Response-adaptive randomization in clinical trials: a whistle-stop tour of myths and barriers

Sofía S. Villar

MRC Biostatistics Unit

UNIVERSITY OF CAMBRIDGE

PSI Scientific Meeting: Adaptive Designs and their Application

Cambridge, 13th October 2022

Acknowledgments

With thanks to

David Robertson

Kim Lee

Boryana Lopez-Kolkovska

and many other colleagues

Outline

Response-Adaptive Randomisation

- Response-adaptive Randomisation (RAR) is perhaps the oldest Adaptive Design.
- First proposed by Thompson (1933) who suggested to "randomise" patients between two treatments in "a Bayesian" fashion using the posterior probability that the response rate of one 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 probability 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

Imbalance and recurrence

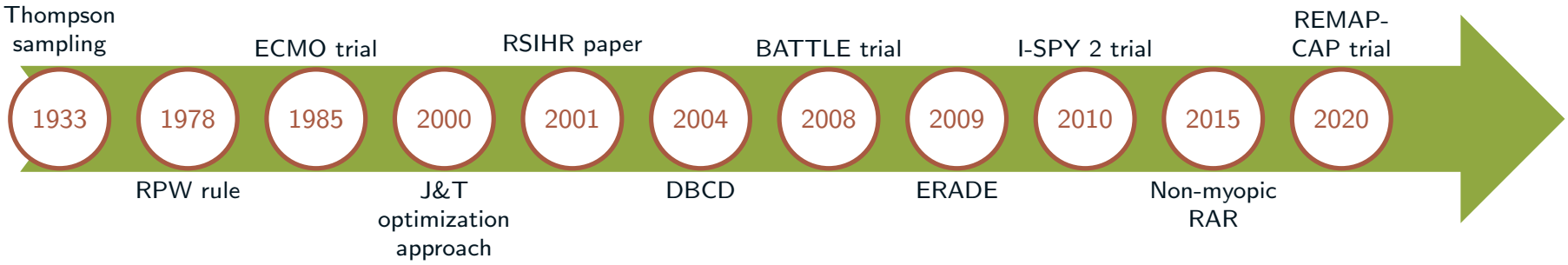

Figure: Timeline summarizing some of the key developments around the theory and practice of RAR in clinical trials. J&T = Jennison and Turnbull (2000), RSIHR = Rosenberger et al. (2001a).

- Large amount of high quality theoretical works paired with few highly influential examples of RAR in practice.
- Persistence of debate and arguments
- Heavy focus on certain aspects, large gaps in others.

Where to read more?

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

Opinion

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 to read this review?

Why (we wanted to) write a review paper in RAR?

- To reconcile apparently conflicting arguments
- To write an updated review (to account for more recent work)
- To classify RAR and provide a non-expert roadmap
- To guide future uptake and research of RAR

Outline

Some basic ideas & notation

- A study with fixed number of patients n (fixed sample) to be randomised to arms: 0 (control) or $k = 1, \ldots, K$ (experimental).
- Some primary outcome variable $Y_k \sim f(\theta_k)$
- Let the treatment 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.

Comparing RAR

- Before we further define RAR, let's explore all relevant dimensions.
- For simplicity, let's do this when $K = 1$ (two-arm study) with $H_0 : p_0 = p_1$ (null) and (some alternative) $H_0 : p_0 \neq p_1$
- A multitude of metrics can be put forward. We focused on 3 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 ignoring some of them).

Classifying RAR

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

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

- RAR as broad class of adaptation, includes 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 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 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:

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

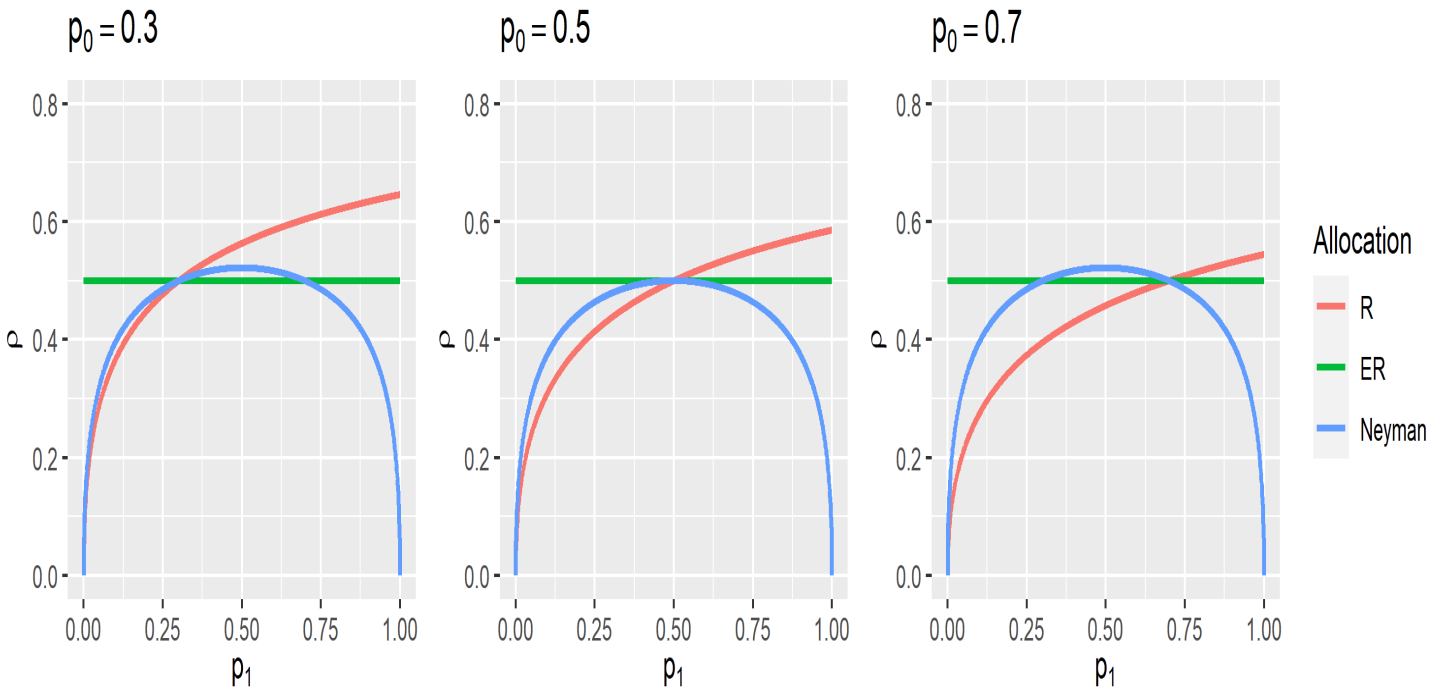

Figure: Optimal ratios ρ^*_{Neyman} and ρ^*_R as a function of p_1, for $p_0 \in \{0.3, 0.5, 0.7\}$

Does RAR reduce power? II

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

- FER 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, more complex considerations but in general impact on power will vary largely for different RAR

Frequently asked questions

- Does using RAR 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 increase chances to receive an inferior arm? I

- New view: RAR has a substantial chance (up to 43%) of producing sample size imbalances in the wrong direction (i.e. towards the inferior arm).

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

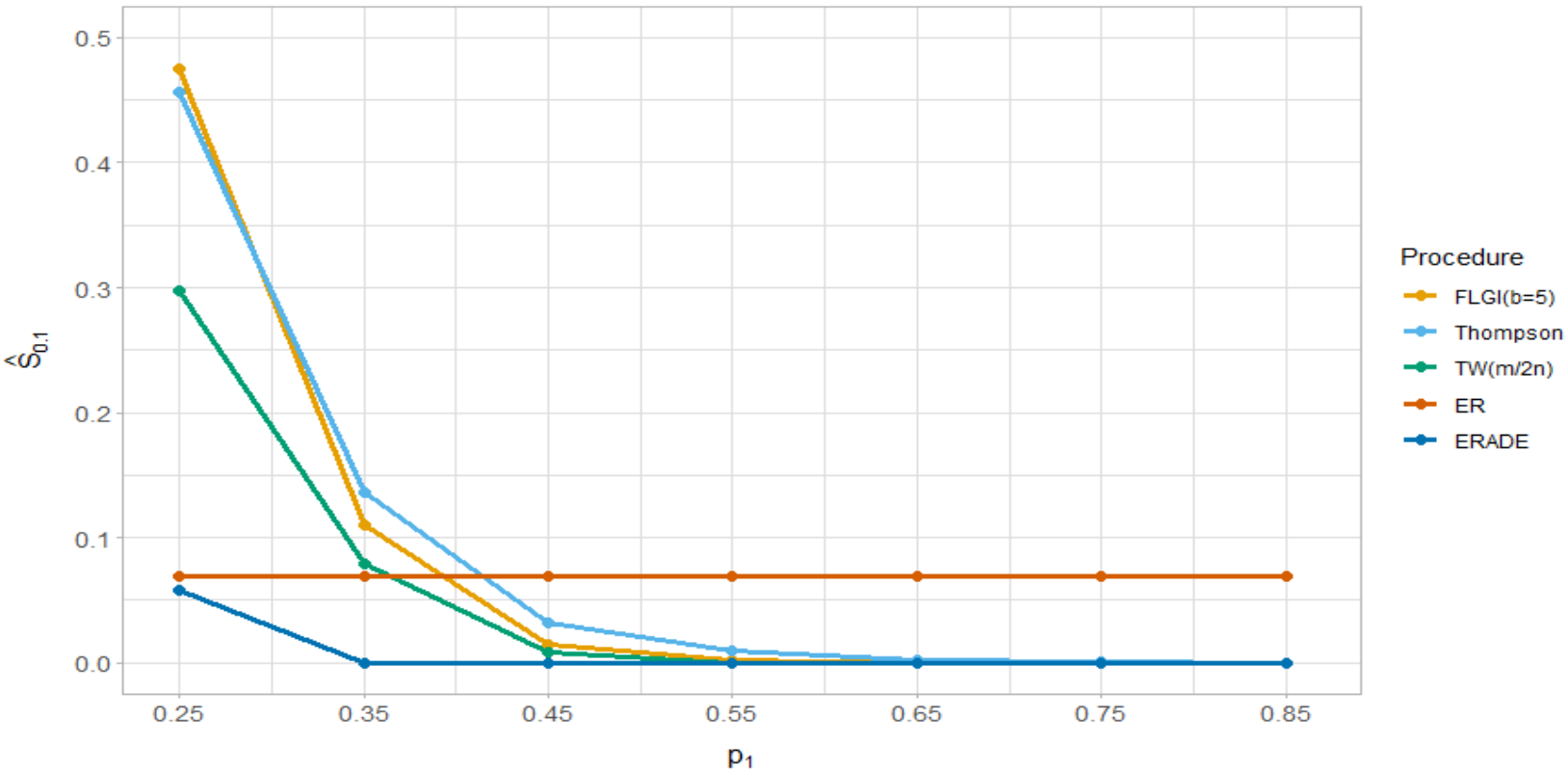

Figure: Plot of $\hat{S}_{0.1}$ for various RAR procedures as a function of p_1, where $p_0 = 0.25$ and $n = 200$. Each data point is the mean of 10^4 trial replicates.

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)

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.
- 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? and more
- Generalisations and broad statements of RAR (in terms of relevant metrics) hardly ever true. Trade-offs are ubiquitous, best strategy is to be aware of them.
- There many ways to implement RAR and the setting should guide both the decision to use it or not and the choice of which one.

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)*,