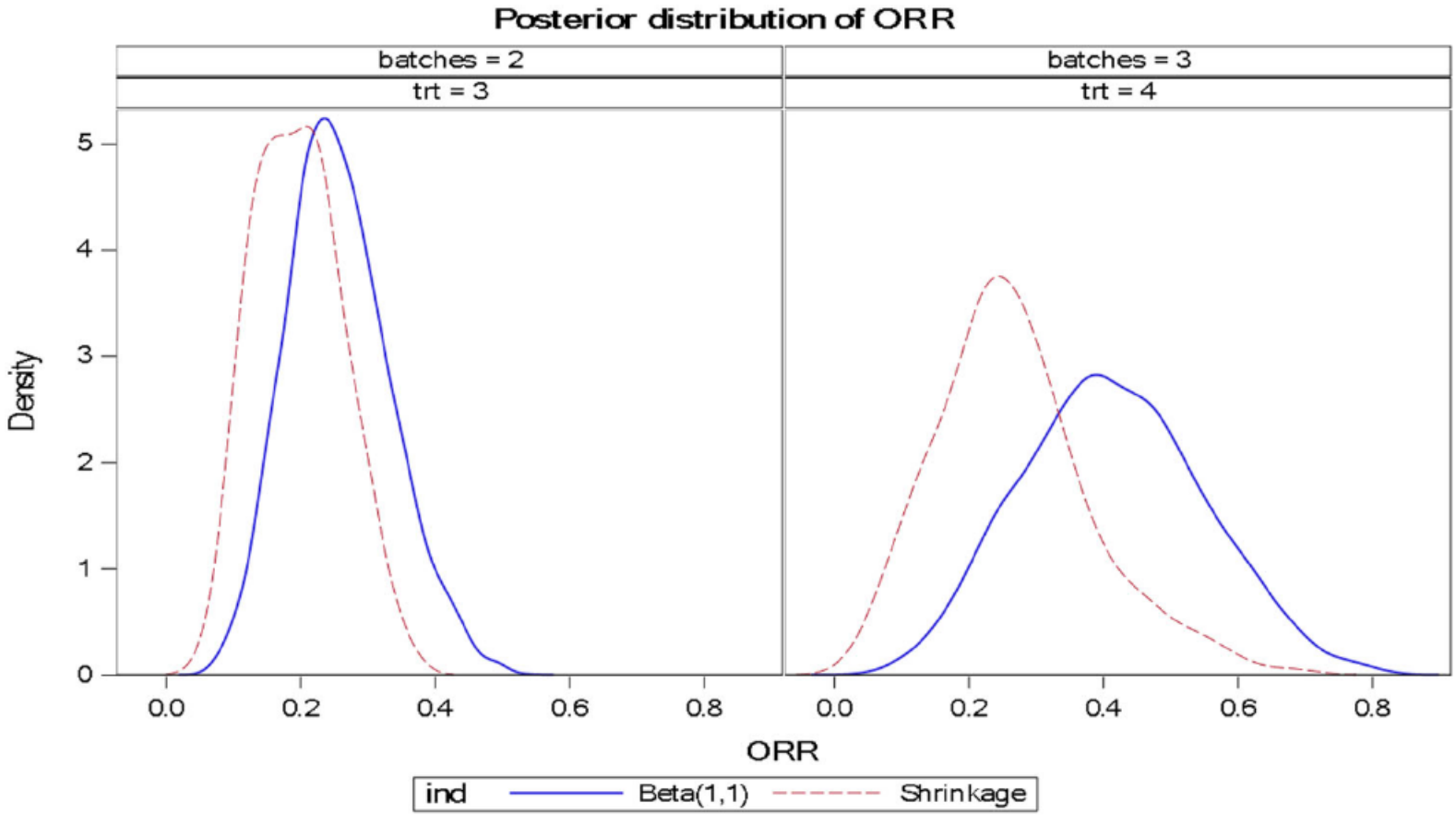
Note: Individual responses are generated using Bernoulli distributions with success probabilities 0.07, 0.10, 0.25, and 0.30 for TRT1, TRT2, TRT3, and TRT4, respectively. It is assumed that enrollment rate is 8 patients per month, and there are 5 months between each interim analysis.

[a] $Pr(\text{Efficacy}) = Pr(ORR > 0.20 \mid \text{data})$.

[b] $Pr(\text{Futility}) = Pr(ORR < 0.15 \mid \text{data})$.

[c] Decisions: Terminate, if $Pr(\text{Futility}) > 0.70$; Expand, if $Pr(\text{Efficacy}) > 0.70$; Continue, if $Pr(\text{Futility}) \leq 0.70$ and $Pr(\text{Efficacy}) \leq 0.70$.

Posterior Distributions of the Shrinkage Estimator



Design Operational Characteristics

Design Parameters & Decision Rules Investigated

Design x (assessment batch n/arm; start with K1 arms, later add K2 arms):

ORR threshold	<x1%	x1-x2%	≥x2%
Decision rule	Prob(ORR<x1%)>p1%		Prob(ORR≥x2%)>p2%
Action	Drop for futility/ stop enrollment	Continue enrollment @	Declare 'winner' *

Design #	Simul. wave	# of arms start/add	n/batch	x1	p1	x2	p2
1	1	5/2	25	10	80	30	90
2	1	5/2	25	15	60	20	60
3	1	5/2	20	15	60	20	60
4	2	3/1/1	20 with 4 batch total	15	70	20	70
5	3	3/1	10 then 8/pts month 4 batch total	15	70	20	70
6	4	4/1/1 ("off times")	**8** pts/mth. with 6 batch total	15	70	20	70

Fine-tuning of operating characteristics

Scenarios to be Tested

Simulation scenario	4 arms			
True ORR	1	2	3	4
Scenario A	7%	10%	25%	30%
Scenario B	10%	15%	25%	35%
Scenario C	10%	10%	10%	10%
Scenario D	5%	7%	10%	10%
Scenario E	25%	25%	25%	25%
Scenario F	25%	28%	30%	35%
Scenario G	7%	10%	25%	25%
Scenario H	7%	15%	25%	25%

'Futile' arms

'Interesting' arms

'Winning' arms

For any given scenario, 10,000 simulation runs were conducted to estimate the design operating characteristics.

For Part 1, we estimated arm-specific decision probabilities (Expand; Terminate; Continue) at different IAs, as well as arm specific sample sizes during Part 1 ($n1$) and percentages for declaring these arms futile or efficacious during Part 1.

For Part 2, we have two sets of results: unconditional (derived across all 10,000 simulation runs), and conditional (derived across those simulation runs for which a decision is made to expand a particular treatment arm into Part 2).

For each set of results, we evaluate for each treatment arm the sample size for Part 2 ($n2$), the total sample size combining Parts 1 and 2 ($n$tot), and the power of the analysis using the frequentist approach based on Clopper–Pearson exact method, based on pooled data from Parts 1 and 2.

Operational Characteristics of Part 1

		Scenario A				Scenario B			
	Decision	TRT1 (0.07)	TRT2 (0.10)	TRT3 (0.25)	TRT4 (0.30)	TRT1 (0.10)	TRT2 (0.15)	TRT3 (0.25)	TRT4 (0.35)
Interim Analysis 1	Expand	2.82	7.4	47.91	N/A	6.9	18.65	47.91	N/A
	Terminate	85.04	73.47	24.02	N/A	73.42	55.22	24.02	N/A
	Continue	12.14	19.13	28.07	N/A	19.68	26.13	28.07	N/A
Interim Analysis 2	Expand	4.1	9.54	25.76	N/A	9.34	18.03	24.85	N/A
	Terminate	7.3	8.21	1.0	N/A	8.38	5.39	0.99	N/A
	Continue	0.74	1.38	1.31	N/A	1.96	2.71	2.23	N/A
Interim Analysis 3	Expand	0.68	1.28	1.3	61.97	1.75	2.59	2.2	73.89
	Terminate	0.06	0.1	0.01	14.29	0.16	0.1	0	8.53
	Continue	0	0	0	23.74	0.05	0.02	0.03	17.58
Interim Analysis 4	Expand	N/A	N/A	N/A	23.48	0.05	0.02	0.03	17.53
	Terminate	N/A	N/A	N/A	0.26	0	0	0	0.05
	Continue	N/A	N/A	N/A	0	0	0	0	0

Note: Probabilities are not cumulative. TRT4 is introduced into the study after IA2. In Scenario A, the true ORRs = 0.07, 0.10, 0.25, and 0.30. In Scenario B, the true ORRs = 0.10, 0.15, 0.25, and 0.35.

Overall Operational Characteristics for the entire study

	Scenario A				Scenario B			
	TRT1 (0.07)	TRT2 (0.10)	TRT3 (0.25)	TRT4 (0.30)	TRT1 (0.10)	TRT2 (0.15)	TRT3 (0.25)	TRT4 (0.35)
Part 1 characteristics								
Decision probabilities (%)								
Pr(Expand)	7.6	18.22	74.97	85.45	18.04	39.29	74.99	91.42
Pr(Terminate)	92.4	81.78	25.03	14.55	81.96	60.71	25.01	8.58
Part 1 sample size (n_1)								
Mean (SD) [IQR]	12 (6) [10-10]	14 (7) [10-10]	15 (8) [10-29]	15 (8) [10-10]	14 (7) [10-10]	15 (8) [10-27]	15 (8) [10-29]	13 (7) [10-10]
Part 2 - unconditional characteristics								
Part 2 sample size (n_2)								
Mean (SD) [IQR]	5 (18) [0-0]	12 (26) [0-0]	37 (29) [0-70]	36 (26) [20-70]	12 (26) [0-0]	24 (32) [0-70]	36 (29) [0-70]	34 (24) [20-70]
Total sample size (n_{tot})								
Mean (SD) [IQR]	17 (22) [10-10]	26 (30) [10-29]	52 (32) [29-80]	51 (27) [30-80]	25 (30) [10-28]	39 (37) [10-80]	52 (32) [29-80]	47 (24) [30-80]
Power (P_{1-2}) (%)	<0.1	0.8	69.7	83.6	0.8	13.1	69.4	90.4
Part 2 - conditional characteristics								
Part 2 sample size (n_2)								
Mean (SD) [IQR]	68 (9) [70-70]	66 (13) [70-70]	49 (22) [20-70]	42 (23) [20-70]	65 (14) [70-70]	61 (18) [70-70]	49 (23) [20-70]	37 (22) [20-70]
Total sample size (n_{tot})								
Mean (SD) [IQR]	90 (15) [80-99]	87 (17) [80-99]	66 (24) [45-80]	56 (24) [30-80]	87 (18) [80-99]	80 (22) [80-99]	65 (24) [45-80]	51 (22) [30-80]
Power (P_{1-2}) (%)	0.4	4.4	93.0	97.4	4.3	33.3	92.5	98.9

Note: Part 1 characteristics: Decision probabilities (%) for the four arms, and Sample size for Part 1 (n_1)—mean (SD) [IQR]. Part 2 characteristics (Unconditional—derived across 10,000 simulation runs, and Conditional—derived across simulation runs for which a decision to expand a treatment into Part 2 is made) for the four arms: Sample size for Part 2 (n_2)—mean (SD) [IQR]; Total sample size (n_{tot})—mean (SD) [IQR]; Power (P_{1-2}) (%) of the final analysis using pooled data from Parts 1 and 2. In Scenario A, the true ORRs = 0.07, 0.10, 0.25, and 0.30. In Scenario B, the true ORRs = 0.10, 0.15, 0.25, and 0.35.
SD = standard deviation; IQR = interquartile range [Q25–Q75].

Comparing to Different Designs

Overview of Comparing Designs

- We conducted additional simulation studies to compare three variants of our OPTIM-ARTS design with a more standard design.

- Due to a platform nature of the trial, it is difficult to find a single design that would serve as a “reference” in this setting.

- Since the study goal is to formally test efficacy of selected promising treatments, we use Simon’s optimal two-stage design for calibrating the total sample size per treatment arm.
  - Assuming$H0$:ORR = 0.10 and$H1$:ORR = 0.30, the Type I error rate of 0.01 and the power of 0.9,
  - Simon’s optimal two stage design initially evaluates 17 patients, and if the observed ORR is 2/17 or fewer, the arm is stopped for futility; otherwise additional 42 patients are evaluated, and in the final analysis $H0$ is rejected if the observed ORR is 12/59 or more.
  - In our trial context, such a design can be viewed as one that has the fixed sample sizes for both Part 1 ($n1 = 17$) and Part 2 ($n2 = 42$), such that $ntot = 17 + 42 = 59$ for any arm.

Sample Size – Selection Phase (Part 1)

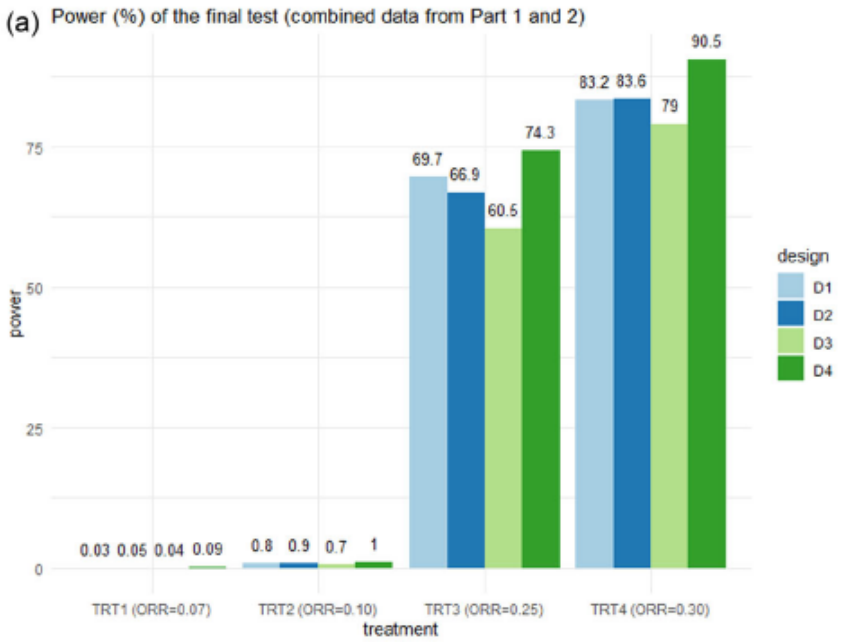
Design	Part 1	Part 2 (only for the "winner" arms from Part 1)
$\mathfrak{D}1$	Platform[a]	Adaptive sample size, to ensure $\geq$70% predictive power in the final analysis
$\mathfrak{D}2$	Platform[a]	Expand to a fixed $n_{tot} = 59$[c]
$\mathfrak{D}3$	Simon's 2-stage for each arm[b]	Expand to a fixed $n_{tot} = 59$[c]
$\mathfrak{D}4$	$n_1 = 17$ patients for each arm, with futility analysis	Expand to a fixed $n_{tot} = 59$[c]

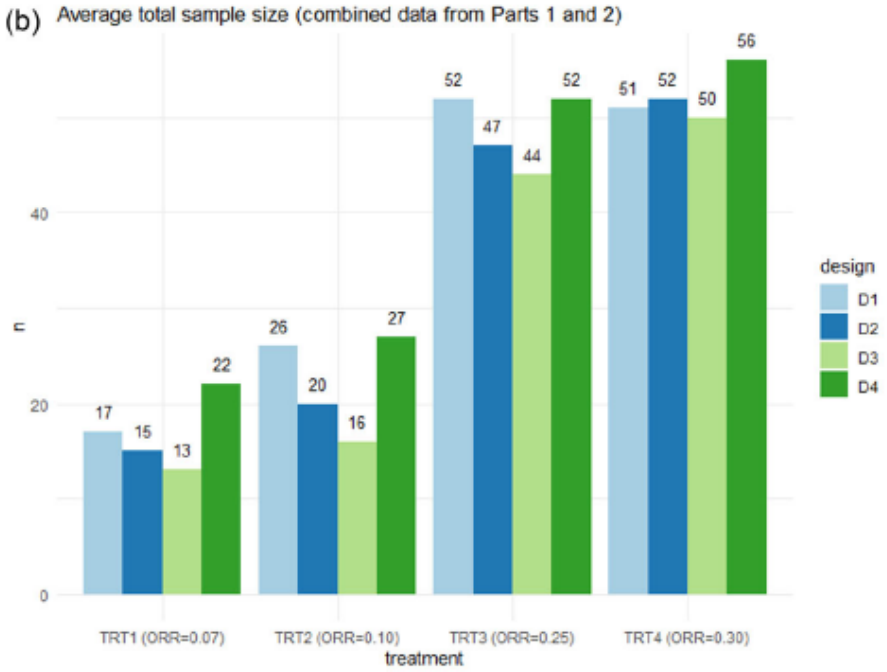
[a]Bayesian decision rules at IAs; maximum number of patients per arm is capped at $n_{max}^{(1)} = 29$.

[b]Futility analysis after first 10 patients; if not futile, enroll 19 more patients and perform an additional analysis based on data from 29 patients.

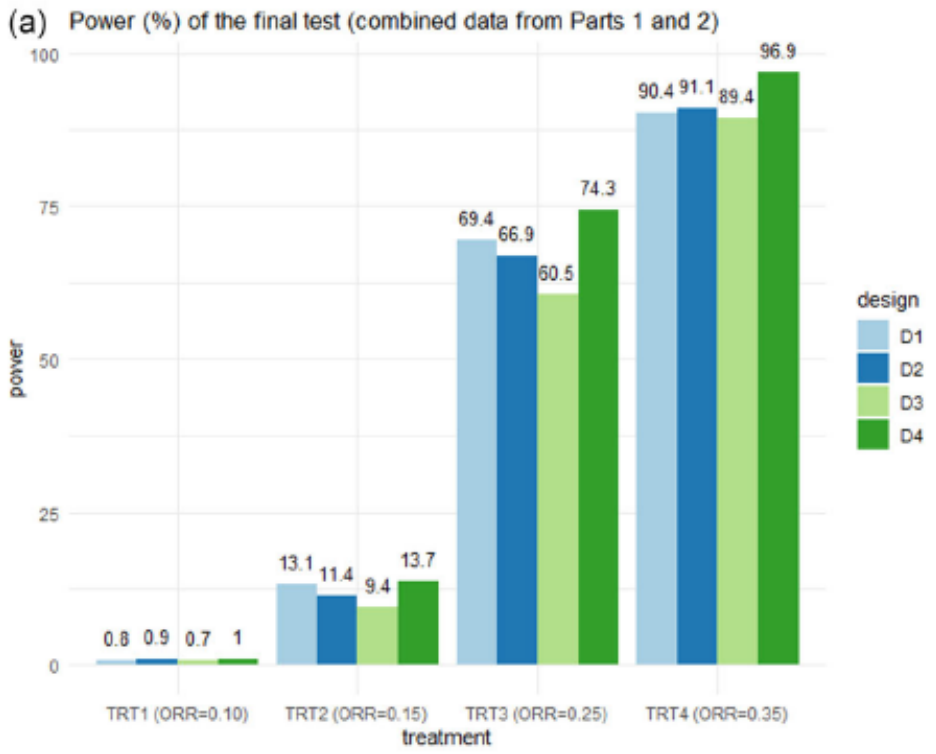
[c] $n_{tot} = 59$ comes from Simon's optimal two-stage design with Type I error of 0.01 (H_0 : ORR=0.10) and power of 0.9 (H_1 : ORR=0.30).

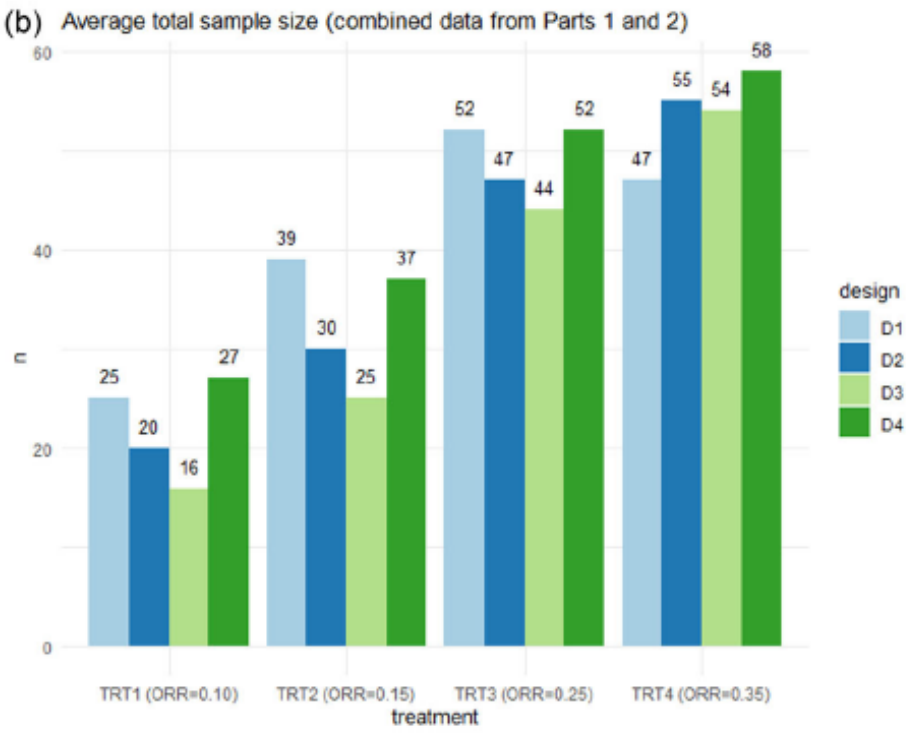
Power and Average Total Sample Size (ATSS) for Scenario A





Power and Average Total Sample Size (ATSS) for Scenario B





Competing design overall assessment

- Among the four designs, D4 has highest power and, in most cases, highest ATSS per arm, whereas D3 has lowest power and lowest ATSS per arm.
- For futile treatments, all four designs have Type I error rate) ≤ 1%, but different values of ATSS;
  - ATSS is 25–26 for D1, 20 for D2, 16 for D3, and 27 for D4.
- For ORR = 0.25 (TRT3 in both scenarios A and B)
  - D1 and D4 have the same ATSS = 52,
  - Different power: 69–70% for D1, and 74.3% for D4.
  - At the same time,D2 and D3 have both lower power (66.9% and 60.5%, respectively) and lower ATSS (47 and 44, respectively).
- Overall, no design seems to be “uniformly best” in terms of cost-efficiency (ATSS/power tradeoff). Higher power naturally comes at the expense of a larger ATSS.
- D1 and D2 have added flexibility due to examining clinical data more frequently compared to D3 and D4.

Conclusions

- Platform trials in the Phase II setting can speed up development of discovering new efficacious compounds/combinations for Phase III development
  - Can test multiple combinations in one master trial
  - Can add new combinations when they become available
  - Framework is very flexible and can be used to construct various designs
  - Can quickly determine combination arms that have either a strong or weak efficacy signal

- Items for further investigation
  - Performance of a Bayesian design may be sensitive to the choice of prior
  - Additional simulations assuming some kind of heterogeneity in the ORR are warranted (especially in a slow enrolling trial)
  - Potential to borrow information across different treatment arms
  - Inclusion of control arm
  - Extension to non-Phase II cancer settings where there may not be binary outcomes

Thank You

Back-Up Slides

Simulation approach for alpha

- Examined scenarios where one or more combination arm(s) had a true ORR of 0.10 (null hypothesis)
  - Assessing type I error for each ‘null hypothesis’ arm defined as the proportion of significant results after completion of part 2 for the given arm; the arm is significant when the lower bound of the 95% CI for ORR is greater than 10%
  - Combines ORR results from Part 1 and Part 2
  - Does not take into account cross-arm assessment
  - Both part 1 and part 2 are fully simulated
- Number of simulations:
  - 1000 simulations are run for part 1
  - additional 100 simulations are run for each simulation for part 2;
    - capture the variance of randomly generating number of responders in expansion phase.
    - ensures that we are not selecting a random high or random low numbers of responders per 1 simulation of the expanded cohort

Overall Alpha Calculation (Simulation-based)

$$Overall\ Alpha = \frac{\sum_{i=1}^{w} f\ (i)}{100000}$$

$$f(i) = \begin{cases} 0\ if arm\ fails\ to\ advance\ to\ part\ 2 \\ 0\ if\ Lower\ 95\%\ CI\ of\ ORR\ is < 0.10 \\ 1\ if\ Lower\ 95\%\ CI\ of\ ORR\ is \geq 0.10 \end{cases}$$

- where the denominator of 100000 represents the following:
  - total number of trials taking into account the number of simulations in part 1 (1000)
  - and the additional number of simulations for part 2 (100) that is conducted for each simulation in part 1

Alpha (type I error) control

- For each arm the potential for alpha inflation arises from the adaptive nature of the design; in particular, from the presence of a selection process
  - Scenario 2: To assess whether alpha is controlled for each combination treatment arm separately, simulations were conducted for several scenarios that include one or more arms with the true ORR of 0.10 corresponding to the null hypothesis threshold
  - Scenario 2 investigated to assess a potential impact of shrinkage estimator on alpha inflation.

Scenario	Arm	True ORR	Probability of Significant Result
1	1	0.10	0.0093
1	2	0.10	0.0114
1	3	0.10	0.0093
2	1	0.10	0.0126
2	2	0.25	NA*
2	3	0.25	NA*

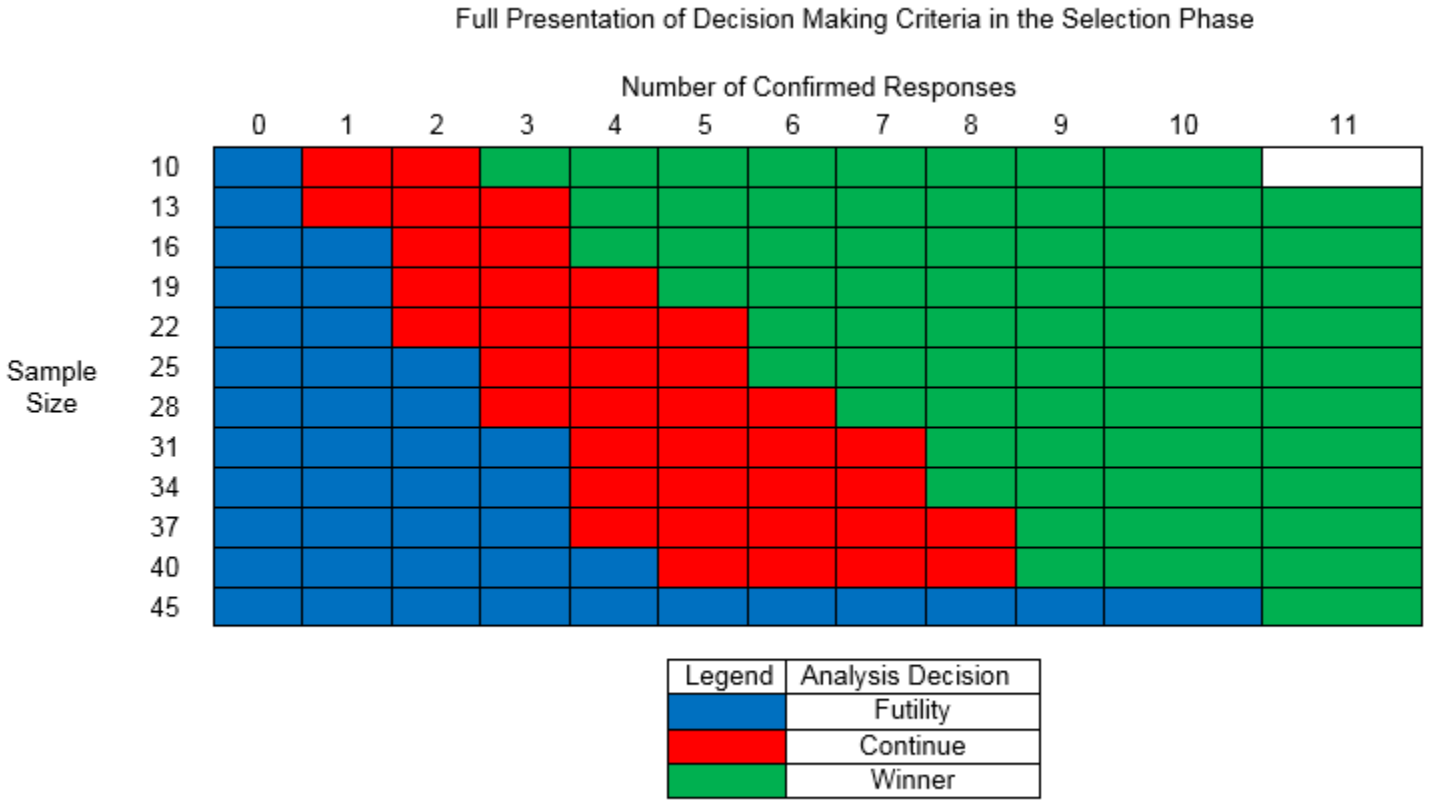
*NA = not applicable in the context of type I error assessment; applicable only to the arms with true ORR = 0.10 corresponding to the null hypothesis

Alpha well controlled

- In simulations, the focus was on assessing type I error (alpha) for each of the 'null hypothesis' arm separately
- Separate type I error (alpha) might be possible if sub-studies / hypotheses are independent

Decision criteria in selection phase

Translated in the # of responders needed to declare ‘winner’ or futility



Note: If we did not cap at 45 patients, 6-10 confirmed responses would be enough to continue with the trial.

Summary for Expansion Phase

- Our “pick the winner” approach in the selection phase may result in the following:
  - Over optimistic estimate of ORR (‘random high’)
  - Could consequently even lead to failure in the extension
- To address these concerns, a hierarchical model that includes shrinkage estimator is used to
  - ‘shrink’ the estimated ORR of the winner
    - Especially in cases when surrounded by futile combinations
  - ‘calculate’ the predictive power of the extension success for choice of sample size in the expansion
  - Provide more robust efficacy calculation to allow for sensitivity analysis
- Through our simulations, it appears that a predictive power of 70% threshold can reduce false positive and provide robust conclusions in terms of power and type I error control