Arm Dropping and Response Adaptive Randomization

Berry Adaptive Design and FACTS Webinar

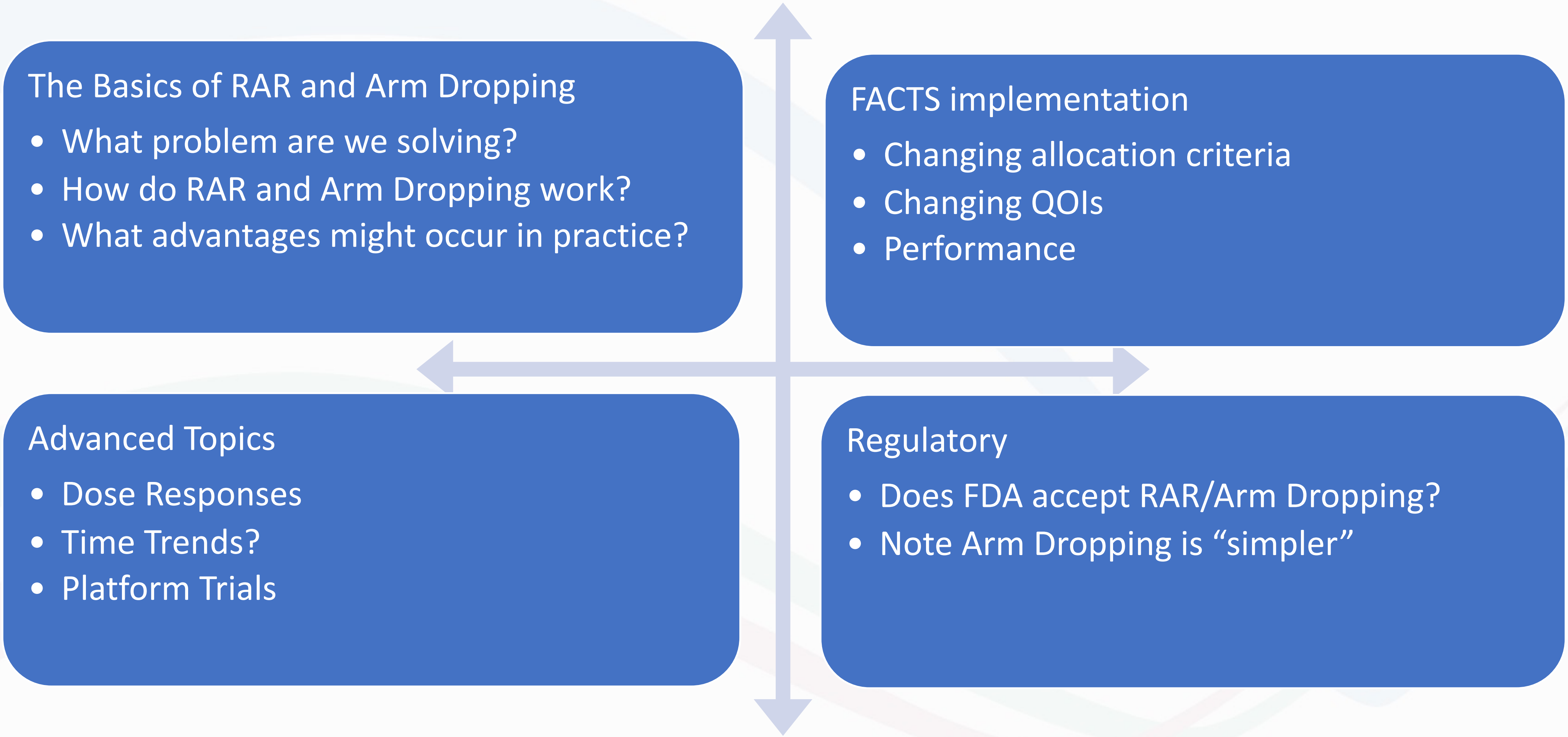
Kert Viele

X/Twitter @KertViele, LinkedIn

Berry Consultants

Statistical Innovation

Outline



The Basics of RAR and Arm Dropping

- What problem are we solving?
- How do RAR and Arm Dropping work?
- What advantages might occur in practice?

FACTS implementation

- Changing allocation criteria
- Changing QOIs
- Performance

Advanced Topics

- Dose Responses
- Time Trends?
- Platform Trials

Regulatory

- Does FDA accept RAR/Arm Dropping?
- Note Arm Dropping is “simpler”

What problem are we solving?

- Many trials have multiple arms
  - dose ranging
  - platform trials investigating multiple therapies for an indication
  - shared control designs

- We may want to determine
  - what is the best arm?
  - identify all effective arms?

- Accumulating data often indicates some arms are poor choices.
  - allocating less to these arms, and more to more promising arms, may allow us to make better, more efficient decisions.

How does RAR/AD work?

- Periodic interim analyses
  - at each interim, evaluate how well each arm is performing
    - Pr(arm is the best active arm)
    - Pr(arm beats control)
    - p-value of each arm vs control
- Adjust allocation in favor of better performing arms
  - increase allocation to arms with the highest performance
  - decrease/eliminate allocation to poorly performing arms
- Arm dropping – if metric falls below a threshold, eliminate arm
- RAR – allocations changes, but can be "smaller" as opposed to 0

Example Interim – Dichotomous data

Arm	Control	Arm A	Arm B	Arm C
Data	6/20	5/20	11/20	9/20
Pr(arm best)	NA	0.0108	0.7304	0.2588
Pr(arm>ctrl)	NA	0.3628	0.946	0.837
pvalue vs ctrl	NA	0.6383	0.0549	0.1636

There are many forms of RAR and Arm Dropping. The details matter.

One common example…going forward, allocate another 80 patients, 20 to control and allocate other 60 to active arms proportionally to Pr(arm best) = 1.08%, 73.04%, 25.88%

Another variant is to threshold small values (<5% for example) to 0 and renormalize
active arms then allocated proportionally to 0.0%, 73.84%, 26.16%

Example Interim – Dichotomous data

Arm	Control	Arm A	Arm B	Arm C
Data	6/20	5/20	11/20	9/20
Pr(arm best)	NA	0.0108	0.7304	0.2588
Pr(arm>ctrl)	NA	0.3628	0.9460	0.8370
pvalue vs ctrl	NA	0.6383	0.0549	0.1636

Another form of RAR might be based on Pr(arm>ctrl)

Allocate active arms proportionally after normalizing (0.3628, 0.9460,0.8370)
Allocation probabilities become 16.91%, 44.09%, 39.01%

Often the choice of metric (arm best or arm>ctrl) depends on the goal, are you looking for the best arm or trying to identify all arms that beat control.

Example Interim – Dichotomous data

Arm	Control	Arm A	Arm B	Arm C
Data	6/20	5/20	11/20	9/20
Pr(arm best)	NA	0.0108	0.7304	0.2588
Pr(arm>ctrl)	NA	0.3628	0.9460	0.8370
pvalue vs ctrl	NA	0.6383	0.0549	0.1636

Arm dropping would simply remove any arm that is performing sufficiently poorly.

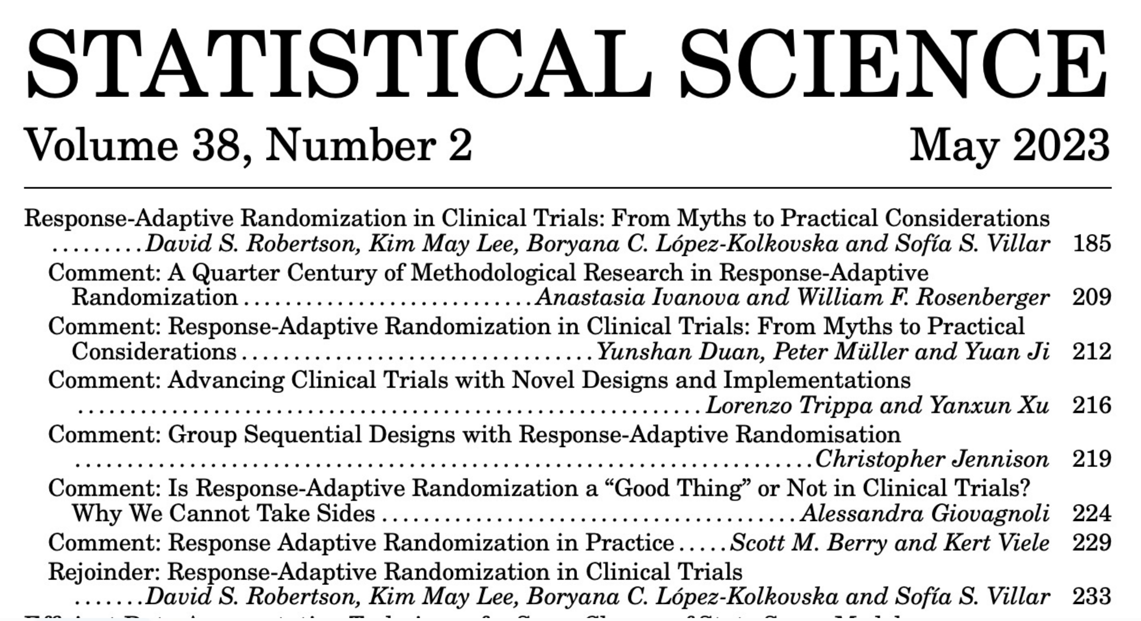
For example, stop any arm with a p-value > 0.5

Future allocation would change the allocation from the 1:1:1:1 previously to 1:0:1:1, dropping arm A

alternatively, could also stop arms with small Pr(arm>ctrl) or small Pr(arm best)

Too many choices!!

- Should we do arm dropping or RAR?
- What should be our metric?
- What should be our thresholds?



STATISTICAL SCIENCE

Volume 38, Number 2 May 2023

- Fortunately, a lot of work exists on this problem!
  - both of what to do, and what NOT to do

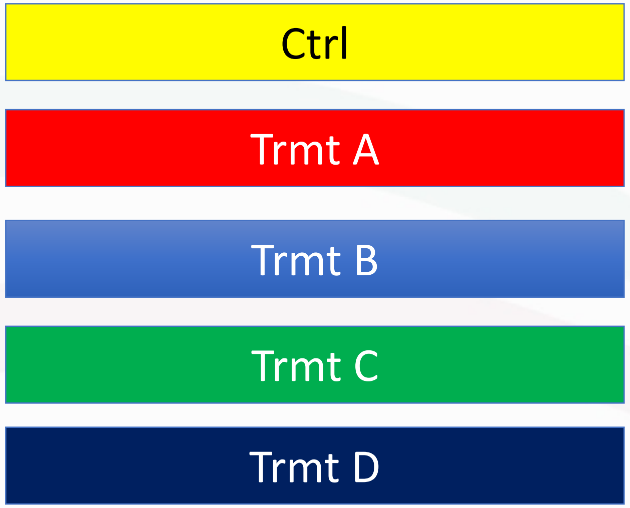
- Review article in Statistical Science (with discussion)

- Robertson et al. Response Adaptive Randomization in Clinical Trials: from Myths to Practical Considerations. *Statistical Science* 2023;38(2);185-209.

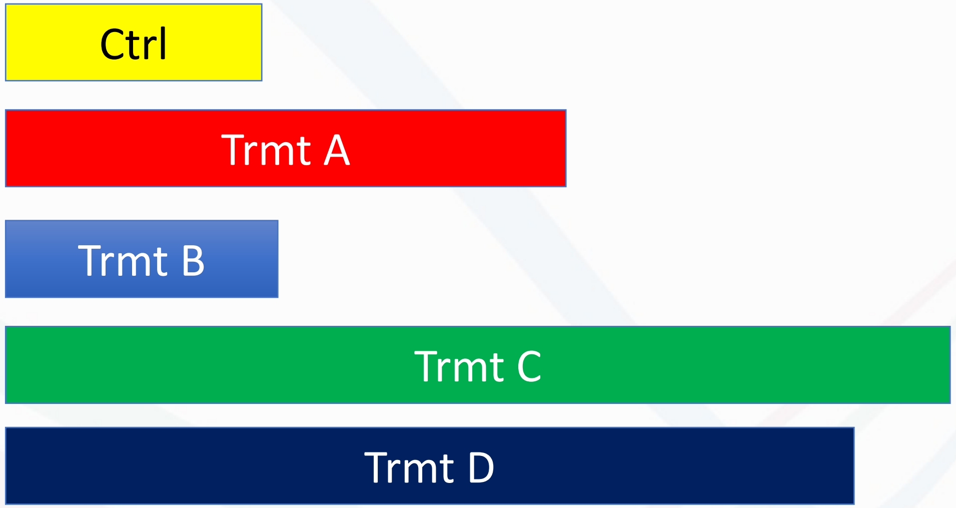
Some important considerations

- If you need to compare to control at some point, then maintaining allocation to control is vital

Equal allocation 100/arm
V(trmt effect) = (1/100)+(1/100)
= 0.02



Potential result after RAR
V(trmt effect) = (1/50) + (1/200) = 0.025
worse…even with more patients in the comparison (250 vs 200). Don’t reduce control!



Some important considerations

- If you need to compare to control at some point, then maintaining allocation to control is vital

- Two arm RAR must violate this. There are only two arms, so altering anything involve altering control allocation
  - Thus, two arm RAR is often problematic
  - May treat patients in the trial better, but at an inferential cost. Worse decisions for patients outside the trial.

- Korn, Freidlin. Outcome adaptive randomization – is it useful? *JCO* 2011;29(6);771-6.
- Thall, Fox, Wathen, Statistical controversies in clinical research… Ann. Oncology. 2015;26(8);1621-8.
- Wathen, Thall. A simulation study of outcome adaptive randomization in multi-arm clinical trials. *Clinical Trials* 2017;14(5);432-440.
- Viele et al Comparison of methods for control allocation…. Clinical Trials 2020;17(1);52-60.

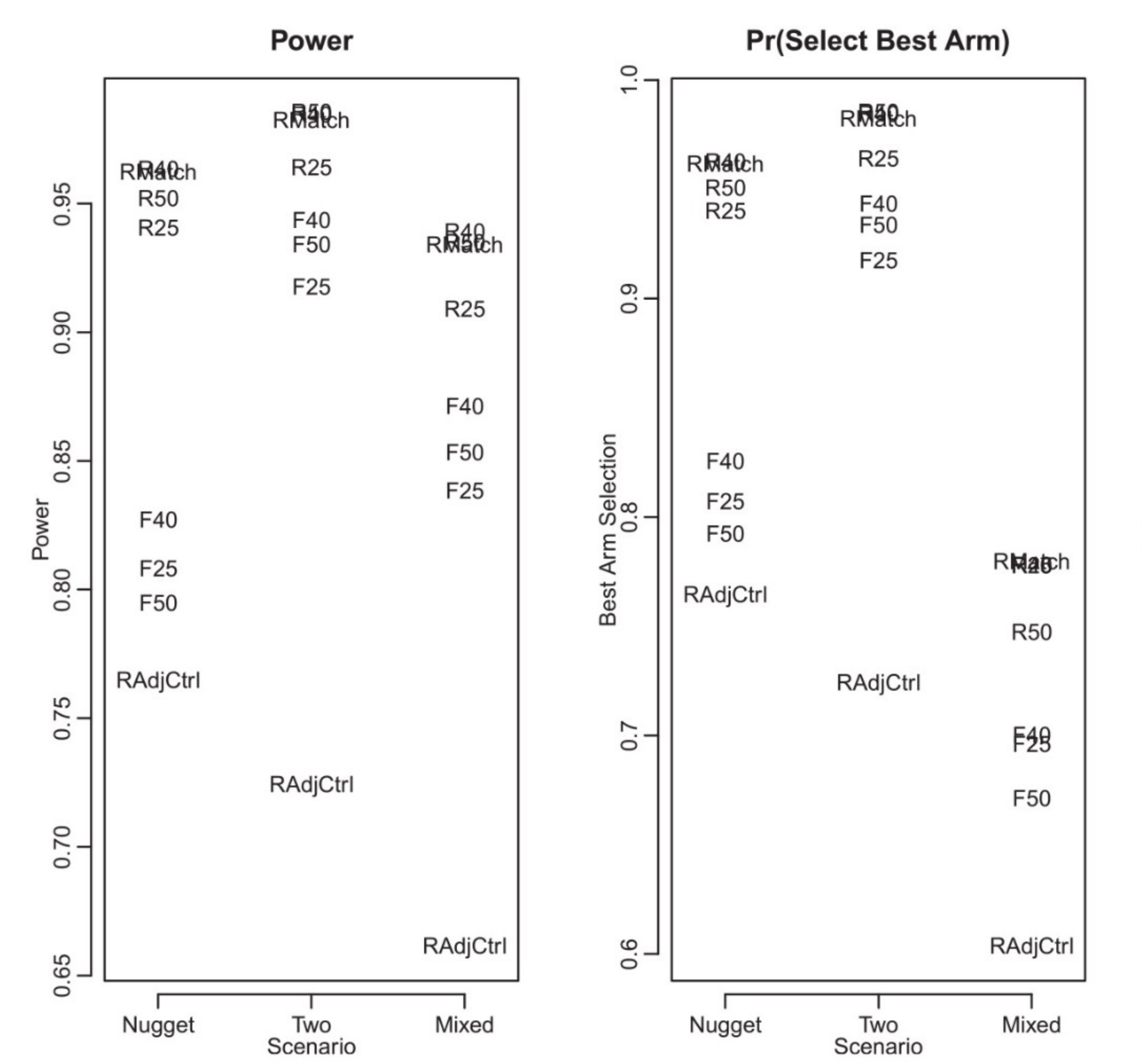
From Viele 2020, comparing RAR variants with different control allocation to different fixed trial variants.

Methods which may lower control allocation (RAdjCtrl) perform poorly overall. Methods which maintain control allocation (R25, R40, R50, Rmatch) perform well.

Also in Viele 2020, RAR maintaining control allocation has

1) improved power compared to non-adaptive
2) better arm selection
3) better estimation of treatment effect (MSE)
4) improved treatment of trial participants

Does not compare to arm dropping



Some important considerations

- Adapting too early or too aggressively leads to bad decisions.

- You can recover from a random high.
  - RAR will put more observations on that arm
  - Random high likely to regress to its true mean.

- Difficult to recover from overreacting to a random low.
  - Few future observations leads to limited chance to recover.
  - Thus, early interims and allocation rules must be carefully calibrated.
  - Fortunately, work here as well.

- Thall, Fox, Wathen, Statistical controversies in clinical research… *Ann. Oncology.* 2015;26(8);1621-8.
- Viele et al. Comparison of response adaptive randomization features…. *Pharm Stat* 2020;19(5);602-612.

Some important considerations

From Viele 2020 (Pharm Stat)

Looked at metric for adaptation
number of interims
starting interim location
thresholding to 0% allocation

by necessity incomplete

Findings…
If identifying the best arm is your goal, use Pr(arm best) instead of Pr(arm>ctrl)
More interims better than fewer (though diminishing returns for more interims)
20% burnin effective (lowest considered)
10% thresholding effective (highest considered)

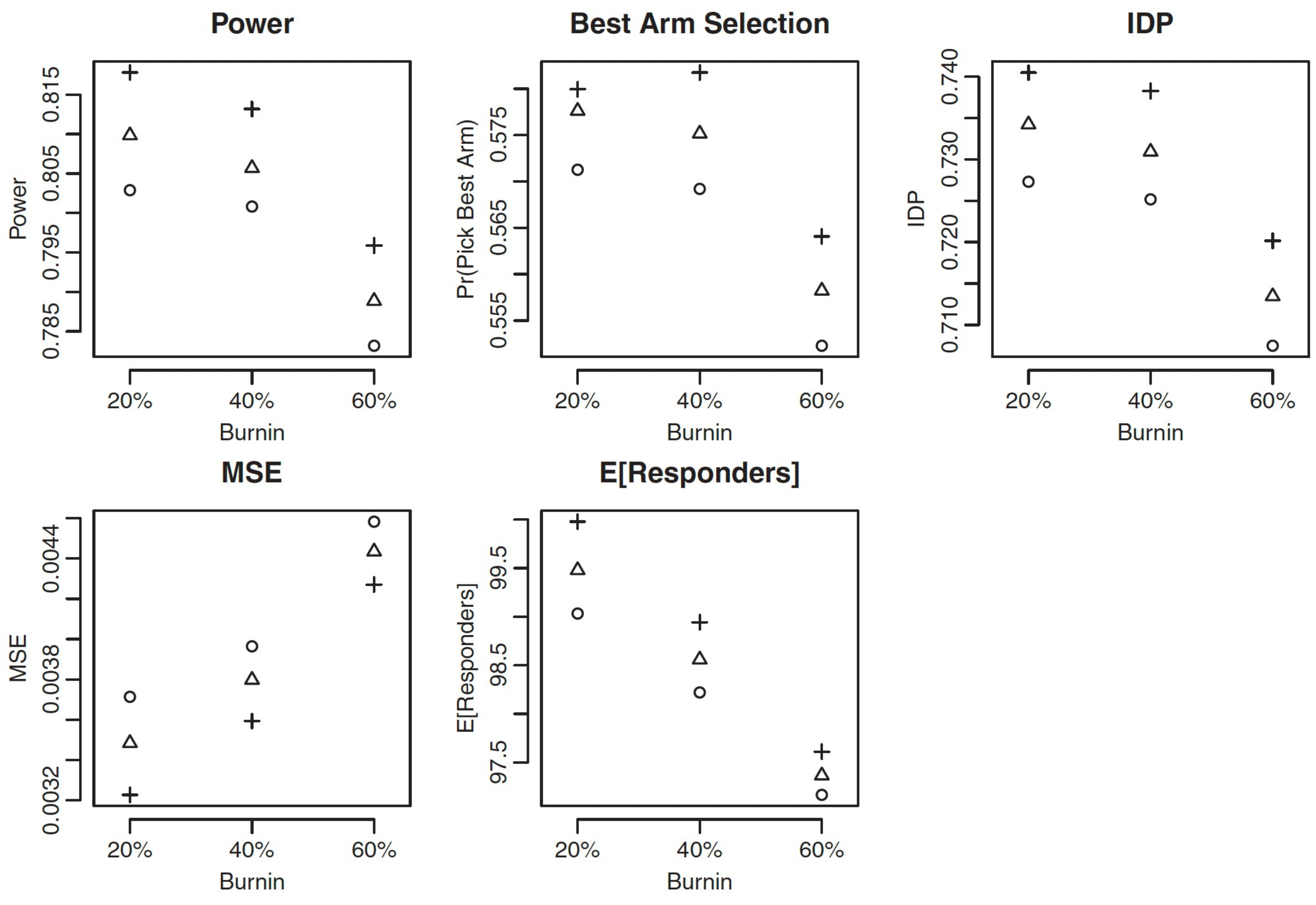


FIGURE 2 Results for each design metric by burnin and dropping threshold for the mixed scenario, focusing on designs driven by Pr(Max) and with many interims. Burnin is shown on the *x*-axis. Point within the plot indicates the arm dropping thresholds (circle = 0%, triangle = 5%, plus = 10%)

Some important considerations

- Match your design choices to your goals

- Driving a design by Pr(arm best) is useful when your goal is to identify the best arm. If you are trying to identify all useful arms, Pr(arm>ctrl) or p-values may be more useful.
    - ”more useful” may just mean “the alternative is bad…"

- Robertson et al. Response Adaptive Randomization in Clinical Trials: from Myths to Practical Considerations. *Statistical Science* 2023;38(2);185-209.
- Trippa et al. Bayesian adaptive randomized clinical trial design… JCO 2012;30(26);3258-63.

Arm Dropping is simply….simpler

- Many of the same principles that apply to RAR apply to arm dropping, but are more automatically satisfied.

- Control allocation maintained in standard arm dropping designs.
- Avoid aggressive early stopping.
- Match the driver of arm dropping (drop on Pr(arm best), Pr(arm>ctrl), etc.) to the goal of the study.

- Much modern arm dropping work found in MAMS designs.
- Wason, Trippa. A comparison of Bayesian adaptive randomization…. *Stat in Med* 2014;33(13);2206-21.

Arm dropping vs RAR

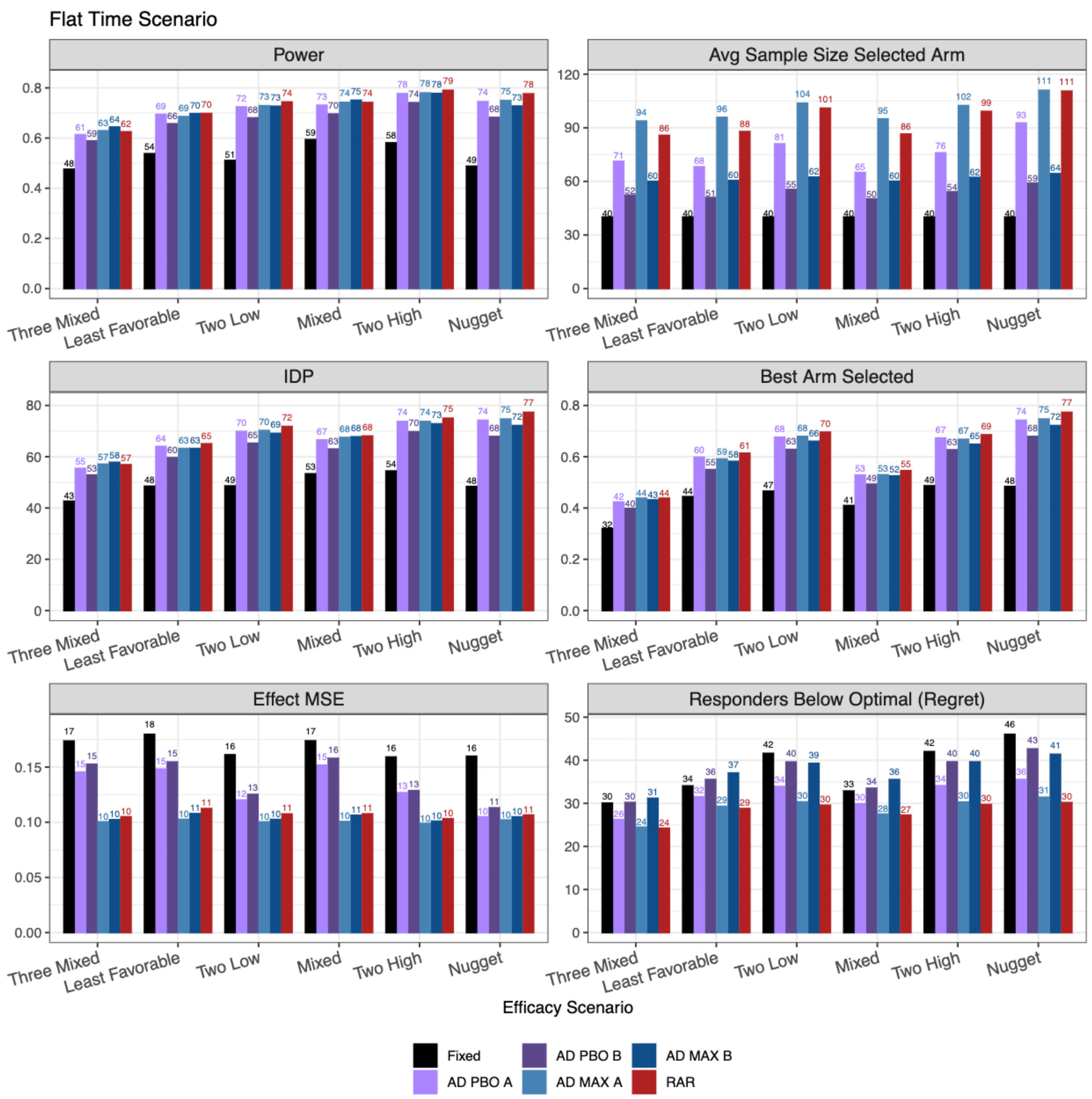
Current work comparing RAR to Arm Dropping (under revision)

Goal is to find the best arm and compare to control (different than “identify all effective arms”)

RAR based on Pr(arm best) and Arm Dropping based on Pr(arm best) are similar

Arm Dropping based on Pr(arm>ctrl) is worse on some metrics (ok on others).
For example sample size on selected arm, MSE, and “responders below optimal”

Everything beats fixed ☺



Berry Consultants

FACTS Implementation

- Template provided outside slide presentation

Dose Response Modeling

- Dose selection was one of the initial primary uses for RAR.

- Given doses are related, we would typically place a model connecting the doses
  - EMAX is quite common
  - NDLM or alternative also valuable if monotonicity is not desired

- RAR or Arm Dropping can focus resources on the optimal dose
  - Note “finding the best arm” is often a primary goal (while often providing “enough” information on the dose response)

- Gajewski et al. Bayesian hierarchical EMAX model…. Stat in Med 2019;38(17);3123-38.

Dose Response Modeling

- FACTS provides a number of dose responses
  - EMAX and related functions
  - Smoothers (NDLM and variants)
  - U-shaped functions
  - Independent models

- Arm Dropping can be implemented with constraints appropriate for doses, for example dropping from highest (or lowest)

- FACTS will also perform a standard phase 2/3 design, where dose finding is followed by a dose selection and confirmatory stage (staged design…upcoming webinar)

Time Trends

- RAR alters the allocation ratio between arms between interims
- Suppose in the 2nd half of a trial outcomes are 20 points higher than the first half, and allocation ratios do NOT change

First half means
Control = 50
Arm A = 70
Arm B = 60

Second half mean
Control = 70
Arm A = 90
Arm B = 80

- Allocate 100 patients per arm per time period (no RAR)
  - Expected mean for control = (50*100) + (70*100) / 200 = 60
  - Expected mean for arm A = (70*100) + (90*100) / 200 = 80
  - Expected mean for observed treatment effect = 80 – 60 = 20. Good!
  - May want to model time to reduce variance, but not needed for bias.

Time Trends

- RAR alters the allocation ratio between arms between interims
- Suppose in the 2nd half of a trial outcomes are 20 points higher than the first half, and allocation ratios do change

First half means
Control = 50
Arm A = 70
Arm B = 60

Second half mean
Control = 70
Arm A = 90
Arm B = 80

- Allocate 100:100:100 in first half, 100:150:50 in second half
  - Expected mean for control = (50*100) + (70*100) / 200 = 60
  - Expected mean for arm A = (70*100) + (90*150) / 250 = 82
  - Expected mean for observed treatment effect = 82 – 60 = 22. Biased!
  - Amount of bias depends on time effect and allocation ratios.

Time Trends

From current work under revision

When a time trend is present and NOT accounted for, biases are present and result in type 1 error inflation

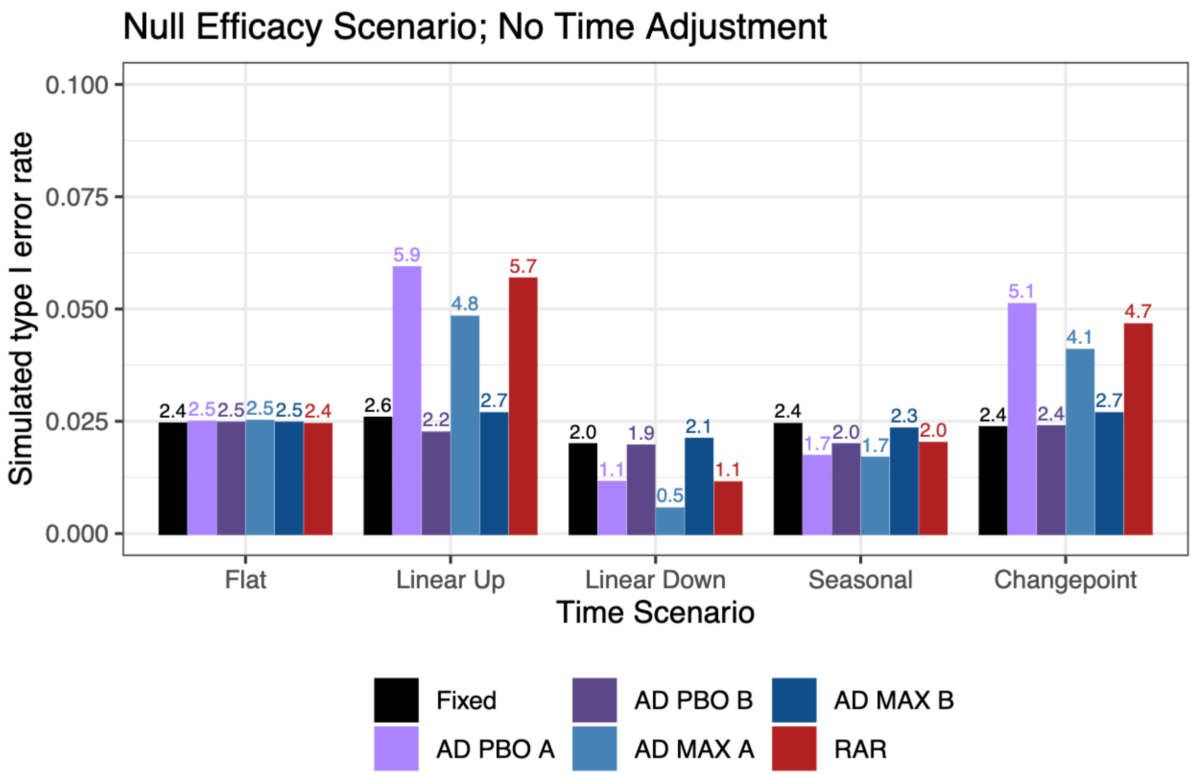


Figure 2: Simulated type I error rates in the null scenario without covariate adjustment for time. The five time trend scenarios are shown on the x-axis.

Time Trends

- Arm Dropping is again simpler, often allocation ratios are maintained.
- Be a little careful
  - Changing from 2:1:1:1:1 to 4:0:2:0:2 avoids need for time adjustment
  - Changing from 2:1:1:1:1 to 2:0:2:0:2 could create time biases
  - Again, key issue is ratios between arms

Time Trends

- This is an additive time trend (all arms move equally)
- It can be corrected with an additive model including time
  - Outcome = Intercept + Arm Effect + Time Effect + Error
- Note maintaining control allocation remains vital, need enough information to estimate time effects (2 arm RAR problematic)

First half means
Control = 50
Arm A = 70
Arm B = 60

Second half mean
Control = 70
Arm A = 90
Arm B = 80

- Korn, Freidlin. Time Trends with response adaptive randomization… Clinical Trials 2022;19(2);158-61.
- Bofill Roig et al. On model based time trend adjustments…. BMC Med Res Method. 2022;22(1);228.
  - Involves nonconcurrent platform controls but math identical for RAR.

Time Trends

From current work, additive model corrects for additive time trend. Type 1 error controlled (some simulation error seen in graph at right)

Remaining performance criteria similar to graph shown previously

RECOMMENDATION…always include an additive time trend for robustness. Minimal cost, can be very helpful if needed.

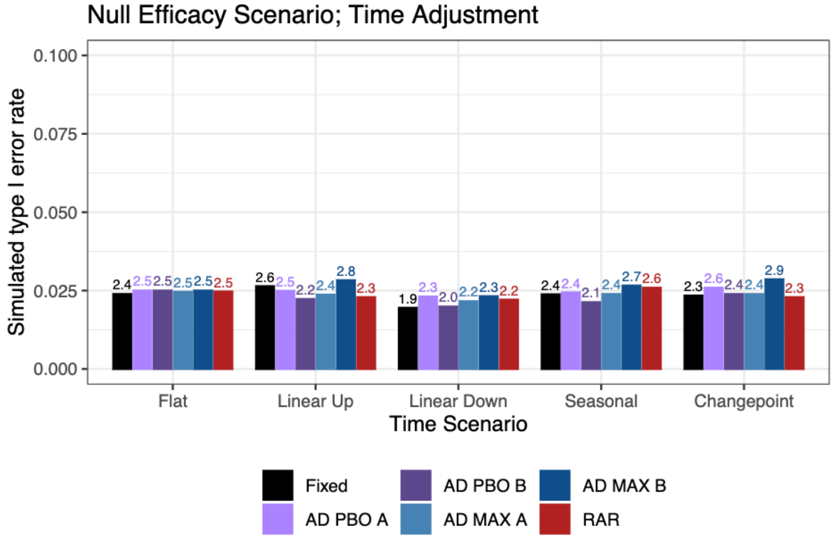


Figure 4: Simulated type I error rates in the null scenario with covariate adjustment for time. The five time trend scenarios are shown on the x-axis.

Time Trends

- Note that additivity is a key assumption
- If treatment effects vary over time (after accounting for covariates, etc.)
  - life is very complex
  - Additive Modeling insufficient
  - desired estimand is truly unclear
  - What are we trying to estimate, and how are we trying to generalize from this trial to a different time?

- Ongoing research….
- Bofill Roig et al. On model based time trend adjustments…. BMC Med Res Method. 2022;22(1);228.

Platform trials

- RAR and Arm Dropping are both used in platform trials
- Ongoing research….these can have different implications in platforms.
- In a fixed horizon trial, RAR reducing allocation can reduce total allocation to an arm.
- In a platform, RAR reducing allocation may simply delay completion of an arm.
    - Thus, RAR might be viewed more as prioritizing arms rather than changing total allocation

Regulatory Notes

- Both RAR and Arm Dropping discussed in the adaptive guidance.
  - Examples of each, including in confirmatory trials

- Both must be calibrated to obtain type 1 error control, often by simulation.

- Arm Dropping is usually the simpler route.

- Do NOT overgeneralize RAR. Difference variants of RAR have different performance.

- Berry, Viele. Response Adaptive Randomization in Practice (discussion of Robertson et al) Statistical Science 2023;38(2);229-33.

My PERSONAL current summary

- RAR and Arm dropping, optimally performed, are both more efficient than fixed trials and similar to each other.

- Arm dropping is simpler, often easier to explain and implement
  - ongoing work in platform trials, where interpretation may change from limiting/increasing allocation to delaying/accelerating allocation

- Additive time trends can be accounted for with additive time models. Interactive time trends are a mess.

Thank you

- Thank You for attending
- Link to Recording will be sent out tomorrow
- Slides will be available via our website at the end of the series
- Any questions please contact us:
  - tom@berryconsultants.com
  - kert@berryconsultants.com
  - facts@berryconsultants.com
  - demo and/or a free evaluation copy of FACTS
- Berry regularly produces blogs and social media posts on adaptive designs
  - @KertViele, Kert Viele on LinkedIn