



# Optimal Sequential Trial Design using Stepwise Monte Carlo for Increased Flexibility and Robustness

**Bradley P. Carlin, PhD**

Senior Director, Data Science and Statistics  
Phase V Trials, Inc.

Bayesian Biostatistics 2025, Leiden, Netherlands  
October 22, 2025

- Background and Introduction
- Simplified Sequential Clinical Trial Decision Boundaries
  - Error spending functions (Hwang-Shih-DeCani)
  - Decision boundaries and theoretical guarantees
  - Parameter optimization
  - *Example:* Binary data Bayesian Multi-arm Trial
  - Comparison with existing alternative approaches
- *Future Work:* Extending the Optimization to Non-Boundary Parameters
  - Sequential Local Optimization with constraints
- Conclusions and References

- Bayesian adaptive clinical trials are becoming increasingly complex, incorporating numerous parameters and degrees of freedom.
- Optimal analytic approaches for these intricate trial designs are often unavailable, necessitating extensive simulation to control the Type I error (false positive) rate, maximize trial power, minimize trial sample size, and ensure other favorable operating characteristics.
- We propose a general method to reduce the number of parameters using group stepwise methods and Monte Carlo simulations, significantly decreasing the number of iterations required to identify near-optimal parameters.
- *Key idea:* Use of a family of **error-spending functions** that use just two parameters (an alpha-spending parameter  $\gamma_\alpha$  and a beta-spending parameter  $\gamma_\beta$ ) which determine sensible stopping boundaries for **efficacy** and **futility**, respectively.
- The algorithm then optimally determines stopping boundaries in such a way that power is maximized and overall Type I error is strictly controlled at a predetermined  $\alpha$  level.
- Our method extends classical group sequential designs, but **does not rely on normality assumptions**, and **can accommodate complex trial designs**.

- Let  $\Theta = \Theta_{base} \cup \Theta_{bounds}$ , where
  - $\Theta_{base}$  represents the parameters that are not the efficacy and futility boundaries (such as maximal sample size, interim timings, etc.)
  - $\Theta_{bounds}$  represents the parameters that dictate the efficacy and futility boundaries
- We reduce  $\Theta_{bounds}$  to  $(\gamma_\alpha, \gamma_\beta)$ , where
  - $\gamma_\alpha$  is an alpha-spending parameter (controls Type I error; fixes efficacy bounds)
  - $\gamma_\beta$  is a beta-spending parameter (controls Type II error; fixes futility bounds)
- Let  $t_i$  be the information time (proportion of total information available) at interim look  $i$
- An **alpha-spending function** (respectively **beta-spending function**) is a non-decreasing function  $f: [0,1] \rightarrow [0, \alpha]$ , with  $f(0) = 0$  and  $f(1) = \alpha$  that determines how we allocate our total **Type I error  $\alpha$**  to the sequential looks at the accumulating data
- The error allocated to interim analysis  $i$  is

$$a_1 = f(t_1), a_i = f(t_i) - f(t_{i-1}), t = 2, \dots, k$$

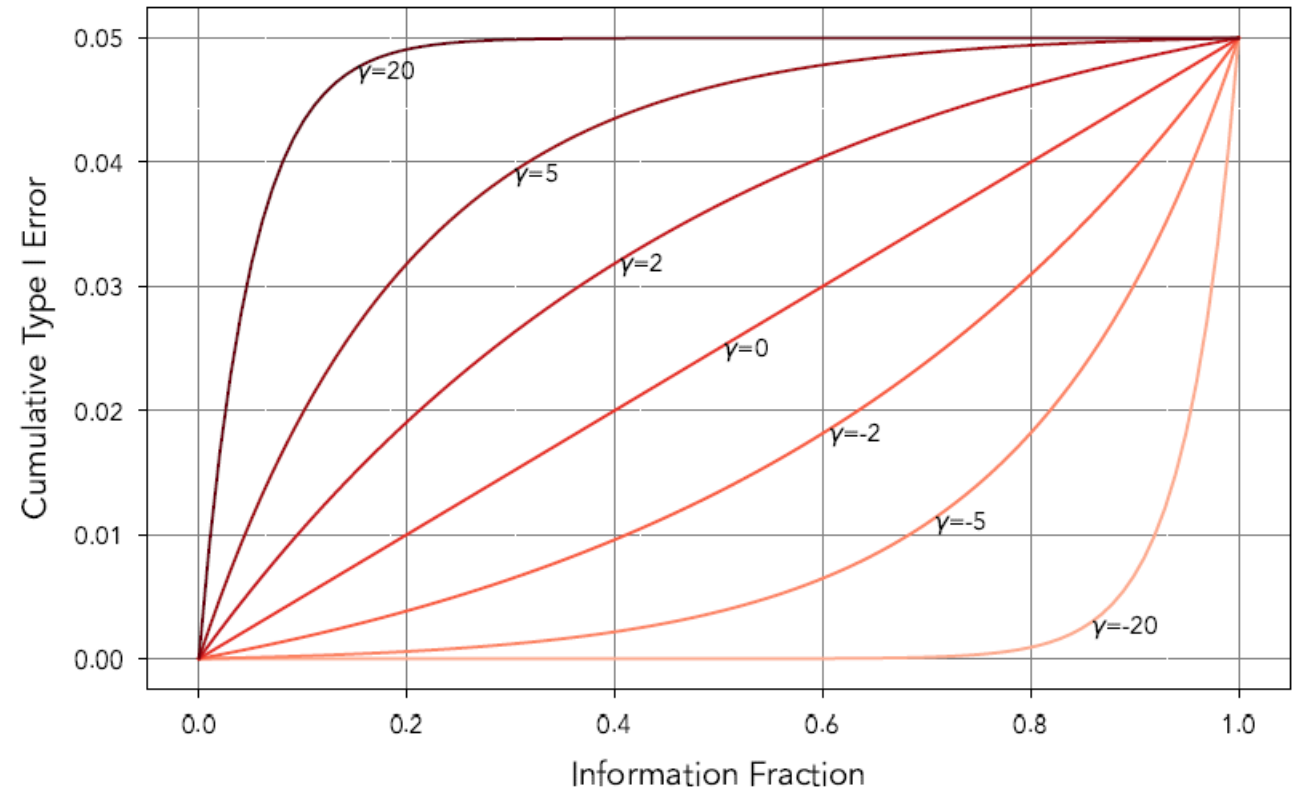
- We focus on the [Hwang-Shih-DeCani \(HSD\)](#) family of error spending functions, i.e.

$$f_{\gamma_\alpha}(t) = \frac{\alpha(1 - e^{-\gamma_\alpha t})}{1 - e^{-\gamma_\alpha}}, \quad \gamma_\alpha \neq 0, \quad \text{and} \quad f_{\gamma_\alpha}(t) = \alpha t, \quad \gamma_\alpha = 0,$$

with similar expressions for the beta-spending function  $f_{\gamma_\beta}(t)$

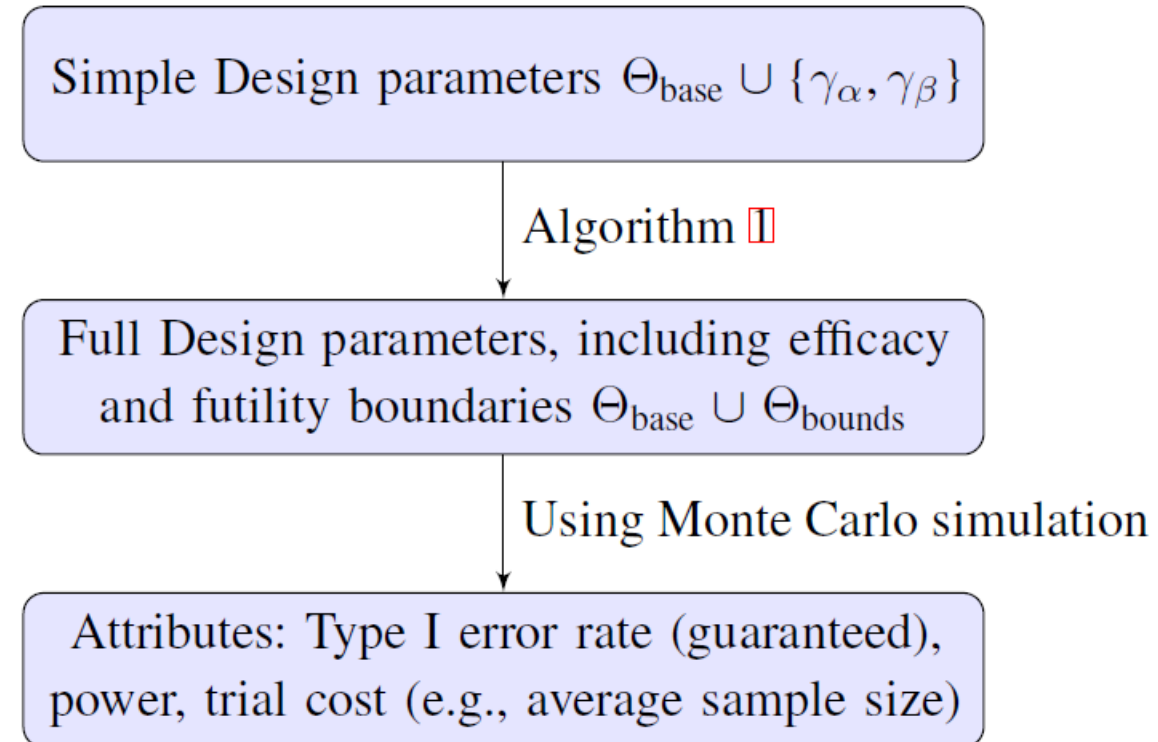
- The parameters  $\gamma_\alpha$  and  $\gamma_\beta$  control the rates at which Type I and II errors are spent
- $\gamma_\alpha = -4$  can be used to approximate an O'Brien-Fleming design
  - Saves more  $\alpha$  for later looks
- $\gamma_\alpha = 1$  approximates a Pocock design well
  - Spends more  $\alpha$  early

Hwang-Shih-DeCani alpha-spending functions,  $\alpha=0.05$



To convert  $\Theta_{base} \cup \{\gamma_\alpha, \gamma_\beta\}$  into  $\Theta_{base} \cup \Theta_{bounds}$  :

- Suppose we have statistics  $Z_1, \dots, Z_k$  calculated at interims  $1, \dots, k$ , with  $k$  being the final analysis
  - **Note:** Despite the use of the  $Z$  notation, the statistics can be arbitrary and need **not** be assumed to follow standard normal distributions
- At interim  $i$  (assuming we get there),
  - We stop for efficacy if  $Z_i > C_i$
  - We stop for futility if  $Z_i < D_i$
  - We set  $D_k = C_k$ , so the trial will always succeed for efficacy or stop for futility at the final look
- We then simulate trials under  $H_0$  and  $H_1$  to determine  $\{C_1, \dots, C_k, D_1, \dots, D_k\}$  for our  $\{\gamma_\alpha, \gamma_\beta\}$  (“Algorithm 1”)
- Given these stopping boundaries, we can now readily simulate Type I error and power for the design



**Theoretical Guarantees:** As the number of artificial trials simulated  $N \rightarrow \infty$ ,

- The probability of stopping for efficacy under  $H_0$  is bounded by  $a_i$  for each interim  $i$ , and the overall Type I error rate is bounded by  $\alpha$ , as desired
- The probability of stopping for futility under  $H_1$  is bounded by  $b_i$  for each interim  $i < k$

## Notes:

- Algorithm 1 does not guarantee the overall Type II error rate (since this may require a **larger sample size**)
- The overall Type I error rate is only guaranteed under **binding** stopping rules (which the DSMB may alter)

**Optimizing the parameters:** Keeping  $\Theta_{base}$  fixed, suppose we wish minimize trial **cost**,  $U$

- **Example:**  $U = \bar{n}_0 + \bar{n}_1$ , the sum of the average sample sizes under  $H_0$  and  $H_1$ , respectively

Given  $\gamma_\alpha$ , increasing  $\gamma_\beta$  will decrease the power (more power spent earlier in the trial). Therefore, there is a maximal  $\gamma_\beta$  such that the power is at least  $1 - \beta$  (assuming  $n$  is large enough that such a  $\gamma_\beta$  exists).

This  $\gamma_\beta$  can be found by binary search applying Algorithm 1 and simulating the trials' power

- The specific implementation of this algorithm depends on the optimization approach and the cost  $U$
- **Specific example coming up next...**

# Example: Binary Data Bayesian Multi-arm Trial

Consider a clinical trial with one control arm and  $m$  treatment arms, where the data from arm  $j$  follows a Bernoulli distribution with unknown parameter  $\theta_j$ . Hypotheses are:

- $H_0: \theta_0 = \theta_1 = \dots = \theta_m$
- $H_1: \theta_j > \theta_0$  for at least one arm  $j$

Under flat priors, we seek to base efficacy and futility decisions on

$$X_{i,j} = P(\theta_j > \theta_0 \mid \text{data at interim look } i) ,$$

If  $X_{i,j} > C_i$ , we stop the trial and declare arm  $j$  the winner;

If  $X_{i,j} < D_i$ , we stop arm  $j$  for futility (but continue with the other arms).

Suppose we further define

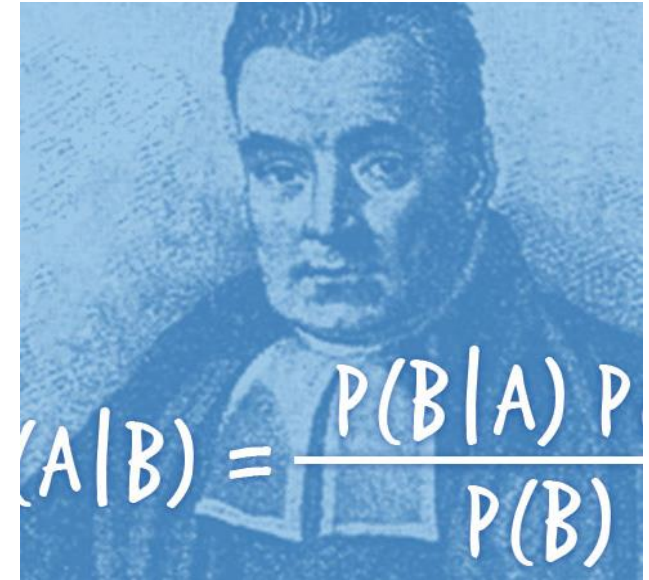
$$Z_i = \max_j (X_{i,j} \mid \text{arm } j \text{ did not stop before interim look } i)$$

Now if  $Z_i > C_i$ , we stop the trial for efficacy, and if  $Z_i < D_i$ , we stop the entire trial for futility.

**Specific example:**  $m = 3$  active arms,  $\alpha = 0.025$ , and target power 80% when  $\theta_0 = 0.5, \theta_1 = 0.7, \theta_2 = 0.6$ , and  $\theta_3 = 0.55$ . A fixed sample size design would require 117 subjects per arm, i.e.  $n = 468$  subjects.

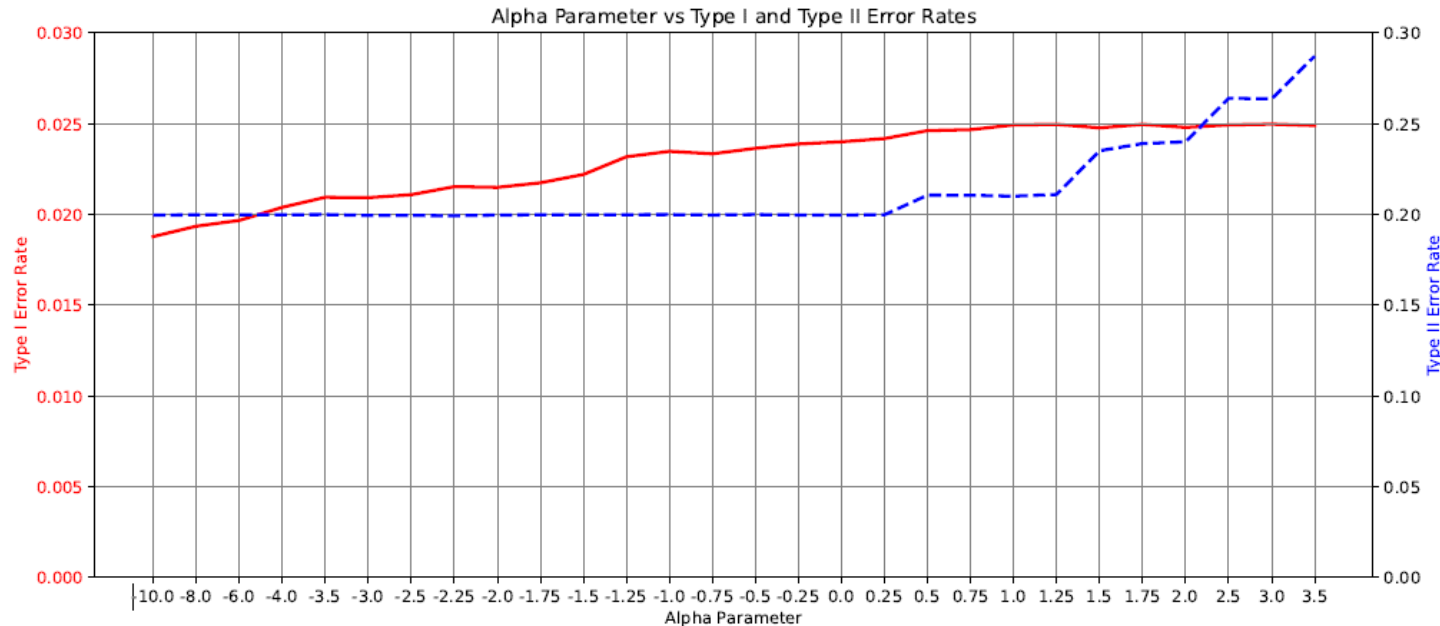
Treatment  $j$  is considered effective if at the final analysis,

$$P(\theta_j > \theta_0 \mid \text{data}) > 0.99$$



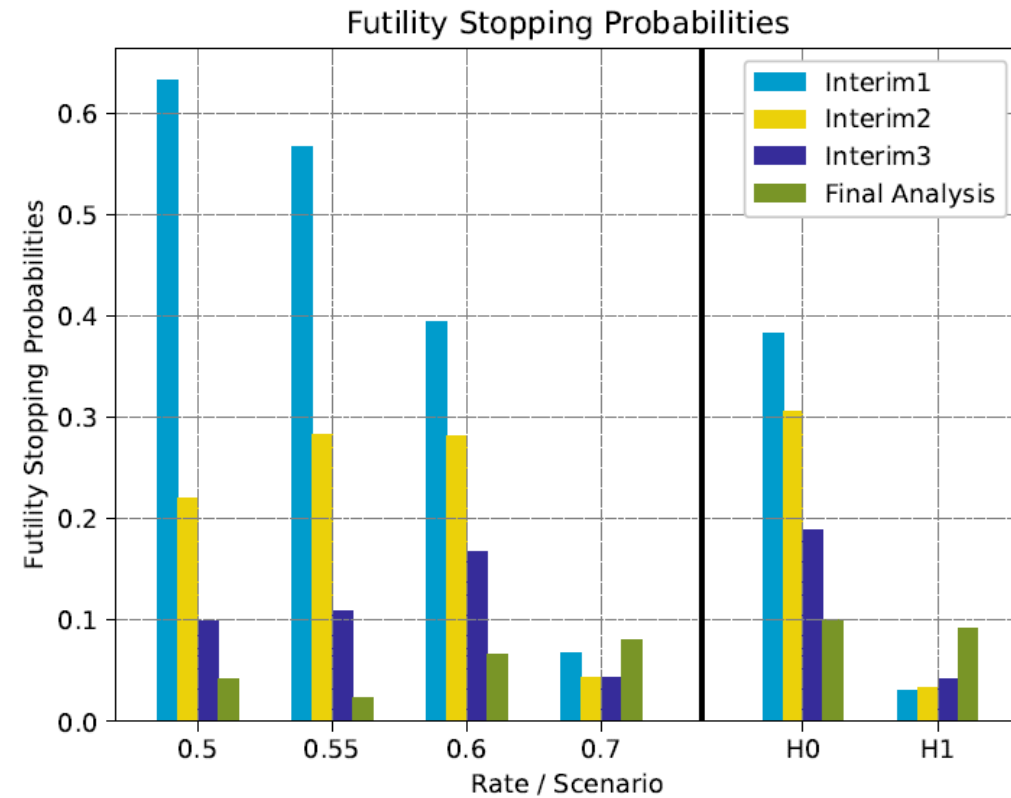
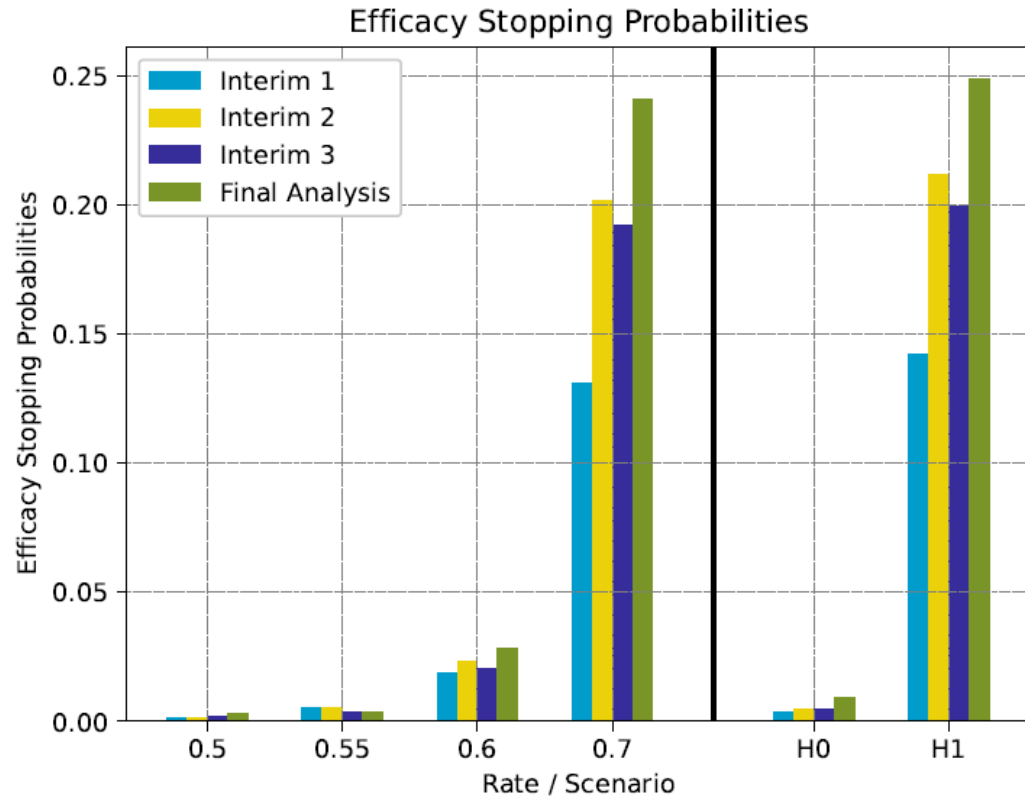
# Results for Binary Data Bayesian Multi-arm Trial

- ❖ For our adaptive design, we set  $n_{max} = 536$  (roughly 15% inflation beyond fixed SS), and make interim analyses at information fractions  $t_i$  of 30%, 50%, and 70%.
- ❖ We used  $N = 1,000,000$  artificial trials, and re-ran our optimization algorithm over a grid of  $\gamma_\alpha$  values.



- ❖ **Type I error** always  $< 0.025$ , but when  $\gamma_\alpha > 0.25$ , there is **no**  $\gamma_\beta$  for which **Type II error**  $< 0.20$
- ❖ Trial cost  $U$  is minimized (and roughly constant) for  $-4 \leq \gamma_\alpha \leq -1.25$
- ❖ **Our final selection:**  $(\gamma_\alpha, \gamma_\beta) = (-1.33, -1.42)$ . This yields:
  - ❖  $\bar{n}_0 = 275$  (**41% savings** over fixed design), and  $\bar{n}_1 = 344$  (**26.5% savings** over fixed design)

Efficacy/futility stopping probabilities for individual arms (L) or full trial under scenarios  $H_0$  or  $H_1$  (R):



- ❖ **Efficacy stopping** is rare for arms with  $\theta_j < 0.7$ , and virtually nonexistent for arms with  $\theta_j = 0.5$  (null)
- ❖ **Efficacy stopping** is common for active arms with  $\theta_j = 0.7$ , especially after the first interim look
- **Futility stopping** for arms having  $\theta_j < 0.7$  (or under the null) is common, often happening at the 1<sup>st</sup> look

- 1. OptGS:** an *R* package for finding near-optimal classical group sequential designs
  - Freely available at <https://www.jstatsoft.org/article/view/v066i02>
  - Extremely fast, but applies only to 2-arm trials with normally distributed outcomes
- 2. General Spending Algorithm:** Instead of using the HSD spending function, try to optimize the error spends  $a_1, \dots, a_k, b_1, \dots, b_k$  directly.
  - In practice, we first transform to  $a'_i = a_i / (\alpha - \sum_{j < i} a_j)$  and  $b'_i = b_i / (\beta - \sum_{j < i} b_j)$ , use a differential evolution algorithm to optimize over  $\{0 \leq a'_i, b'_i \leq 1\}$ , and then transform back
  - We are now optimizing over a parameter space of size  $2(k - 1)$ , whereas for HSD it was only 2
  - This leads to longer runtimes (especially for larger numbers of looks  $k$ ), but possibly better results
- 3. Direct Boundaries Algorithm:** As closely related alternative, we could run an optimization algorithm directly on the stopping boundaries  $\{C_1, \dots, C_{k-1}, D_1, \dots, D_{k-1}, D_k = C_k\}$  themselves.
  - Again we would use differential evolution to perform the optimization
  - Again the parameter space is much larger than with HSD ( $2k - 1$  versus 2), hence we expect longer runtimes but possibly better accuracy

We considered a simulation study across multiple different trial scenarios:

1. The number of looks  $k$  is between 3 and 6 (4 possible values in total)
2. The number of arms  $m$  is between 2 and 4 (3 possible values in total)
3. The futility stops are considered either binding or non-binding (2 possible values).

For each of the above 24 possibilities, we set or randomized the other trial parameters 5 times:

1. The interim looks happen after 50% of the data accumulate for 2 looks, and uniformly between 30% and 100% of the data for more than 2 looks
2. The true response rate of each arm under  $H_0$  (and the control arm under  $H_1$ ) is 0.5
3. The true rate of each active treatment arm under  $H_1$  is chosen randomly between (0.55, 0.6, 0.7, 0.8)
4. The increase of the maximal sample size, relative to the fixed sample size trial, is randomly chosen between 5% and 25%

The statistical properties we measure are:

- **Optimality:** the cost ( $\bar{n}_0 + \bar{n}_1$ ) relative to that of the HSD-based algorithm (HSD cost in denominator)
- **Precision:** the absolute deviation from the desired Type I error rate (2.5%) and power (80%)
- **Efficiency (Runtime):** the average number of function evaluations used by each optimization algorithm

# Results of the Simulation Study

- Results for **optimality**:
  - General spending algorithm is only a bit better (about 2%) than the HSD-based algorithm
  - Direct boundaries algorithm is more or less equivalent to the HSD-based algorithm
- Results for **precision**:
  - General spending algorithm and HSD-based algorithm are equally precise
  - Direct Boundaries algorithm is about half as precise as HSD-based

Algorithm type	Average and Std of cost relative to HSD-based algorithm (Optimality)	Average and Std of Type I error rate deviation from 2.5% (Precision)	Average and Std of power deviation from 80% (Precision)
HSD-based Algorithm	-	$0.03\% \pm 0.02\%$	$0.1\% \pm 0.16\%$
General Spending Algorithm	$0.981 \pm 0.018$	$0.03\% \pm 0.02\%$	$0.1\% \pm 0.17\%$
Direct Boundaries Algorithm	$0.995 \pm 0.018$	$0.06\% \pm 0.04\%$	$0.26\% \pm 0.39\%$

- Results for **efficiency**:
  - The number of function evaluations for our HSD algorithm is nearly constant across all numbers of looks
  - For the other two algorithms, the number of function evaluations grows substantially with the number of looks

Algorithm type	3 looks	4 looks	5 looks	6 looks
HSD-based Algorithm	$601 \pm 153$	$614 \pm 263$	$604 \pm 173$	$577 \pm 144$
General Spending Algorithm	$1957 \pm 782$	$4579 \pm 3500$	$6421 \pm 2184$	$8071 \pm 3183$
Direct Boundaries Algorithm	$9941 \pm 3613$	$20882 \pm 8643$	$37877 \pm 15245$	$50854 \pm 23906$

**Summary:** By reducing  $\Theta_{bounds}$  to two key spending parameters, we have gained flexibility to handle a wide range of trial designs (e.g., **non-standard endpoints, multiple arms, adaptive features**, etc.)

**Reference for this work:** Kamber, A., Berkman, E., Frostig, T., Pryluk, R., Racah, Y., and Carlin, B.P. (2025). Group sequential trial design using stepwise Monte Carlo for increased flexibility and robustness. Technical report, PhaseV Trials, Inc. To appear *Statistics in Medicine* (<http://dx.doi.org/10.1002/sim.70249>).

**Next step:** Adaptive trials include many other degrees of freedom in  $\Theta_{base}$  (e.g., rate of **alpha/beta spending**, precise **locations of the interim looks**, etc.)

- **Goal:** Over a vector of unknown trial design parameters  $\theta \in \Theta \subset \mathbb{R}^d$ , minimize trial cost  $U(\theta)$  subject to  $\alpha(\theta) \leq \alpha_0$  and  $\beta(\theta) \leq \beta_0$ , where  $\alpha_0$  and  $\beta_0$  are the desired upper bounds on Type I and Type II error rates.
- **Complication:** a constraint may not be satisfiable in the current local neighborhood.
- **Approach:** Optimize  $\log(\bar{n}) + \sum_i (C_0 I_{V_i > 0} + C_1 V_i + C_2 V_i^2)$ , where  $V_i$  is degree to which constraint  $i$  is violated

**Reference:** Racah, Y., Kamber, A., Frostig, T., Cohen, R., Levy, A., Pryluk, R., Berkman, E., and Carlin, B.P. (2025). Model-guided parameter optimization for complex innovative trial design. Technical report, PhaseV Trials, Inc.

# Thank you for your attention

*Happy to have questions or comments!*

Bradley P. Carlin, PhD  
Senior Director, Statistics and Data Science  
[brad@phasevtrials.com](mailto:brad@phasevtrials.com)

