The Properties of Multi-Arm Response Adaptive Designs

Tom Parke

Adaptive Designs and Multiple Testing Procedures 2025

Introduction

- Confession: the first ever clinical trial I worked on (The ASTIN Trial) used RAR to find the ED90 from amongst 16 different doses.
  - Pfizer (UK)
  - Neuroprotectant in Stroke
  - Trial ran from 1999 to 2001
  - Designed by Don Berry and Peter Mueller
  - PI was Dr Michael Krams
  - Pfizer statistician was Prof Andy Grieve
  - Grieve AP, Krams M. ASTIN: a Bayesian adaptive dose-response trial in acute stroke. Clin Trials. 2005;2(4):340-51; discussion 352-8, 364-78. doi: 10.1191/1740774505cn094oa. PMID: 16281432.
- It was a thrilling and exciting ride, and I've been a fan of adaptive trials and RAR ever since.

Perplexed

- I continue to be surprised that adaptive trials and RAR are still not used more often.
- Advantages not widely perceived?
- There are authors opposed to them
  - Korn & Friedlin:
    - Korn EL, Freidlin B. Outcome--adaptive randomization: is it useful? J Clin Oncol. 2011 Feb 20;29(6):771-6. doi: 10.1200/JCO.2010.31.1423. Epub 2010 Dec 20. PMID: 21172882; PMCID: PMC3056658.
    - Korn EL, Freidlin B. Time trends with response-adaptive randomization: The inevitability of inefficiency. Clinical Trials. 2022;19(2):158-161. doi:10.1177/17407745211065762
  - Thall and Wathen
    - Wathen JK, Thall PF. A simulation study of outcome adaptive randomization in multi-arm clinical trials. Clin Trials. 2017 Oct;14(5):432-440. doi: 10.1177/1740774517692302. Epub 2017 Feb 1. PMID: 28982263; PMCID: PMC5634533.
  - Proschan
    - Proschan M, Evans S. Resist the Temptation of Response-Adaptive Randomization. Clin Infect Dis. 2020 Dec 31;71(11):3002-3004. doi: 10.1093/cid/ciaa334. PMID: 32222766; PMCID: PMC7947972.

- Countervailing arguments have been put forward by my colleague Kert Viele
  - Viele K, Saville BR, McGlothlin A, Broglio K. Comparison of response adaptive randomization features in multiarm clinical trials with control. Pharm Stat. 2020 Sep;19(5):602-612. doi: 10.1002/pst.2015. Epub 2020 Mar 21. PMID: 32198968.
  - Viele K, Broglio K, McGlothlin A, Saville BR. Comparison of methods for control allocation in multiple arm studies using response adaptive randomization. Clin Trials. 2020 Feb;17(1):52-60. doi: 10.1177/1740774519877836. Epub 2019 Oct 19. PMID: 31630567.
- For a review see:
  - Robertson DS, Lee KM, López-Kolkovska BC, Villar SS. Response-adaptive randomization in clinical trials: from myths to practical considerations. Stat Sci. 2023 May;38(2):185-208. doi: 10.1214/22-STS865. PMID: 37324576; PMCID: PMC7614644.
- But in recent advert for an "**Ethics and Innovative Clinical Trial Designs**" event:
  - The controversy surrounding these methods raises both ethical and methodological questions. There is ongoing debate about whether, when and which of such methods are more efficient, meaning they can answer research questions with fewer participants without compromising reliability or generalizability. This debate is fueled by competing simulation studies that present conflicting assessments of these designs, as well as concern that certain approaches, particularly Bayesian designs, may introduce unacceptable bias.

Main Accusations

- From Michael Proschan:
  - “Unfortunately, RAR causes many problems, including
    (1) bias from temporal trends,
    (2) inefficiency in treatment effect estimation,
    (3) volatility in sample-size distributions that can cause a nontrivial proportion of trials to assign more patients to an inferior arm,
    (4) difficulty of validly analyzing results, and
    (5) the potential for selection bias and other issues inherent to being unblinded to ongoing results.”

Purpose of this talk

- To explore the criticisms of multi-arm RAR through as simple example as possible

Simulation Example

- Control plus 3 treatment arms
- Continuous endpoint, normally distributed, SD 2 points, target treatment improvement 1 point.
- Select arm with greatest treatment effect.
- Target type-1 error: 0.05, power: 0.8 [This is a phase 2 trial]
- Sample size: 288. (Fixed design: 72 per arm)
- Consider response scenarios: (0, 0, 0, 0), (0, 0, 0, 1), (0, 0.33, 0.67, 1), (0, 0.5, 1, 0.8)
- Principal operating characteristics (OCs): type-1 error, power, when successful was the correct arm selected?
- Secondary OCs: Bias and Error in estimation of response of the selected arm
- Secondary OCs: Mean number of subjects allocated to selected arm, number of times allocation is < 72
- Consider 1 stage (fixed), 2 stage (1 interim), 5 stage (4 interims) and 10 stage (9 interims).

Example

- We will simulate a fixed design and 3 RAR designs in FACTS using a v simple Bayesian design:
  - Analysis is a simple pairwise Normal comparison of each treatment arm with Control
    - The prior for the estimate of response θ_d is $N(0, 10^2)$ – centered at the expected control response, effective sample size $1/25^{th}$ of a subject.
    - The prior for the estimate of a common variance σ^2 is IG(10, 1) (central value, weight) [this was a mistake, should have been centered at 2, but the weight is small and resulting bias is 0.08 – will be higher at earlier interims]
  - Final success will be by testing the posterior probability that the response on the treatment arm is greater than the response on control against a critical threshold.
    - E.g. $Pr(\theta_d > \theta_{d=ctl}) > 0.975$
    - We choose the actual critical thresholds by simulating the Null scenario in each design 100,000 times and choose a critical threshold that limits the observed type-1 errors to less than 0.05.
- Fixed design uses equal allocation 1:1:1:1
- RAR designs have 1, 4, or 9 equally spaced interims (giving 2, 5 or 10 stages) which result is different critical thresholds, but are otherwise the same.

Example cont'd

- We start with equal allocation
- Designs will have 0, 1, 4 or 9 interims equally spaced throughout the trial
  - After 288 subjects
  - After 144, 288 subjects
  - After 58, 115, 173, 230, 288 subjects
  - After 29, 58, 87, 115, 144, 173, 202, 230, 259, 288 subjects
- We simulate a v short time to endpoint (0.1 of a week) so at any time the number of subjects enrolled is v close to the number complete.
- At each interim we perform the Bayesian analysis,
  - Post interim the allocation between the treatment arms is proportional to the posterior probability that the arm has the maximum response
- Allocation to control is either:
  - Fixed allocation, e.g. 1:T or Sqrt(T):T where T is the number of treatment arms
  - 1:T initially then matching the allocation to the arm with the highest allocation ratio. So Control:Best-Arm allocation tends to 1:1.

Example simulation from the simulations with 2 stages of the Null scenario.

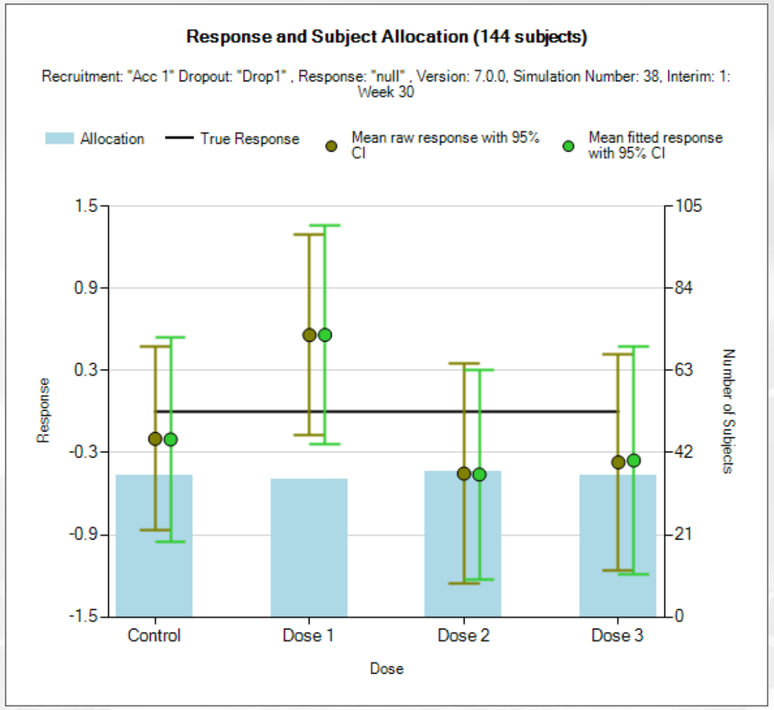
Blue bars show the number of subjects enrolled per arm (right hand y-axis).

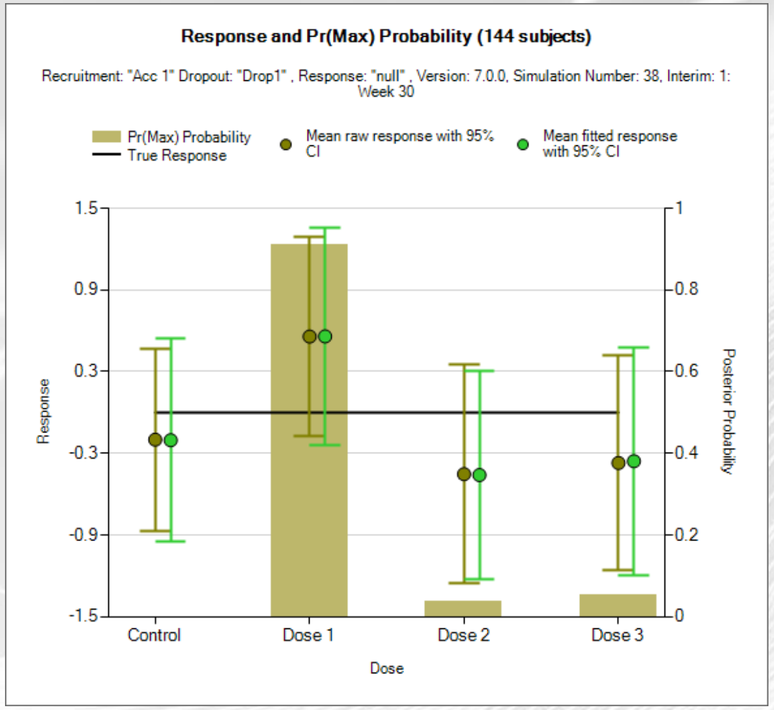
Brown dot and bars show raw response & 95% CI. Green dot and bars show fitted response (left hand y-axis).

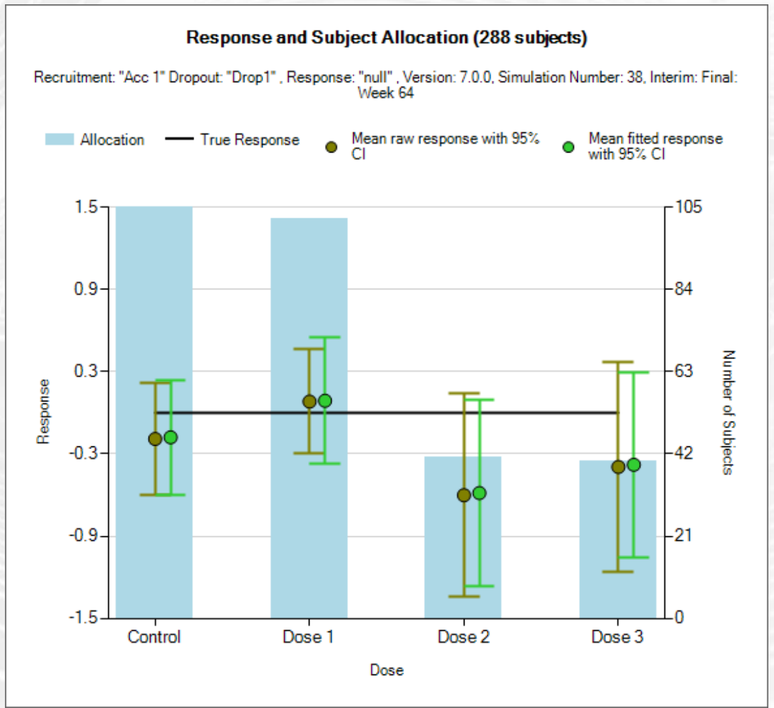
Top graphs show data and corresponding pr(Max) (brown bars far graph, right hand y-axis) at the interim.

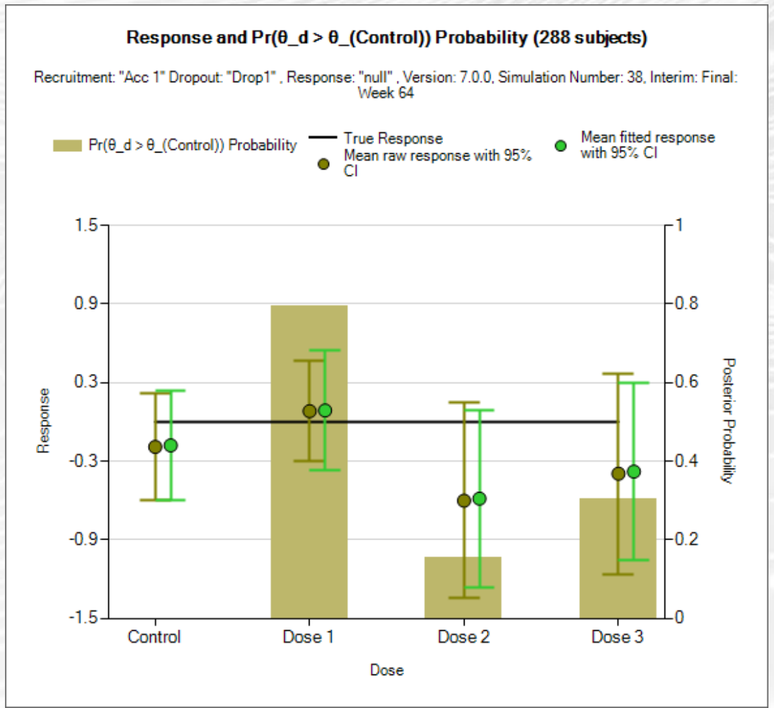
Bottom graphs show data and corresponding $Pr(\theta_d > \theta_{d=ctl})$ (brown bars far graph, right hand y-axis) at the final analysis .

See how in this case a possible type-1 error was rescued by some regression to the mean on Control and “Dose 1” aided by the higher allocation to those arms.









Fixing the final alpha level

- Initial run of 100,000 simulations of the global Null for each design, allowed critical value for $Pr(\theta_d > \theta_{d=ctl})$ to be derived that control type-1 error rates to 0.05.
  - 1 stage: 0.9775
  - 2 stages: 0.974
  - 5 stages: 0.9725
  - 10 stages: 0.972
- 10,000 simulations of each design in the other scenarios were then run

Compare Fixed to 1 Interim: power, arm selection, allocation

Scenario	Design	P(success)	Select arm 1	Select arm 2	Select arm 3	Alloc Ctl	Alloc arm 1	Alloc arm 2	Alloc arm 3
(0, 0, 0, 0)	1 stage	0.049	0.017	0.016	0.016	72.0	72.0	72.0	72.0
	2 stages	0.049	0.016	0.016	0.016	93.0	87.7	88.4	88.2
(0, 0, 0, 1)	1 stage	0.823	0.000	0.000	0.822	72.0	72.0	72.0	72.0
	2 stages	0.920	0.000	0.000	0.919	103.3	76.5	99.0	103.3
(0, 0.33, 0.67, 1)	1 stage	0.854	0.007	0.118	0.729	72.0	72.0	72.0	72.0
	2 stages	0.908	0.008	0.128	0.771	95.6	75.4	84.1	92.9
(0, 0.5, 1, 0.8)	1 stage	0.884	0.029	0.630	0.225	72.0	72.0	72.0	72.0
	2 stages	0.920	0.032	0.652	0.236	93.5	75.0	89.0	83.9

Scenario	Design	Under 72 Ctl	Under 72 Arm 1	Under 72 Arm 2	Under 72 Arm 3
(0, 0, 0, 0)	1 stage	0.000	0.000	0.000	0.000
	2 stages	0.001	0.003	0.002	0.003
(0, 0, 0, 1)	1 stage	0.000	0.000	0.000	0.000
	2 stages	0.001	0.000	0.000	0.008
(0, 0.33, 0.67, 1)	1 stage	0.000	0.000	0.000	0.000
	2 stages	0.006	0.003	0.032	0.080
(0, 0.5, 1, 0.8)	1 stage	0.000	0.000	0.000	0.000
	2 stages	0.010	0.013	0.103	0.057

- Power has gone up
- Type-1 error is controlled
- Probability of selecting the best arm has increased
- Allocation to control and the selected arm has increased
- Probability of under allocating is low

BERRY

Scenario	Design	Bias Ctl	Bias Arm 1	Bias Arm 2	Bias Arm 3	MSE Ctl	MSE Arm 1	MSE Arm 2	MSE Arm 3
(0, 0, 0, 0)	1 stage	-0.38	0.43	0.44	0.43	0.17	0.21	0.22	0.21
	2 stages	-0.33	0.39	0.39	0.39	0.13	0.17	0.17	0.17
(0, 0, 0, 1)	1 stage	-0.06	0.53	0.83	0.05	0.05	0.31	0.71	0.05
	2 stages	-0.02	0.64	0.49	0.02	0.04	0.42	0.24	0.03
(0, 0.33, 0.67, 1)	1 stage	-0.05	0.50	0.29	0.08	0.05	0.28	0.12	0.05
	2 stages	-0.03	0.49	0.26	0.05	0.04	0.26	0.09	0.03
(0, 0.5, 1, 0.8)	1 stage	-0.04	0.44	0.11	0.24	0.05	0.22	0.05	0.09
	2 stages	-0.02	0.40	0.07	0.20	0.04	0.18	0.04	0.07

- Bias and MSE on selected arm have decreased

Summary

- Power, decision making and estimation are all BETTER.
- On average we will have 25-50% MORE data on the selected arm
- With ~10% chance we have less than with the fixed trial

Effect of more interims

BERRY

Power, correct arm selection, E(n) on selected arm when successful

Scenario	Design	P(success)	Select arm 1	Select arm 2	Select arm 3	Alloc Ctl	Alloc arm 1	Alloc arm 2	Alloc arm 3
(0, 0, 0, 0)	1 stage	0.049	0.017	0.016	0.016	72.0	72.0	72.0	72.0
	2 stages	0.049	0.016	0.016	0.016	93.0	87.7	88.4	88.2
	5 stages	0.050	0.016	0.017	0.017	104.4	97.6	96.8	97.7
	10 stages	0.050	0.017	0.016	0.016	108.0	99.2	99.9	98.9
(0, 0, 0, 1)	1 stage	0.823	0.000	0.000	0.822	72.0	72.0	72.0	72.0
	2 stages	0.920	0.000	0.000	0.919	103.3	76.5	99.0	103.3
	5 stages	0.950	0.001	0.000	0.949	117.5	73.6	98.8	119.0
	10 stages	0.952	0.001	0.000	0.951	121.7	85.7	56.5	123.0
(0, 0.33, 0.67, 1)	1 stage	0.854	0.007	0.118	0.729	72.0	72.0	72.0	72.0
	2 stages	0.908	0.008	0.128	0.771	95.6	75.4	84.1	92.9
	5 stages	0.923	0.012	0.131	0.780	107.5	69.6	91.7	103.4
	10 stages	0.935	0.012	0.143	0.780	111.1	76.3	94.5	106.7
(0, 0.5, 1, 0.8)	1 stage	0.884	0.029	0.630	0.225	72.0	72.0	72.0	72.0
	2 stages	0.920	0.032	0.652	0.236	93.5	75.0	89.0	83.9
	5 stages	0.936	0.033	0.652	0.251	104.7	76.0	98.3	92.2
	10 stages	0.943	0.035	0.663	0.244	108.3	80.6	101.2	94.8

Power, selection of correct arm, and allocation to selected arm increases with increased number of interims, but there are diminishing returns.

BERRY

Probability of allocating fewer than the fixed trial to the selected arm

Scenario	Design	Under 72 Ctl	Under 72 Arm 1	Under 72 Arm 2	Under 72 Arm 3
(0, 0, 0, 0)	1 stage	0.000	0.000	0.000	0.000
	2 stages	0.001	0.003	0.002	0.003
	5 stages	0.000	0.002	0.002	0.002
	10 stages	0.000	0.002	0.002	0.002
(0, 0, 0, 1)	1 stage	0.000	0.000	0.000	0.000
	2 stages	0.001	0.000	0.000	0.008
	5 stages	0.000	0.000	0.000	0.005
	10 stages	0.000	0.000	0.000	0.005
(0, 0.33, 0.67, 1)	1 stage	0.000	0.000	0.000	0.000
	2 stages	0.006	0.003	0.032	0.080
	5 stages	0.000	0.006	0.024	0.057
	10 stages	0.000	0.005	0.026	0.046
(0, 0.5, 1, 0.8)	1 stage	0.000	0.000	0.000	0.000
	2 stages	0.010	0.013	0.103	0.057
	5 stages	0.001	0.015	0.070	0.043
	10 stages	0.001	0.013	0.068	0.037

Probability of allocating fewer than the fixed trial decreases with greater number of interims, if using more that one interim its less than 7% and that’s’ even when there’s an arm that is very close to the ”best” arm in terms of response.

BERRY

Bias and MSE on selected arm when successful

Scenario	Design	Bias Ctl	Bias Arm 1	Bias Arm 2	Bias Arm 3	MSE Ctl	MSE Arm 1	MSE Arm 2	MSE Arm 3
(0, 0, 0, 0)	1 stage	-0.38	0.43	0.44	0.43	0.17	0.21	0.22	0.21
	2 stages	-0.33	0.39	0.39	0.39	0.13	0.17	0.17	0.17
	5 stages	-0.30	0.37	0.38	0.37	0.11	0.15	0.16	0.16
	10 stages	-0.29	0.37	0.37	0.37	0.11	0.15	0.15	0.15
(0, 0, 0, 1)	1 stage	-0.06	0.53	0.83	0.05	0.05	0.31	0.71	0.05
	2 stages	-0.02	0.64	0.49	0.02	0.04	0.42	0.24	0.03
	5 stages	-0.01	0.51	0.56	0.01	0.03	0.28	0.32	0.03
	10 stages	-0.01	0.51	0.81	0.01	0.03	0.29	0.77	0.03
(0, 0.33, 0.67, 1)	1 stage	-0.05	0.50	0.29	0.08	0.05	0.28	0.12	0.05
	2 stages	-0.03	0.49	0.26	0.05	0.04	0.26	0.09	0.03
	5 stages	-0.02	0.47	0.23	0.04	0.04	0.25	0.08	0.03
	10 stages	-0.02	0.46	0.22	0.04	0.03	0.24	0.07	0.03
(0, 0.5, 1, 0.8)	1 stage	-0.04	0.44	0.11	0.24	0.05	0.22	0.05	0.09
	2 stages	-0.02	0.40	0.07	0.20	0.04	0.18	0.04	0.07
	5 stages	-0.02	0.39	0.06	0.18	0.04	0.17	0.03	0.06
	10 stages	-0.02	0.37	0.06	0.17	0.03	0.16	0.03	0.05

Bias and MSE decreased number of interims, but again there are diminishing returns.

Time Trends?

- Here the “Allocate to control is the same proportion as the best dose” ensures there is a degree of balance between control and the selected arm throughout the trial – though this is not guaranteed. The finally selected arm may not be the arm that appeared best at the outset or midpoint.
- If the risk of a time trend is a major concern, then control data in stages where the allocation to control was > the eventually selected arm, could be down weighted. So the data on control and the selected arm is the same in every stage. Obviously this would bring some small loss in power.

Operational Bias?

- I’ve worked on a number of RAR trials and never seen operational bias.
- I think more interims reduces the risk of investigator behavior changing.
  - If there is 1 mid point interim when things switch from fixed to favoring the best, perhaps this point has some effect on their thinking and behavior
  - But if adaptation is almost from the outset, and continuous, then it seems reasonable to me that this will influence them much less: “the trial is always adapting”.
- **Operational notes:**
- add a very early interim with no adaptation just to test the data management & data processing.
- Make sure interims will be no more frequent than the operation can handle!
- Make sure interim processing is set up to be swift

Delay to endpoint

- If endpoint was 33% of accrual time, in our example this would be ~19 weeks, at any interim we'd have ~95 enrolled but not complete. So there are 193 subjects we can adapt to.
  - If 1 interim, its at 96 complete
  - If 4 interims, they are at: 38, 76, 114, 152 complete
  - If 9 interims, they are at: 19, 38, 57, 76, 96, 115, 134, 153 and 173 complete.
- If endpoint was 50% of accrual time, in our example this would be ~29 weeks, at any interim we'd have ~144 enrolled but not complete. So there are 144 subjects we can adapt to.
  - If 1 interim, its at 72 complete
  - If 4 interims, they are at: 29, 58, 87, 116 complete
  - If 9 interims, they are at: 15, 29, 44, 58, 72, 87, 101, 116 and 130 complete.
- The improvements are reduced but they are still there.

Last observations

- I initially prepared the talk using R and frequentist analysis, but benefits were not so clear. I think
  - Using p-value combination started to lose power with larger number of stages
  - My approximation of the Bayesian “probability of having the maximum response” was too certain and lead to an over aggressive adaptation. I haven’t had time to explore this.
- Doing simulations in R using Bayesian analysis is too slow for my taste so I reverted to FACTS.

- FACTS is free for academic and government institutions
- FACTS is free for evaluation for commercial organizations
- Email me (tom@berryconsultants.com) if you’d like to be on our webinar emailing list.

Berry Consultants

Statistical Innovation

CONTACT INFORMATION

info@berryconsultants.com | (512) 213-6428

WEBSITE
berryconsultants.com

YOUTUBE
youtube.com/berryconsultants

TWITTER
@berryconsultant