MRC Biostatistics Unit

UNIVERSITY OF CAMBRIDGE

Short course on Response-Adaptive Methods for Clinical Trials

Lecture 3: Further methodological considerations in RAR – early stopping, missing data and estimation

Sofía S. Villar

MRC Biostatistics Unit

November 30th, 2025 - Perth

Outline

- Early stopping
  - Early stopping rules
  - Group sequential RAR
- Missing data
  - Impact of missing data
  - Use of mean imputation
- Estimation
  - Bias of the MLE
  - Bias-correction

Early stopping

- Assumption in many papers on RAR is that total sample size $n = n_{\max}$ is fixed.
- However, both in theory and practice, this assumption can be relaxed to allow for early stopping of individual arms or indeed the whole trial
- Early stopping of the trial could be for safety reasons, but here we focus on stopping early for either efficacy or lack of benefit
- Advantages of early stopping
  - Fewer patients needed on average $\rightarrow$ savings in terms of time and cost
  - Ethical reasons: help ensure patients are not exposed to ineffective/inferior treatments

Early stopping rules

- In Bayesian framework, natural to think of early stopping rules in terms of posterior probabilities of the parameter(s) of interest
- **Example:** Bayesian RAR (BRAR) in Lecture 1
- Recall allocation probabilities are (proportional to) $P(p_1 > p_0 | \boldsymbol{A}^{(i-1)}, \boldsymbol{Y}^{(i-1)})$
- Natural to use early stopping rules also based on this posterior probability:
  - Stop early for efficacy if $P(p_1 > p_0 | \boldsymbol{A}^{(i-1)}, \boldsymbol{Y}^{(i-1)}) > \xi$
  - Stop early for lack of benefit if $P(p_1 > p_0 | \boldsymbol{A}^{(i-1)}, \boldsymbol{Y}^{(i-1)}) < \eta$
  - Here ξ and η are chosen to ensure acceptable frequentist operating characteristics (for example)

Groups sequential designs

- Alternative approach is to use well-established group sequential designs (particularly for confirmatory trials)
- Here, interim analyses take place when groups of patients have had their outcomes observed
- At jth analysis, test statistic Z_j is calculated using patients assessed so far.
- A general one-sided group sequential test is defined by constants (l_j, u_j) with $l_j < u_j$ for $j = 1, \ldots, J$ and $l_J = u_J$

Groups sequential designs

- Alternative approach is to use well-established group sequential designs (particularly for confirmatory trials)
- Here, interim analyses take place when groups of patients have had their outcomes observed
- At jth analysis, test statistic Z_j is calculated using patients assessed so far.
- A general one-sided group sequential test is defined by constants (l_j, u_j) with $l_j < u_j$ for $j = 1, \ldots, J$ and $l_J = u_J$

After group $j = 1, \ldots, J-1$

if $Z_j \geq u_j$ stop, reject H_0 (*early stopping for efficacy*)

if $Z_j \leq l_j$ stop, do not reject H_0 (*early stopping for lack of benefit*)

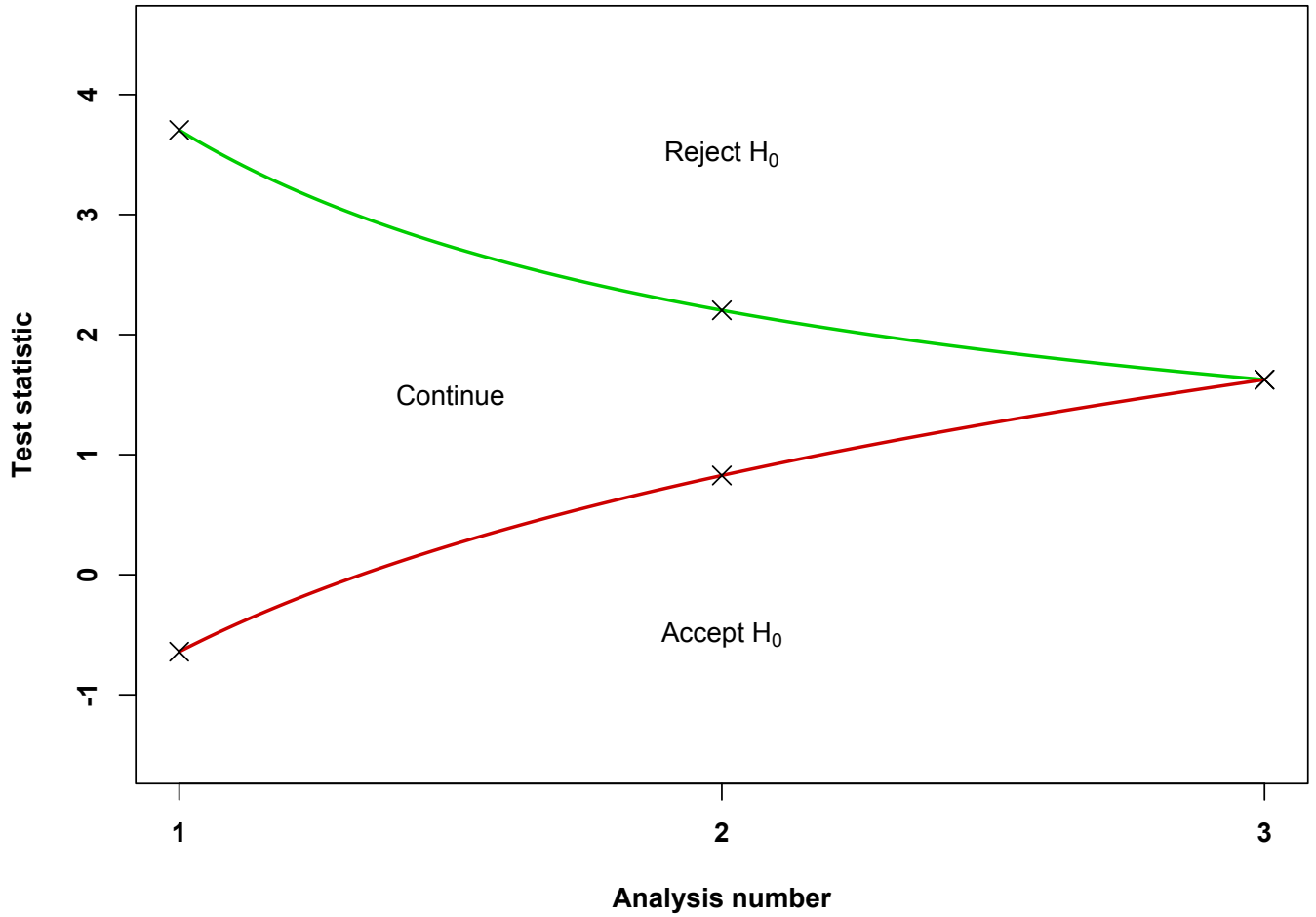
otherwise continue to group $j+1$

After group J

if $Z_J \geq u_J$ stop, reject H_0

if $Z_J < l_J$ stop, do not reject H_0

Group sequential design schematic



Calculating error probabilities

- Let $\boldsymbol{l} = (l_1, \ldots, l_J)$ be the *lack of benefit boundaries* and $\boldsymbol{u} = (u_1, \ldots, u_J)$ the *efficacy boundaries*.

- Choose $\boldsymbol{l}$, $\boldsymbol{u}$ and the (per-group) sample sizes so that design has overall type I error rate α and power $(1 - \beta)$

Group sequential RAR

- We can combine RAR with a group sequential design
- **Example**: Consider the RAR design from *Jennison and Turnbull (2000)*
- Two-arm trial with normally-distributed responses with means μ_0, μ_1 and known and common variance σ^2
- Treatment effect (mean difference) $\theta = \mu_1 - \mu_0$
- Hypothesis test $H_0 : \theta \leq 0$ vs $H_1 : \theta > 0$
- Trial conducted with type I error probability $\alpha = 0.025$ and power $1 - \beta = 0.9$ when $\theta = \delta$

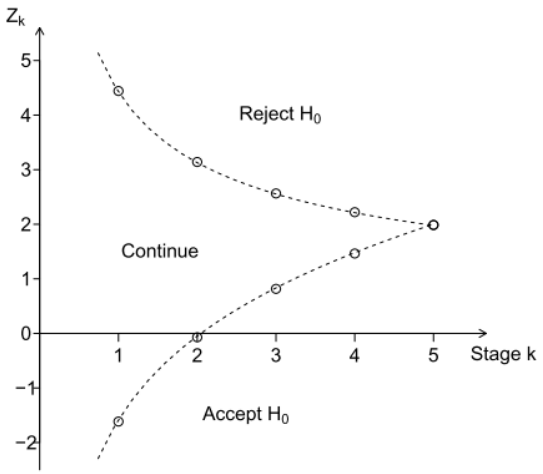
Group sequential RAR

- RAR design chosen to minimise a loss function of the form

$$L(\theta) = \begin{cases} N_1/\sigma^2 + 4^{\theta/\delta} N_0/\sigma^2 & \text{when } \theta \geq 0 \\ 4^{-\theta/\delta} N_1/\sigma^2 + N_0/\sigma^2 & \text{when } \theta \leq 0 \end{cases}$$

  where N_0, N_1 is the number of patients assigned to the control and experimental arm, respectively

- Consider *Pampallona and Tsiatis* stopping boundaries



Jennison (2023)

Group sequential RAR

RAR without early stopping

- Five stages with four updates to the allocation ratios
- First stage used as burn-in with equal allocation

	GS RAR (fixed $n_{\max}$)	
θ	$E(N_1)$	$E(N_0)$
$-\delta/2$	86.4	122.8
0	101.4	101.5
$\delta/2$	122.8	86.4
δ	153.0	75.7
$3\delta/2$	195.9	68.2

Jennison (2023)

As comparison, non-adaptive (ER) design always requires 100 observations per treatment

RAR with early stopping

	ER		GS RAR (variable $n_{\max}$)	
θ	$E(N_1)$	$E(N_0)$	$E(N_1)$	$E(N_0)$
$-\delta/2$	43.2	43.2	41.3	47.7
0	59.3	59.3	64.2	57.5
$\delta/2$	79.8	79.8	99.9	68.2
δ	74.0	74.0	110.3	57.3
$3\delta/2$	55.6	55.6	105.6	39.2

Jennison (2023)

Binary endpoint example

- Aims: **Type 1 error** rate of 5% two sided, and **power** of 90%, assuming success rates of 12% vs. 37% in the 2 groups.
- What does RAR achieve that early stopping cannot?

	Scenario	Prob reject null	$E(N)$	$E(N_1)$	$E(N_0)$	$n - E(N)$	$n/2 - E(N_0)$
GS-RAR	Null	0.048	148.5	74.2	74.3	-12	-6
GS-RAR	Alternative	0.905	81.6	52.5	29.2	54	39
GS-ER	Null	0.050	148.8	74.4	74.4	-13	-6
GS-ER	Alternative	0.930	83.2	41.6	41.6	53	26
ER	Null	0.050	136	68	68	0	0
ER	Alternative	0.950	136	68	68	0	0

Table: Simulated operating characteristics of two adaptive designs against the non-adaptive RCT.

$E(N)$ expected total sample size, $E(N_1)$ expected sample size for ECMO and $E(N_0)$ expected sample size for ACLS.

Group sequential RAR

- Incorporating RAR into a group sequential design can be effective in further reducing the expected number of patients receiving the inferior treatments
- Potential increase in total expected sample size when common variance applies
- These group sequential RAR designs can also be made robust to a time trend and delays in observing patient responses (*Jennison and Turnbull, 2000*)
- Still an open question how best to design and when to use group sequential RAR designs

References for Group sequential RAR

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

Jennison, C. (2023). Group Sequential Designs with Response-Adaptive Randomisation. *Statistical Science* **38(2)** 219–223.

Outline

- Early stopping
  - Early stopping rules
  - Group sequential RAR
- Missing data
  - Impact of missing data
  - Use of mean imputation
- Estimation
  - Bias of the MLE
  - Bias-correction

Missing data and RAR

- Problem with the use of many RAR designs in practice is that the response data required to update the allocation of subsequent patients might be missing:
  - Adverse reactions/Death
  - Loss to follow-up
- Missing data can reduce statistical power and produce biased estimates (if not accounted for)
- Simplest analysis approach is to only use patients that do not have any missing data ('complete case design' analogous to a 'complete case analysis')
- For RAR designs, their (fully) sequential nature and the cumulative impact of the allocation procedure means missing data is not just an analysis issue at the end of a trial, but also a design problem

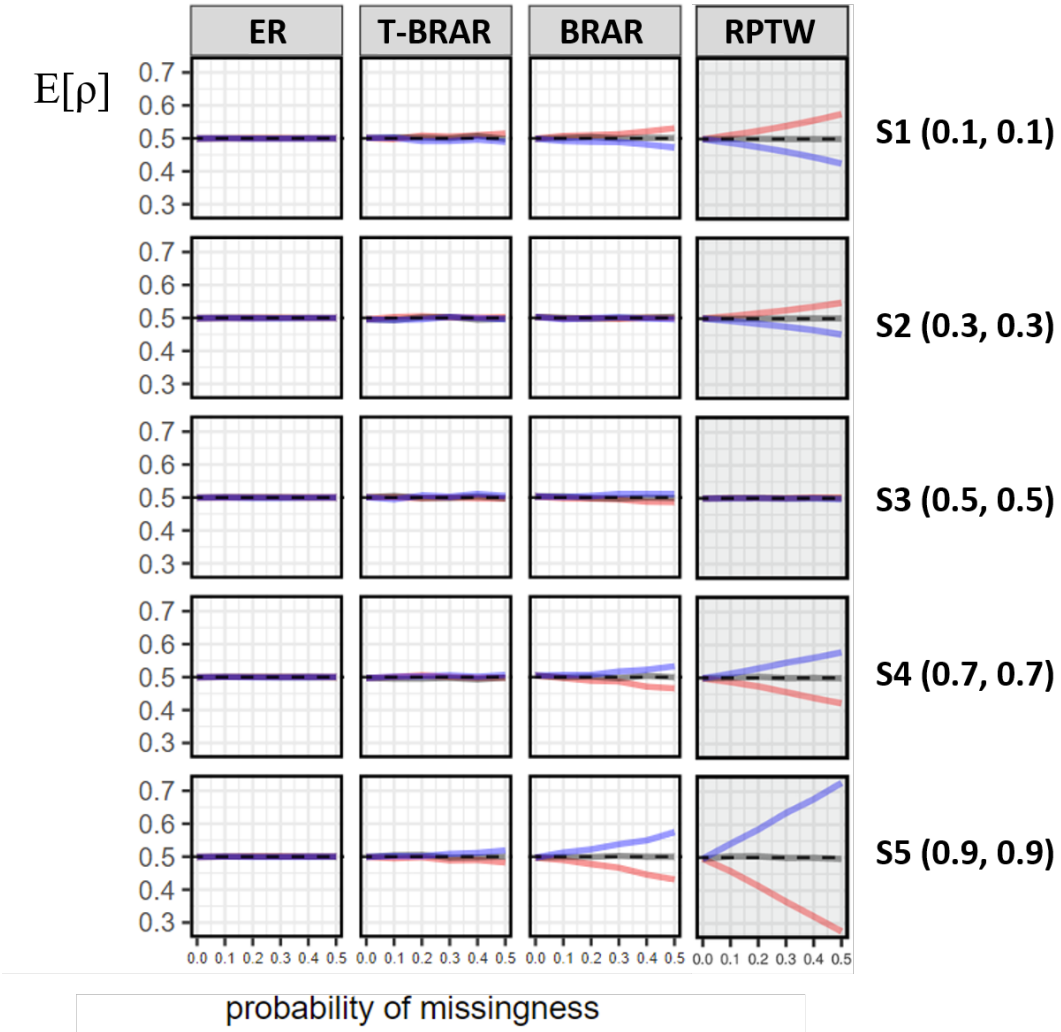
Missing data and RAR framework

- Consider two-armed trial with binary outcomes using RAR
- Assume responses are missing at random (MAR): probability of being missing is the same within groups defined by the observed data (i.e. allocated treatment)
- Focus on effect on proportion of patients assigned to the experimental arm ρ (notation as in Lecture 2)

Missing data and RAR algorithms

- *Chen et al. (2022)* consider the following RAR algorithms:
  - BRAR (see Lecture 1) – both untuned and tuned (T-BRAR)
  - RPTW (see Lecture 1)
- Compare with fixed, equal randomisation (FR)
- Let p_0^m and p_1^m denote the (fixed) probability of missingness in arm 0 (control) and arm 1 (experimental) respectively
- How does differential probabilities of missingness affect $E[\rho]$?

Simulation results under the null ($p_0 = p_1$)

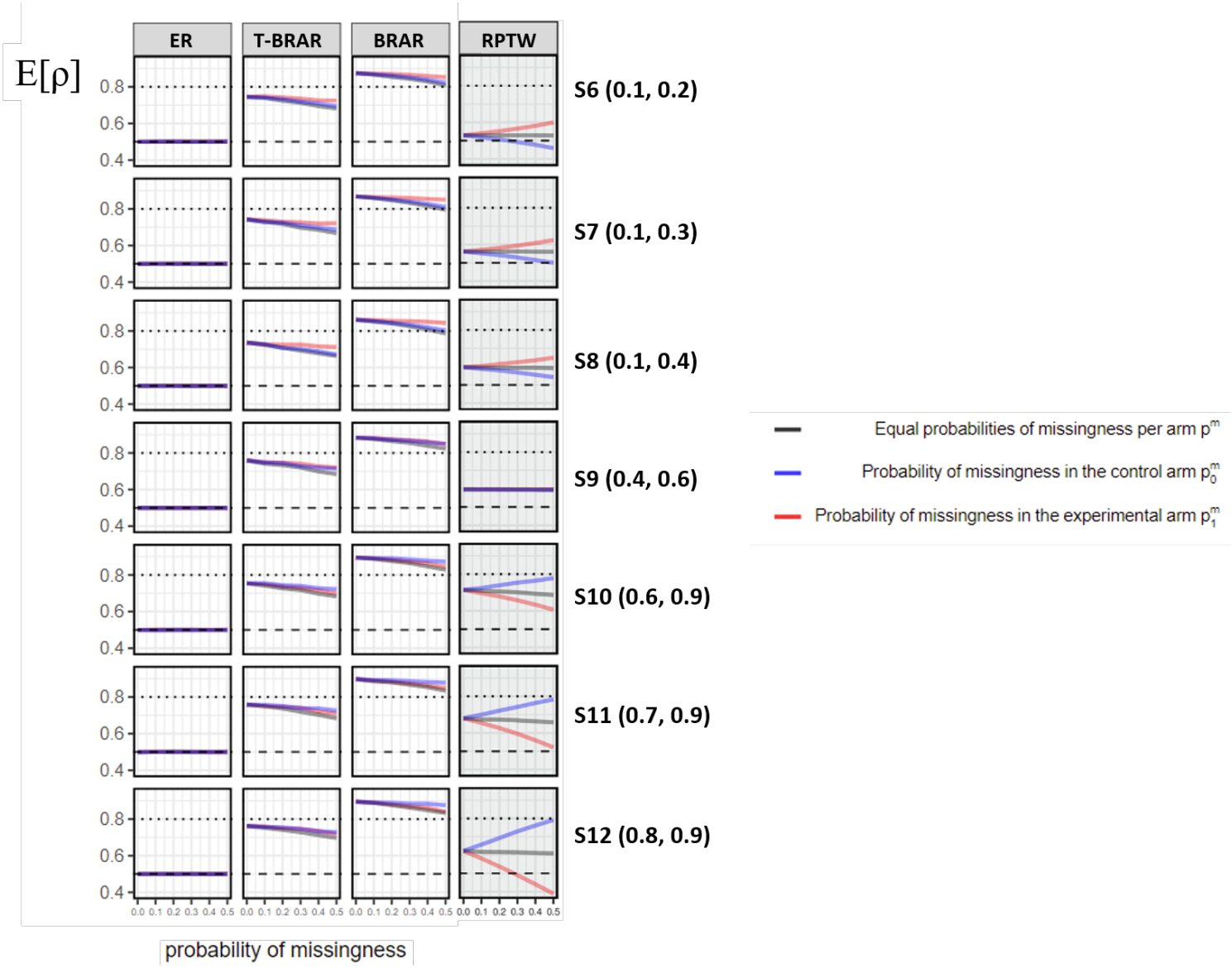


Equal probabilities of missingness per arm p^m

Probability of missingness in the control arm p_0^m

Probability of missingness in the experimental arm p_1^m

Simulation results under the alternative ($p_0 \neq p_1$)



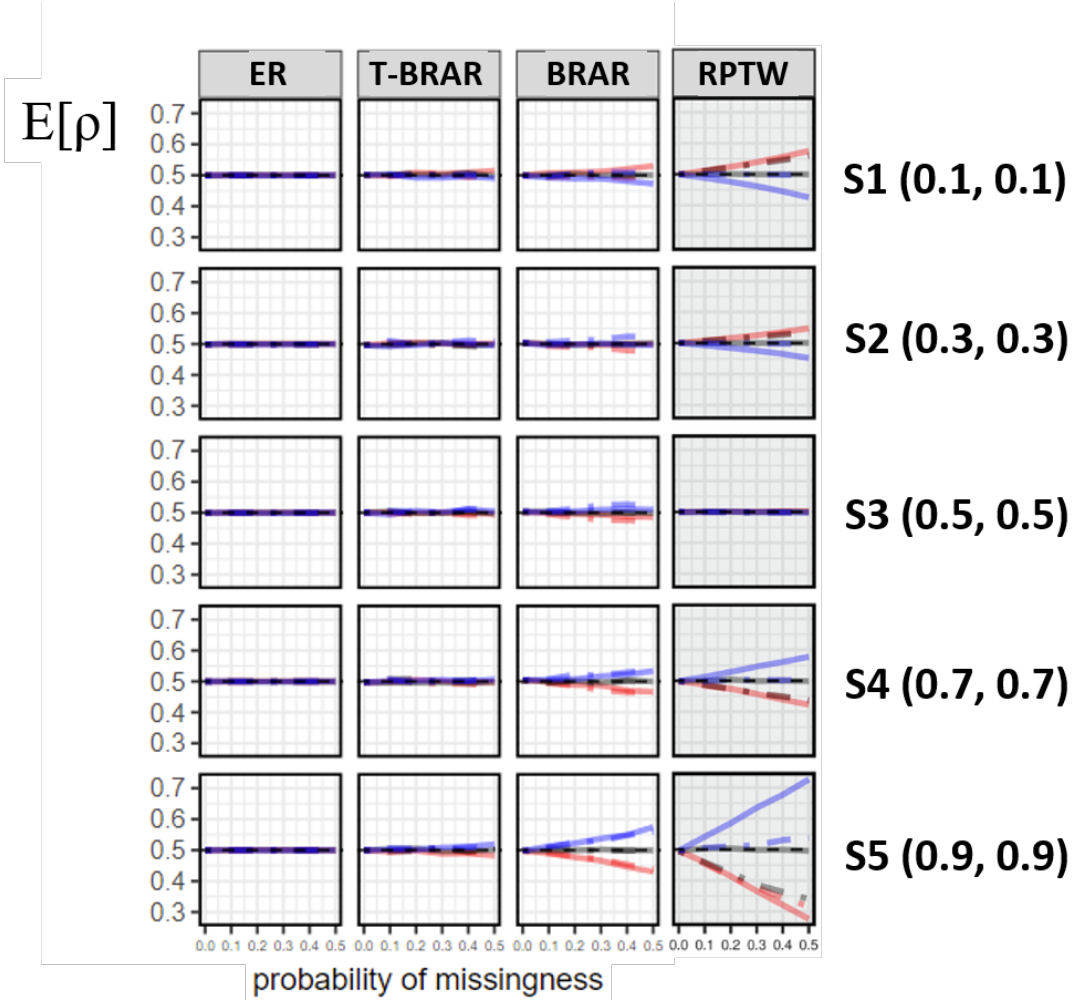
Simulation results

- **Under the null**
  - (T-)BRAR hardly affected by missing data, except for BRAR under high or low null scenarios S1 (0.1, 0.1), S4 (0.7, 0.7), S5 (0.9, 0.9)
  - RPTW is strongly affected by missing data except for S3 (0.5, 0.5). The arm with missing data is less likely to be 'selected', and thus fewer patients are assigned to this arm.
- **Under the alternative**
  - (T-)BRAR is affected by missing data, with fewer patients assigned to the superior arm
  - RPTW is strongly affected by missing data except for S9 (0.4, 0.6). The arm with missing data is less likely to be 'selected', and thus fewer patients are assigned to this arm.

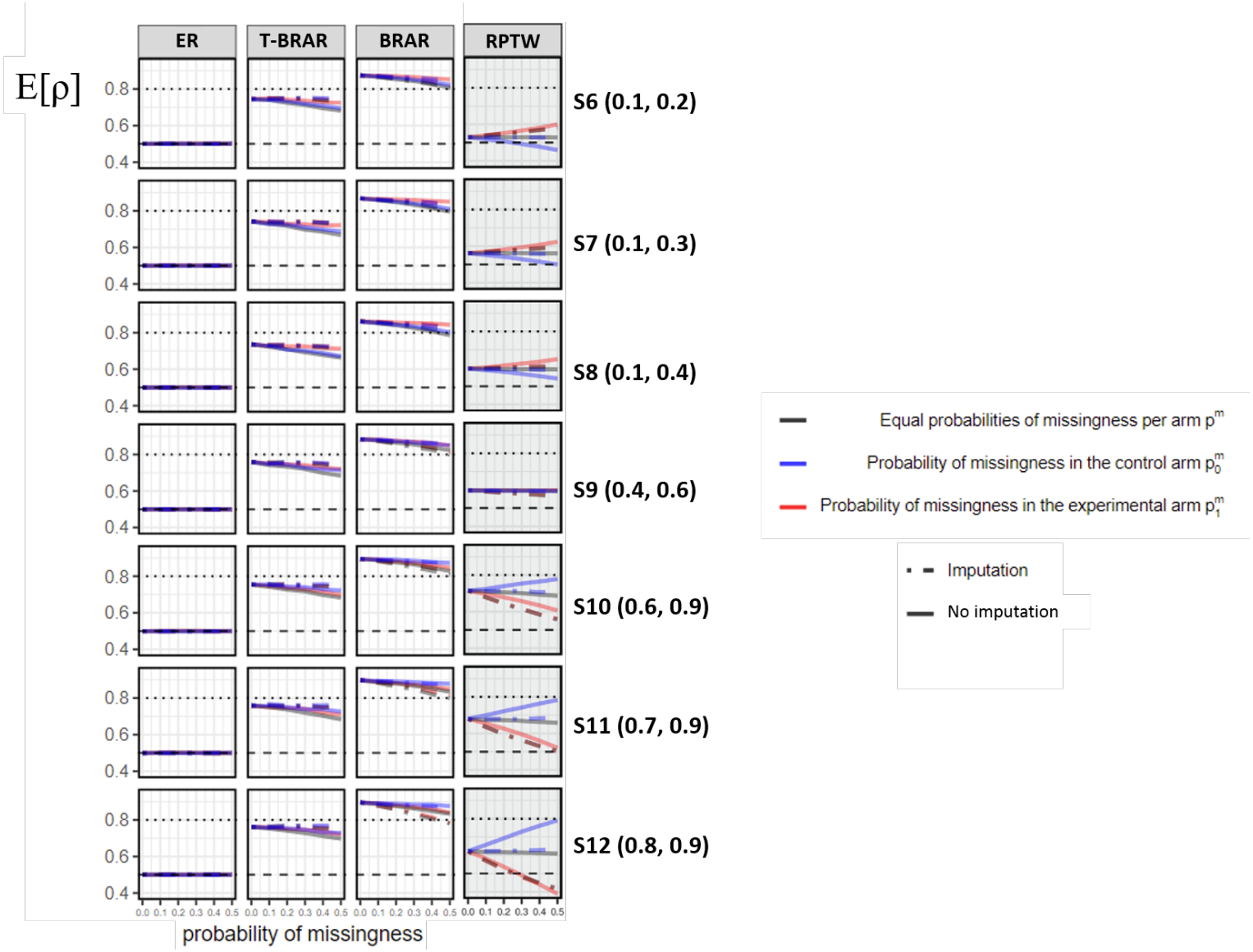
Missing data imputation

- Possible simple solution considered in *Chen et al. (2022)*: single mean imputation to replace missing value with the current mean value of the variable
- More precisely, if there is a missing value (outcome) for patient i on treatment k, replace it with an imputed value from a $Bern(\hat{p}_{i,k})$ where $\hat{p}_{i,k}$ is the current empirical success probability of treatment k

Imputation results under the null



Imputation results under the alternative



Imputation results

- **Under the null**
  - For (T-)BRAR, single mean imputation has little value
  - For RPTW, single mean imputation sometimes helps, but sometimes makes things worse!
- **Under the alternative**
  - For (T-)BRAR, single mean imputation does help mitigate some of the effects of missing data, although not for high or low success probabilities.
  - For RPTW, single mean imputation again sometimes helps, but sometimes makes things worse!

Missing data and RAR

- Missing data can have large impacts on performance of RAR designs
- Magnitude of impact depends on which RAR design is used
- Simple mean imputation does not seem to mitigate the effects well
- Recent work at BSU [Tackney and Villar2025]
  Novel Online Imputation Strategy: possible re-imputation of previously imputed missing responses using updated estimates of success probabilities as the trial.
  Identification of Problematic Scenarios: when treatment arms have very different rates of missingness
- Further research in this area is still needed! progresses.

References for Missing data

Chen, X., May Lee, K., Villar, S.S., and Robertson, D.S. (2021). Some performance considerations when using multi-armed bandit algorithms in the presence of missing data. *PLOS ONE* **17(9)** e0274272.

Tackney, M.S., and Villar, S.S. (2025). Implementing response-adaptive designs when responses are missing: Impute or ignore? *Statistical Methods in Medical Research* **0(0)**. doi:10.1177/09622802251366843.

Outline

- Early stopping
  - Early stopping rules
  - Group sequential RAR
- Missing data
  - Impact of missing data
  - Use of mean imputation
- Estimation
  - Bias of the MLE
  - Bias-correction

Estimation for RAR designs

- The usual/standard estimators, i.e. the maximum likelihood estimators (MLEs), for the parameters of interest for a trial using RAR will typically be biased in small samples
  - Bias of an estimator $\hat{\theta}$ for a parameter θ is given by $E(\hat{\theta}) - \theta$
- This is illustrated for a number of RAR procedures for binary outcomes through simulation in *Villar et al. (2015)* and *Thall et al. (2015)*. Later studied theoretically for some designs.
- Important to distinguish bias induced by early stopping from that induced by the RAR procedure
- Setting without early stopping: *Bowden and Trippa (2007)*
- More general case with early stopping: *Shin et al. (2019)*

Estimation for RAR designs

- Consider RAR trial with K treatment arms and binary outcomes
- Usual MLE for p_k is simply the sample mean
- MLE is consistent (i.e. $\hat{p}_k \to p_k$ as $n \to \infty$) if there is a non-zero probability of being allocated to the treatment during the trial.
- Asymptotic distribution of MLE under a very large class of RAR schemes given in *Hu and Rosenberger (2006)*.

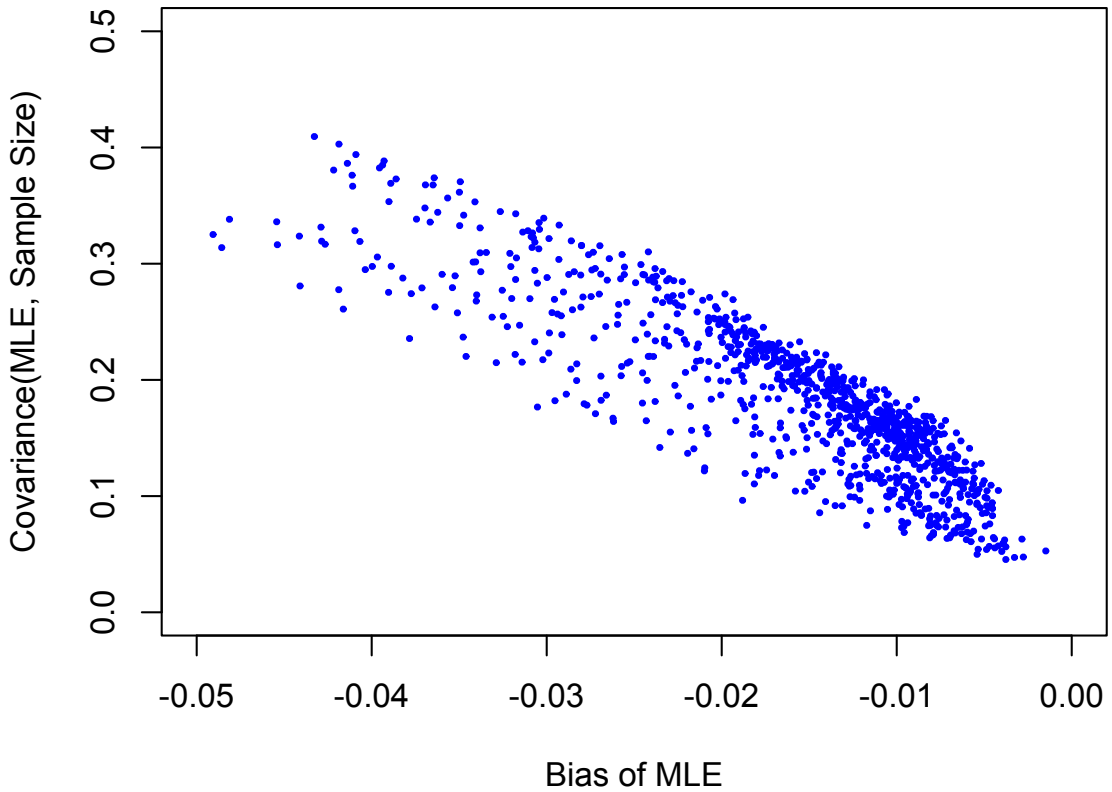
Estimation for RAR designs

- Consider RAR trial with K treatment arms and binary outcomes
- Usual MLE for p_k is simply the sample mean
- MLE is consistent (i.e. $\hat{p}_k \to p_k$ as $n \to \infty$) if there is a non-zero probability of being allocated to the treatment during the trial.
- Asymptotic distribution of MLE under a very large class of RAR schemes given in *Hu and Rosenberger (2006)*.
- RAR leaves asymptotic properties of the MLE intact, but for a finite sample the MLE will be biased in general.
- *Bowden and Trippa (2017)* show that bias of the MLE for finite samples is given by

$$\text{bias}(\hat{p}_k) = -\frac{\text{Cov}(N_k, \hat{p}_k)}{E(N_k)}$$

Bias of the MLE

Example: RPTW rule (see Lecture 1)
$n = 25$, $(p_0, p_1) \in (0.1, 0.9)$
Bias for $\hat{p}_0$



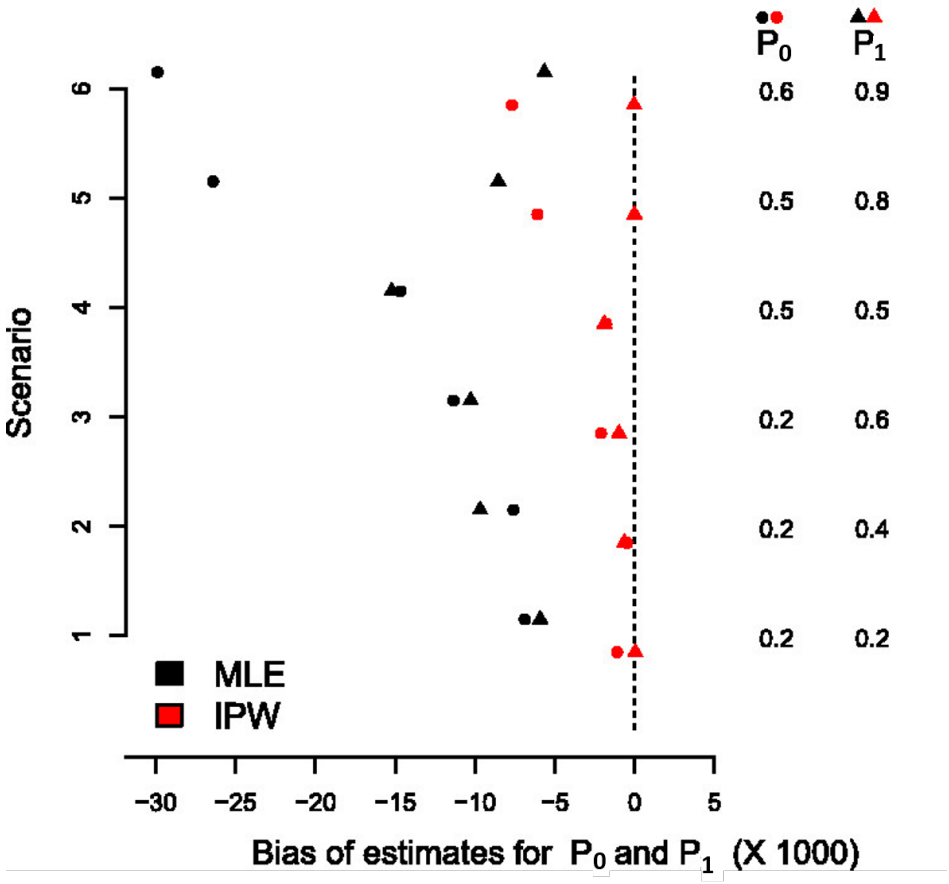
Bias adjusted estimation

Bowden and Trippa (2017) proposed three bias-adjusted estimators:

- **Horvitz-Thompson (HT)**
- **Inverse Probability Weighted (IPW)**
- **Rao-Blackwellised HT (RBHT)**

Bias adjusted estimation

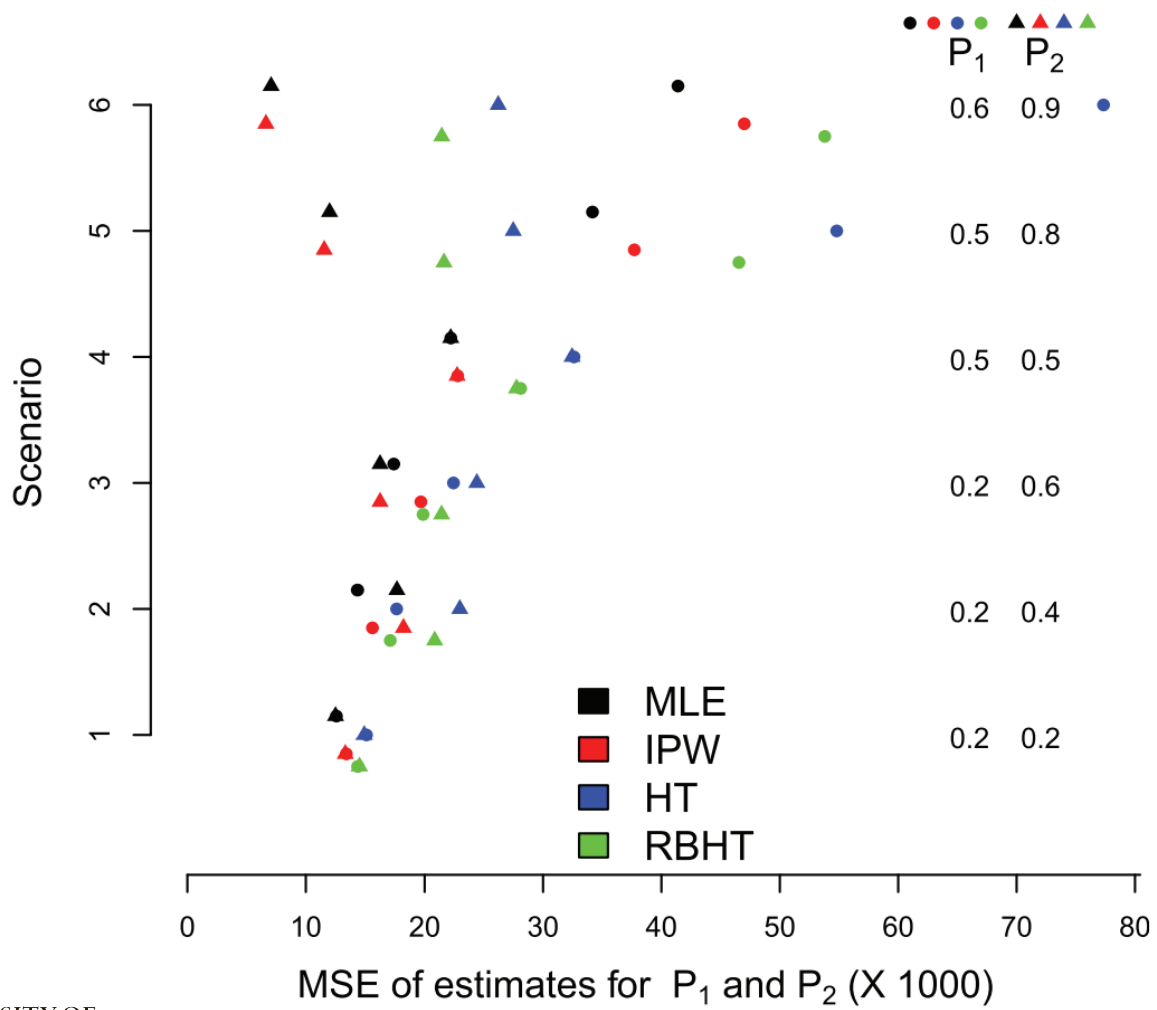
RPTW rule, $n = 25$



Bowden and Trippa (2017)

Bias adjusted estimation

RPTW, $n = 25$. MSE of an estimator $\hat{\theta}$ for a parameter θ is defined as $E[(\hat{\theta} - \theta)^2]$



Bowden and Trippa (2017)

Bias adjusted estimation

- HT estimator is unbiased but has high MSE
- IPW estimator improves on HT estimator but is unstable when weights approach zero
- RBHT is unbiased and has lower MSE than the HT estimator, but can be very computationally intensive
- Note: when estimating the treatment difference the bias can be either negative or positive!
- Confidence intervals: see *Hadad et al. (2021)*

References for Estimation

- Bowden, J. and Trippa, L. (2017). Unbiased estimation for response adaptive clinical trials. *Statistical Methods in Medical Research*. **26(5)** 2376–2388.
- Hadad, V., Hirshberg, D.A., Zhan, R., Wager, S. and Athey, S. (2021). Confidence intervals for policy evaluation in adaptive experiments. *PNAS*. **118(15)** e2014602118.
- Hu, F. and Rosenberger, W.F. (2006). *The Theory of Response-Adaptive Randomization in Clinical Trials*. *Wiley Series in Probability and Statistics*.
- Shin, J., Ramdas, A. and Rinaldo, A. (2015). Are sample means in multi-armed bandits positively or negatively biased? *NeurIPS 2019*
- 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.
- Villar, S.S., Bowden, J. and Wason, J. (2015). Multi-armed Bandit Models for the Optimal Design of Clinical Trials: Benefits and Challenges. *Statistical Science* **30(2)** 199–215.