Response-adaptive Randomization in Clinical Trials: a broad view of methods

Sofía S. Villar

MRC Biostatistics Unit

UNIVERSITY OF CAMBRIDGE

Invited Session 10: Bringing Balance to Response Adaptive Randomization

Baltimore, 5th May 2023
SCT 44th Annual Meeting

Acknowledgments

Response-adaptive randomization in clinical trials: from myths to practical considerations

David S. Robertson[1], Kim May Lee[1], Boryana C. López-Kolkovska[1], and Sofía S. Villar *[1]

[1]MRC Biostatistics Unit, University of Cambridge, Cambridge, UK

With thanks to

David Robertson

Kim Lee

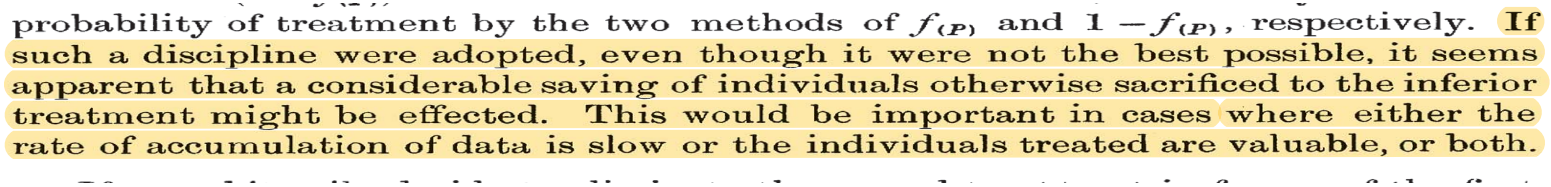
Boryana Lopez-Kolkovska

(and many other colleagues)

Outline

What is Response-Adaptive Randomisation?

- Response-adaptive Randomisation (RAR) is perhaps the oldest form of Adaptive Design of experiments.
- First proposed by Thompson (1933) in the context of clinical trials treatment is greater than the other


probability of treatment by the two methods of $f_{(P)}$ and $1 - f_{(P)}$, respectively. If such a discipline were adopted, even though it were not the best possible, it seems apparent that a considerable saving of individuals otherwise sacrificed to the inferior treatment might be effected. This would be important in cases where either the rate of accumulation of data is slow or the individuals treated are valuable, or both.

+ The paper derives formulae to compute that posterior probability P by hand (redundant for today's computing standards!)
+ Advocated for using data ("*however meagre*") to guide action (or *adaptivity*), specifically with an "ethical" goal

RAR timeline since 1933

Imbalance and recurrence

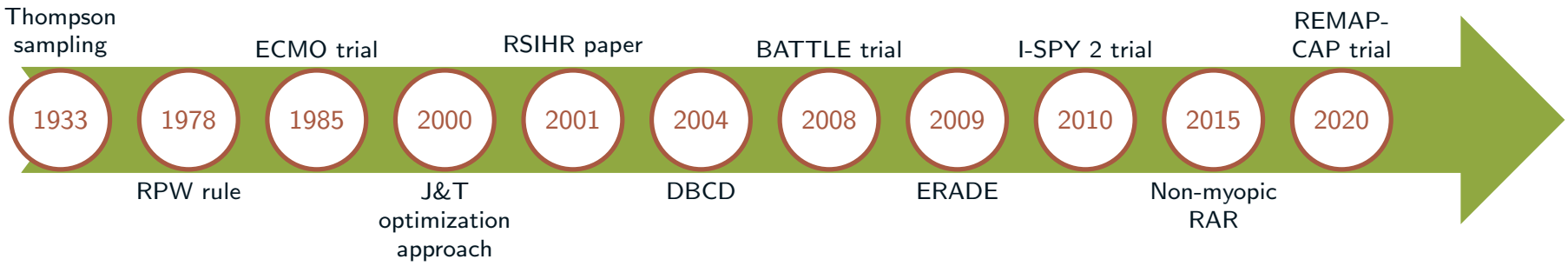


Figure: Timeline summarizing some of the key developments around the theory and practice of RAR in clinical trials. J&T = Jennison and Turnbull (2000), RSIHR = Rosenberger et al. (2001a). See Section 2 of Robertson et al (2023)

- **Imbalance I** Abundant (high quality) theoretical works paired with few highly influential examples of RAR in practice.
- **Imbalance II** Heavy focus on certain aspects (e.g., design, binary outcome), large gaps in others (e.g. analysis, non binary endpoints) .
- **Recurrence** Persistence of debate and arguments

Where to read about the debate?

VIEWPOINT

Optimizing the Trade-off Between Learning and Doing in a Pandemic

Derek C. Angus, MD, MPH
University of Pittsburgh and UPMC Health System, Pittsburgh, Pennsylvania; and Associate Editor, *JAMA*.

The world is united regarding the goal of ending the coronavirus disease 2019 (COVID-19) pandemic but not the strategy to achieve that goal. One stark example is the debate over whether to prescribe available therapies, such as quinine-based antimalarial drugs (eg, chloroquine or hydroxychloroquine), or test these drugs in randomized clinical trials (RCTs). At the heart of the problem is one of the oldest dilemmas in human organizations: the "exploitation-exploration" trade-off.[1] Exploitation refers to acting on current knowledge, habits, or beliefs despite uncertainty. This is the "just do

Three Major Challenges to Learning While Doing

The chief tool in the learning toolkit is the RCT, primarily because randomization is such a powerful mechanism for inferring causal effects. It is not perfect, and there are alternatives, but in the absence of a miracle drug that dramatically eradicates the disease, randomization will be crucial to determine what therapies work. There are, however, 3 major challenges.

Randomization is profoundly uncomfortable. Kalil has suggested that a clinician who wishes to administer chloroquine (rather than defer to randomized assignment)

Clin Infect Dis. 2020 Dec 31;71(11):3002-3004. doi: 10.1093/cid/ciaa334.

Resist the Temptation of Response-Adaptive Randomization

Michael Proschan [1], Scott Evans [2]

Affiliations

PMID: 32222766 DOI: 10.1093/cid/ciaa334

Abstract

Response-adaptive randomization (RAR) has recently gained popularity in clinical trials. The intent is noble: minimize the number of participants randomized to inferior treatments and increase the amount of information about better treatments. Unfortunately, RAR causes many problems, including

The Temptation of Overgeneralizing Response-adaptive Randomization

Sofía S Villar ✉, David S Robertson, William F Rosenberger

Clinical Infectious Diseases, ciaa1027,
https://doi.org/10.1093/cid/ciaa1027
Published: 22 July 2020 **Article history** ▾

TO THE EDITOR—We read with interest the recent article by Proschan and Evans [1] on the use of response-adaptive randomization (RAR) and its potential problems; however, these problems are neither new nor applicable in general to all

Why this review paper now?

These are some reasons why (we wanted to) write a review paper in RAR:

- To reconcile apparently conflicting arguments (particularly in medical/applied journals)
- To have an updated review (to account for more recent work)
- To (try to) classify RAR and provide a non-expert roadmap
- To guide both future uptake in practice and research directions in RAR

Outline

Some basic ideas & notation

- A study with **fixed number**[1] of patients n (fixed sample) to be randomised to arms: 0 (control) or $k = 1, \ldots, K$ (experimental).
- Some primary outcome variable $Y_k \sim f(\theta_k)$
- Allocations during the trial be $a_{k,i} = 1$ iff arm k is allocated to patient i
- $\pi_{k,i} = P(a_{k,i} = 1)$ is the probability of patient i receiving the arm k.

  Traditional (fixed and equal) randomisation is such that $\pi_{k,i} = c \ \forall i, k$, where usually $c = 1/(K+1)$.

(D) RAR defines $\pi_{k,i}$ as function of past data and actions:

$$\pi_{k,i} = P(a_{k,i} = 1 | \overline{Y_{i-1}}, \overline{a_{i-1}})$$

$\overline{Y_{i-1}}$ and $\overline{a_{i-1}}$ outcomes and allocations up to patient $i-1$ respectively. Note that RA**R** requires $\pi_{k,i} \in (0,1) \, \forall k, i$

Comparing RAR

- What are the relevant dimensions (for clinical trials)?

  For simplicity, let's do this when $K = 1$ (two-arm study) with $H_0 : p_0 = p_1$ (null) and (some alternative) $H_1 : p_0 \neq p_1$

- Many metrics can be put forward. We focused on 3 main classes.
  1. Testing metrics: type I error $\alpha = P(\textit{reject } H_0 | H_0 \textit{ true})$ and power $(1 - \beta) = P(\textit{reject } H_0 | H_1 \textit{ true})$
  2. Estimation metrics: mean bias $= E(\hat{\theta}_k) - \theta_k$, variance of estimator $= V(\hat{\theta}_k)$ or the mean squared error of an estimator $= E[(\hat{\theta}_k - \theta_k)^2]$
  3. Patient benefit metrics: the proportion of patients allocated to the best arm$= p^* = \frac{\sum_{i=1}^{n} a_{k,i}}{n}$
  4. Other metrics:: sample size (minimum n to achieve power and control type I error).

(!) Many conflicting views are explained by a focus on conflicting metrics (or simply by ignoring some of them).

Classifying RAR

- Some papers critise (or praise) the use of RAR (in general) with arguments that apply to a specific procedure (in particular)

  e.g. RAR is still heavily criticised after the Randomised Play the winner (RPTW) ECMO trials.

- RAR as broad class of adaptation, includes very different "families" of procedures

e.g. 1 *Optimal* (e.g. Rosenberger et al. (2001a)) versus *Design-driven* (e.g., RPTW, Wei and Durham (1978)) RAR

e.g. 2 *Single* (e.g., Neyman ratio) versus *Multi* objective RAR (e.g. Rosenberger et al. (2001a))

Outline

Frequently asked questions

- Does using RAR reduce statistical power?
- Does using "patient-driven" RAR lead to a substantial chance to allocate patients to an inferior treatment?
- Can RAR be used if there is potential for temporal trends?
- Is implementing RAR in practice more challenging?
- Is RAR (more) ethical?

Frequently asked questions

- Does using RAR (inevitably) reduce statistcal power?
- Does using "patient-driven" RAR lead to a substantial chance to allocate patients to an inferior treatment?
- Can RAR be used if there is potential for temporal trends?
- Is implementing RAR in practice more challenging?
- Is RAR (more) ethical?

Does RAR (inevitably) reduce power? I

- Established view: $\pi_{k,i} = 1/2$ (FER) maximises power, thus RAR must reduce power. Many publications saying this without any caveats.

(!) Y_k binary, $\theta_k = p_k \in (0,1)$ and fixed n. Two optimal allocation ratios:

$$\rho^*_{\text{Neyman}}(p_0, p_1) = \frac{\sqrt{p_1(1-p_1)}}{\sqrt{p_0(1-p_0)} + \sqrt{p_1(1-p_1)}}, \quad \rho^*_R(p_0, p_1) = \frac{\sqrt{p_1}}{\sqrt{p_0} + \sqrt{p_1}}$$

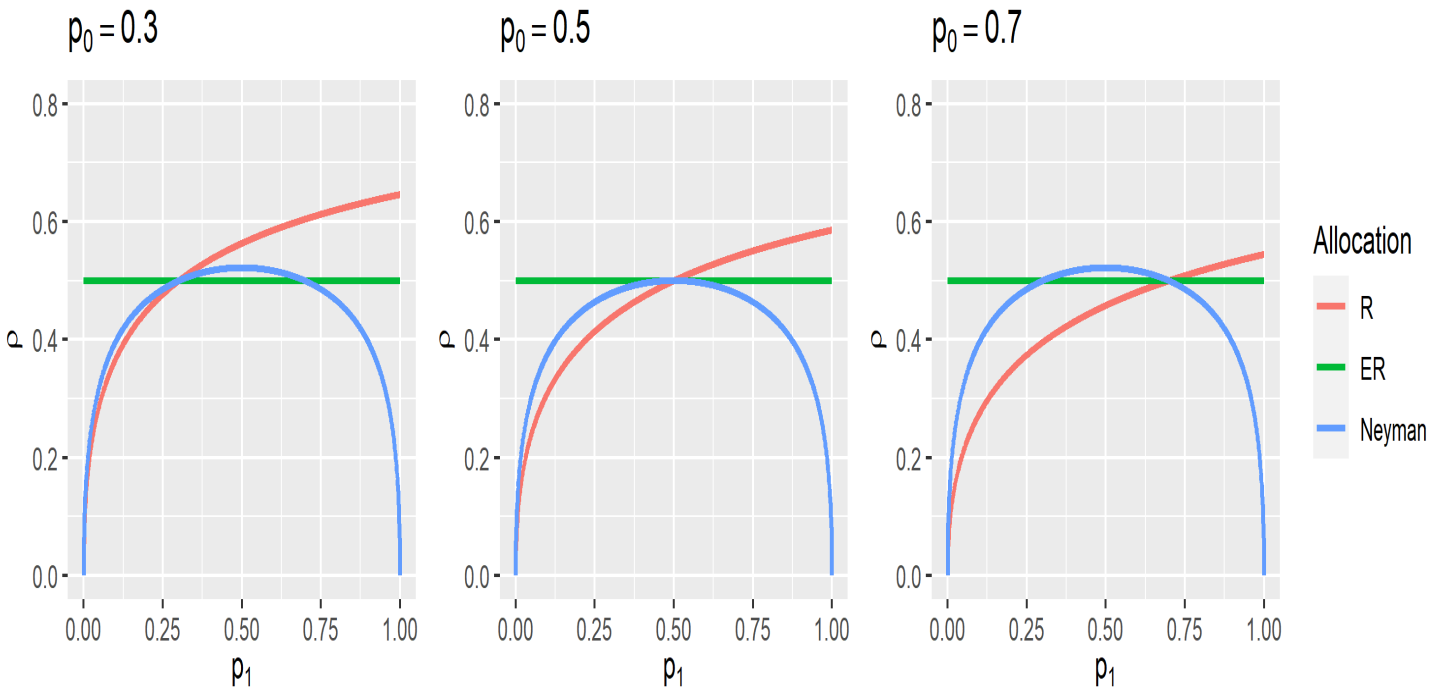


Figure: Optimal ratios ρ^*_{Neyman} and ρ^*_R as a function of p_1, for $p_0 \in \{0.3, 0.5, 0.7\}$

Optimal ratios: minimise sample size for given power

- Wald Test: $Z = \frac{\hat{p}_1 - \hat{p}_0}{\sqrt{s^2_{\Delta\hat{p}}(N_0, N_1)}}$ $s^2_{\Delta\hat{p}}(n) = \frac{\hat{p}_0(1-\hat{p}_0)}{N_0} + \frac{\hat{p}_1(1-\hat{p}_1)}{N_1}$.

Q What is **the minimum sample size** $n = N_0 + N_1$ given a power constraint (or a fixed variance level)?

Let $\rho = \frac{N_1}{N_0+N_1}$ (and $1 - \rho = \frac{N_0}{N_0+N_1}$),

$\min_\rho N_0 + N_1 \quad$ s. t. $\sigma^2_{\Delta\hat{p}}(N_0, N_1) = C$

Solution (a.k.a., Neyman allocation): ρ^*_{Neyman}

Q What is **the minimum expected number of failures in** n given a power constraint (or a fixed variance level)?

$\min_\rho (1 - p_0) N_0 + (1 - p_1) N_1 \quad$ s. t. $\sigma^2_{\Delta\hat{p}}(N_0, N_1) = C$

Solution (a.k.a., Rosenberger et al allocation): ρ^*_R

Does RAR (inevitably) reduce power? II

For a binary endpoint (reward) setting (which is the most common for RAR literature)

- Equal randomisation maximisies power only when the success rates are equal
- For low success rates, both optimal ratios differ from ER and in the same direction.
- For high success rates, the two optimal ratios will deviate in contrary directions (ethical conflict)

In other settings (other endpoints & $K > 1$), more complex considerations but in general impact on power will vary largely for different RAR.
There is a question of how achievable power differences are (or how meaningful)?[2]

Frequently asked questions

- Does using RAR reduce statistical power?
- Does using "patient-driven" RAR lead to a substantial chance to allocate patients to an inferior treatment?
- Can RAR be used if there is potential for temporal trends?
- Is implementing RAR in practice more challenging?
- Is RAR (more) ethical?

Does RAR increase chances to receive an inferior arm? I

- RAR has a substantial chance (up to 43%) of sample size imbalances in the wrong direction (i.e. towards the inferior arm) (Thall et al., 2015, 2016).

$$\hat{S}_{0.1} = \Pr[(N_0 - N_1 > 0.1n)|(p_1 > p_0)];$$

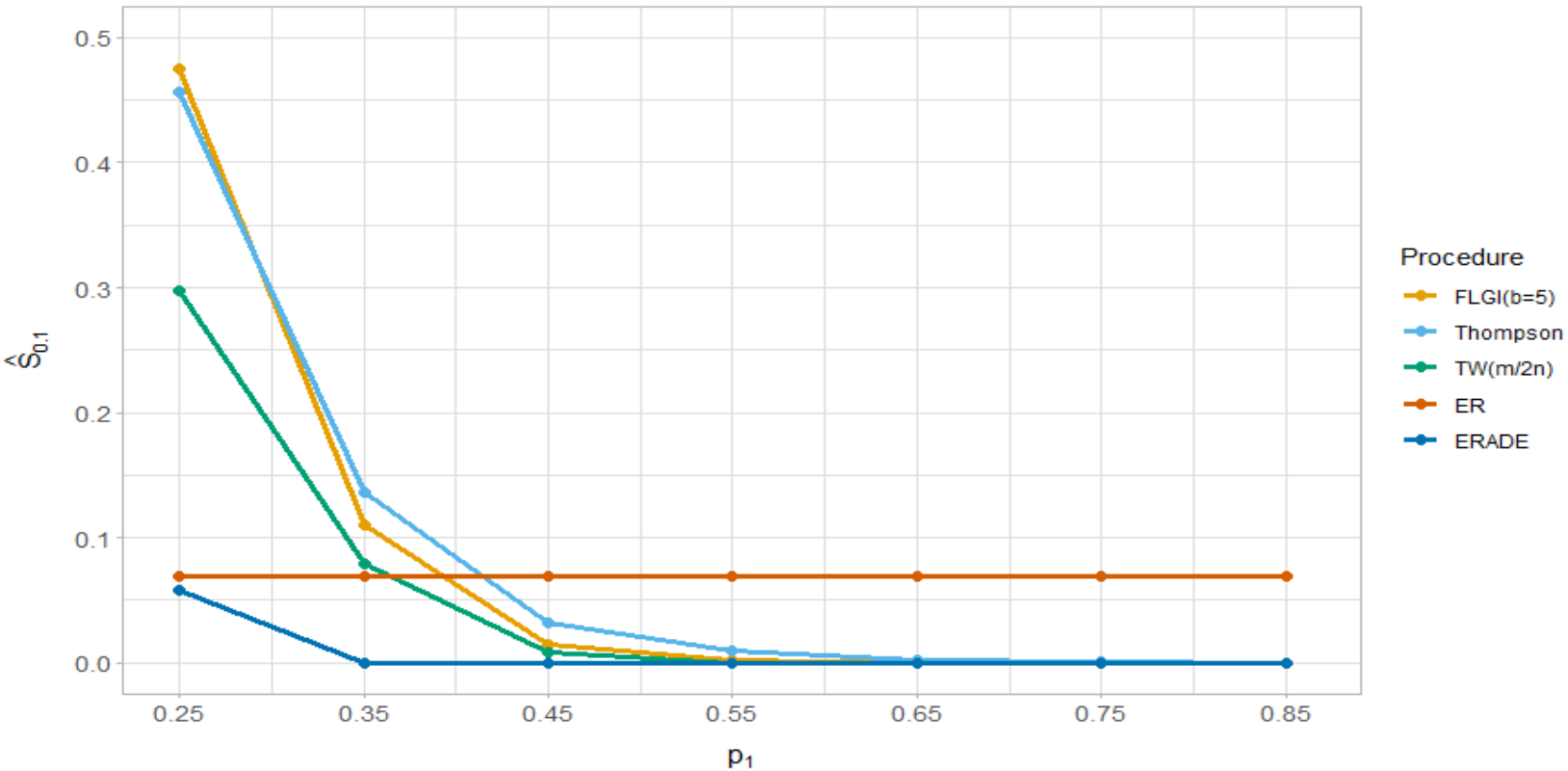


Figure: Plot of $\hat{S}_{0.1}$ for various RAR procedures as a function of p_1, where $p_0 = 0.25$ and $n = 200$. Each data point is the mean of 10^4 trial replicates.

Does RAR increase chances to receive an inferior arm? II

In a binary endpoint setting (which is the most common for RAR literature)

- "Aggressive" RAR (Thompson Sampling or Bandit rules) tend to have values of $\hat{S}_{0.1}$ larger than that of *FER* (simple randomisation)
- For lower differences in the success rates, larger values of $\hat{S}_{0.1}$ (also less real difference to patients)
- For higher differences in the success rates, particularly when n ensures power, smaller values of $\hat{S}_{0.1}$ (also larger difference to patients in terms of expected effect)

So while the reported value of 43% was correct (we replicated it) this value is very different for other RAR and more importantly, affected by expected treatment effect and sample size.

Outline

Concluding Remarks

- The pace of methodological work has sped up recently (and so has the uptake in practice).
- *Future work*: there are areas that remain under explored (mostly linked with important practical aspects), e.g. RAR for other endpoints than binary, how to best use RAR on a surrogate endpoint?, how to best deal with missing data (online), how to do efficient/robust inference in small samples? and more
- *Our own take-away*: Generalisations and broad statements of RAR (in terms of relevant metrics) hardly ever true. Trade-offs for Adaptive Designs (including those with RAR) are ubiquitous, best strategy is to be aware of them.
- Several ways to design and implement RAR and the setting should guide both the decision to use it or not and the choice of which one.

A broad look at RAR

Response-adaptive randomization in clinical trials: from myths to practical considerations	David S Robertson, Kim May Lee, Boryana C Lopez-Kolkovska, and Sofia S Villar
Comment: A Quarter Century of Methodological Research in Response-Adaptive Randomization	Anastasia Ivanova and William F Rosenberger
Discussion of "Response-adaptive randomization in clinical trials: from myths to practical considerations"	Yunshan Duan, Peter Mueller, and Yuan Ji
Advancing Clinical Trials with Novel Designs and Implementations	Lorenzo Trippa and Yanxun Xu
Group sequential designs with response-adaptive randomisation	Christopher Jennison
Is Response-Adaptive Randomization a "good thing" or not in clinical trials? Why we cannot take sides	Alessandra Giovagnoli
Response Adaptive Randomization in Practice: A Discussion of Robertson et al.	Scott M Berry and Kert Viele

Submitted to Statistical Science

Rejoinder: Response-adaptive randomization in clinical trials

David S. Robertson, Boryana C. López-Kolkovska, Kim May Lee and Sofía S. Villar

Rejoinder (some points for discussion)

- How does a fixed n assumption: RAR as single adaptation versus RAR as one of many play here?

1 RAR with no early stopping $N_0 = 153.0, N_1 = 75.7$ so $N = 228.7$, $(\mu_0 - \mu_1 = \delta > 0)$ (Table 1) whereas RAR with early stopping $N_0 = 110.3, N_1 = 57.3$ $N = 167.6$ (Table 3).
As a baseline a non-adaptive ER trial would have $N_0 = 100.4, N_1 = 100$ so $N = 200$. Jennision's commentary.

- What other flexibility can RAR offer? RAR as approximations to optimal designs - Duan, Muller & Ji
- Examples in practice illustrate the above principle - Berry & Viele.
- Key role to good (RAR) design: Objective, metrics, simulations and data sharing (all discussants).

2 RAR & Power: Ivanova & Rosenberger $\theta = \log\left(\frac{p_0(1-q_1)}{(1-q_0)p_1}\right)$ with $p_0 = 0.941, p_1 = 0.991$ results in $\rho^*_{\text{Neyman}} = 0.72$ (power gain only achieved by one procedure to target it) and $\rho^*_{\text{R}} = 0.87$ (equal power to ER but 40% less failures).

Questions?

Thank you for listening! sofia.villar@mrc-bsu.cam.ac.uk

Questions?

References

Thompson, W.R. (1933). On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika* **25** 285–294.

Jennison, C. and Turnbull, B.W. (2000). *Group Sequential Methods with Applications to Clinical Trials*. Chapman and Hall/CRC Press, Boca Raton.

Wei, L.J. and Durham, S. (1978). The Randomized Play-the-winner Rule in Medical Trials. *Journal of Medical Statistics Association* **73** 840–843.

Rosenberger, W.F., Stallard, N., Ivanova, A., Harper, C.N. and Ricks, M.L. (2001a). Optimal Adaptive Designs for Binary Response Trials. *Biometrics* **57** 909–913.

Thall, P.F., Fox P. and Wathen, J. (2015). Statistical Controversies in Clinical Research: Scientific and Ethical Problems with Adaptive Randomization in Comparative Clinical Trials. *Annals of Oncology* **26** 1621–1628.

Thall, P.F., Fox, P.S. and Wathen, J.K. (2016). Some Caveats for Outcome Adaptive Randomization in Clinical Trials. In *Modern Adaptive Randomized Clinical Trials. Statistical and Practical Aspects*, edited by Sverdlov, O. Chapman and Hall/CRC Press, Boca Raton.

Proschan, M. and Evans, S. (2020). Resist the Temptation of Response-Adaptive Randomization. *Clinical Infectious Diseases*. **71(11)** 3002–3004.

Villar, S.S., Robertson, D.S. and Rosenberger, W.F. (2021). The Temptation of Overgeneralizing Response-Adaptive Randomization. *Clinical Infectious Diseases*, **73(1)** e842.

Robertson, D.S.,López-Kolkovska B., Lee K. and Villar S.S., (2022).Response-adaptive randomization in clinical trials: from myths to practical considerations. *Statistical Science (To appear)*,

Clinical considerations of RAR in trials

Will Meurer
University of Michigan

Disclosures

Twitter: @goalsofhair @willmeurer

insta: @goals.of.hair

- Consulting fees for medicolegal cases / Berry Consultants (various commercial clients – I am not on any projects related to cooling)
- Funding from NINDS
  - -PI of clinical trials methodology course
  - Funded co-investigator on NETT/SIREN-CCC and BOOST trial
- NIH-NIDCD (cluster RCT studying dizziness intervention)
- NIH-NIMHD - PI of trial of hypertension
  - Co-investigator to improve stroke care in Flint
- NIH-NHLBI (ICECAP PI and P-ICECAP PI/ also co-investigator in cardiac arrest expedited transport/ECMO trial)

- AHRQ (prehospital stroke care / dizziness self management)
- Massey Foundation (studying pupillary response in TBI) - past
- FDA and NIH (reviewer)
- Meth Stats Reviewer
  - Annals Of Emergency Medicine
  - Academic Emergency Medicine
- Senior Editor (paid)
  - BMC Journal - Trials
- PCORI (co-I Kidney Stone Trial)

Objectives

-Consider RAR in real world trials

-Consider patient perspective

-Consider practical aspects

Four parts

1. Why do people volunteer for clinical trials?
2. Does RAR increase or decrease power / # of patients with a good outcome?
3. Could RAR improve recruitment in an emergency trial?
4. What do we say to people in real-life RAR trials?

1. What do people think they are getting out of research, anyway?

VIEWPOINT

Adaptive Clinical Trials

A Partial Remedy for the Therapeutic Misconception?

William J. Meurer, MD, MS
Roger J. Lewis, MD, PhD
Donald A. Berry, PhD

THERE IS A COMMON "THERAPEUTIC MISCONCEPTION" among patients considering participation in clinical trials.[1] Some trial participants and family members believe that the goal of a clinical trial is to improve their outcomes—a misperception often reinforced by media advertising of clinical research.[2] Clinical trials have

Although knowledge regarding the relative effectiveness of the treatments involved accumulates over the course of a clinical trial, beginning with a state of equipoise and having high confidence near the end, fixed assignment ensures that this information is ignored. The result is that a fixed proportion of patients will receive potentially inferior therapy—whichever therapy that turns out to be—assuming there are differences in efficacy of the treatments in the trial. The primary scientific goal of a clinical trial should not be compromised, but interim information available in a trial could be used to improve the outcomes of trial par-

JAMA 2012 Jun 13;307(22):2377-8. doi: 10.1001/jama.2012.4174.

2. RAR can increase OR decrease power (and number of favorable outcomes within study)

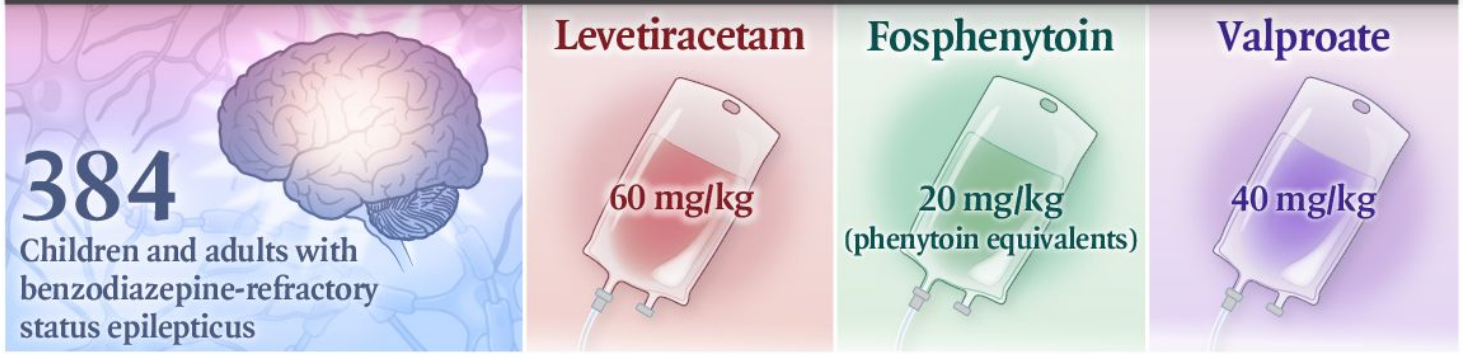
Scenario	Adaptive Randomization			Fixed Randomization		
	Power	Mean N	% to Best	Power	Mean N	% to Best
Null 0.50 – 0.50 – 0.50	0.031	507	N/A	0.029	499	N/A
One Good 0.50 – 0.50 – 0.65	0.90	483	48	0.88	497	33
Two Good 0.50 – 0.65 – 0.65	0.76	679	84	0.86	687	67
One Middle One Good 0.50 – 0.575 – 0.65	0.68	586	47	0.69	599	33
All Bad 0.25 – 0.25 – 0.25	0.044	524	N/A	0.030	509	N/A
All Really Bad 0.10 – 0.10 – 0.10	0.006	400	N/A	0.028	400	N/A

J Clin Epidemiol. 2013 Aug; 66(8 0): S130–S137.

The NEW ENGLAND JOURNAL of MEDICINE

Trial of Three Anticonvulsant Medications for Status Epilepticus

MULTICENTER, RANDOMIZED, DOUBLE-BLIND TRIAL



No significant difference in rates of seizure cessation or in safety

J. Kapur et al. 10.1056/NEJMoa1905795

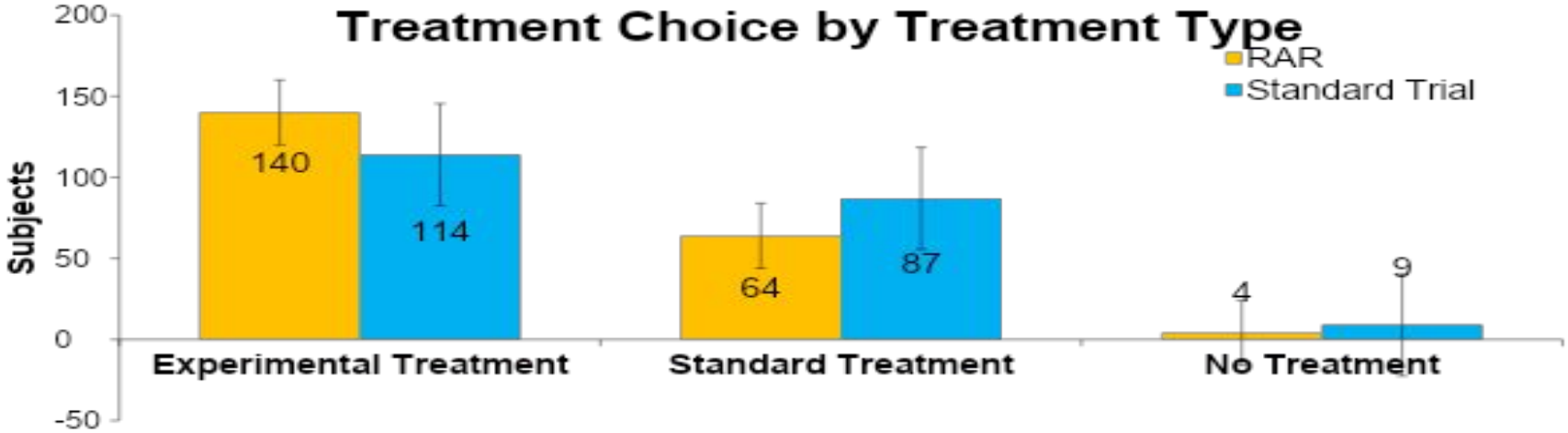
3. Do more people agree to an emergency trial with RAR, hypothetically?

Methods

- Cross-sectional survey of adult patients in the University of Michigan Hospital Emergency Department.
  - Patients were over 18 years old, were presenting without stroke, without altered mental status, and not in the trauma bay.
- Subjects were randomized and shown a video describing a scenario of either an RAR or ”standard” hypothetical acute stroke trial.
- Primary outcome measure was participation in the hypothetical research trial.
  - Also assessed understanding and recall of the assigned trial design, and evaluated the adequacy of the simulated informed consent process (ICQ-4)

Main Results

- **Primary Outcome:** 67.3% research participation of the RAR group versus 54.3% research participation in the standard group
  - Absolute difference of 13% (95% CI: 3.7 to 22.1 %).



Results – continued

	Standard (n = 210)	RAR (n=208)
Recall of how XPA/tPA was determined		
- Computer coin flip	85%	5%
- Study algorithm (RAR)	3%	62%
- Doctor's decision	7%	13%
- Don't know/Don't remember	5%	19%
Correctly Identified allocation method		
- Yes	85%	62%

- As a whole, the RAR trial was more difficult for subjects to understand.
- RAR subjects correctly identified the allocation method 62.4% of the time compared to 85.3% of the standard.

Main Results – continued

- RAR group subjects had **1.89 odds*** of agreeing to participate in the hypothetical acute stroke trial compared to subjects in the standard research group (O.R. 1.89, 95% CI: [1.2 – 2.9]).
  - *Adjusted for age, gender, race/ethnicity, education, self-reported understanding of protocol, correct identification of allocation technique, stroke risk factor awareness. (None of these covariates were statistically significant)

Stroke. 2014;45:2131–2133

Follow up study!

https://www.biorxiv.org/content/10.1101/091819v1

Optimizing Communication of Emergency Response Adaptive Randomization Clinical Trials to Potential Participants

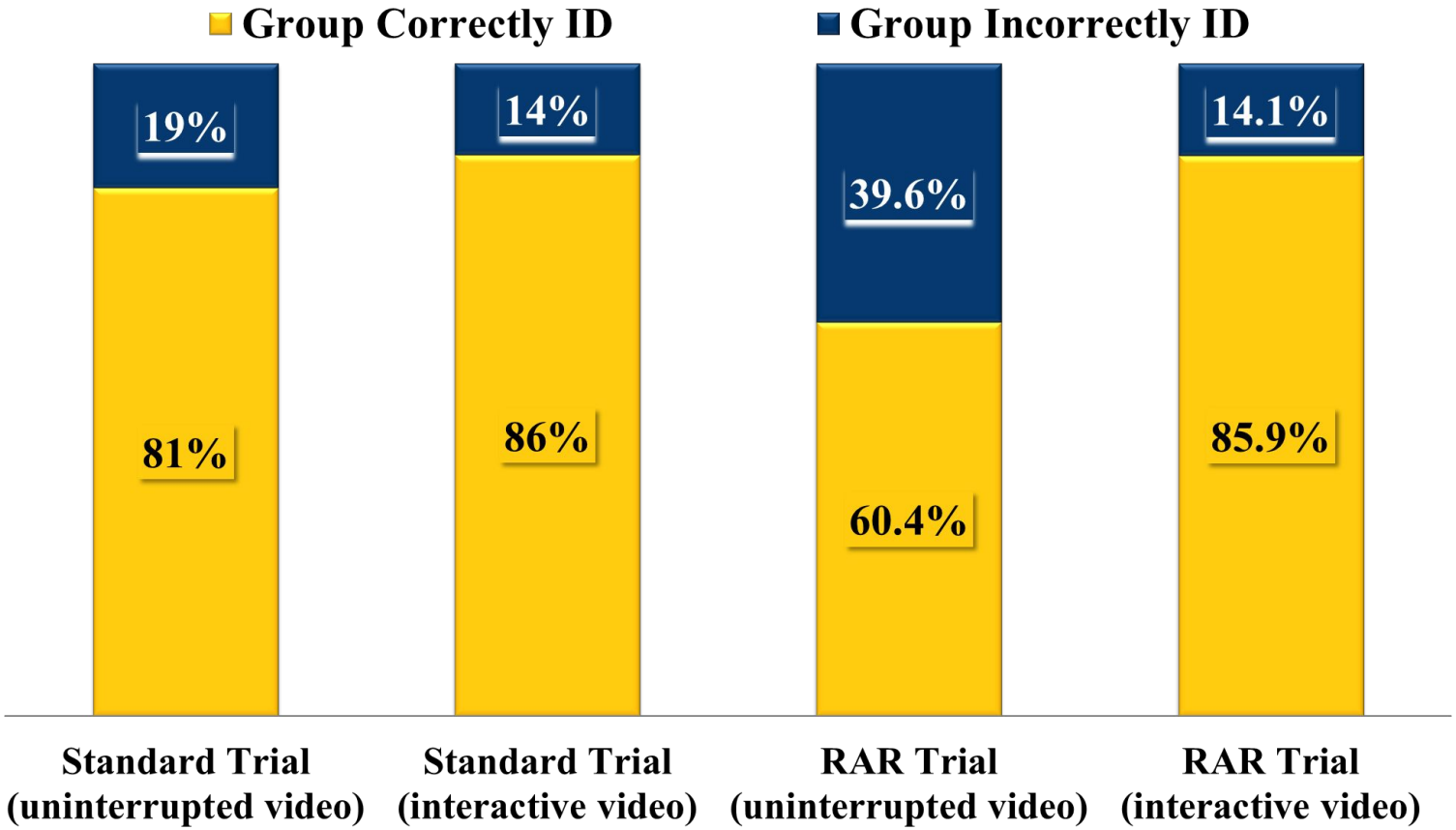
Bredan McEvoy, David Haidar, Jason Tehranisa,

View ORCID Profile

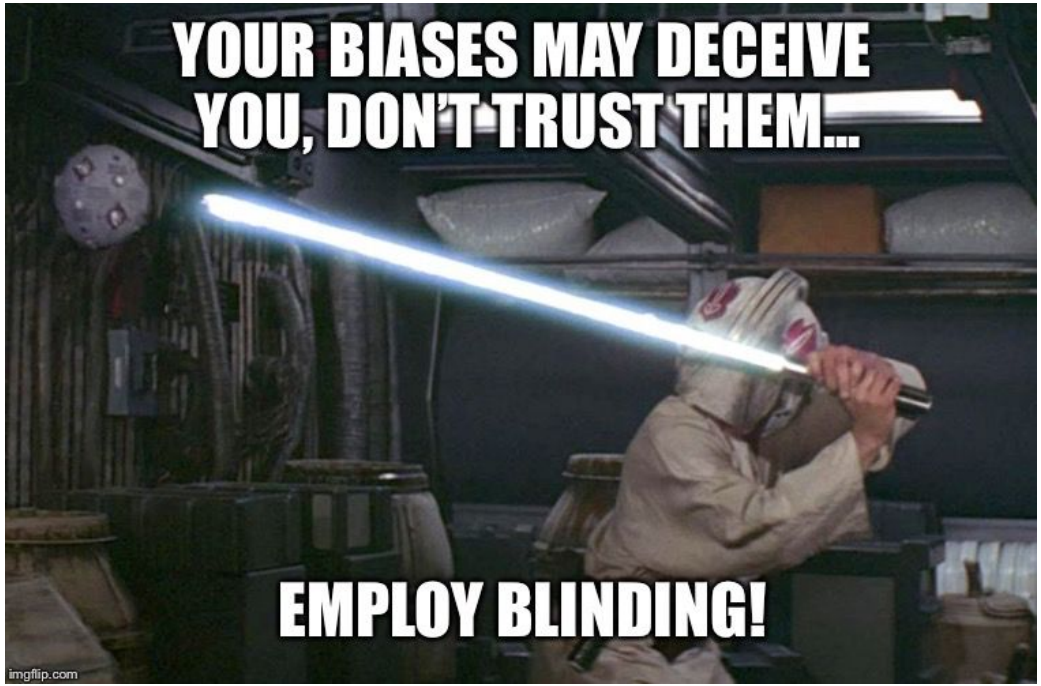
William J. Meurer

doi: https://doi.org/10.1101/091819

	Standard uninterrupted video N=236	Standard interactive video N=50	RAR uninterrupted video N=285	RAR interactive video N=149
Agreed to Trial	129 (54.3%)	31 (62%)	192 (67.3%)	124 (83.2%)
95% CI for Agree	48% to 61%	48% to 74%	62% to 73%	76% to 88%
Refused Trial	107 (45.7%)	19 (38%)	93 (33.7%)	25 (16.8%)


Group Correctly ID
Group Incorrectly ID
19%
81%
14%
86%
39.6%
60.4%
14.1%
85.9%
Standard Trial (uninterrupted video)
Standard Trial (interactive video)
RAR Trial (uninterrupted video)
RAR Trial (interactive video)

4. What do we say, these days?



SIREN Clinical Trials Network

4 Ongoing Trials (Multicenter

3 Use RAR (Adult and Peds Hypothermia Cardiac Arrest Trials, Hyperbaric Oxygen in TBI trial)

Max N: 1800, 900, 200

Adult Consent Form ICECAP

Key information - “Durations are assigned **mostly** by chance, using a computer program …”

Adult Consent Form ICECAP

More information later - “The duration each participant receives is determined by a computer program. The duration is picked mostly at random…As the study progresses patients are more likely to be randomized to durations of cooling where patients are doing better.”

More information www.icecaptrial.org

Pediatric ICECAP Consent Form

Key Information: “The length of cooling for each child is chosen by a computer mostly by chance, like flipping a coin.”

Pediatric ICECAP Consent Later Section

“In this study, the length of cooling is assigned by chance, like by rolling dice. The first 150 children in the study have the same chance to be cooled for 24, 48 or 72 hours. After 150 children are in the study, children may be assigned by chance to 0, 12, 18, 24, 36, 48, 60, 72, 84 or 96 hours of cooling.”

HOBIT

ICF Language doesn’t mention RAR

(Smaller Phase II trial)

Summary

- Therapeutic misconception is real
- RAR can have different impact depending on situation (power might increase power might decrease)
- In certain circumstances, patients may prefer a trial design that “has their back” on average
- In emergency trials, the subtle point of a moving randomization ratio is not likely to be most important detail

On to discussant!

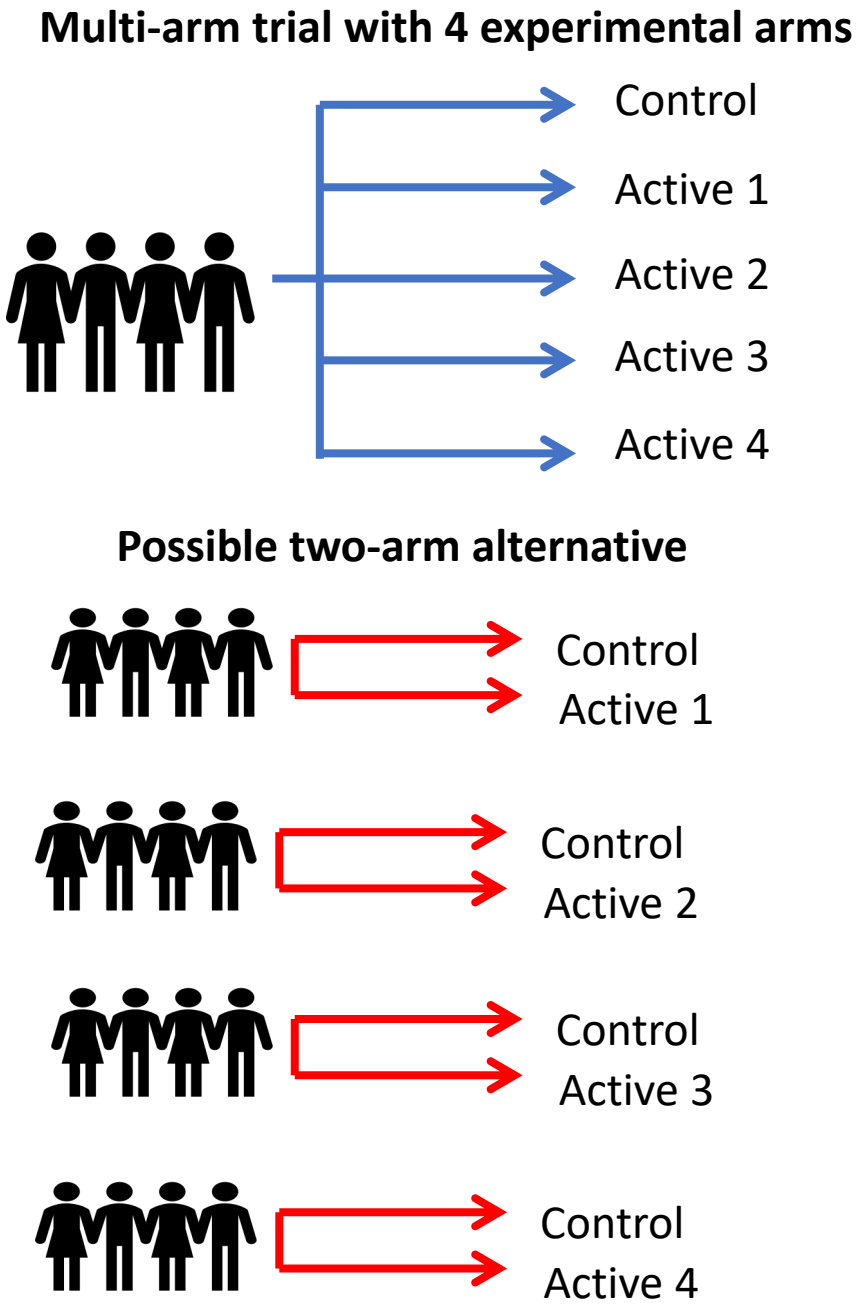
Allocation in Multi-Arm Trials

Lindsay Berry

SCT 2023

Multi-arm clinical trials

- Dose-finding studies, adaptive platform trials, pandemic/hard-to-treat settings
- Efficiency from shared control
- Head-to-head comparison
- Many possible goals



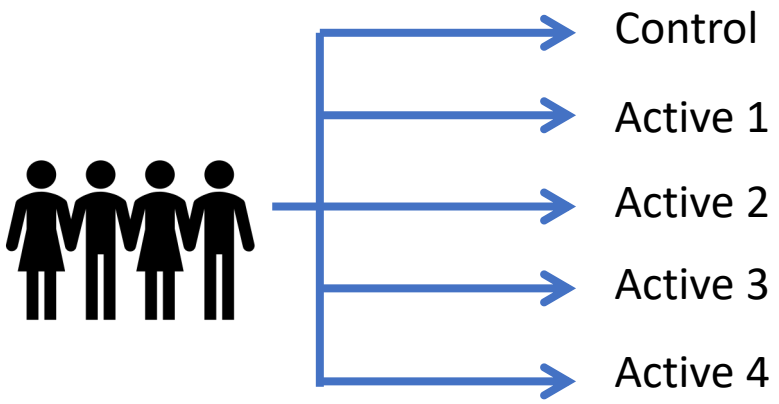
Multi-arm clinical trials

Our Primary Goals:

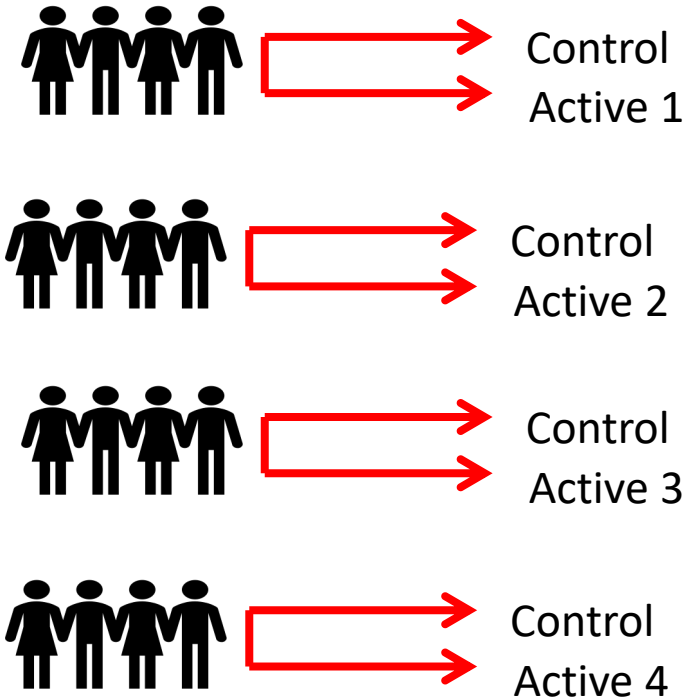
- Identify **best arm** and demonstrate **superiority to control**
- Accurately estimate treatment effect of best arm

Less interest in estimates/tests of non-optimal arms

Multi-arm trial with 4 experimental arms

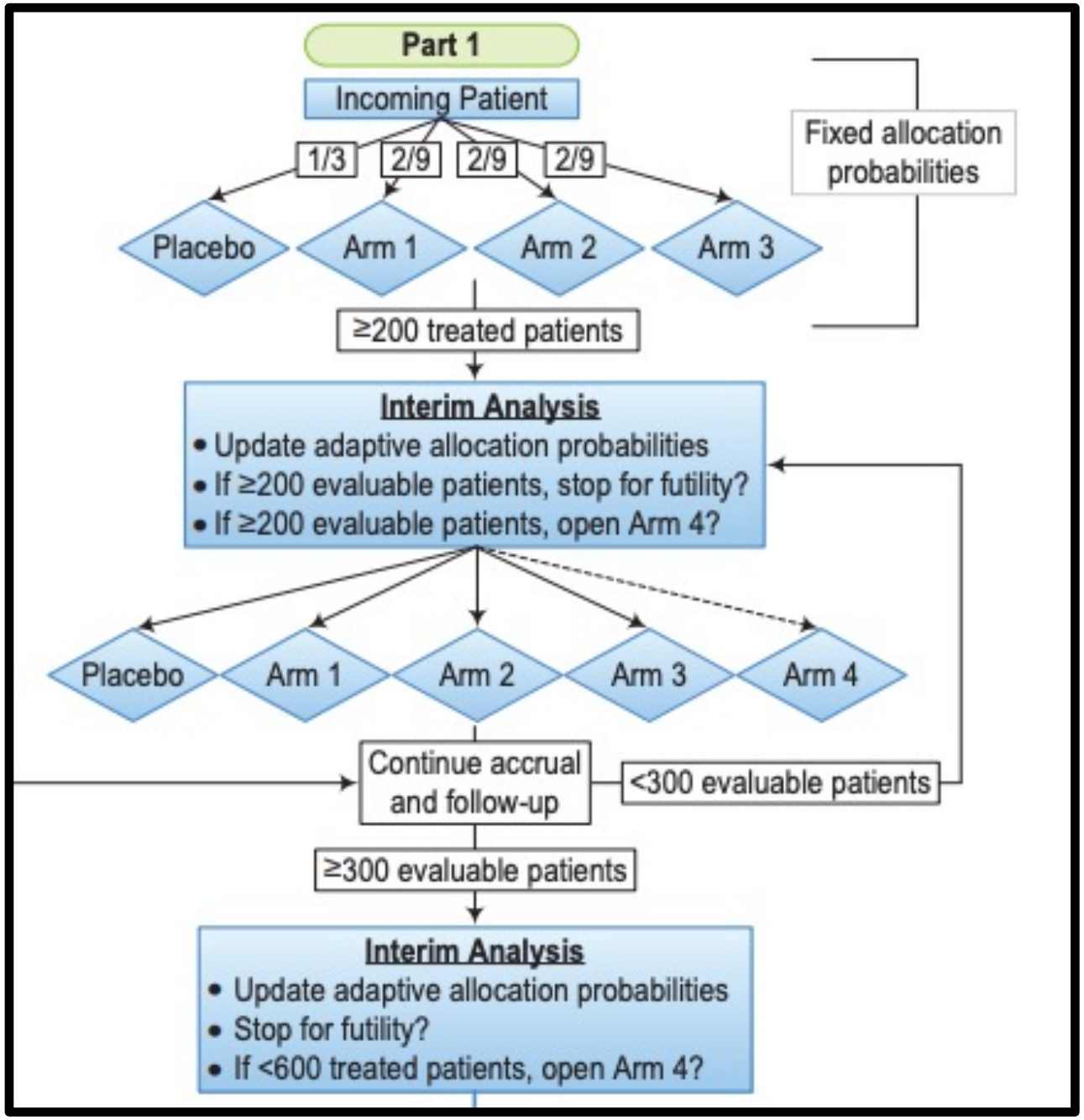


Possible two-arm alternative



SEPSIS-ACT

- Patients with septic shock requiring vasopressors
- 4 dosing regimens of selepressin and placebo
- Part 1: Dose finding with RAR
  - "**preferentially allocate** patients to the dosing regimens that appear to have the maximum benefit"
  - "optimise the efficiency of **selecting the optimal** dosing regimen"
- Part 2:
  - 1:1 comparison of best arm to placebo



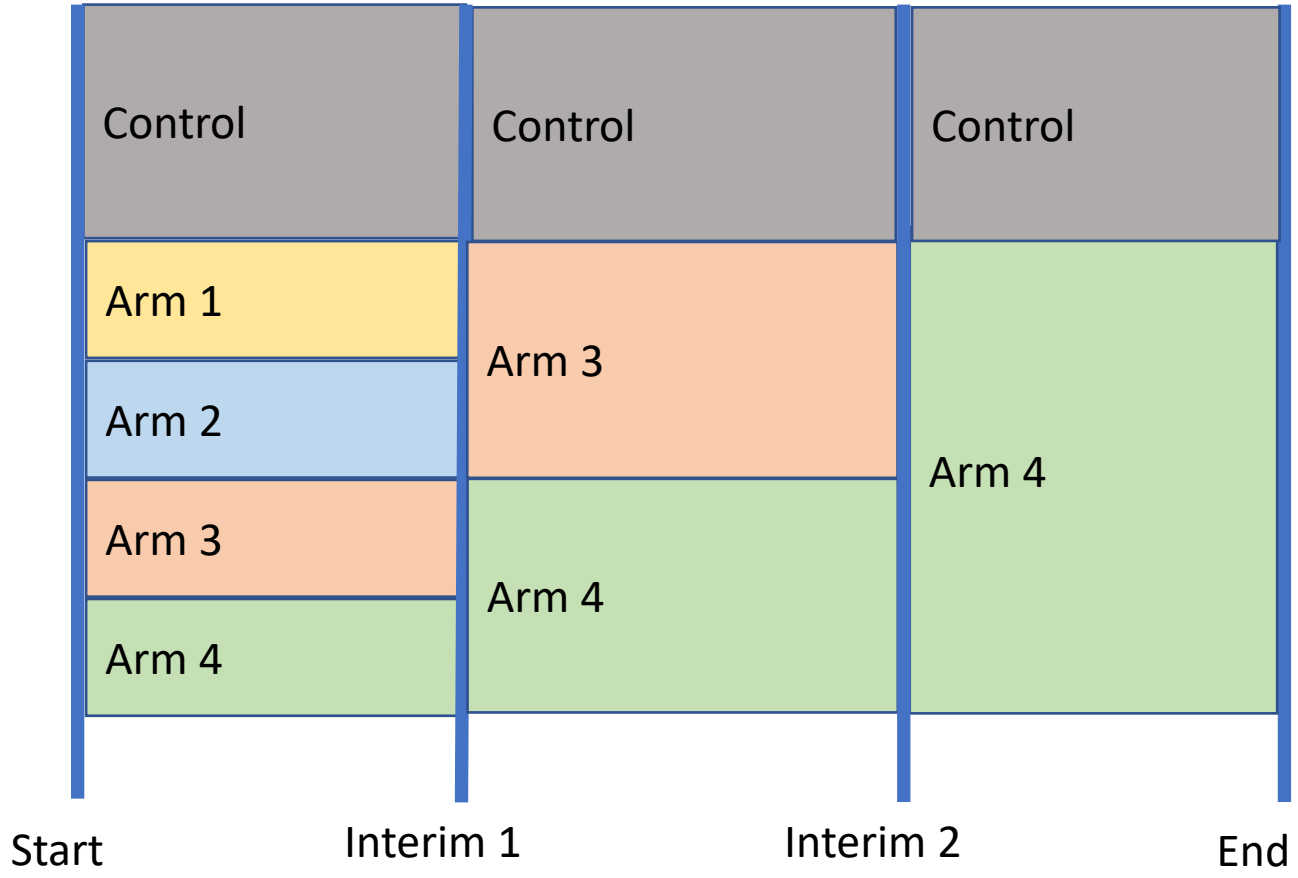
[*JAMA.* 2019;322(15):1476-1485. doi:10.1001/jama.2019.14607]

How to allocate?

- Intuitively appealing to “update” allocation as we gather data
  - **Stop allocating** to non-optimal arms
  - Prioritize potentially optimal arms
- Adaptive allocation (arm dropping/response adaptive allocation)
  - May improve performance compared to fixed allocation
- What type of allocation is best suited for our goals?

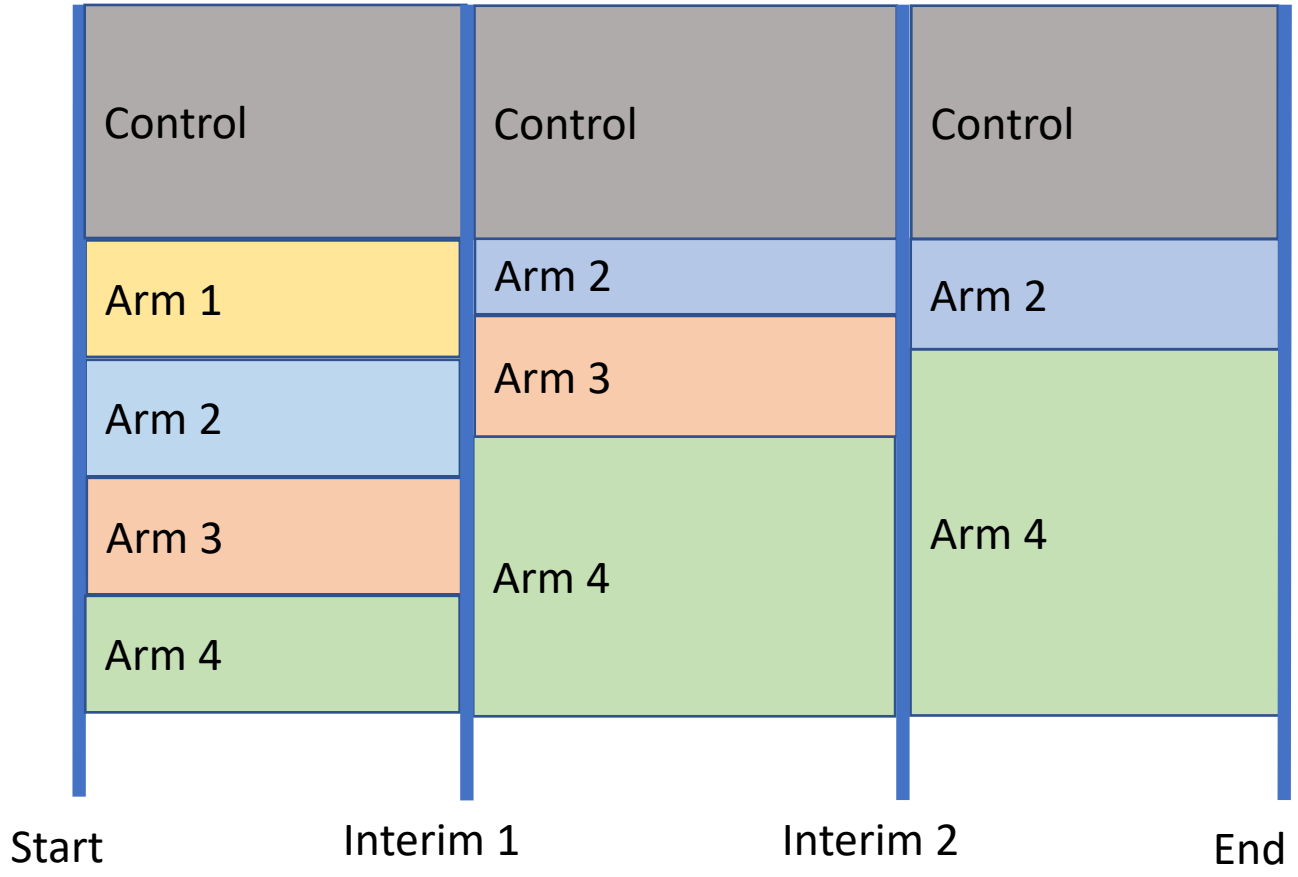
Multi-arm allocation strategies

- Arm dropping
  - Can permanently stop randomization to experimental arms
  - Often based on a p-value comparison of each arm to control
  - Thresholds pre-defined for dropping arms by interim



Multi-arm allocation strategies

- Response adaptive randomization (RAR)
  - Allocation proportions updated based on interim data
  - Proportions often driven by Bayesian analysis model (ex: posterior probability each arm is best)
  - Many versions of RAR exist – should be tailored to trial goals



Response Adaptive Randomization

- Controversial in two-arm settings (decreased power, other risks)
- Nuanced in multiple arm settings. Viele 2020ab examined
  - Control allocation
  - Allocation driver (Pr(beat control) vs Pr(best arm))
  - Interim frequency, burn-in length, thresholding to 0
- Conclusions
  - Some RAR variants perform quite poorly (Thall 2015)
  - Other variants perform far better (particularly fixed control allocation)
    - RAR outperforms fixed randomization (Trippa et al 2012)

RAR or Arm Dropping?

- Limited literature comparisons of “best RAR” to arm dropping
- The “best” arm dropping is unclear
- We compare the best RAR in Viele 2020ab to Arm Dropping
- Viele 2020ab did not consider time trends
  - Risk of “non-comparable” treatment groups if allocation changes?
  - We evaluate the performance of allocation methods when there are additive time trends

Overview

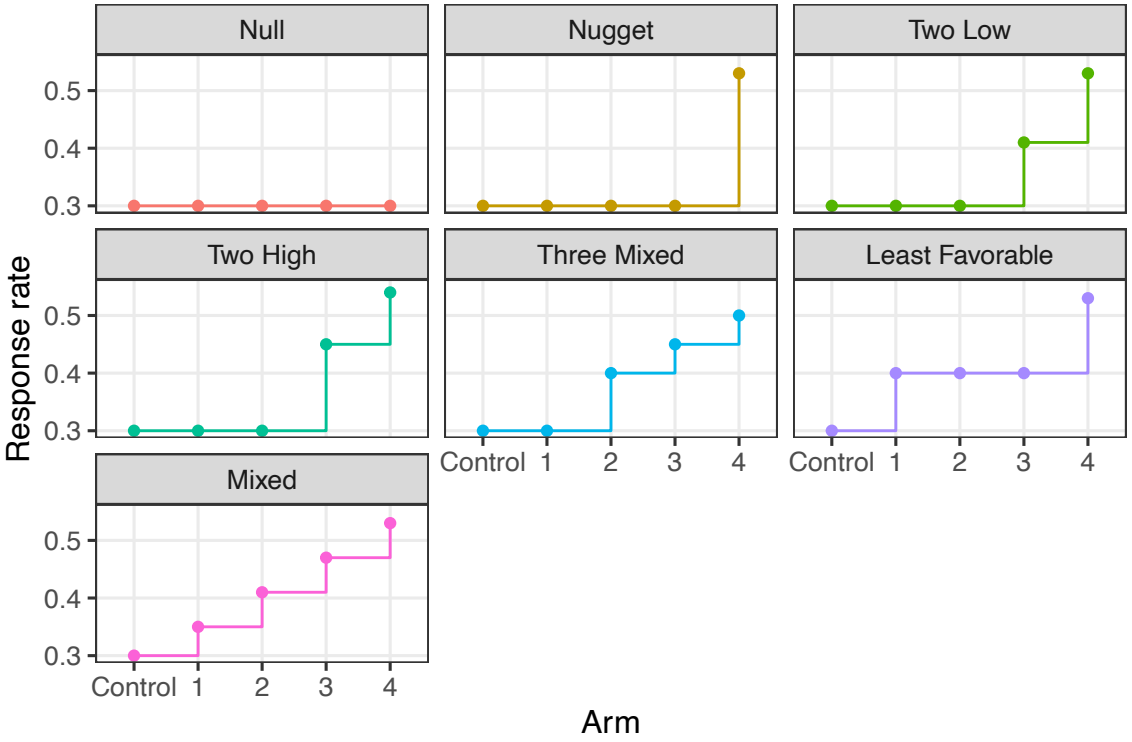
- Sample size of N=240
- 4 experimental and 1 control arm
- Dichotomous responder endpoint
- Common features:
  - Burn-in of 48
  - Interims every 24
  - Control allocation fixed to 1/3 for all designs
- **Cherry-picking** of best arm affects all designs and causes bias
  - One-sided type I error rate fixed to 2.5% for all designs

Allocation strategies

- Fixed allocation (2:1:1:1:1)
- Arm dropping
  - **AD-PBO**: Based on p-value comparison with control
    - Drop arm if p-value falls above pre-specified threshold
  - **AD-MAX**: Based on Pr(best active arm)
    - Drop arm if Pr(best active arm) falls below 10%
  - Equal allocation of non-dropped arms between interims
- **RAR**
  - Optimal method from Viele et al (2020)
  - Allocation proportional to Pr(best active arm)
  - Threshold to zero if Pr(best) < 12.5%

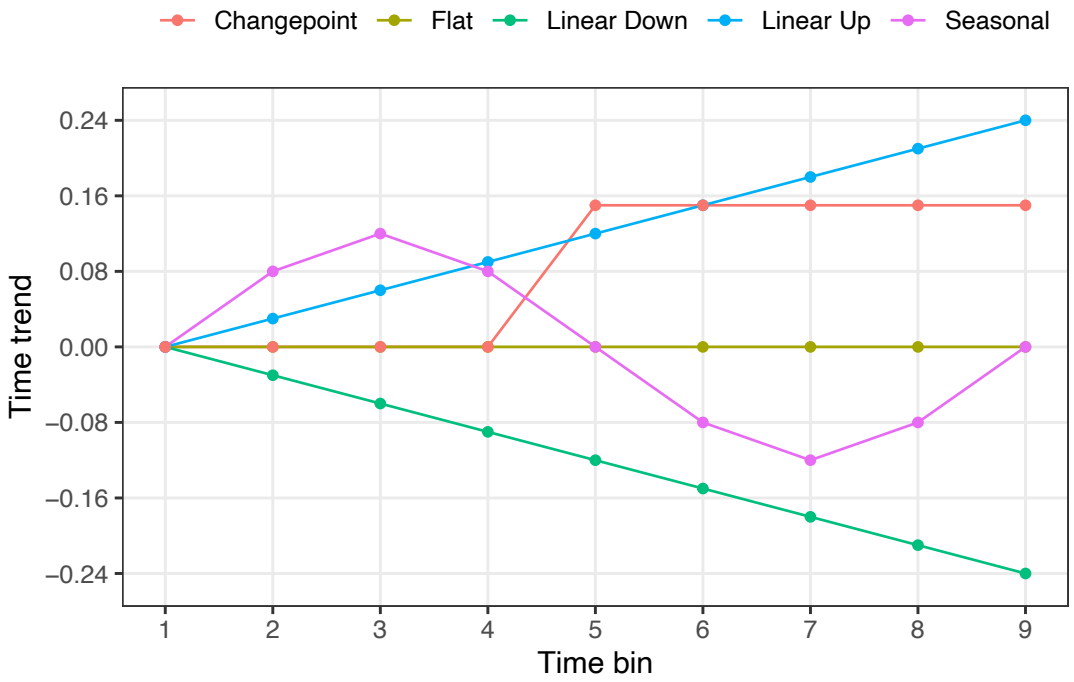
Grid of 35 simulation scenarios

7 Efficacy scenarios:



Scenario	Control	Arm 1	Arm 2	Arm 3	Arm 4
Null	0.30	0.30	0.30	0.30	0.30
Nugget	0.30	0.30	0.30	0.30	**0.53**
Two Low	0.30	0.30	0.30	**0.41**	**0.53**
Two High	0.30	0.30	0.30	**0.45**	**0.54**
Three Mixed	0.30	0.30	**0.40**	**0.45**	**0.50**
Least Favorable	0.30	**0.40**	**0.40**	**0.40**	**0.53**
Mixed	0.30	**0.35**	**0.41**	**0.47**	**0.53**

5 Additive time trend scenarios:



	Time trend bins								
	1	2	3	4	5	6	7	8	9
Flat	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Linear Up	0.00	0.03	0.06	0.09	0.12	0.15	0.18	0.21	0.24
Linear Down	0.00	-0.03	-0.06	-0.09	-0.12	-0.15	-0.18	-0.21	-0.24
Seasonal	0.00	0.08	0.12	0.08	0.00	-0.08	-0.12	-0.08	0.00
Changepoint	0.00	0.00	0.00	0.00	0.00	0.15	0.15	0.15	0.15

Final analysis model

- Bayesian logistic regression model with weakly informative priors
- Compare model with and without adjustment for time

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \mathbf{x}_i'\boldsymbol{\theta} \qquad (1)$$

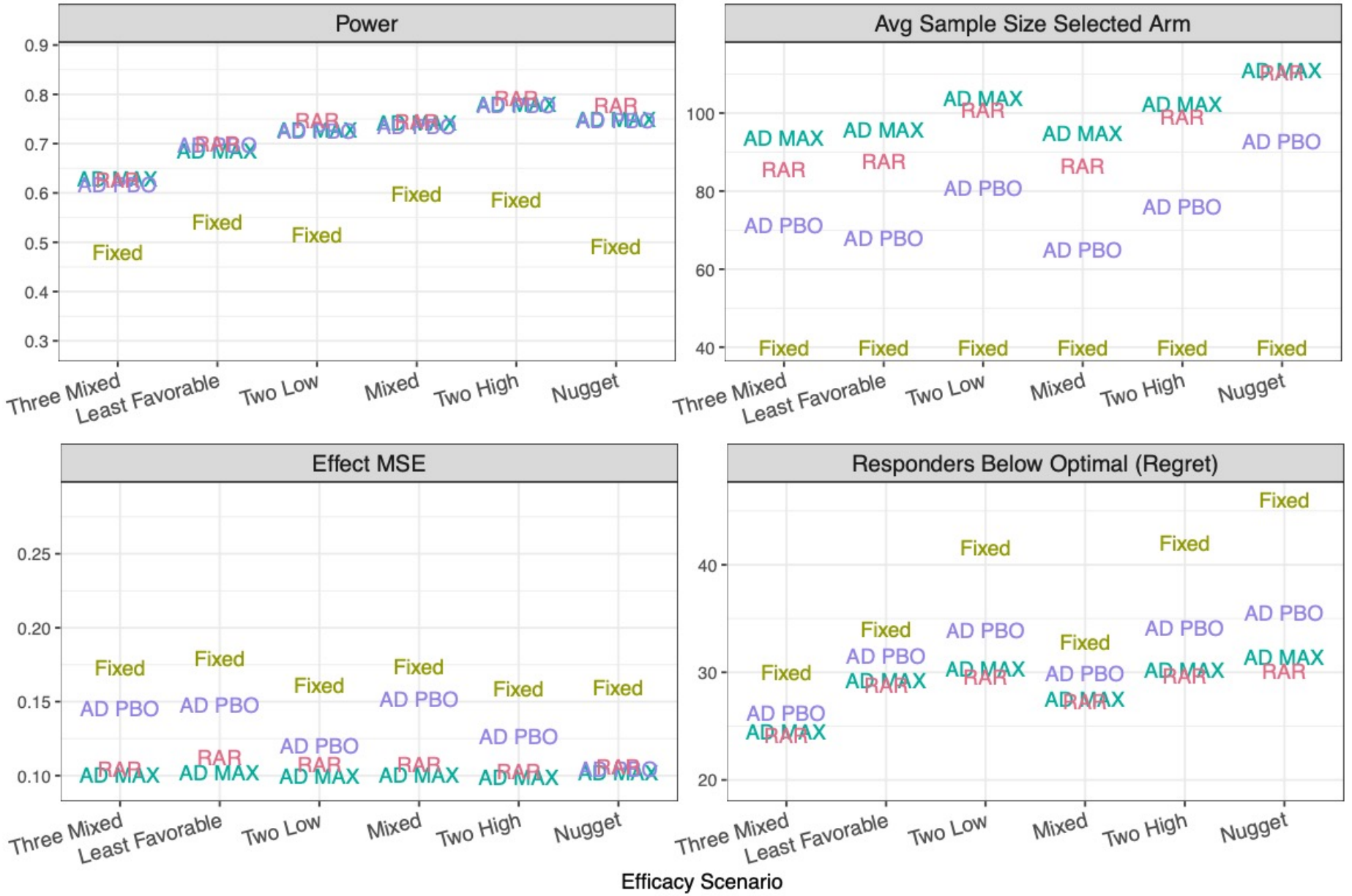
$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \mathbf{x}_i'\boldsymbol{\theta} + \mathbf{t}_i'\boldsymbol{\beta} \qquad (2)$$

- Success criteria:

$$\Pr(\theta_{max} > 0) > \delta_{design}$$

each δ_{design} achieves 2.5% type I error in the joint null

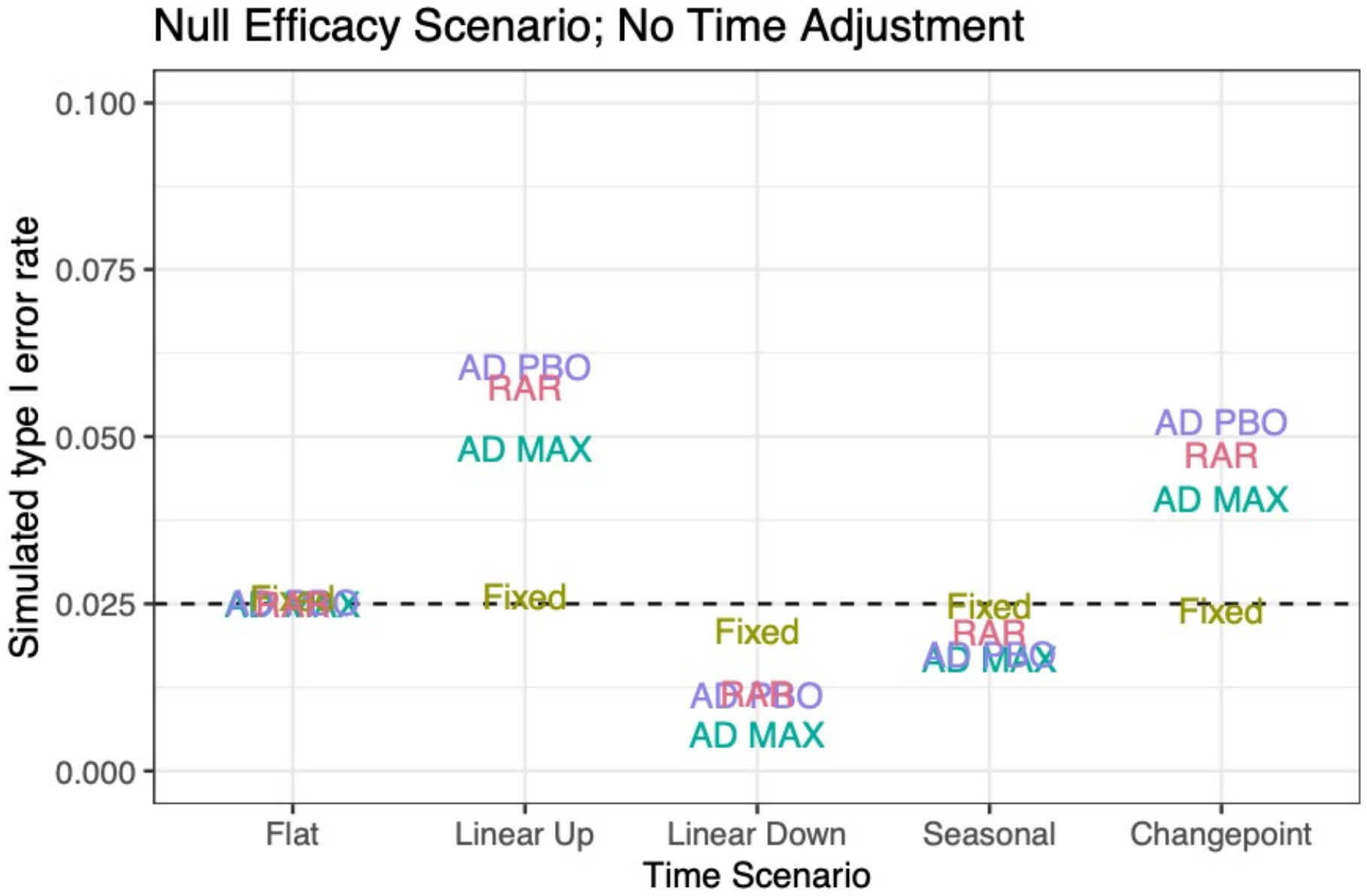
Flat time trend, No time adjustment



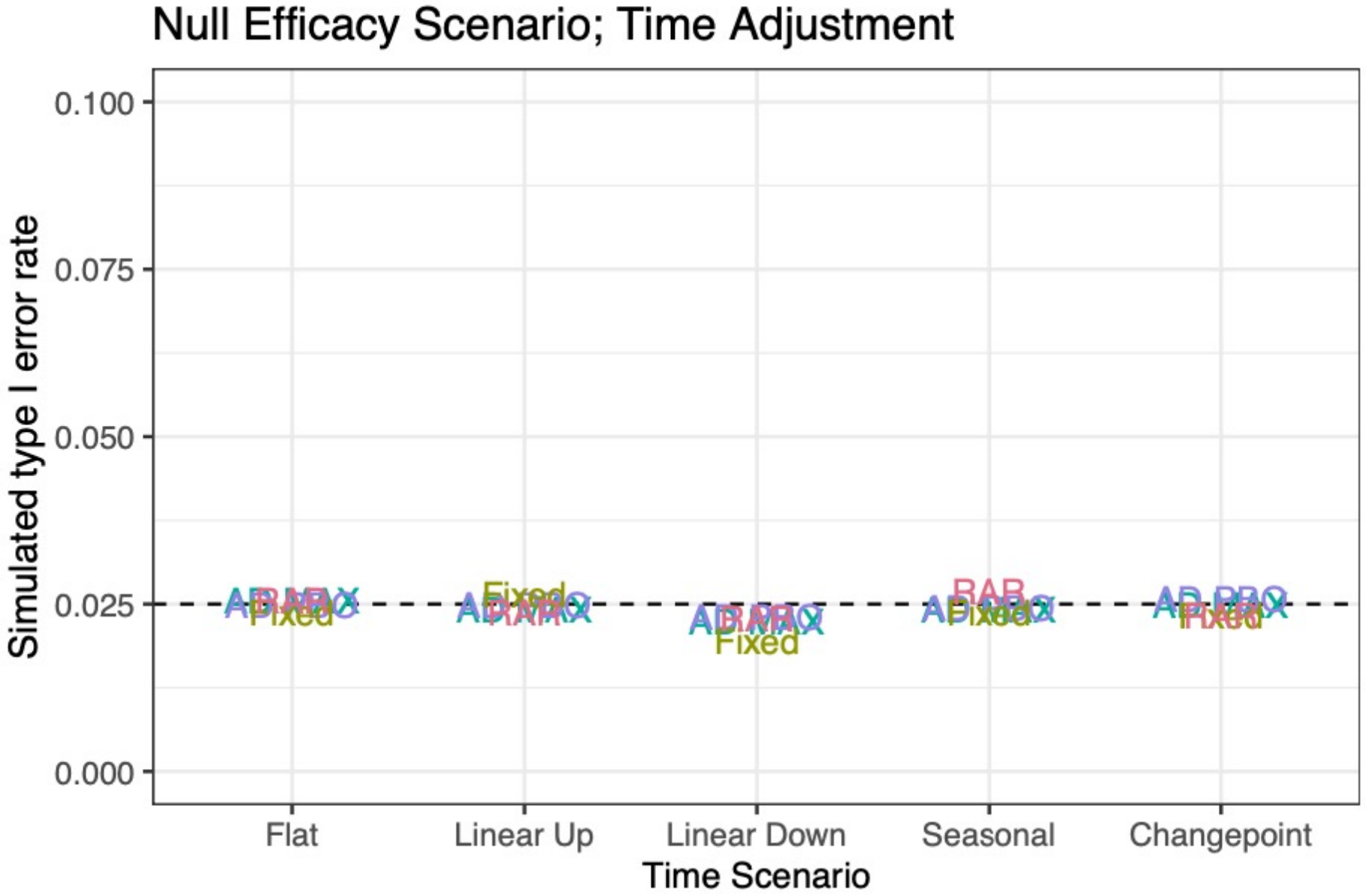
Higher better

Lower better

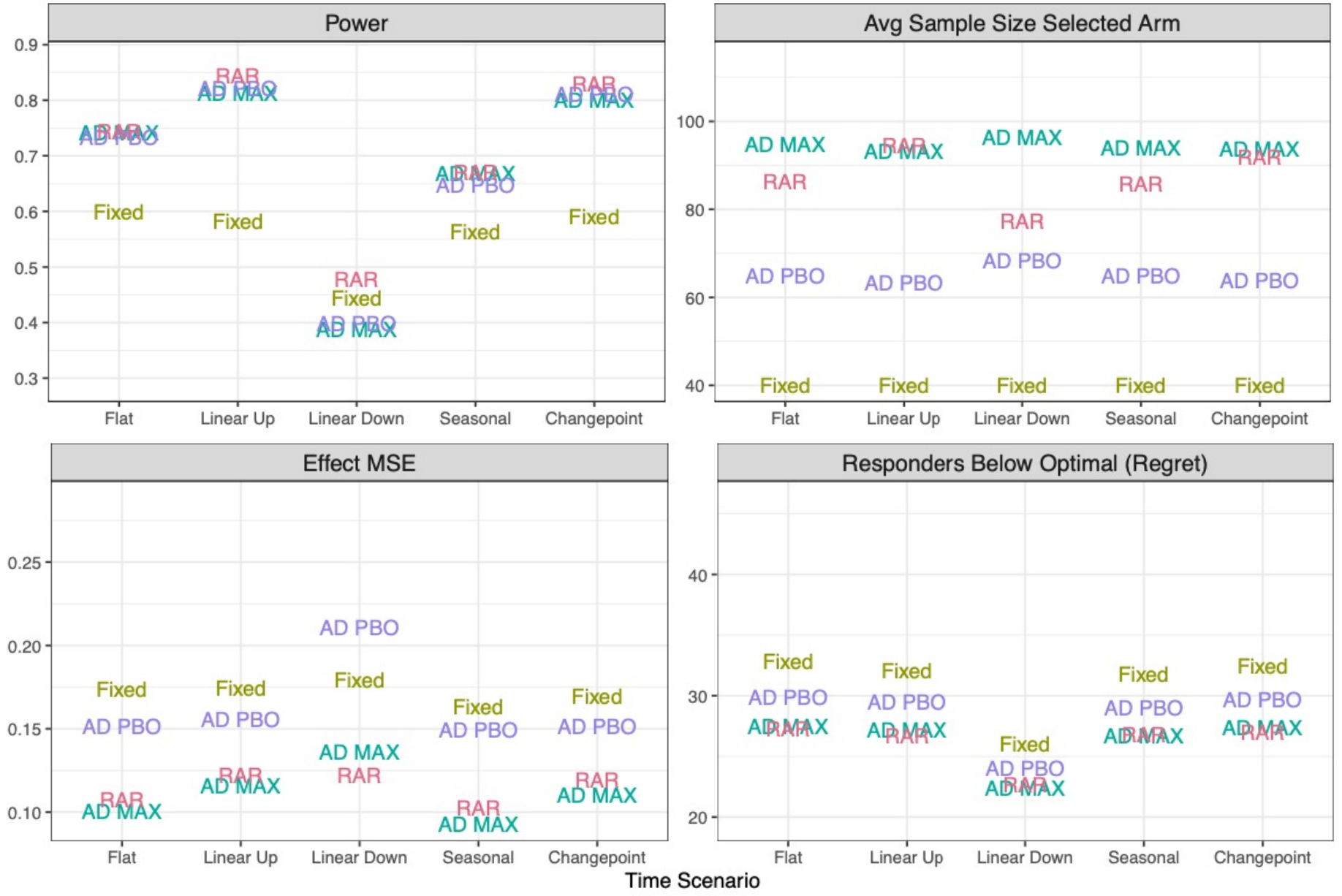
Type I error rates without time covariate



Type I error rates with time covariate



Mixed scenario: Time trends, no time adjustment



Higher better

Lower better

Mixed scenario: Time trends, yes time adjustment

Higher better

Lower better

Conclusion/Discussion

- Adaptive allocation outperforms fixed on all metrics/scenarios
- RAR and AD MAX are comparable across all metrics
  - AD PBO similar on power but has downsides
- If time trends are present, time adjustment avoids type I error inflation in adaptive designs
- Future work will expand to continuous endpoints
  - Linear regression may “soak up” more variability in treatment effect estimates than logistic regression when time trends are real

Acknowledgements

- The publication is currently in progress, and is joint work with other Berry Consultants:
  - Amy Crawford
  - Kert Viele
  - Liz Lorenzi
  - Nick Berry
  - Peter Jacko

Selected References

K Viele, K Broglio, A McGlothlin, and BR Saville. Comparison of methods for control allocation in multiple arm studies using response adaptive randomization. *Clinical Trials*, 17(1):52–60, 2020a.

K Viele, BR Saville, A McGlothlin, and K Broglio. Comparison of response adaptive randomization features in multiarm clinical trials with control. *Pharmaceutical Statistics*, 19(5):602–612, 2020b.

P Thall, P Fox, and J Wathen. Statistical controversies in clinical research: scientific and ethical problems with adaptive randomization in comparative clinical trials. *Annals of oncology : official journal of the European Society for Medical Oncology,* 26(8):1621-8, 2015.

Trippa, L et al. “Bayesian adaptive randomized trial design for patients with recurrent glioblastoma.” *Journal of Clinical Oncology : official journal of the American Society of Clinical Oncology,* 30(26): 3258-63, 2012.

JMS Wason and L Trippa. “A comparison of Bayesian adaptive randomization and multi-stage designs for multi-arm clinical trials.” *Statistics in Medicine,* 33(13): 2206-21, 2014.

Bringing Balance to RAR Discussion

Kert Viele

SCT 2023

Thanks to all the speakers!

- 3 wonderful talks, all on different aspects of RAR!

- RAR remains a complex topic with active research programs
  - theoretical
  - applied
  - inside and outside of clinical trials

Sofia Villar

- 80 years of RAR in 20 minutes!

- Many many many…many…many….variants of RAR
  - some extensively studied with theoretical optima available
  - others still under active research

- Thompson sampling and variants thereof
  - randomize based on a target
- Play the winner/Bandit solutions
  - often not randomized, but randomized options exist

Sofia Villar

- Don’t overgeneralize results!

- Characteristics of RAR X don’t necessarily generalize to RAR Y
- Often behavior depends on “auxillary” choices
  - for example control allocation

- RAR methods need to be matched to the problem
- Are you after the best arm? all good arms?

Will Meurer

- How do patients view RAR?
  - Enrollment and Understanding
- Is RAR helpful for enrollment?
  - Yes!
- Is RAR well understood by potential patients
  - not as much….
  - difficult balance between technical accuracy/good communication
  - ICF information is often simplified

Lindsay Berry

- How does RAR compare to Arm dropping?
- What about time trends?

- Remember arm dropping isn't one design either!
- Difficult to create complete "apples to apples" comparisons
  - control allocation matters to RAR. Match control allocation in AD?
  - control allocation interacts with 2:…:1 or 1:…:1 starting allocation, etc.
- Like previous results in RAR, when chasing the best arm, better to make decisions based on the best arm
  - standard theme…match methods to goals

Lindsay Berry

- Only additive time trends considered (in reality and in model)
- Minimal cost to including time in the model
  - recommend always including time in the model!

- Inclusion of time into the model accurately account for additive time trends, both in RAR and arm dropping
  - not needed for fixed allocation to avoid biases, but still useful

- What is a time/treatment interaction anyway?
  - covariate distribution change over time?
  - what is the estimand?
  - what assumptions can we apply into the future, if any?

RAR in Practice (Berry and Viele 2023)

- Lecanamab phase 2
- SEPSIS-ACT
- AWARD-5
- PROSPECT
- ICECAP

- Our current practice
  - multiple arm trials, not two arm (control and treatment)
  - searching for the best arm
  - maintenance of control allocation