

Bayesian Response Adaptive Randomization with Point Null Bayesian Hypothesis Testing

Samuel Pawel 

Leonhard Held 

Epidemiology, Biostatistics and Prevention Institute (EBPI)
Center for Reproducible Science and Research Synthesis (CRS)
University of Zurich
`{samuel.pawel,leonhard.held}@uzh.ch`

September 25, 2025

Abstract

Response adaptive randomization methods use accumulated data to adapt randomization probabilities in order to increase the chance of allocating patients to more effective treatments. A popular approach is Thompson sampling, which randomizes patients proportionally to the Bayesian posterior probability that each treatment is the most effective. However, a common problem is that these methods can be highly variable in the early stages of a study, potentially leading to an increased risk of randomizing patients to less effective treatments. We propose a principled approach based on Bayesian hypothesis testing that addresses this issue. Specifically, we introduce a point null hypothesis that postulates equal effectiveness of treatments. This induces shrinkage toward equal randomization probabilities, the amount of which is controlled by the prior probability of the null hypothesis. Equal randomization and Thompson sampling correspond to special cases where the prior probability is set to one or zero, respectively. Simulated and real-world examples illustrate that this approach provides experimenters with a principled approach to balancing highly variable Thompson sampling with static equal randomization. A simulation study demonstrates that the method has comparable statistical properties to Thompson sampling with ad hoc modifications such as power transformation and probability capping. We implement the method in the free and open-source R package `brar`, enabling experiments to easily perform point null Bayesian response adaptive randomization and support more effective randomization of patients.

Keywords: Adaptive trials, A/B testing, Bayes factor, Bayesian model averaging, sharp null hypothesis

1 Introduction

Response adaptive randomization (RAR) methods, sometimes also called outcome adaptive randomization methods, randomly allocate experimental units (e.g., patients or animals) to treatments in a manner that is informed by accumulating data (Thall and Wathen, 2007; Berry et al., 2010; Grieve, 2016; Robertson et al., 2023). A popular approach is Thompson sampling (Thompson, 1933), which randomizes participants proportionally to the Bayesian posterior probability that each treatment is the most effective. RAR methods are attractive because they naturally balance gathering information on treatment effectiveness and assigning subjects to effective treatments.

Despite its benefits, there are also various challenges with Thompson sampling. For example, the method can exhibit high variability, particularly in the early stages of a study when posterior uncertainty is high (Thall et al., 2015). This can lead to erratic allocation probabilities, ethical concerns (e.g., exposing participants to inferior treatments with high probability), and inferential challenges (e.g., reduction of statistical power or biased effect estimates). Consequently, there has been substantial interest in modifying Thompson sampling to make it more reliable.

For example, Thall and Wathen (2007) propose to use the randomization probability $\pi = p^c / \{p^c + (1 - p)^c\}$ where p is the posterior probability that the experimental treatment is more effective than the control treatment, and c is an additional parameter that controls the variability of the method. Setting $c = 1$ produces Thompson sampling, while $c < 1$ reduces variability with $c = 1/2$ being often recommended. Another approach to reduce extreme randomization probabilities is to cap them, for instance, setting them to 0.1 or 0.9 if a method assigns more extreme probabilities (Thall and Wathen, 2007; Lee and Lee, 2021). Finally, RAR methods are often combined with “burn-in” periods at the start of the study, during which units are randomized with equal probabilities to mitigate high variability (Thall and Wathen, 2007; Wathen and Thall, 2017; Robertson et al., 2023).

While such ad hoc modifications can address some of the limitations of RAR methods, they conflict with the principles of coherent Bayesian learning. For example, a transformed or capped posterior probability does no longer correspond to an actual posterior probability, and it cannot be used as a genuine prior for future data. This raises the question of whether it is possible to devise a RAR method with desirable properties, such as reduced variability compared to Thompson sampling, that is coherent with Bayesian principles, and if so, how it relates to these ad hoc modifications.

In this paper, we propose a novel RAR method that reduces variability in a coherent Bayesian manner, which we term “point null Bayesian RAR”. The idea is to consider a point null hypothesis

postulating that treatments are equally effective. This is equivalent to using a “spike-and-slab” prior (Raftery et al., 1997), also known as “lump-and-smear” prior, which is a mixture of a point mass at equal effectiveness and a probability density elsewhere (Spiegelhalter, 2004, Section 5.5.4). The prior probability of the point null hypothesis determines the mixture weight. As we will show, setting this prior probability to zero produces equal randomization, whereas setting it to one produces Thompson sampling. The proposed method thus interpolates between equal randomization and Thompson sampling in a coherent Bayesian way. As a by-product, posterior probabilities and Bayes factors are also obtained. These can be used to monitor evidence of the effectiveness of each treatment, in a manner that is aligned with the randomization probabilities.

In the following Section 2 we introduce the general idea of the method in more detail, followed by tailoring it to the setting of approximately normal effect estimates (Section 3) and binary outcomes (Section 4). In Section 5, we then illustrate the method on data from the infamous ECMO trial (Bartlett et al., 1985), followed by evaluating its statistical properties in a simulation study (Section 6). The paper ends with concluding remarks on advantages, limitations, and opportunities for future research (Section 7). Appendix A illustrates our R package `brar` for performing point null Bayesian RAR.

2 Point null Point null Bayesian RAR

We will now explain the general idea of point null Bayesian RAR, without going into specifics such as data distribution or computation (these will follow in the subsequent sections). Throughout we will assume that we have observed data y and we want to use these to randomize a future experimental unit. We start with the basic but important setting of one control and one treatment group, and extend it afterwards to multiple treatment groups.

2.1 Two group comparisons

In case there is a control group and only one treatment group, we may consider the hypotheses:

H_- : Treatment is less effective than control

H_0 : Treatment and control are equally effective

H_+ : Treatment is more effective than control

How exactly these statements are translated into statistical hypotheses related to parameters depends on the type of data and model used, but often relates to an effect size parameter being less, equal, or greater than zero. There may also be situations where there is no control group but only two competing treatments. In this case, the methods detailed here are still applicable but with the control group replaced by a reference group (the choice of the reference may be somewhat arbitrary).

The Bayesian posterior probability of a hypothesis $H_i \in \{H_-, H_0, H_+\}$ can then be calculated by

$$\Pr(H_i | y) = \frac{p(y | H_i) \Pr(H_i)}{\sum_{j \in \{-, 0, +\}} p(y | H_j) \Pr(H_j)} = \left\{ \sum_{j \in \{-, 0, +\}} \text{BF}_{ji}(y) \frac{\Pr(H_j)}{\Pr(H_i)} \right\}^{-1} \quad (1)$$

where $\Pr(H_i)$ is the prior probability of hypothesis H_i ,

$$p(y | H_i) = \int_{\Theta} p(y | \theta) p(\theta | H_i) d\theta$$

is the marginal likelihood of the data y under H_i obtained from marginalizing the likelihood $p(y | \theta)$ with respect to the prior distribution $p(\theta | H_i)$ assigned to the model parameters $\theta \in \Theta$ under H_i , and

$$\text{BF}_{ji}(y) = \frac{\Pr(H_j | y)}{\Pr(H_i | y)} \bigg/ \frac{\Pr(H_j)}{\Pr(H_i)} = \frac{p(y | H_j)}{p(y | H_i)} \quad (2)$$

is the Bayes factor contrasting H_j to H_i (Jeffreys, 1939; Good, 1958; Kass and Raftery, 1995). The Bayes factor (2) is the updating factor of the prior odds of H_j to H_i to the corresponding posterior odds (first equality), which is equivalent to the ratio of marginal likelihoods of the data under H_j and H_i (second equality). The posterior probabilities (1) can thus be computed from the marginal likelihoods of the data under each considered hypotheses along with their prior probabilities, or from the set of Bayes factors and prior hypothesis odds.

Regardless in which way they are computed, the question is how to translate posterior probabilities into randomization probabilities. We propose to randomize a future unit to the treatment group with probability

$$\pi = \frac{\Pr(H_0 | y)}{2} + \Pr(H_+ | y), \quad (3)$$

while the probability to randomize to the control group is consequently

$$1 - \pi = \frac{\Pr(H_0 | y)}{2} + \Pr(H_- | y). \quad (4)$$

Figure 1 shows such randomization probabilities for different combinations of posterior probabilities. We can see that the randomization probability π shrinks towards 50% as the posterior probability of the null hypothesis $\Pr(H_0 | y)$ increases.

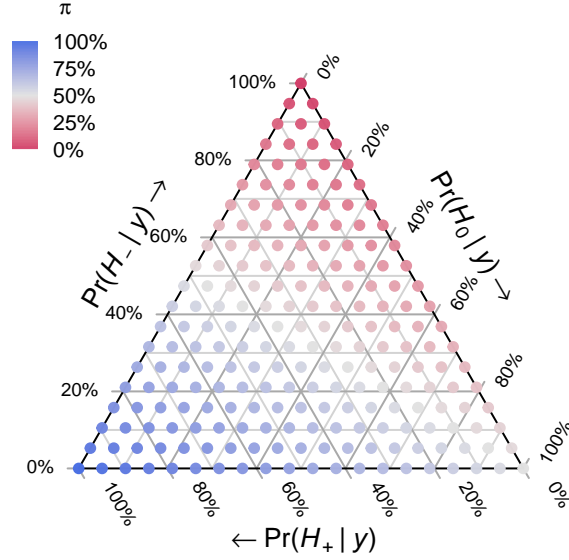


Figure 1: Ternary plot where each point depicts the treatment randomization probability π for a particular combination of posterior probabilities of H_- , H_0 , and H_+ .

This scheme can be motivated as Bayesian hypothesis averaged randomization probability where the hypothesis-specific randomization probabilities are 0, $1/2$, and 1, under H_- , H_0 , and H_+ , respectively. That is, if we would know that H_+ or H_- is true, we should assign the next patient to the treatment or control group, respectively, with probability one to maximize utility (patient benefit). On the other hand, if we knew that H_0 is true, randomizing assignment with probability $1/2$ is sensible.

Interestingly, the randomization scheme (3) reduces to Thompson sampling when the prior probability of H_0 is set to zero since then $\Pr(H_0 | y) = 0$ regardless of the data, and consequently $\pi = \Pr(H_+ | y)$. However, the scheme induces shrinkage toward randomization probabilities of $\pi = 1/2$ otherwise. In the most extreme case when $\Pr(H_0) = 1$, equal randomization is obtained as then $\Pr(H_0 | y) = 1$ regardless of the data, and consequently $\pi = 1/2$. The scheme thus interpolates between Thompson sampling and equal randomization in a coherent Bayesian way.

2.2 More than two groups

Suppose there are $K > 1$ treatment groups in addition to the control group. In this case, we may modify the procedure and consider the hypotheses:

H_- : All treatments are less effective than control

H_0 : All treatments are equally effective as control

H_{+1} : Treatment 1 is more effective than control and all other treatments

\vdots

H_{+K} : Treatment K is more effective than control and all other treatments

Posterior probabilities of each hypothesis can be computed from the marginal likelihoods and prior hypotheses probabilities with the summation in (1) extended to encompass all hypotheses (i.e., summing over $j \in \{-, 0, +1, \dots, +K\}$). Similarly, they can be translated into randomization probabilities

$$\pi_i = \frac{\Pr(H_0 | y)}{K + 1} + \Pr(H_{+i} | y)$$

with the corresponding control randomization probability

$$1 - \sum_{i=1, \dots, K} \pi_i = \frac{\Pr(H_0 | y)}{K + 1} + \Pr(H_- | y),$$

which reduce to the randomization probabilities (3) and (4) for $K = 1$.

Also in the multi-treatment case, the randomization probabilities are shrunk towards equal randomization $\pi = 1/(K + 1)$ by introducing the null hypothesis H_0 . Similarly, Thompson sampling and equal randomization are obtained by setting $\Pr(H_0) = 0$ and $\Pr(H_0) = 1$, respectively, since then the posteriors $\Pr(H_0 | y) = 0$ and $\Pr(H_0 | y) = 1$ are obtained for any observed data y .

3 Point null Bayesian RAR under approximate normality

We will now develop the aforementioned ideas of point null Bayesian RAR for the setting where the data are summarized by an asymptotically normally distributed effect estimate. This does not mean that the raw data (e.g., a vector of outcomes and a matrix of covariates) from which the estimate is computed have to be normally distributed. For example, regression coefficients from generalized

linear models estimated with maximum likelihood satisfy asymptotic normality. This is also useful because computation of posterior and randomization probabilities can be performed efficiently without the need for simulation. While this framework is widely applicable, improvements may be possible by considering the exact distribution of the data, which will be detailed for binary outcomes in Section 4.

3.1 Two group comparisons

Assume that the data are summarized by $y = \{\hat{\theta}, \sigma\}$, where $\hat{\theta}$ is an effect estimate of the true treatment effect θ that quantifies the effect of a treatment over the control and σ is the standard error of the estimate. For example, $\hat{\theta}$ could be an estimated (standardized) mean difference, log odds/rate/hazard ratio, risk difference, or regression coefficient. Suppose further that the estimate is (at least approximately) normally distributed, i.e., $\hat{\theta} \mid \theta \sim N(\theta, \sigma^2)$.

If the effect θ is oriented such that a positive effect indicates treatment benefit, the three hypotheses from Section 2.1 translate into:

$$H_- : \theta < 0 \quad \text{versus} \quad H_0 : \theta = 0 \quad \text{versus} \quad H_+ : \theta > 0$$

The point null hypothesis H_0 is a simple hypothesis with no free parameters, or equivalently, a point (Dirac) mass at zero. The H_- and H_+ hypotheses are composite hypotheses and require prior distributions for the effect θ . A natural choice is a $\theta \sim N(\mu, \tau^2)$ normal distribution whose support is truncated to the negative and positive side, respectively. This distribution may be specified based on prior knowledge (e.g., previous studies). In the absence of prior knowledge, it seems sensible to center the prior on zero ($\mu = 0$) to represent clinical equipoise, as a zero-centered prior gives equal probability to harmful and beneficial effects. Averaging the prior over the three hypotheses leads to a spike-and-slab prior, as illustrated in Figure 2.

To compute posterior hypothesis probabilities, specification of prior hypothesis probabilities is required. The prior probability of the null hypothesis $\Pr(H_0)$ represents the a priori plausibility of an absent effect, and also acts as a tuning parameter that controls the degree of variability of the adaptive randomization. An intuitive default is $\Pr(H_0) = 0.5$ as it represents the equipoise position of equal probability of an absent effect relative to a present (either harmful or beneficial) effect. For a given $\Pr(H_0)$, it is then natural to set the prior probabilities of the other two hypotheses to $\Pr(H_+) = \{1 - \Pr(H_0)\} \times \Phi(\mu/\tau)$ and $\Pr(H_-) = \{1 - \Pr(H_0)\} \times \Phi(-\mu/\tau)$ so that to the prior

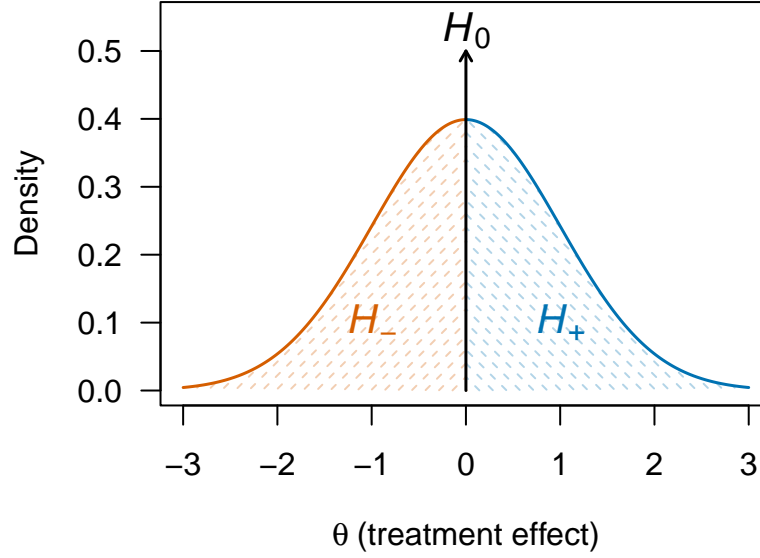


Figure 2: Illustration of spike-and-slab prior for the effect θ . A point mass prior at 0 is assumed under H_0 . A normal prior $\theta \sim N(0, 1)$ with support truncated to the positive or negative side is assumed under H_+ and H_- , respectively. These priors are averaged assuming prior hypothesis probabilities $\Pr(H_0) = 0.5$, $\Pr(H_+) = 0.25$, and $\Pr(H_-) = 0.25$.

distribution averaged over H_- and H_+ is again the $N(\mu, \tau^2)$ normal distribution that was truncated in the first place. For example, when setting $\Pr(H_0) = 0.5$ and specifying a zero-centered prior ($\mu = 0$) as in Figure 2, we obtain $\Pr(H_+) = \Pr(H_-) = 0.5 \times 0.5 = 0.25$.

With the prior densities and prior hypothesis probabilities specified, we can compute the marginal likelihood of the observed effect estimate under each hypothesis, and in turn obtain posterior probabilities (1). In the conjugate normal likelihood and prior framework, all of them can be straightforwardly derived in closed-form. Denoting by $N(x \mid m, v)$ the normal density function with mean m and variance v evaluated at x , the marginal likelihoods are given by

$$p(\hat{\theta} \mid H_-) = N(\hat{\theta} \mid \mu, \sigma^2 + \tau^2) \times \frac{\Phi(-\mu_*/\tau_*)}{\Phi(-\mu/\tau)} \quad (5a)$$

$$p(\hat{\theta} \mid H_0) = N(\hat{\theta} \mid 0, \sigma^2) \quad (5b)$$

$$p(\hat{\theta} \mid H_+) = N(\hat{\theta} \mid \mu, \sigma^2 + \tau^2) \times \frac{\Phi(\mu_*/\tau_*)}{\Phi(\mu/\tau)} \quad (5c)$$

with posterior mean and variance

$$\mu_* = \frac{\hat{\theta}/\sigma^2 + \mu/\tau^2}{1/\sigma^2 + 1/\tau^2} \quad \text{and} \quad \tau_*^2 = \frac{1}{1/\sigma^2 + 1/\tau^2}.$$

Taking ratios of marginal likelihoods produces the Bayes factors

$$\begin{aligned} \text{BF}_{+0}(\hat{\theta}) &= \exp \left[-\frac{1}{2} \left\{ \frac{(\hat{\theta} - \mu)^2}{\sigma^2 + \tau^2} - \frac{\hat{\theta}^2}{\sigma^2} \right\} \right] \times \frac{\Phi(\mu_*/\tau_*)}{\Phi(\mu/\tau)} \bigg/ \sqrt{1 + \frac{\tau^2}{\sigma^2}} \\ \text{BF}_{+-}(\hat{\theta}) &= \frac{\Phi(\mu_*/\tau_*)}{\Phi(\mu/\tau)} \bigg/ \frac{\Phi(-\mu_*/\tau_*)}{\Phi(-\mu/\tau)}, \end{aligned}$$

and the Bayes factors for other hypothesis comparisons can be obtained by transitivity and reciprocity, for example, $\text{BF}_{-0} = \text{BF}_{+0} / \text{BF}_{+-}$. Posterior probabilities can now be obtained by plugging the Bayes factors and prior odds into (1), leading to

$$\begin{aligned} \Pr(H_- | \hat{\theta}) &= \left(1 + \frac{\Phi(\mu_*/\tau_*)}{\Phi(-\mu_*/\tau_*)} + \frac{\Pr(H_0)}{1 - \Pr(H_0)} \times \exp \left[-\frac{1}{2} \left\{ \frac{\hat{\theta}^2}{\sigma^2} - \frac{(\hat{\theta} - \mu)^2}{\sigma^2 + \tau^2} \right\} \right] \times \frac{\sqrt{1 + \tau^2/\sigma^2}}{\Phi(-\mu_*/\tau_*)} \right)^{-1} \\ \Pr(H_0 | \hat{\theta}) &= \left(1 + \frac{1 - \Pr(H_0)}{\Pr(H_0)} \times \exp \left[-\frac{1}{2} \left\{ \frac{(\hat{\theta} - \mu)^2}{\sigma^2 + \tau^2} - \frac{\hat{\theta}^2}{\sigma^2} \right\} \right] \bigg/ \sqrt{1 + \frac{\tau^2}{\sigma^2}} \right)^{-1} \\ \Pr(H_+ | \hat{\theta}) &= \left(1 + \frac{\Phi(-\mu_*/\tau_*)}{\Phi(\mu_*/\tau_*)} + \frac{\Pr(H_0)}{1 - \Pr(H_0)} \times \exp \left[-\frac{1}{2} \left\{ \frac{\hat{\theta}^2}{\sigma^2} - \frac{(\hat{\theta} - \mu)^2}{\sigma^2 + \tau^2} \right\} \right] \times \frac{\sqrt{1 + \tau^2/\sigma^2}}{\Phi(\mu_*/\tau_*)} \right)^{-1}. \end{aligned}$$

These Bayes factors and posterior probabilities can be monitored as data accumulate to see how the evidence for the hypotheses change. Moreover, they can be used to define the randomization probabilities via (3) leading to

$$\begin{aligned} \pi &= \frac{\Pr(H_0 | y)}{2} + \Pr(H_+ | y) \\ &= \left(1 + \frac{1 - \Pr(H_0)}{\Pr(H_0)} \times \exp \left[-\frac{1}{2} \left\{ \frac{(\hat{\theta} - \mu)^2}{\sigma^2 + \tau^2} - \frac{\hat{\theta}^2}{\sigma^2} \right\} \right] \bigg/ \sqrt{1 + \frac{\tau^2}{\sigma^2}} \right)^{-1} \bigg/ 2 \\ &\quad + \left(1 + \frac{\Phi(-\mu_*/\tau_*)}{\Phi(\mu_*/\tau_*)} + \frac{\Pr(H_0)}{1 - \Pr(H_0)} \times \exp \left[-\frac{1}{2} \left\{ \frac{\hat{\theta}^2}{\sigma^2} - \frac{(\hat{\theta} - \mu)^2}{\sigma^2 + \tau^2} \right\} \right] \times \frac{\sqrt{1 + \tau^2/\sigma^2}}{\Phi(\mu_*/\tau_*)} \right)^{-1}. \quad (6) \end{aligned}$$

As expected, the randomization probability (6) approaches equal randomization as the prior probability of H_0 increases to one (i.e., $\pi \rightarrow 1/2$ as $\Pr(H_0) \nearrow 1$), while it approaches the ordinary Bayesian posterior tail probability of $\theta > 0$ based on a normal prior $\theta \sim N(\mu, \tau^2)$ as the prior probability of H_0 decreases to zero (i.e., $\pi \rightarrow \Phi(\mu_*/\tau_*)$ as $\Pr(H_0) \searrow 0$).

Figure 3 shows sequences of randomization probabilities computed from simulated normal data. We can see that the probability to randomize to the treatment group based on $\Pr(H_0) = 1$ remains static at 50% (yellow line). In contrast, the randomization probabilities based on $\Pr(H_0) < 1$ tend towards 100% as more data accumulate, as expected, given that the data were simulated under a

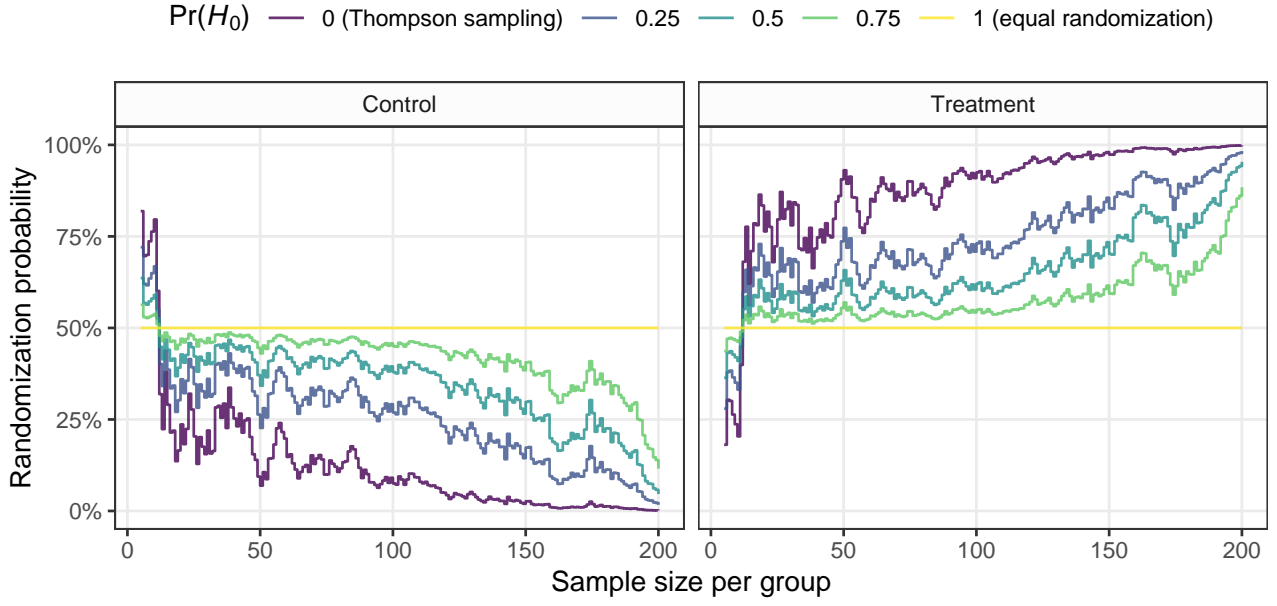


Figure 3: Evolution of Bayesian RAR probabilities for one treatment and control group. In each step one observation from the control and one from the treatment group are simulated from a normal distribution with a true standard deviation of 1 and assuming a true mean difference $\theta = 0.25$. Randomization probabilities are then computed assuming a normal spike-and-slab prior centered at zero with standard deviation $\tau = 1$, and different prior probabilities $\Pr(H_0)$.

beneficial treatment effect. Moreover, a clear ordering is visible: Probabilities under $\Pr(H_0) = 0$, which corresponds to Thompson sampling, are the most extreme and may even go strongly in the “wrong” direction. For example, the probability to randomize to control is higher than 75% at some point during the earlier stages of the study. In contrast, probabilities based on $0 < \Pr(H_0) < 1$ show the same qualitative behavior but are less extreme. Setting a higher prior probability $\Pr(H_0)$ thus reduces the variability of randomization probabilities but also makes convergence to a probability of 100% for the more effective treatment slower.

3.2 More than two groups

Suppose now there are $K > 1$ treatment groups and consequently K effect estimates, each estimate quantifying the effect of the corresponding treatment relative to the control. A natural generalization is to stack the estimates into a vector $\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_K)^\top$ and assume an approximate K -variate normal distribution $\hat{\theta} \mid \theta \sim N_K(\theta, \Sigma)$, where $\theta = (\theta_1, \dots, \theta_K)^\top$ is the vector of true effects and Σ is the covariance matrix of $\hat{\theta}$. For example, $\hat{\theta}$ could be a vector of estimated regression coefficients and Σ the corresponding covariance matrix.

The hypotheses from Section 2.2 then translate into

$$\begin{aligned}
H_- &: \theta_i < 0, i = 1, \dots, K \\
H_0 &: \theta_i = 0, i = 1, \dots, K \\
H_{+1} &: \theta_1 > 0 \text{ and } \theta_1 > \theta_i, i = 2, \dots, K \\
&\vdots \\
H_{+K} &: \theta_K > 0 \text{ and } \theta_K > \theta_i, i = 1, \dots, K - 1
\end{aligned}$$

All hypotheses apart from H_0 are composite and require the specification of a prior distribution. In analogy to the one treatment case, we specify a K -variate normal prior $\theta \sim N_K(\mu, \mathcal{T})$ and truncate its support to the region of corresponding hypothesis (i.e., for hypothesis H_{+i} , the space in \mathbb{R}^K where the i th component is greater than zero and all other components). Similarly, the prior hypothesis probabilities $\Pr(H_{+i})$ for $i = 1, \dots, K$ may again be specified so that the $N_K(\mu, \mathcal{T})$ distribution is recovered when the prior is averaged over the hypotheses.

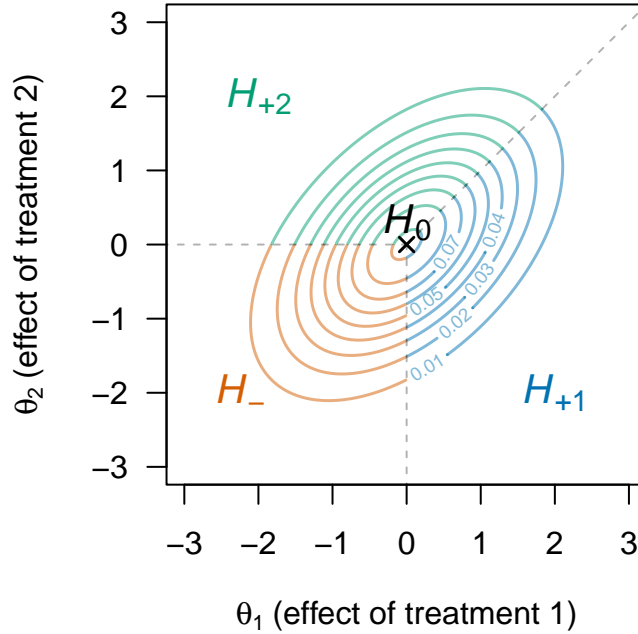


Figure 4: Illustration of a spike-and-slab prior for a two-dimensional effect $\theta = (\theta_1, \theta_2)^\top$. A point mass prior at $(0,0)^\top$ is assumed under H_0 . A normal prior $\theta \sim N((0,0)^\top, \mathcal{T})$ with $\mathcal{T}_{ij} = 0.5$ for $i \neq j$ and $\mathcal{T}_{ij} = 1$ for $i = j$, with support truncated to the space of the corresponding hypothesis is assumed under H_- , H_{+1} , and H_{+2} . The correlation ensures that all treatments receive equal prior probability $\Pr(H_-) = \Pr(H_{+1}) = \Pr(H_{+2}) = \{1 - \Pr(H_0)\}/3$.

Figure 4 illustrates such a spike-and-slab prior for the two treatment groups scenario ($K = 2$). Specifying a prior covariance matrix \mathcal{T} with uniform correlation of $1/2$ ensures that the prior prob-

abilities of all hypotheses but H_0 are equal, which seems a sensible default. This can also be motivated by the fact that for normal outcomes with a shared control group and equal allocation, mean difference effect estimates are correlated by a factor $1/2$ due to the common control group for all treatments.

As in the two-group case (Section 3.1), the normal-normal conjugate framework allows us to derive marginal likelihoods in closed-form. The marginal likelihood under H_0 is

$$p(\hat{\boldsymbol{\theta}} \mid H_0) = N_K(\hat{\boldsymbol{\theta}} \mid \mathbf{0}, \boldsymbol{\Sigma}),$$

while the marginal likelihood under H_- is

$$p(\hat{\boldsymbol{\theta}} \mid H_-) = N_K(\hat{\boldsymbol{\theta}} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma} + \boldsymbol{\mathcal{T}}) \times \frac{\Phi_K(\mathbf{0} \mid \boldsymbol{\mu}_*, \boldsymbol{\mathcal{T}}_*)}{\Phi_K(\mathbf{0} \mid \boldsymbol{\mu}, \boldsymbol{\mathcal{T}})}$$

with $N_K(\boldsymbol{x} \mid \boldsymbol{m}, \boldsymbol{V})$ and $\Phi_K(\boldsymbol{x} \mid \boldsymbol{m}, \boldsymbol{V})$ the density and cumulative distribution functions of the K -variate normal distribution with mean vector \boldsymbol{m} and covariance matrix \boldsymbol{V} evaluated at \boldsymbol{x} , and posterior mean $\boldsymbol{\mu}_* = (\boldsymbol{\Sigma}^{-1} + \boldsymbol{\mathcal{T}}^{-1})^{-1}(\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\theta}} + \boldsymbol{\mathcal{T}}^{-1}\boldsymbol{\mu})$ and covariance $\boldsymbol{\mathcal{T}}_* = (\boldsymbol{\Sigma}^{-1} + \boldsymbol{\mathcal{T}}^{-1})^{-1}$. Finally, the marginal likelihood under H_{+i} is given by

$$p(\hat{\boldsymbol{\theta}} \mid H_{+i}) = N_K(\hat{\boldsymbol{\theta}} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma} + \boldsymbol{\mathcal{T}}) \times \frac{\Phi_K(\mathbf{0} \mid \boldsymbol{A}_{i,K}\boldsymbol{\mu}_*, \boldsymbol{A}_{i,K}\boldsymbol{\mathcal{T}}_*\boldsymbol{A}_{i,K}^\top)}{\Phi_K(\mathbf{0} \mid \boldsymbol{A}_{i,K}\boldsymbol{\mu}, \boldsymbol{A}_{i,K}\boldsymbol{\mathcal{T}}\boldsymbol{A}_{i,K}^\top)}$$

where $\boldsymbol{A}_{i,K}$ is a $K \times K$ contrast matrix that maps $\boldsymbol{\theta}$ to the space where the negative orthant corresponds to the space of hypothesis H_{+i} , for example, for $i = 2$ and $K = 3$, the matrix is

$$\boldsymbol{A}_{2,3} = \begin{bmatrix} 0 & -1 & 0 \\ 1 & -1 & 0 \\ 0 & -1 & 1 \end{bmatrix}$$

with the first row encoding the constraint of θ_1 being positive, and the second and third rows encoding the constraints of θ_1 being larger than θ_2 and θ_3 , respectively. As expected, for $K = 1$, these marginal likelihoods reduce to (5). Moreover, since they are all available in closed-form, Bayes factors, posterior probabilities, and randomization probabilities are also available in closed-form. Crucially, only the exact evaluation of multivariate normal densities and cumulative distribution functions is required, and both are efficiently implemented in statistical software (e.g., in the `mvtnorm` R

package, [Genz and Bretz, 2009](#)). This thus leads to an efficient Bayesian RAR method that allows experimenters to compute randomization probabilities in complex settings, for instance, multiple regression, where a full Bayesian analysis may involve various complexities, such as priors for nuisance parameters and Markov chain Monte Carlo for the computation of posteriors.

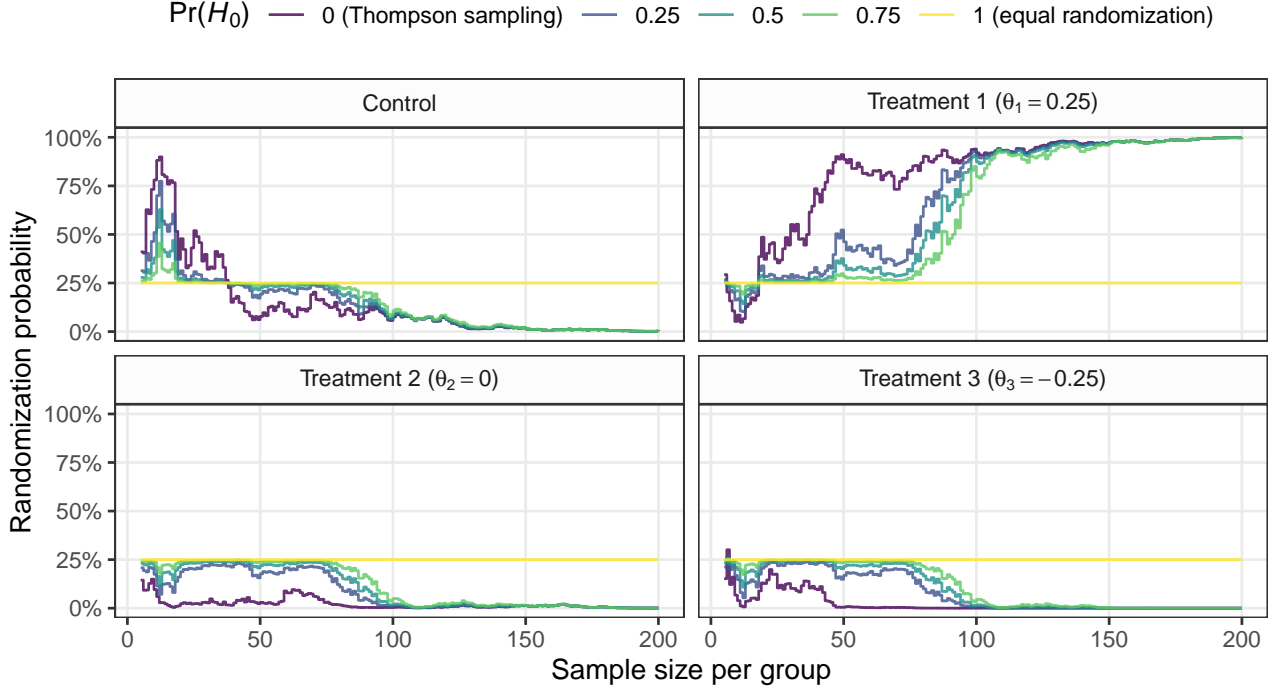


Figure 5: Evolution of Bayesian RAR probabilities for 3 treatments and a control group. In each step, one observation from each group is simulated from a normal distribution with a standard deviation of 1 and corresponding mean differences as indicated in the panel titles (treatment 1 is the most effective treatment). Randomization probabilities are computed using a normal spike-and-slab prior centered at the origin and with covariance matrix \mathcal{T} with $\mathcal{T}_{ij} = 0.5$ for $i \neq j$ and $\mathcal{T}_{ij} = 1$ for $i = j$, and for different prior probabilities $\Pr(H_0)$.

Figure 5 shows sequences of randomization probabilities computed from simulated normal data with $K = 3$ treatment groups. We see again that assigning a prior probability $\Pr(H_0) = 1$ leads to static equal randomization at $\pi_i = 1/(K + 1) = 0.25$, while assigning $\Pr(H_0) = 0$ (Thompson sampling) leads to the most variable randomization probabilities. Since data are simulated assuming that treatment 1 is the most effective, randomization probabilities based on $\Pr(H_0) < 1$ converge towards $\pi_1 = 100\%$ and toward 0% for the remaining treatments. While convergence is the fastest for $\Pr(H_0) = 0$, this prior probability also accidentally produces rather high randomization probabilities for the control group at the start of the study, which is less pronounced for positive prior probabilities $\Pr(H_0) > 0$.

4 Point null Bayesian RAR for binary outcomes

We will now consider the setting with binary outcomes, as such outcomes frequently occur in applications of Bayesian RAR (e.g., in clinical trials or A/B testing settings). Suppose that we observe data of the form $y = \{y_C, y_1, \dots, y_K, n_C, n_1, \dots, n_K\}$ where y_i denotes the number of successes out of n_i trials in group $i \in \{C, 1, \dots, K\}$ coming from a control group (index C) and K treatment groups. All success counts are assumed to be binomially distributed with probabilities $\theta_C, \theta_1, \dots, \theta_K$, respectively, and higher values are assumed to indicate a higher benefit (e.g., a higher probability of disease recovery). The hypotheses of interest from Section 2.2 then translate into

$$\begin{aligned} H_- : \theta_C &> \theta_i, i \in \{1, \dots, K\} \\ H_0 : \theta_C &= \theta_1 = \dots = \theta_K \\ H_{+1} : \theta_1 &> \theta_i, i \in \{C, 2, \dots, K\} \\ &\vdots \\ H_{+K} : \theta_K &> \theta_i, i \in \{C, 1, \dots, K-1\} \end{aligned}$$

Under the previously considered approximate normal framework, these hypotheses could be translated into hypotheses related to log odds ratios $\psi_i = \log\{\theta_i(1 - \theta_C)\} / \{(1 - \theta_i)\theta_C\}$, which can be estimated with logistic regression or other methods (we will provide a comparison with this approach below). However, such normal approximations can be inaccurate for small sample sizes and/or extreme probabilities close to zero/one. It is therefore preferable to compute randomization probabilities via the exact binomial distribution.

To compute posterior and randomization probabilities, it is necessary to compute the marginal likelihood of the observed data y under the different hypotheses. The null hypothesis H_0 is no longer a simple hypothesis but requires specification of a prior for the common probability θ_C . Assuming a beta prior $\theta_C \mid H_0 \sim \text{Beta}(a_0, b_0)$, the marginal likelihood of the observed data is

$$\Pr(y \mid H_0) = \prod_{j \in \{C, 1, \dots, K\}} \binom{n_j}{y_j} \times \frac{B(a_0 + \sum_{j \in \{C, 1, \dots, K\}} y_j, b_0 + \sum_{j \in \{C, 1, \dots, K\}} n_j - \sum_{j \in \{C, 1, \dots, K\}} y_j)}{B(a_0, b_0)}$$

with $B(a, b) = \int_0^1 t^{a-1}(1-t)^{b-1}dt$ the beta function. Under the remaining hypotheses, it is natural to assume independent beta priors $\theta_i \sim \text{Beta}(a_i, b_i)$ for $i \in \{C, 1, \dots, K\}$, and truncate their support to

the space of the corresponding hypothesis. This leads to the marginal likelihoods

$$\begin{aligned} \Pr(y \mid H_{+i}) &= \prod_{j \in \{C, 1, \dots, K\}} \binom{n_j}{y_j} \times \frac{B(a_j + y_j, b_j + n_j - y_j)}{B(a_j, b_j)} \\ &\times \frac{Q_i(a_C + y_C, a_1 + y_1, \dots, a_K + y_K, b_C + n_C - y_C, b_1 + n_1 - y_1, \dots, b_K + n_K - y_K)}{Q_i(a_C, a_1, \dots, a_K, b_C, b_1, \dots, b_K)} \end{aligned} \quad (7)$$

with

$$\begin{aligned} Q_i(a_C, a_1, \dots, a_K, b_C, b_1, \dots, b_K) &= \Pr(\theta_i = \max\{\theta_C, \theta_1, \dots, \theta_K\} \mid a_C, a_1, \dots, a_K, b_C, b_1, \dots, b_K) \\ &= \int_0^1 p(\theta_i \mid a_i, b_i) \times \prod_{j \in \{C, 1, \dots, K\} \setminus \{i\}} \Pr(\theta_j < \theta_i \mid a_j, b_j) d\theta_i \\ &= \int_0^1 \frac{t^{a_i-1} (1-t)^{b_i-1}}{B(a_i, b_i)} \times \prod_{j \in \{C, 1, \dots, K\} \setminus \{i\}} I_t(a_j, b_j) dt, \end{aligned}$$

where $I_x(a, b) = \{\int_0^x t^{a-1} (1-t)^{b-1} dt\} / B(a, b)$ is the regularized incomplete beta function. These can be efficiently computed with numerical integration and standard implementations of the regularized incomplete beta function (e.g., `stats::pbeta` in R). The marginal likelihood of the data under H_- can similarly be obtained from (7) with $i = C$.

For a specified prior probability of the null hypotheses $\Pr(H_0)$, we may again distribute the remaining prior probability among the other hypotheses by

$$\Pr(H_{+i}) = \{1 - \Pr(H_0)\} \{1 - Q_i(a_C, a_1, \dots, a_K, b_C, b_1, \dots, b_K)\}$$

to ensure that the averaged prior is again the beta prior that was truncated in the first place. Plugging these marginal likelihoods and prior probabilities into equation (1) produces posterior probabilities, which in turn can be used for obtaining randomization probabilities. Similarly, ratios of marginal likelihoods can be taken to obtain Bayes factors for monitoring accumulating evidence.

Figure 6 illustrates RAR probabilities for simulated binomial data with a control and $K = 3$ treatment groups. Randomization probabilities were computed assuming exact binomial likelihoods with uniform priors for the probabilities or a normal approximation to the vector of log odds ratios obtained from logistic regression along with a multivariate normal prior. We can see that the exact (solid lines) and approximate (dashed lines) probabilities are not too far from each other in most cases even though the true probabilities from which the data are simulated are rather small. The RAR probabilities with $\Pr(H_0) = 0.5$ remain considerably closer to 25% (equal randomization) compared to the

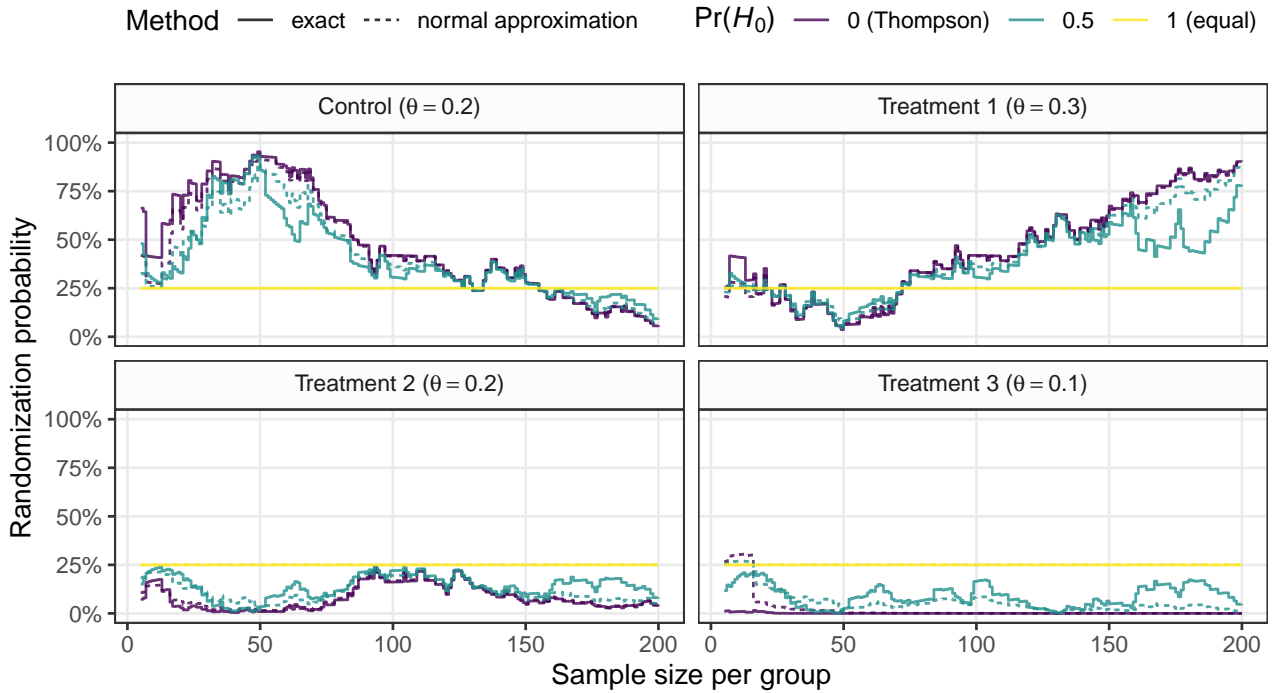


Figure 6: Evolution of Bayesian RAR probabilities for 3 treatments and a control group. In each step, one observation from each group is simulated from a binomial distribution with probability as indicated in the panel titles (treatment 1 is the most effective). Randomization probabilities are computed assuming binomial likelihoods with independent uniform priors or a normal approximation to the vector of log odds ratios obtained from logistic regression along with a multivariate normal prior with variances 1 and correlations 0.5.

RAR probabilities with $\Pr(H_0) = 0$ (Thompson sampling), which shows the largest variability. Since the true probability in treatment group 1 is the highest, all randomization probabilities converge towards 100% for treatment 1, as expected, while they decrease to 0% for the other treatments.

5 Reanalyzing the ECMO trial

The ECMO trial (Bartlett et al., 1985) investigated the efficacy of ECMO (extracorporeal membrane oxygenation) treatment in the critical care of newborns. It was the first prominent clinical trial to use a “randomized play-the-winner” (RPW) RAR design (Wei and Durham, 1978). While some earlier non-randomized studies had shown a substantial treatment effect, a randomized study was required to confirm this. However, this presented an ethical dilemma, as the investigators were convinced that the risk of death would be much higher in the control group than in the ECMO group. To mitigate this, the RPW design was chosen. The outcome of the trial was bizarre: The first newborn was randomized to receive ECMO treatment and survived. The second newborn was randomized to receive the control treatment and died. The ten subsequently enrolled newborns were all randomized

to ECMO, and all survived (Bartlett et al., 1985). The trial was then stopped for efficacy and its results published. However, the unusual outcome sparked intense debates about whether the trial was even a proper randomized clinical trial. Several subsequent trials were conducted, all re-establishing the effectiveness of ECMO, which is now a standard treatment (Bartlett, 2024). In the following, we will reanalyze the ECMO data.

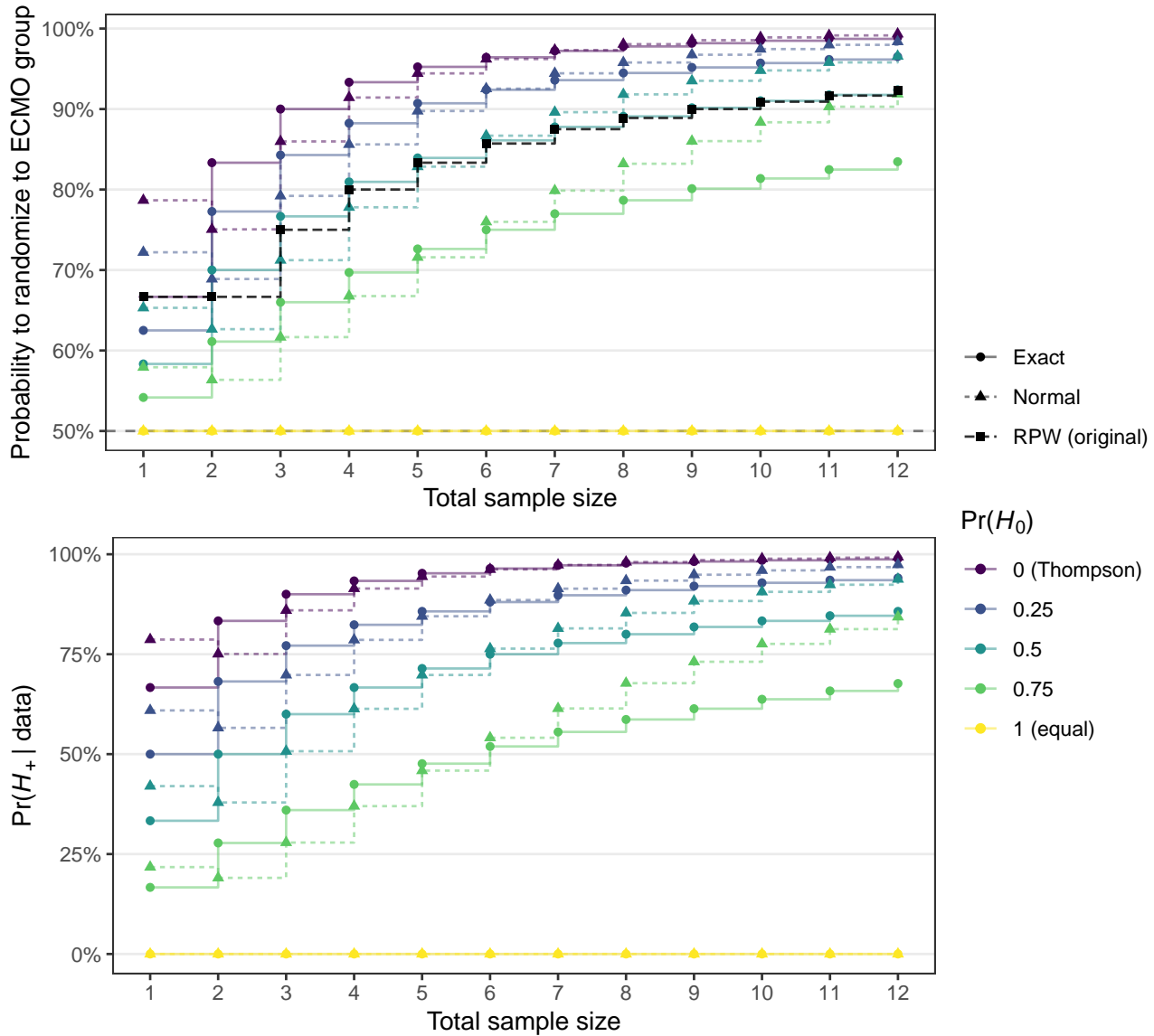


Figure 7: Evolution of Bayesian RAR randomization probabilities (top plot) and posterior probability of a beneficial ECMO treatment effect (bottom plot) for data from the ECMO trial.

The top plot in Figure 7 shows Bayesian RAR probabilities computed from the ECMO data sequence. The normal approximation and the exact binomial methods were used, as well as the RPW method that was originally used in the ECMO trial. A standard normal prior was assigned to the log odds ratio for the normal method, while uniform priors were assigned to the probabilities for the

binomial method. For the normal method, log odds ratio estimates were computed with a Yates' correction (i.e., adding a half to each cell) to avoid issues with zero cells. We can see that this correction induces a slight anomaly at the start of the study as the probabilities from the normal method slightly decrease after the first observation despite the first patient surviving the ECMO treatment. This does not happen for the exact method whose randomization probability increase after every additional patient, which only happens for the normal approximation after the second patient.

Comparing the randomization probabilities for different prior probabilities of the null hypothesis for each method, we can see that $\Pr(H_0) = 0$ (Thompson sampling) shows the most extreme randomization probabilities that rapidly increase to 100%, whereas $\Pr(H_0) = 1$ (equal randomization) leaves the probabilities completely static at 50%. In between, the randomization probabilities gradually shift from 50% to the Thompson sampling probabilities. The RPW probabilities (black squares) are relatively close to the probabilities from $\Pr(H_0) = 0.5$ (exact) at the beginning of the study and become closer to the probabilities from $\Pr(H_0) = 0.75$ (normal approximation) at later time points.

The bottom plot in Figure 7 shows the corresponding posterior probability of the ECMO treatment being effective. Depending on the prior probability assigned to the null hypothesis of equal effectiveness, we can see that the posterior probability at the end of the trial may be very high (e.g., 98.9% for $\Pr(H_0) = 0$), or only moderately in favour of ECMO (e.g., 67.7% for $\Pr(H_0) = 0.75$). From a Bayesian perspective, stopping the trial seems thus a sensible decision if the prior probability of equal effects was low. Conversely, more evidence would have been needed if the prior probability had been higher.

6 Simulation study

We conducted a simulation study to evaluate the performance of point null Bayesian RAR for different values of the prior probability of the null hypothesis $\Pr(H_0)$, and compare it to Thompson sampling (potentially modified with burn-in periods, randomization probability capping, power transformations) and equal randomization. Considered patient performance measures were the rate of successes, the rate of extreme randomization probabilities, and sample size imbalance. Additionally, bias and coverage were used to evaluate the performance of rate difference point estimates and confidence intervals under RAR, while the type I error rate and power of the corresponding tests were used to quantify hypothesis testing performance under RAR. A binomial data-generating mechanism was used which was partially based on the simulation studies from [Robertson et al. \(2023\)](#), [Thall and](#)

Wathen (2007), and Wathen and Thall (2017). Detailed description of the design and results of the simulation study following the structured ADEMP approach (Morris et al., 2019; Siepe et al., 2024) are provided in Appendix B. A supplemental website provides an interactively explorable results dashboard (<https://brar-42c2c0.pages.uzh.ch/simulation-dashboard.html>).

Across all simulations, we observed a trade-off between patient benefit and parameter estimation / hypothesis testing operating characteristics. Some methods performed better in terms of patient benefit but had worse bias, coverage, type I error rate, and power while others performed better in terms of bias, coverage, type I error rate, and power but had worse patient benefit. This trade-off is well described in the RAR literature (Hu and Rosenberger, 2003; Robertson et al., 2023).

The main result of the simulation study regarding the newly proposed method was that, under most conditions, point null Bayesian RAR with a high prior probability of the null hypothesis $\Pr(H_0) = 0.75$ showed similar operating characteristics to Thompson sampling with capped randomization probabilities at 0.1 and 0.9 and a power transformation with $c = i/(2n)$, where i is the current sample size and n is the maximum sample size. Both point null Bayesian RAR with a prior probability of $\Pr(H_0) = 0.75$ and modified Thompson sampling could mitigate some issues with ordinary Thompson sampling. For instance, they exhibited less negative sample size imbalance, biased parameter estimates, undercoverage, and inflated type I error rates. However, this improvement came at the cost of worse patient benefit performance compared to unmodified Thompson sampling. For instance, the mean success rate was lower, though still considerably better than equal randomization in most cases. Most importantly, the variability of randomization probabilities was much reduced with $\Pr(H_0) = 0.75$, producing rarely negative imbalances (i.e., a large proportion of patients randomized to an inferior treatment). Setting a lower but positive prior probability than $\Pr(H_0) = 0.75$ produced operating characteristics comparable to those from less extreme modifications of Thompson sampling, such as a power transformation with $c = 1/2$. In sum, the simulation study demonstrated that point null Bayesian RAR has comparable statistical properties to Thompson sampling with ad hoc modifications.

7 Discussion

In this paper we have proposed a modification of standard Bayesian RAR (Thompson sampling) via the introduction of a point null hypothesis, which leads to a spike-and-slab prior, and a recasting of the problem in a Bayesian hypothesis testing framework. While the plausibility of point

null hypotheses is often a matter of philosophical debate (see e.g., [Berger and Delampady, 1987](#); [Ly and Wagenmakers, 2022](#)), this method has desirable properties in the RAR setting, as it can interpolate between equal randomization and Thompson sampling by changing the prior probability of the null hypothesis $\Pr(H_0)$. This allows experimenters to balance patient benefit with classical operating characteristics, such as power, type I error rate, bias, and coverage. For large values of $\Pr(H_0)$, we observed behaviors and operating characteristics similar to those obtained with ad hoc modifications of Thompson sampling, such as capping, burn-ins, and power transformations. Our method is implemented in the free and open source R package `brar` for binomial outcomes and for data summarized by approximately normal effect estimates. The latter makes the method applicable to many settings, for example, settings where treatment effects are estimated with regression analyses.

One advantage of our framework is that the randomization probabilities correspond coherently with the available statistical evidence (Bayes factors) and beliefs (posterior probabilities). In principle, both could be used as decision-making tools instead of relying on frequentist test criteria. For instance, if the posterior probability of a treatment’s superiority is greater than 99%, say, a study could be stopped. This also makes sense from the perspective that it is unnatural to randomize participants with extreme randomization probabilities that are associated with such high posterior probabilities.

Although we conducted a simulation study to understand the method’s basic behaviour, more realistic and comprehensive evaluations are needed to understand its applicability in real-world conditions. For example, the method needs to be evaluated in combination with futility stopping, as well as taking into account the fact that outcomes are not observed immediately, but usually take weeks to years to observe (e.g. survival). Another issue to consider is how to select the prior probability of the null hypothesis. In our simulation study, we found that setting a value of $\Pr(H_0) = 0.75$ mitigated many of the issues with Thompson sampling. However, other choices could be considered, such as setting a higher value. Similar considerations apply to the prior distributions of the parameters, as we did not assess the effect of varying these on the operating characteristics. Future work may therefore investigate whether a more efficient randomization procedure can be obtained by specifying a certain prior distribution. Finally, in many clinical trial settings, RAR methods are not directly applicable because outcomes such as death may only be observed after long follow-up periods, by which time recruitment and randomized allocation will already have finished. An alternative could be to perform point null Bayesian RAR with an informative surrogate outcome that is sooner observed than the primary outcome ([Gao et al., 2024](#)).

Acknowledgments

We thank František Bartoš and Małgorzata Roos for valuable comments on drafts of the manuscript. The acknowledgment of these individuals does not imply their endorsement of the paper.

Conflict of interest

We declare no conflict of interest.

Software and data

Code and data to reproduce our analyses are openly available at <https://github.com/SamCH93/brar>. A snapshot of the repository at the time of writing is available at <https://doi.org/10.5281/zenodo.XXXXXX>. We used the statistical programming language R version 4.4.1 (2024-06-14) for analyses (R Core Team, 2017) along with the ggplot2 (Wickham, 2016), dplyr (Wickham et al., 2023), SimDesign (Chalmers and Adkins, 2020), mvtnorm (Genz and Bretz, 2009), ggh4x (van den Brand, 2024), ggpubr (Kassambara, 2023), and knitr (Xie, 2015) packages.

References

- Bartlett, R. H. (2024). The story of ECMO. *Anesthesiology*, 140(3):578. doi:[10.1097/ALN.0000000000004843](https://doi.org/10.1097/ALN.0000000000004843).
- Bartlett, R. H., Roloff, D. W., Cornell, R. G., Andrews, A. F., Dillon, P. W., and Zwischenberger, J. B. (1985). Extracorporeal Circulation in Neonatal Respiratory Failure: A Prospective Randomized Study. *Pediatrics*, 76(4):479–487. doi:[10.1542/peds.76.4.479](https://doi.org/10.1542/peds.76.4.479).
- Berger, J. O. and Delampady, M. (1987). Testing precise hypotheses. *Statistical Science*, 2(3). doi:[10.1214/ss/1177013238](https://doi.org/10.1214/ss/1177013238).
- Berry, S. M., Carlin, B. P., Lee, J. J., and Muller, P. (2010). *Bayesian Adaptive Methods for Clinical Trials*. Chapman & Hall/CRC.
- Chalmers, R. P. and Adkins, M. C. (2020). Writing effective and reliable Monte Carlo simulations with the SimDesign package. *The Quantitative Methods for Psychology*, 16(4):248–280. doi:[10.20982/tqmp.16.4.p248](https://doi.org/10.20982/tqmp.16.4.p248).

- Gao, J., Hu, F., and Ma, W. (2024). Response-adaptive randomization procedure in clinical trials with surrogate endpoints. *Statistics in Medicine*, 43(30):5911–5921. doi:[10.1002/sim.10286](https://doi.org/10.1002/sim.10286).
- Genz, A. and Bretz, F. (2009). *Computation of Multivariate Normal and t Probabilities*. Lecture Notes in Statistics. Springer-Verlag, Heidelberg.
- Good, I. J. (1958). Significance tests in parallel and in series. *Journal of the American Statistical Association*, 53(284):799–813. doi:[10.1080/01621459.1958.10501480](https://doi.org/10.1080/01621459.1958.10501480).
- Grieve, A. P. (2016). Response-adaptive clinical trials: case studies in the medical literature. *Pharmaceutical Statistics*, 16(1):64–86. doi:[10.1002/pst.1778](https://doi.org/10.1002/pst.1778).
- Heinze, G., Boulesteix, A., Kammer, M., Morris, T. P., and White, I. R. (2023). Phases of methodological research in biostatistics—building the evidence base for new methods. *Biometrical Journal*, 66(1). doi:[10.1002/bimj.202200222](https://doi.org/10.1002/bimj.202200222).
- Hu, F. and Rosenberger, W. F. (2003). Optimality, variability, power: Evaluating response-adaptive randomization procedures for treatment comparisons. *Journal of the American Statistical Association*, 98(463):671–678. doi:[10.1198/016214503000000576](https://doi.org/10.1198/016214503000000576).
- Jeffreys, H. (1939). *Theory of Probability*. Clarendon Press, Oxford, first edition.
- Kass, R. and Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795. doi:[10.1080/01621459.1995.10476572](https://doi.org/10.1080/01621459.1995.10476572).
- Kassambara, A. (2023). *ggpubr: 'ggplot2' Based Publication Ready Plots*. URL <https://CRAN.R-project.org/package=ggpubr>. R package version 0.6.0.
- Lee, K. M. and Lee, J. J. (2021). Evaluating Bayesian adaptive randomization procedures with adaptive clip methods for multi-arm trials. *Statistical Methods in Medical Research*, 30(5):1273–1287. doi:[10.1177/0962280221995961](https://doi.org/10.1177/0962280221995961).
- Ly, A. and Wagenmakers, E.-J. (2022). Bayes factors for peri-null hypotheses. *TEST*, 31(4):1121–1142. doi:[10.1007/s11749-022-00819-w](https://doi.org/10.1007/s11749-022-00819-w).
- Morris, T. P., White, I. R., and Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11):2074–2102. doi:[10.1002/sim.8086](https://doi.org/10.1002/sim.8086).

- Pawel, S., Bartoš, F., Siepe, B. S., and Lohmann, A. (2025). Handling missingness, failures, and non-convergence in simulation studies: A review of current practices and recommendations. *The American Statistician*, pages 1–27. doi:[10.1080/00031305.2025.2540002](https://doi.org/10.1080/00031305.2025.2540002).
- R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92(437):179–191. doi:[10.1080/01621459.1997.10473615](https://doi.org/10.1080/01621459.1997.10473615).
- Robertson, D. S., Lee, K. M., López-Kolkovska, B. C., and Villar, S. S. (2023). Response-adaptive randomization in clinical trials: From myths to practical considerations. *Statistical Science*, 38(2). doi:[10.1214/22-sts865](https://doi.org/10.1214/22-sts865).
- Siepe, B. S., Bartoš, F., Morris, T. P., Boulesteix, A.-L., Heck, D. W., and Pawel, S. (2024). Simulation studies for methodological research in psychology: A standardized structure for planning, preregistration, and reporting. *Psychological Methods*. doi:[10.1037/met0000695](https://doi.org/10.1037/met0000695). To appear.
- Spiegelhalter, D. J. (2004). Incorporating Bayesian ideas into health-care evaluation. *Statistical Science*, 19(1):156–174. <https://www.jstor.org/stable/4144380>.
- Thall, P., 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(8):1621–1628. doi:[10.1093/annonc/mdv238](https://doi.org/10.1093/annonc/mdv238).
- Thall, P. F. and Wathen, J. K. (2007). Practical Bayesian adaptive randomisation in clinical trials. *European Journal of Cancer*, 43(5):859–866. doi:[10.1016/j.ejca.2007.01.006](https://doi.org/10.1016/j.ejca.2007.01.006).
- Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285. doi:[10.2307/2332286](https://doi.org/10.2307/2332286).
- van den Brand, T. (2024). *ggh4x: Hacks for 'ggplot2'*. URL <https://CRAN.R-project.org/package=ggh4x>. R package version 0.3.0.
- Wathen, J. K. and Thall, P. F. (2017). A simulation study of outcome adaptive randomization in multi-arm clinical trials. *Clinical Trials*, 14(5):432–440. doi:[10.1177/1740774517692302](https://doi.org/10.1177/1740774517692302).

- Wei, L. J. and Durham, S. (1978). The randomized play-the-winner rule in medical trials. *Journal of the American Statistical Association*, 73(364):840–843. doi:[10.1080/01621459.1978.10480109](https://doi.org/10.1080/01621459.1978.10480109).
- Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. URL <https://ggplot2.tidyverse.org>.
- Wickham, H., François, R., Henry, L., Müller, K., and Vaughan, D. (2023). *dplyr: A Grammar of Data Manipulation*. URL <https://CRAN.R-project.org/package=dplyr>. R package version 1.1.4.
- Xie, Y. (2015). *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition. URL <https://yihui.org/knitr/>. ISBN 978-1498716963.

Appendix A The R package brar

Our R package can be installed by running `remotes::install_github(repo = "SamCH93/brar", subdir = "package")` in an R session (requires the `remotes` package, which is available on CRAN). The main functions of the package are `brar_normal` and `brar_binomial`, which implement the approximate normal method from Section 3 and the exact binomial method from Section 4. The following code chunk illustrates how both functions can be used.

```
library(brar) # load package

## observed successes and trials in control and 3 treatment groups
y <- c(10, 9, 14, 13)
n <- c(20, 20, 22, 21)
group <- c("control", paste("treatment", seq(1, 3)))

## conduct exact point null Bayesian RAR
brar_binomial(y = y, n = n,
              ## uniform prior for common probability under H0
              a0 = 1, b0 = 1,
              ## uniform priors for all probabilities
              a = c(1, 1, 1, 1), b = c(1, 1, 1, 1),
              ## prior probability of the null hypothesis
              pH0 = 0.5)

## DATA
##           Events Trials Proportion
## Control      10     20      0.500
## Treatment 1   9     20      0.450
## Treatment 2  14     22      0.636
## Treatment 3  13     21      0.619
```



```
##
## PRIOR PROBABILITIES
##   H-   H0   H+1   H+2   H+3
## 0.125 0.500 0.125 0.125 0.125
##
## BAYES FACTORS (BF_ij)
##       H-   H0   H+1   H+2   H+3
## H-   1.000 0.0341  2.16 0.1837 0.223
## H0   29.335 1.0000 63.45 5.3891 6.533
## H+1   0.462 0.0158  1.00 0.0849 0.103
## H+2   5.443 0.1856 11.77 1.0000 1.212
## H+3   4.490 0.1531  9.71 0.8249 1.000
##
## POSTERIOR PROBABILITIES
##       H-   H0   H+1   H+2   H+3
## 0.00777 0.91148 0.00359 0.04228 0.03488
##
## RANDOMIZATION PROBABILITIES
##   Control Treatment 1 Treatment 2 Treatment 3
##   0.236      0.231      0.270      0.263

## get data into shape for approximate point null Bayesian RAR
fit <- glm(cbind(y, n - y) ~ group, family = binomial) # fit logistic regression
estimate <- fit$coefficients[-1] # remove intercept to get logOR estimates
sigma <- vcov(fit)[-1,-1] # remove intercept to get logOR covariance matrix

## conduct approximate point null Bayesian RAR
pm <- c(0, 0, 0) # set prior mean to zero
## set 0.5 correlated prior covariance so that equal prior probabilities
psigma <- matrix(0.5, nrow = 3, ncol = 3)
diag(psigma) <- 1
brar_normal(estimate = estimate, sigma = sigma, pm = pm, psigma = psigma,
            pH0 = 0.5)

## DATA
##           Effect estimate      SE
## Treatment 1          -0.201 0.634
## Treatment 2           0.560 0.630
## Treatment 3           0.486 0.634
##
## PRIOR PROBABILITIES
##   H-   H0   H+1   H+2   H+3
## 0.125 0.500 0.125 0.125 0.125
##
## BAYES FACTORS (BF_ij)
##       H-   H0   H+1   H+2   H+3
## H-   1.000 0.1338  1.88 0.231 0.274
## H0   7.472 1.0000 14.06 1.726 2.051
## H+1  0.531 0.0711  1.00 0.123 0.146
## H+2  4.331 0.5795  8.15 1.000 1.189
## H+3  3.644 0.4876  6.86 0.841 1.000
```

```

##
## POSTERIOR PROBABILITIES
##      H-      H0      H+1      H+2      H+3
## 0.0254 0.7587 0.0135 0.1099 0.0925
##
## RANDOMIZATION PROBABILITIES
##      Control Treatment 1 Treatment 2 Treatment 3
##      0.215      0.203      0.300      0.282

```

Appendix B Simulation study

We will now describe the design and results of a simulation study following the structured ADEMP approach (Morris et al., 2019; Siepe et al., 2024). Our simulation study was not preregistered as it constitutes early-phase methodological research where the properties of a new method are explored without the intention to give recommendations for practitioners (Heinze et al., 2023). A website with additional details and results is provided at <https://samch93.github.io/brar/>.

B.1 Aims

The aim of the simulation study is to evaluate the design characteristics of the newly proposed point null Bayesian RAR approach, and compare it to existing methods.

B.2 Data-generating mechanism

The data-generating mechanism was inspired by the simulation studies from Robertson et al. (2023), Thall and Wathen (2007), and Wathen and Thall (2017). In each repetition, a data set with n binary outcomes is simulated through RAR: A patient i is randomly allocated to the control group or one of the K treatment groups based on randomization probabilities computed from the $1, \dots, i-1$ preceding outcomes. Depending on the allocation, an outcome is either simulated from a Bernoulli distribution with probability θ_C in the control group, θ_1 in the first treatment group, or θ_2 for the remaining treatment groups (in case $K > 1$).

Parameters were chosen similar to the simulation study from Robertson et al. (2023). We vary the sample size $n \in \{200, 654\}$ to represent low and high powered studies, the number of treatment groups $K \in \{1, 2, 3\}$, and probability in the first treatment group $\theta_1 \in \{0.25, 0.35, 0.45\}$. The probability in the control group and the remaining groups is always fixed at $\theta_C = 0.25$ and $\theta_2 = \theta_3 = 0.3$,

respectively. All these parameters are varied fully-factorially, leading to $2 \times 3 \times 3 = 18$ parameter conditions.

Since treatment allocation determines from which true probability an outcome is simulated, data generation is directly influenced by the RAR methods described below. These come with additional parameters that are, however, considered as method tuning-/hyper-parameters rather than true underlying parameters.

B.3 Estimands and other targets

The primary interest of this simulation study lies in assessing the patient benefit characteristics of different RAR methods. Additionally, the estimand of interest is the rate difference $RD_1 = \theta_1 - \theta_C$ and the target of interest is the null hypothesis of $RD_1 = 0$.

B.4 Methods

We consider the above described point null RAR methods. The prior probability of H_0 is a tuning parameter and controls the variability of the randomization probabilities. Setting $\Pr(H_0) = 1$ produces equal randomization, whereas $\Pr(H_0) = 0$ produces Thompson sampling. We consider values of $\Pr(H_0) \in \{0, 0.25, 0.5, 0.75, 1\}$, as well as the normal approximation and exact binomial version of RAR. For approximate normal RAR, a normal prior with mean 0, variance 1, and in case of $K > 1$ a correlation of $1/2$, is considered. Independent uniform priors are assigned for binomial RAR. Log odds ratios along with their covariance are estimated with logistic regression and then used as inputs for the normal RAR method, while the exact method uses success counts and sample sizes only. In case a method fails to converge, equal randomization is applied as a back-up strategy, as this mimics what an experimenter might do in practice when a RAR method fails to converge ([Pawel et al., 2025](#)).

We also consider three modifications of these methods: In some conditions, a “burn-in” phase is carried out during which the first 50 patients are always randomized with equal probability $1/(K + 1)$ to each group. For Thompson sampling ($\Pr(H_0) = 0$), we additionally consider conditions with power transformations of randomization probabilities, i.e., if π_k is the randomization probability of group k , we take $\pi_k^* = \pi_k^c / \sum_{j \in \{C, 1, \dots, K\}} \pi_j^c$. We consider $c = 1/2$ and $c = i/(2n)$ with i the current and n the maximal sample size, which are two popular choices of the tuning parameter c ([Wathen and Thall, 2017](#)). Additionally, in some conditions “capping” is applied to Thompson sampling. That is, randomization probabilities outside the $[0.1, 0.9]$ interval are set to either 0.1 or 0.9. After

capping has been performed, randomization probabilities are re-normalized to sum to one (Wathen and Thall, 2017; Lee and Lee, 2021). This re-normalization is only performed for randomization probabilities greater than 0.1, as these would otherwise be reduced again to probabilities less than 0.1. In case, a re-normalized probability becomes less than 0.1, it is also capped at 0.1 and second re-normalization performed. Finally, for equal randomization ($\Pr(H_0) = 1$), no burn-in, capping, or power transformation conditions are simulated as these manipulations have no effect.

B.5 Performance measures

Different performance measures were used. Patient benefit was quantified with:

- The mean rate of successes per study

$$\overline{\text{RS}} = \frac{1}{n_{\text{sim}}} \sum_{i=1}^{n_{\text{sim}}} \sum_{j=1}^n \frac{y_{ij}}{n}$$

where y_{ij} denotes the 0/1 success indicator of patient j in simulation i , n is the sample size, and n_{sim} is the number of simulation repetitions.

- The mean rate of extreme randomization probabilities (less than 10% or greater than 90%)

$$\overline{\text{REP}} = \frac{1}{n_{\text{sim}}} \sum_{i=1}^{n_{\text{sim}}} \sum_{j=1}^n \frac{\mathbb{1}(\text{any randomization probability at time } j < 10\% \text{ or } > 90\%)}{n}$$

with indicator function $\mathbb{1}(\cdot)$.

- The proportion of simulations where the number of allocations to treatment 1 was at least 10% of the total sample size n less than the average sample size in the remaining groups

$$\hat{S}_{0.1} = \frac{1}{n_{\text{sim}}} \sum_{i=1}^{n_{\text{sim}}} \mathbb{1} \left(\frac{n - n_{1i}}{K} - n_{1i} > 0.1n \right)$$

where n_{1i} is the number of allocations to treatment group 1 in simulation i . For $K = 1$, this reduces to the $\hat{S}_{0.1}$ sample size imbalance measure from Robertson et al. (2023), which in turn was inspired by the performance evaluation in the simulation study from Thall et al. (2015).

Parameter estimation and hypothesis testing performance was quantified with:

- The empirical bias of the estimate of the rate difference RD_1

$$\text{Bias}(RD_1) = \frac{1}{n_{\text{sim}}} \sum_{i=1}^{n_{\text{sim}}} \hat{\theta}_{1i} - \hat{\theta}_{0i} - RD_1$$

where $\hat{\theta}_{1i}$ and $\hat{\theta}_{0i}$ are the maximum likelihood estimates of the probabilities θ_1 and θ_0 in simulation repetition i .

- Empirical coverage of the 95% Wald confidence intervals of RD_1

$$\text{Coverage}(RD_1) = \frac{1}{n_{\text{sim}}} \sum_{i=1}^{n_{\text{sim}}} \mathbb{1} \{95\% \text{ CI includes } RD_1 \text{ in simulation } i\}.$$

- Empirical rejection rate (type I error rate or power depending on the condition) related to the Wald test of the rate difference RD_1

$$\text{RR}(RD_1) = \frac{1}{n_{\text{sim}}} \sum_{i=1}^{n_{\text{sim}}} \mathbb{1} \{ \text{Test rejects } H_0: RD_1 = 0 \text{ in simulation } i \}.$$

Each condition was simulated 10000 times. This ensures a MCSE (Monte Carlo Standard Error) for the Type I error rate and power of at most 0.5%. MCSEs were calculated using the formulae from [Siepe et al. \(2024\)](#) and are provided for all measures in the following figures and the supplemental website. In most cases the MCSEs are hardly visible due to them being tiny relative to the range of the figures.

B.6 Computational aspects

The simulation study was run on a server running Debian GNU/Linux 13 (trixie) and R version 4.5.0 (2025-04-11). The SimDesign R package was used to organize and run the simulation study ([Chalmers and Adkins, 2020](#)). Our newly developed brar R package was used to perform RAR (see Appendix A for an illustration). Code and data to reproduce this simulation study are available at <https://github.com/SamCH93/brar>.

B.7 Results

Convergence Non-convergence happened only rarely for the normal approximation method due to logistic regression not converging at the start of the study when no events were observed for some

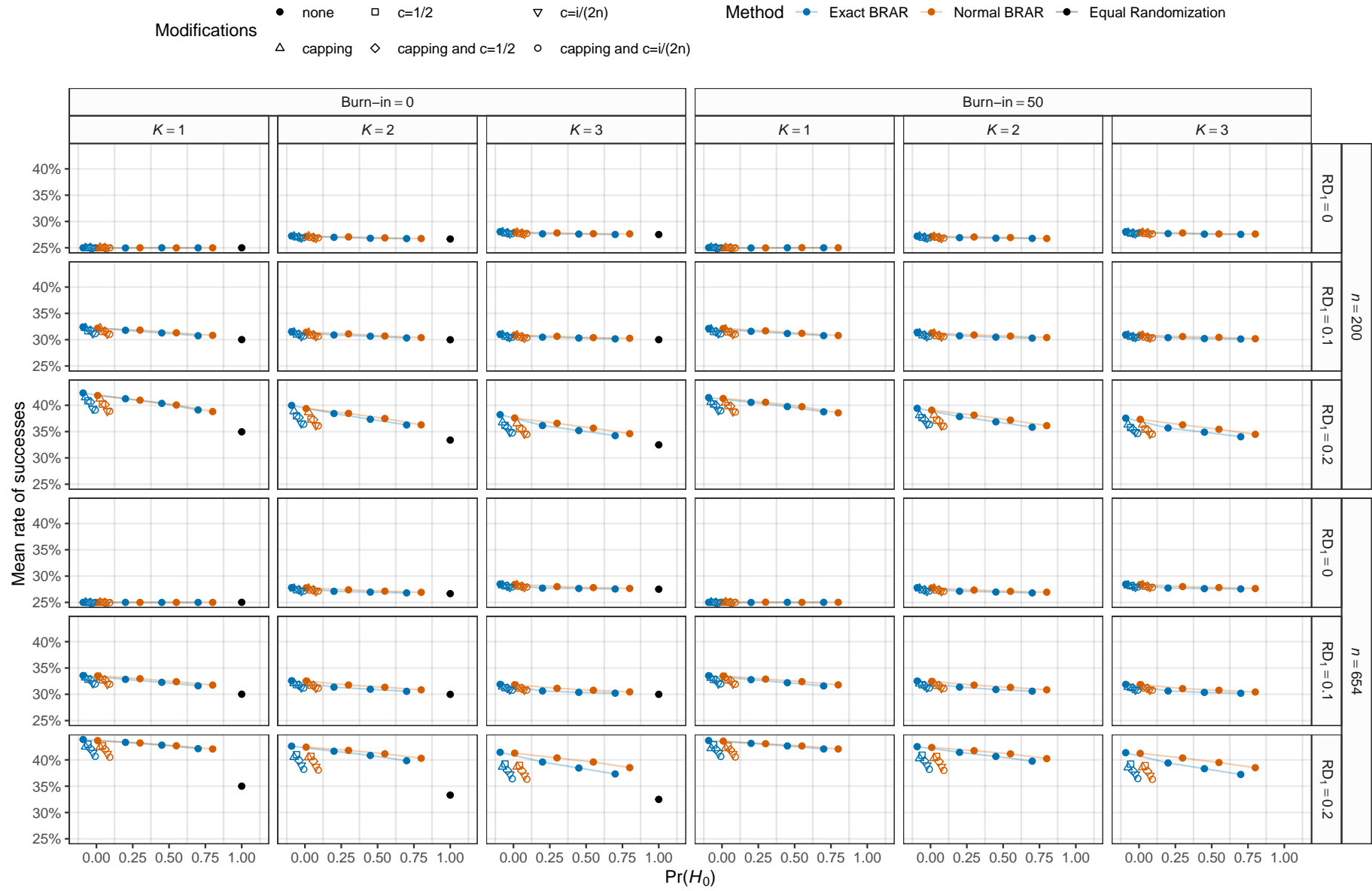


Figure 8: Mean rate of successes (i.e., the number of successes in a study divided by its sample size averaged over all 10000 simulation repetitions). The maximum MCSE is 0.047%.

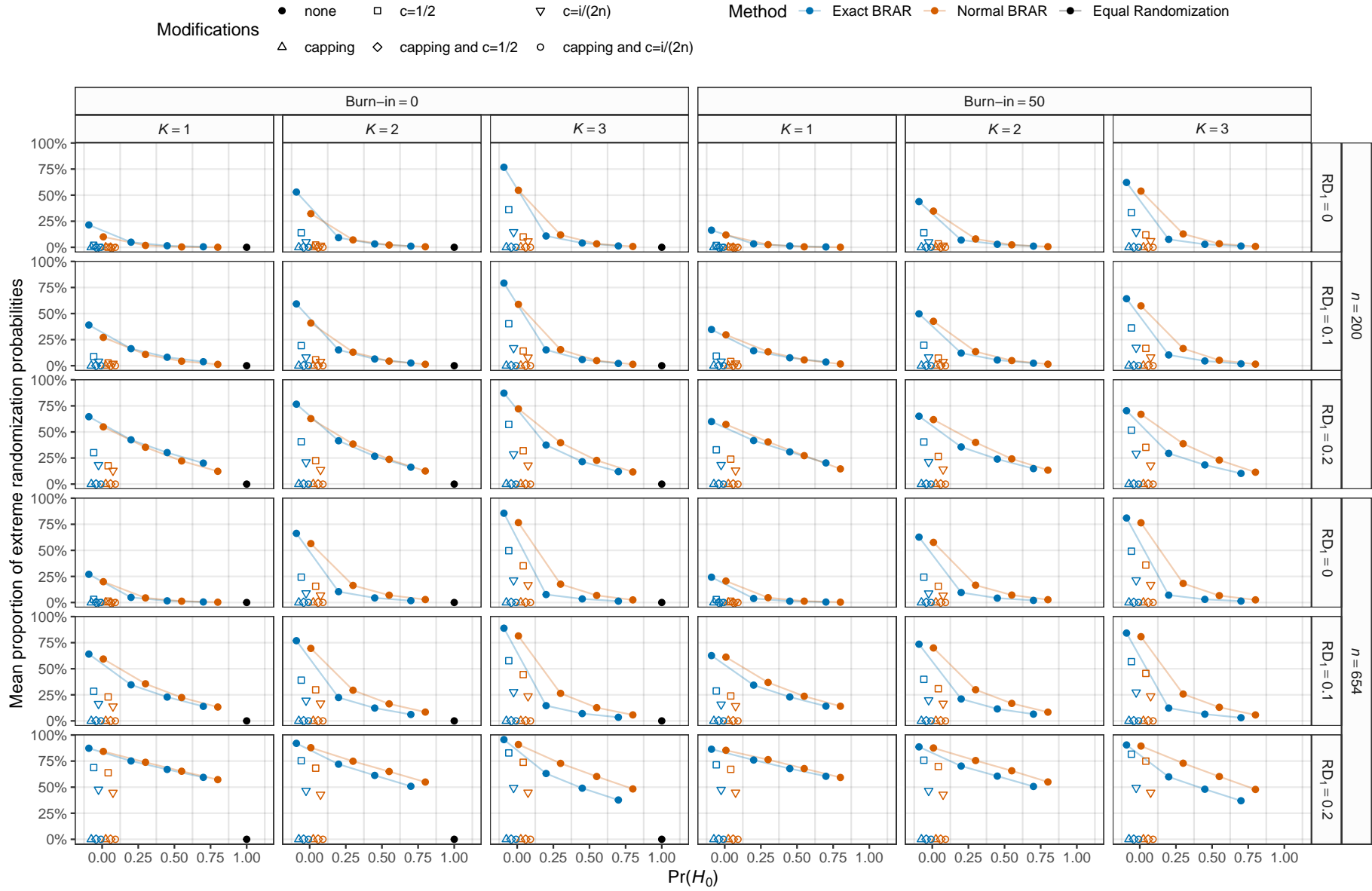


Figure 9: Mean proportion of 10000 simulations with randomization probabilities either less than 0.1 or greater than 0.9. The maximum MCSE is 0.33%.

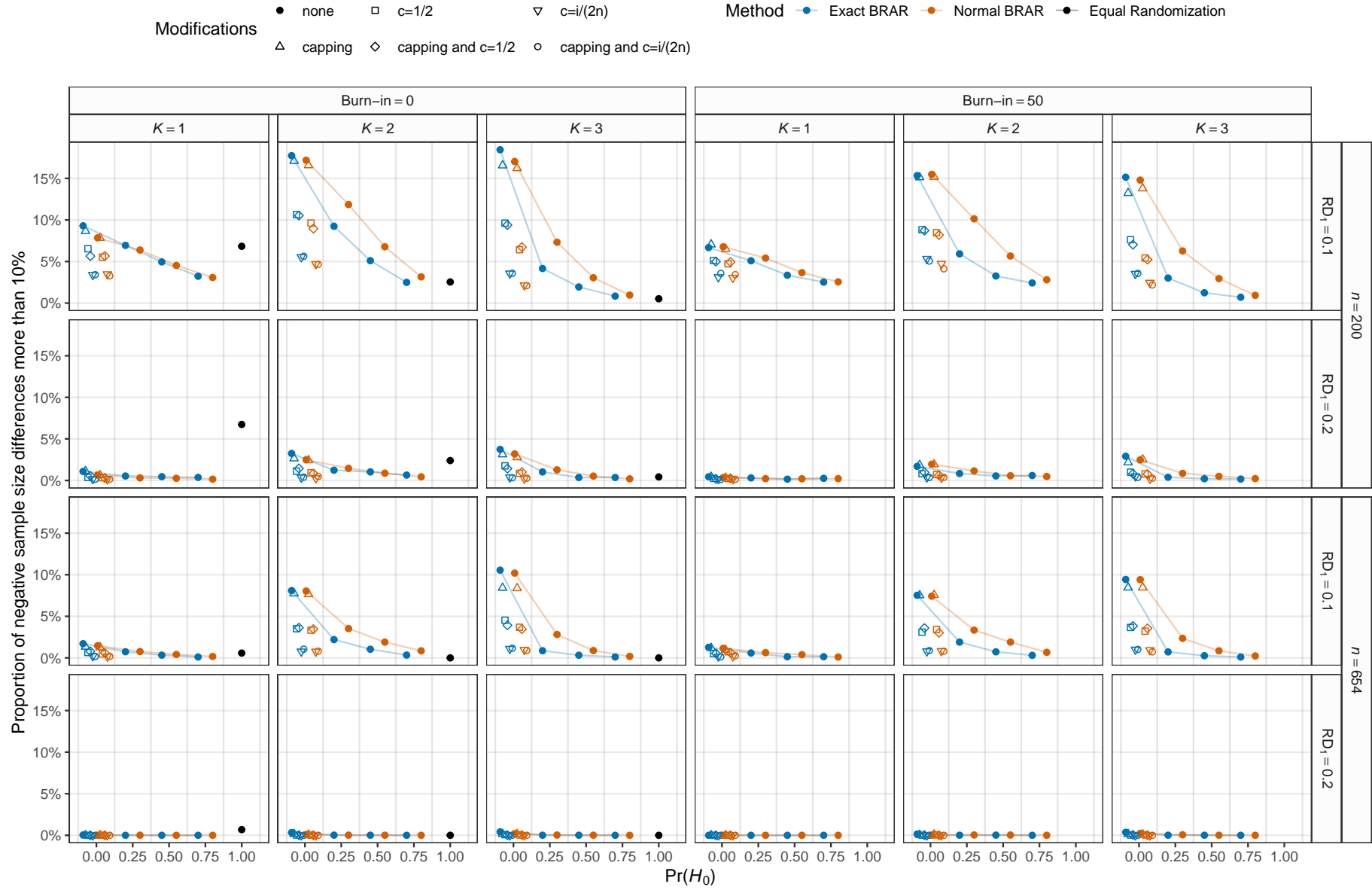


Figure 10: Proportion of 10000 simulations with more than 10% of the sample size randomized to other groups than treatment group 1. The maximum MCSE is 0.5%.

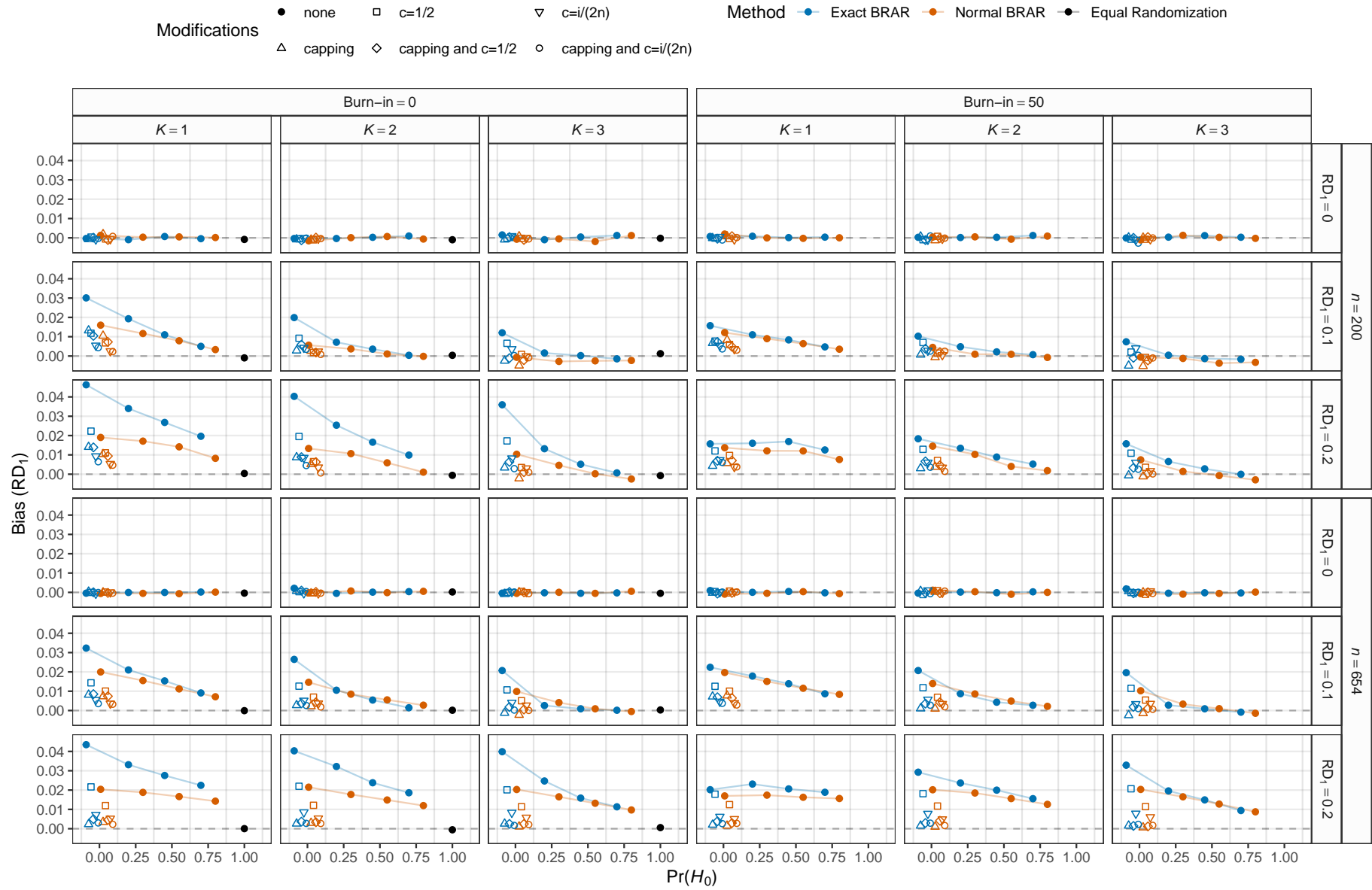


Figure 11: Empirical bias of the estimate of the rate difference RD_1 between the first treatment group and the control group based on 10000 simulation repetitions. The maximum MCSE is 0.0013.

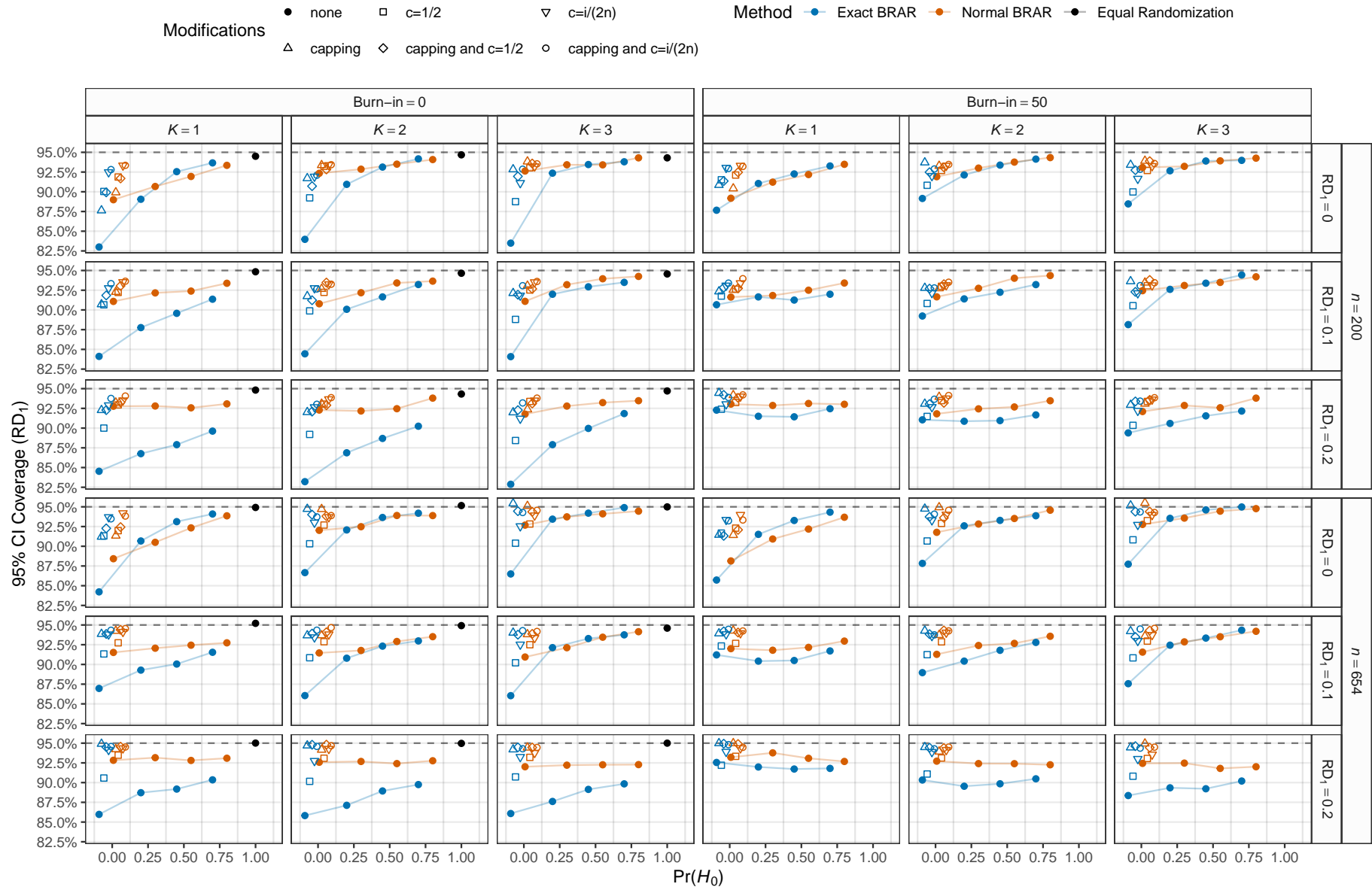


Figure 12: Empirical coverage of the 95% Wald confidence interval for the risk difference RD_1 based on 10000 simulation repetitions. The maximum MCSE is 0.38%.

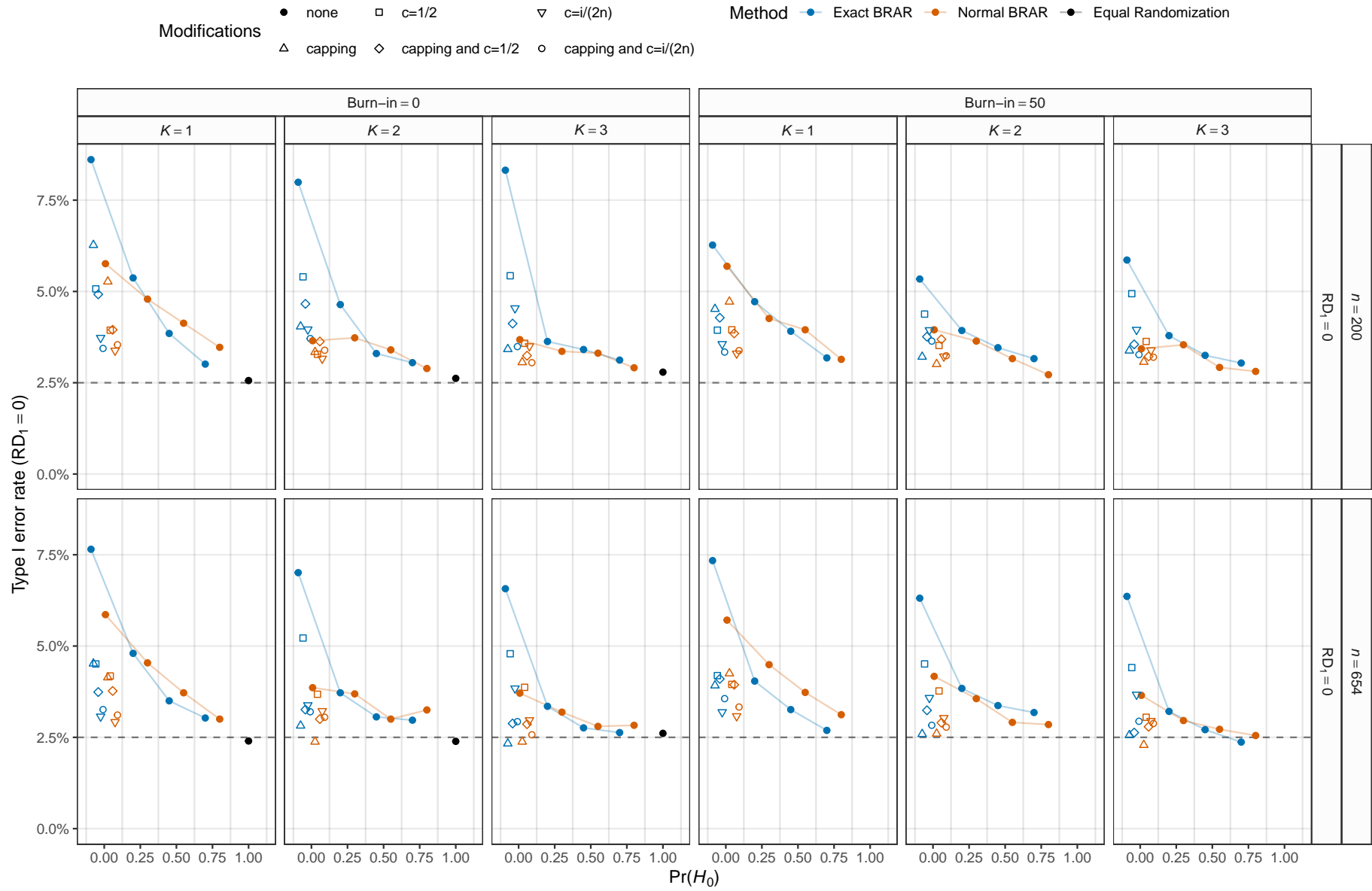


Figure 13: Empirical type I error rate of the Wald test of $RD_1 = 0$ based on 10000 simulation repetitions. The maximum MCSE is 0.28%.

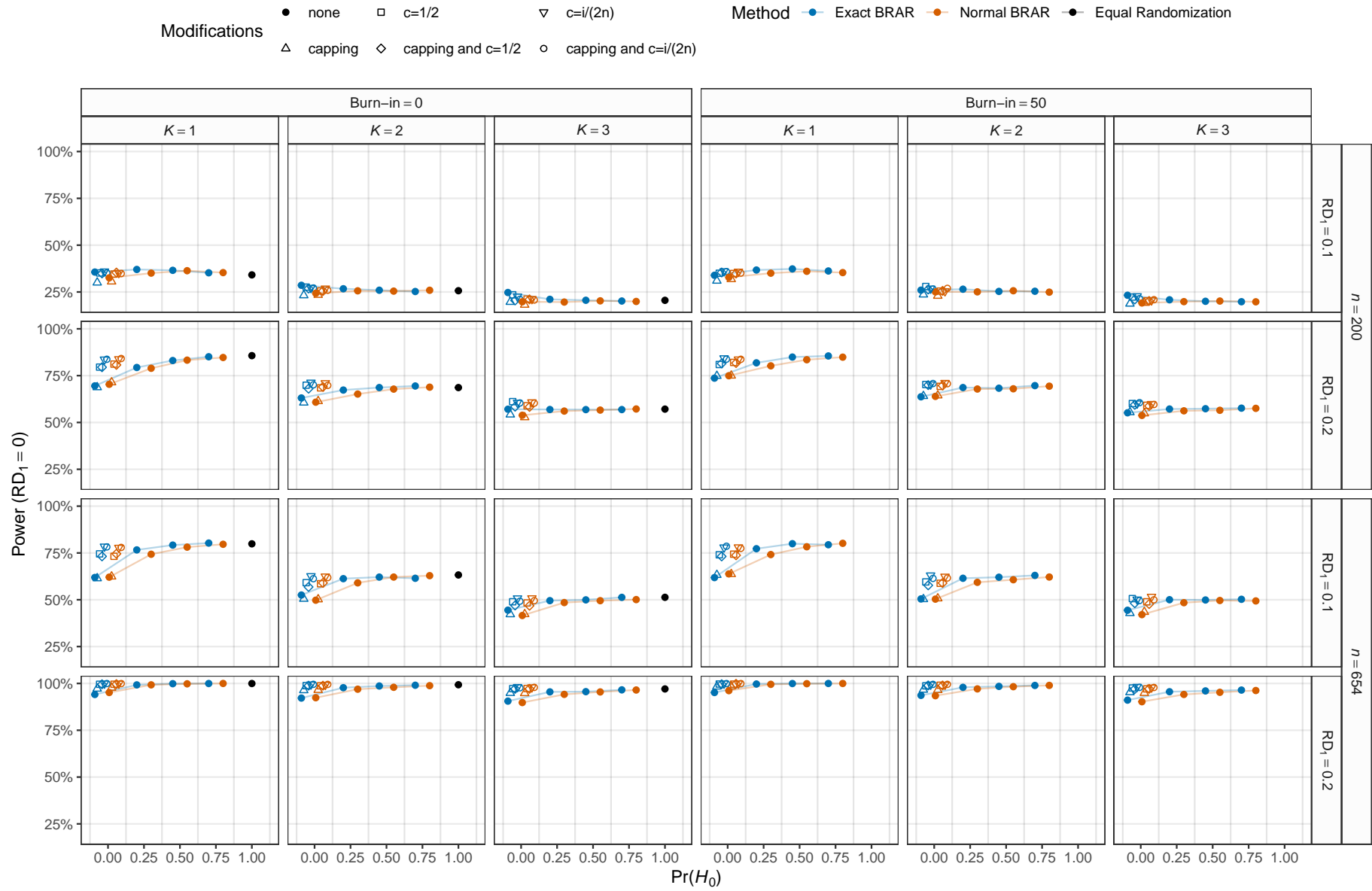


Figure 14: Empirical power of the Wald test of $RD_1 = 0$ based on 10000 simulation repetitions. The maximum MCSE is 0.5%.

groups. In this case, equal randomization was applied. The highest rate of such non-convergence was in 4.6% of the n allocations in a condition with 3 treatment groups. No other forms of missingness were observed. Figures and tables with per-condition-method non-convergence rates are available at <https://brar-42c2c0.pages.uzh.ch/simulation-dashboard.html>.

Rate of successes and extreme randomization probabilities The mean rate of successes is shown in Figure 8. It was generally the highest for Thompson sampling and lowest for equal randomization. The normal approximation and exact method produced mostly similar rates, with the normal method sometimes showing slightly higher rates (e.g., for $K = 3$ and $RD_1 = 0.2$). The Thompson sampling modifications generally reduced the mean success rate. In conditions with small sample size these rates were similar as when the prior probability was $H_0 = 0.75$, while they were lower when the sample size was larger. This makes sense as for larger sample sizes, uncapped randomization probabilities are more likely to converge to extreme ones. This is also visible in Figure 9 where more extreme randomization probabilities are observed with increasing sample size and rate difference for all methods but the ones with capping.

Negative sample size imbalance Figure 10 shows negative sample size imbalance as quantified by the $S_{0.1}$ metric. Imbalance was the greatest for Thompson sampling and reduced when modifications were applied. Similarly, increasing the prior probability of H_0 decreased negative imbalance, in some cases even below Thompson sampling with modifications.

Bias and coverage Figures 11 and 12 show empirical bias and coverage related to estimates of the rate difference between the first treatment group and the control group. In conditions where there was no difference, all methods produced unbiased point estimates, although all methods but equal randomization had undercoverage. Bias occurred in conditions with non-zero rate differences, the bias being the greatest for Thompson sampling and decreasing to some extent when modifications were introduced or the prior probability of H_0 increased. Similarly, modifications or increasing prior probabilities improved coverage, although they still remained suboptimal in most conditions. For large rate differences, coverage was much better for the normal than the exact version of RAR. Similarly, bias was in some cases larger for the exact compared to the normal version.

Type I error rate and power Figures 13 and 14 show empirical type I error rate and power associated with the Wald test of $RD_1 = 0$. We see that standard Thompson sampling shows an inflated type I error rate above the nominal 2.5% which is reduced to some extent by increasing either the prior probability of H_0 or applying modifications. In the same way, Thompson sampling exhibits reduced power compared to equal randomization, which is again alleviated by modifications or positive prior probability of the null hypotheses. In small sample sizes and large rate differences, power was slightly increased for the exact compared to the normal version of RAR, while for Thompson sampling the type I rate was slightly higher for the exact compared to the normal version.

Computational details

```
cat(paste(Sys.time(), Sys.timezone(), "\n"))

## 2025-09-25 10:45:49.751144 Europe/Zurich

sessionInfo()

## R version 4.4.1 (2024-06-14)
## Platform: x86_64-pc-linux-gnu
## Running under: Ubuntu 24.04.3 LTS
##
## Matrix products: default
## BLAS: /usr/lib/x86_64-linux-gnu/blas/libblas.so.3.12.0
## LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.12.0
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8      LC_NUMERIC=C
##  [3] LC_TIME=de_CH.UTF-8      LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=de_CH.UTF-8  LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=de_CH.UTF-8     LC_NAME=C
##  [9] LC_ADDRESS=C             LC_TELEPHONE=C
## [11] LC_MEASUREMENT=de_CH.UTF-8 LC_IDENTIFICATION=C
##
## time zone: Europe/Zurich
## tzcode source: system (glibc)
##
## attached base packages:
## [1] stats      graphics  grDevices  utils      datasets  methods    base
##
## other attached packages:
## [1] Ternary_2.3.5      ggpubr_0.6.0      ggh4x_0.3.0      ggplot2_3.5.2
## [5] mvtnorm_1.3-3      brar_0.1          SimDesign_2.14.1 dplyr_1.1.4
## [9] knitr_1.50
##
## loaded via a namespace (and not attached):
##  [1] gtable_0.3.6      xfun_0.52         PlotTools_0.3.1
##  [4] rstatix_0.7.2     lattice_0.22-6    vctrs_0.6.5
##  [7] tools_4.4.1       generics_0.1.3    curl_5.2.1
## [10] parallel_4.4.1    tibble_3.2.1      highr_0.11
## [13] pkgconfig_2.0.3   RPushbullet_0.3.4 lifecycle_1.0.4
## [16] farver_2.1.2      compiler_4.4.1    stringr_1.5.1
## [19] munsell_0.5.1     codetools_0.2-19  carData_3.0-5
## [22] httpuv_1.6.15     htmltools_0.5.8.1 later_1.3.2
## [25] pillar_1.10.1     car_3.1-2         tidyr_1.3.1
## [28] sessioninfo_1.2.2 iterators_1.0.14   abind_1.4-5
## [31] foreach_1.5.2     mime_0.12         parallelly_1.42.0
## [34] tidyselect_1.2.1  digest_0.6.37     stringi_1.8.4
## [37] future_1.34.0     purrr_1.0.4       listenv_0.9.1
## [40] labeling_0.4.3    cowplot_1.1.3     fastmap_1.2.0
```

```

## [43] grid_4.4.1           colorspace_2.1-1    cli_3.6.4
## [46] beeper_1.3           magrittr_2.0.3      broom_1.0.6
## [49] future.apply_1.11.3 withr_3.0.2         promises_1.3.0
## [52] scales_1.3.0         backports_1.5.0     sp_2.1-3
## [55] globals_0.16.3       audio_0.1-11        gridExtra_2.3
## [58] progressr_0.15.1     ggsignif_0.6.4      shiny_1.8.1.1
## [61] pbapply_1.7-2         evaluate_0.24.0     RcppHungarian_0.3
## [64] viridisLite_0.4.2    rlang_1.1.5         Rcpp_1.1.0
## [67] xtable_1.8-4         glue_1.8.0          jsonlite_1.8.8
## [70] R6_2.6.1

```