Effects of allocation method and time trends on identification of the best arm in multi-arm trials

Journal:	*Statistics in Biopharmaceutical Research*
Manuscript ID	SBR-22-091.R2
Manuscript Type:	RRDART
Keywords:	
Classifications:	arm dropping, response adaptive randomization, Adaptive design

SCHOLARONE™
Manuscripts

Effects of Allocation Method and Time Trends on Identification of the Best Arm in Multi-arm Trials

For Peer Review Only

Abstract

Many trial designs, such as dose-finding trials, shared-control designs, or adaptive platform trials, investigate multiple therapies simultaneously. Often these trials seek to identify the best arm and compare it to a control. Adaptive trials are commonly considered in this space, focusing on methods that drop arms or adjust allocation in response to accumulating information. These methods continue to be compared in the literature, most recently with an emphasis on the effect of time trends during the experiment. Here we compare several methods, considering their performance with and without time trends present. The four procedures are: 1) fixed allocation, 2) arm dropping based on p-values (two variants), 3) arm dropping based on the posterior probability each arm is best (two variants), and 4) response-adaptive randomization. These procedures are compared in terms of their ability to identify the best arm, statistical power, accuracy of estimation, and potential benefit to participants inside the trial. We find arm dropping based on the probability each arm is best and RAR among the best options from the methods considered. Arm dropping based on p-values performs moderately worse, and fixed allocation is much worse on all metrics within this context.

Keywords: adaptive design, arm dropping, response-adaptive randomization

1 Introduction

Many modern clinical trials investigate multiple therapies simultaneously. These include dose-finding trials, shared-control designs comparing multiple independent treatments to a single control arm, and adaptive platform trials. These trials are often designed with adaptive features such as early stopping of individual arms for success or futility and/or response-adaptive randomization (RAR). In either of these methods, arm dropping or RAR, allocation is altered throughout the trial to favor better performing arms and away from poorly performing arms. These adaptations allow resources to be focused on the arms with the most promise.

In this paper we compare adaptive allocation strategies for multi-arm trials with the goal of identifying the best available arm. Identifying the 2nd, 3rd, or worse arms, even if they are effective, is less important in this setting. For a number of real world examples that have an objective of finding the best arm, see Berry and Viele [2023]. The methods we consider have also been studied in applications with different goals such as identifying all effective arms. Contrasting trial objectives may, of course, be best solved by different adaptive strategies. For example, aggressively pursuing the best arm may make it difficult to determine if the 2nd best arm is effective. In an indication area such as antibiotics, we might anticipate the current best arm might not be the future best arm due to the development of resistance. Adaptive designs focused solely on success of the best arm are likely to have lower sample sizes (and thus lower power) or delayed conclusions on the "lesser" arms. It is important to align the adaptive strategy with the goal of the multi-arm experiment. We proceed with the goal of identifying the best arm only.

There have been many recent articles demonstrating the importance of time trends in both hypothesis testing and estimation. As such, we assess the adaptive strategies under scenarios including additive time trends. As noted in Roig et al. [2022], who provide examples of both additive and non-additive time trends, non-additive time trends can introduce bias in standard covariate adjustments for time. We provide a review in section 2. A good adaptive method should be robust to time trends in the experiment.

In this paper, we focus on making the following comparisons:

- We evaluate two forms of arm dropping designs. These include designs that drop arms

For Peer Review Only

on the basis of a p-value and designs that drop arms on the basis of the posterior probability the arm is best. We find that arm dropping designs based on the posterior probability each arm is best perform better on all metrics (identification of best arm, power, and estimation). The quantity is reflective of the specified goal of identifying the best arm only. Previous work has found that arm dropping designs based on p-values perform adequately when interest centers on identifying all active arms. [Freidlin and Korn, 2013]. We consider two variants of these procedures based on whether control allocation is held constant throughout the trial or whether the ratio of control allocation to each active arm allocation is constant throughout the trial (both cannot be maintained simultaneously while dropping arms).

- We compare arm dropping based on the posterior probability each arm is best to a response-adaptive randomization design previously shown to be a good match for this problem [Viele et al., 2020a]. We find both methods perform similarly.
- We compare the adaptive allocation methods in scenarios where there is additive temporal drift in outcomes for all participants in the trial. In these scenarios, the statistical models driving the adaptive designs are implemented with and without a covariate adjusting for time. With time included in the model, arm dropping based on the probability each arm is best and RAR perform similarly, both far outclassing non-adaptive trials. This result, specific to multi-arm trials, contrasts with the behavior seen in two arm (control versus a single treatment) trials, for example in Korn and Freidlin [2022].
- We compare all adaptive allocation methods to non-adaptive (fixed allocation) designs. The non-adaptive design has inferior performance.

The rest of the paper is organized as follows. In section 2 we discuss the proposed allocation procedures from existing literature. In section 3 we give formal definitions of the allocation procedures used in this paper together with detailed descriptions of the models employed and the simulation scenarios and metrics used for comparison of the procedures. In section 4 we present the results of a simulation study comparing the allocation methods without and with time trends. Finally, section 5 provides conclusions and a discussion of

For Peer Review Only

the results. In the supplementary material we supply several more comparisons (varying stopping rules and the inclusions of an MAMS design) that may be of interest.

2 Related work

Response-adaptive randomization is the adjustment of allocation probabilities at interim analyses based on the accumulating outcome data in the trial. There are many variants depending on the exact function used to relate the data to the allocation probabilities. Most generally, arm dropping can be viewed as a special case of response-adaptive randomization where the only change in allocation allowed is to “zero out” the allocation to poorly performing arms, combined with some renormalization of the remaining arms. Typically arm dropping is implemented as permanent, where any dropped arm cannot return under any circumstances. In contrast, many variants of RAR temporarily drop arms, truncating any sufficiently small allocation probability to 0, with the possibility of returning if the future data so indicate.

For Peer Review Only

2.1 Fixed allocation in multi-arm trials

Some multiple arm trials used fixed allocation that may be equal or unequal across arms. Although fixed 1:1 allocation is often considered in two-arm trials, equal allocation may be suboptimal in trials with more than two arms. Dunnett [1955, Section 5] considered J experimental arms and one control arm, all with normally distributed endpoints with equal variance; using the Student’s t-test, the optimal allocation to the control arm (in terms of maximizing power) is slightly lower than $\sqrt{J}$ times the allocation to each of the experimental arms, and $\sqrt{J}$ was suggested as a practical approximate allocation proportion of the control arm. Thus, if the goal of a multiple arm trial is maximizing power of each pairwise comparison to control, the optimal fixed allocation scheme involves allocating a higher proportion of participants to the control arm. Others have suggested that, despite small decreases in power, equal allocation may be more appealing than the optimal allocation ratio since participants have a higher chance of being randomized to non-control arms [Freidlin et al., 2008].

2.2 Arm dropping

In multi-arm arm dropping designs, several experimental arms are enrolled along with a shared control and experimental arms may be dropped from further allocation based on efficacy and/or futility. Interim analyses and stopping boundaries are pre-specified to evaluate whether experimental arms can be dropped.

A well-studied and commonly implemented family of arm dropping designs are group-sequential designs in which stopping boundaries are typically based on a test statistic (or p-value) comparison of each experimental arm to control. One type of group-sequential designs is the pick-the-winner (or drop-the-losers) design in which K experimental arms are compared to a shared control in stage 1, and the optimal arm (with highest test statistic) is selected to continue to stage 2 with the control arm [Stallard and Todd, 2003, Thall et al., 1988]. Group-sequential designs have been generalized to select multiple experimental arms to continue to stage 2 [Stallard and Friede, 2008, Kelly et al., 2005], and to perform arm selection across multiple stages [Wason et al., 2012, Wason and Trippa, 2014, Wason et al., 2017, Magirr et al., 2012]. In this paper, we evaluate group sequential designs in which multiple arms can be dropped at each interim for futility (and not efficacy) similar to the group sequential approach in Magirr et al. [2012], which we refer to as the Arm dropping comparing to the placebo control (AD PBO) designs.

Another type of group-sequential designs are the multi-arm multi-stage (MAMS) designs. In this paper, the MAMS design differs from the other arm dropping designs in two main ways. First, the interim analyses in the MAMS design are timed based on a fixed number of participants enrolled per arm rather than a fixed number of total participants across all active arms [Wason and Trippa, 2014]. Second, in the MAMS design, the trial may stop for futility if all experimental arms are dropped whereas the other arm dropping designs in this paper keep at least one experimental arm for the duration of the trial.

In this paper we introduce a family of arm-dropping designs which are not based on a test statistic, but instead, employ the posterior probability that each experimental arm is the best, which we refer as the AD MAX designs. To the best of our knowledge, there is no existing literature discussing such designs.

For Peer Review Only

2.3 Response-adaptive randomization

A wide variety of RAR procedures have been proposed and studied in the academic literature. In general, RAR is an intuitive procedure that randomizes participants to optimise an operating characteristic (e.g. statistical power, type I error, participant benefit) or a combination of operating characteristics. RAR procedures can be based on Bayesian quantities (e.g., posterior or predictive probability), frequentist quantities (e.g., confidence bound of a maximum likelihood estimator), or ad hoc quantities (e.g., urn models). RAR procedures can be used to design experiments regardless of whether the data analyses (interim and final) are Bayesian or frequentist. Williamson [2020, Chapter 2], Grieve [2017], Robertson et al. [2023] include reviews of RAR procedures. Response-adaptive procedures have also been developed in other disciplines, as solutions to the multi-armed bandit problem under a variety of performance measures, see, e.g., Jacko [2019], and the resulting randomised variants referred to as bandit-based RAR procedures, see, e.g., Villar et al. [2015], Williamson et al. [2022].

There are two common approaches for setting RAR allocation probabilities in the Bayesian setting: (1) using the posterior probability each arm is the best arm or better than another arm, often referred to as Bayesian response-adaptive randomization (BRAR) [Thompson, 1933], and (2) obtained by optimizing the expected reward using Bayesian decision theory, which originate from Bradt et al. [1956], Bellman [1956]. In this paper, the focus is on the former, as the latter is not well developed for trials with more than two arms.

In any Bayesian approach, prior distributions are specified for unknown parameters. BRAR procedures depend on posterior estimation of treatment effect parameters and, as a result, may be influenced by choice of priors. In the BRAR procedure, a Bayesian model is fit at pre-specified interim analyses and parameters summarising the treatment effect on each arm are estimated. Allocation probabilities for each arm are updated using posterior probabilities derived from the model. Examples of BRAR procedures include Thompson sampling (TS) in which participants are allocated to arms in proportion to the posterior probability each arm is best or in proportion to the probability each arm is superior to control, and their generalizations (such as using clipping or exponentiation of the posterior

For Peer Review Only

probability to control the rate at which the randomisation probabilities are allowed to change over the duration of the trial).

These generalizations of BRAR are important for practical implementation in clinical trials [e.g. Thall and Wathen, 2007, Trippa et al., 2012]. The methodology for this class of procedures is well developed to address a variety of practicalities, including delayed responses, the use of interim analyses instead of a fully-sequential updating, incorporation of models with covariates, continuous or time-to-event outcomes [Viele et al., 2020a], and different trial types, e.g., for rare disease trials [Wang, 2021] or N-of-1 trials [Shrestha and Jain, 2021]. BRAR procedures have become increasingly popular in practice [see Biswas et al., 2009, Lee et al., 2010] and have been successfully implemented in several clinical trials, particularly cancer trials, to allocate more participants to treatments that have performed well for similar participants following the recent surge in personalised medicine [Wason et al., 2015]. Notable examples in oncology include the I-SPY 2 [Barker et al., 2009], BATTLE [Kim et al., 2011] and BATTLE-2 [Papadimitrakopoulou et al., 2016, Gu et al., 2016] trials. In this paper, we implement a generic variant of BRAR similar to those implemented in the above trials, which we call RAR for simplicity; see subsection 3.1.

For Peer Review Only

2.4 Inferential risks in this setting

Our goal of identifying the best arm and comparing to control results in some inherent multiplicities. Even in non-adaptive settings, point estimates of the best arm are upwardly biased, and hypothesis tests must account for the selection or risk inflated type I error. For fixed trials, fortunately many methods exist to correct for this inflation due to multiple testing and control the family-wise error rate at desired levels [Dunnett, 1955, Bauer, 1991]. In arm dropping and RAR, we rely on simulation to obtain a cutoff which maintains type I error control.

If the allocation ratio changes during a trial, then there is the risk of potential bias from non-comparable treatment groups. For example, if allocation is 1:1 (experimental:control) in time period 1 and 3:1 in time period 2 with 100 participants enrolled in each period, then the proportion of participants from each time period will differ between groups. We find 67% of the control patients are from time period 1, while only 40% of the treatment patients are from time period 1. If there is a change in the outcome distribution between time periods, this change in the allocation ratio could result in non-comparable groups and bias in the treatment effect estimate. This risk of non-comparable groups applies to any change in the experimental:control randomization ratio including arm dropping, RAR, and adaptive platform trials where arms are added/removed [Roig et al., 2022, Saville et al., 2022, Thall et al., 2015]. Villar et al. [2018] discuss the issue of type I error inflation due to unknown time trends with several variants of RAR including a Thompson Sampling approach, and demonstrate that, when linear time trends are present, covariate adjustment for time trends can avoid type I error inflation in a two-arm trial using a bandit-based variant of RAR. The RAR design we evaluate has two key differences from the Villar et al. [2018] TS method: 1) simulation-based control of type I error when no time trends are present to allow for a direct comparison across allocation procedures and 2) maintenance of the control allocation throughout the trial as recommended in Villar et al. [2018] and Viele et al. [2020a] to avoid reduction of power and poor estimation of the treatment effect. We extend the Villar et al.'s exploration of the impact of covariate adjustment for time to the multi-arm context for several adaptive allocation procedures (including our variant of Thompson Sampling-based RAR) and a broad range of time trend and efficacy scenarios.

For Peer Review Only

3 Methods

We consider a trial exploring four experimental arms compared to a control arm with a dichotomous primary outcome. We denote the set of arms by $\mathcal{J} := \{0, 1, 2, 3, 4\}$, where arm $j = 0$ corresponds to the control arm. The primary goal of the trial is to determine if any experimental arm is superior to control, with important secondary goals of correctly identifying the best arm and accurately estimating the treatment effect for that best arm relative to control. After these goals, where possible, we would like to treat participants within the trial as effectively as possible. We conduct a simulation study within this general trial structure and compare multiple allocation procedures. We explore a range of treatment effect and additive time trend scenarios. We evaluate the methods with metrics including statistical power, identification of the best arm, accuracy of treatment effect estimation, and benefit to participants within the trial.

3.1 Allocation procedures to be compared

We compare 10 procedures for allocating participants to arms in the trial; 6 are presented in the main paper and 4 in Appendix B. We include a fixed design with constant allocation ratios to provide a non-adaptive reference method. Our primary interest is to compare three adaptive allocation methods, considering 9 particular procedures due to different choices of methods' tuning parameters. Adjustments to the allocation ratio for the adaptive procedures will occur at interim analyses. Each trial enrolls the first 48 participants during a run-in period with fixed 2:1:1:1:1 allocation to control and the four experimental arms. After the initial run-in period, the first interim analysis occurs and allocation ratios may be adjusted or arms dropped based on the accrued data and adaptive design. Each subsequent analysis is performed every 24 participants up to the maximum sample size of 240, resulting in a total of 8 interim analyses and a single final analysis (described in subsection 3.3). The adaptive designs may allow for individual arms to be dropped at interim analyses, but do not incorporate early stopping of study enrollment for efficacy/futility (except the MAMS design, see Appendix A). As a result, all trials enroll to the maximum sample size with the control arm and at least one experimental arm throughout the trial. Since the metrics under consideration value inferential accuracy, we have focused on making use of all available patients to maximize information. Designs which have superior performance for a specific fixed sample size can usually also be formulated, via stopping rules, to obtain equivalent performance at smaller sample sizes. We compare the following allocation procedures:

1. **Fixed:** Fixed allocation ratio of 2:1:1:1:1 throughout the entire trial. No interim analyses are performed in this design. The control allocation ratio is the square root of the number of experimental arms ($\sqrt{4} = 2$), which is approximately optimal in some specific settings as discussed in subsection 2.1.

2. **AD PBO A and AD PBO B:** Arm dropping based on p-value comparison to placebo control (PBO), calculated at each interim analysis. We utilize group sequential futility bounds from a Hwang-Shih-DeCani spending function with a parameter $\gamma = -0.5$ [Hwang et al., 1990]. These futility bounds approximate the commonly used futility bounds suggested by Freidlin and Korn [2013]. If the p-value for an active arm exceeds the futility p-value bound, the arm is permanently dropped from

For Peer Review Only

the trial (and is not active anymore). If the p-values for all active arms exceed the futility threshold at an interim analysis, then the active arm with the lowest p-value is kept active for the remainder of the trial. In Version A, randomization between interim analyses entails 8 of 24 participants allocated to control and the remaining 16 allocated equally to the active (non-dropped) arms. In Version B, randomization between interim analyses is 2:1 to control and each active arm. Maintenance of the control allocation provides a more apples to apples comparison of the allocation of the active arms, while maintenance of the allocation ratio provides protection from time trends when adjustment is not employed. Both cannot be obtained simultaneously.

For Peer Review Only

3. **AD MAX A and AD MAX B:** Arm dropping based on the posterior probability, "Pr(max)" (calculated at each interim analysis), that each active arm j is the best out of the four experimental arms:

$$\Pr(\max)_j := \mathbb{P}(\theta_j = \max_{i \in \mathcal{J} \setminus \{0\}} \theta_i \,|\, \text{data}) \quad \text{for } j \in \mathcal{J} \setminus \{0\} \tag{1}$$

where θ_j is the treatment effect of the jth experimental arm relative to control (see subsection 3.3). As this probability is always referenced by the available data at decision time (allocation update or final analysis), in what follows we simply use $\Pr(\max)_j$ to indicate the posterior probability with the current data, and thus $\Pr(\max)_j$ changes as the data accumulate. If $\Pr(\max)_j$ falls below the threshold of 0.10, we permanently drop arm j. We considered alternative thresholds ranging between 0.05 and 0.15 and selected 0.10 based on a variety of metrics including statistical power, average sample size on best arm, and estimation accuracy (see subsection 3.5). Analogously to the AD PBO procedure, we consider Version A (with 8/24 control participants between interims) and Version B (with a fixed 2:1 ratio between control and each active arm) of the AD MAX design.

4. **RAR:** RAR based on $\Pr(\max)_j$ (as defined in (1)) is used to allocate participants in this design. Control allocation is fixed to 1/3 (exactly 8 out of every 24 participants between interim analyses), and allocation to the active arms is proportional to $\Pr(\max)_j$ and normalized to sum to 2/3. If $\Pr(\max)_j$ falls below 0.125 for an experimental arm, the allocation probability for that arm is set to zero until the next

interim analysis occurs (i.e., the arm is not active for 1 randomization period). The threshold of 0.125 was chosen so that, between two interim analyses, each active arm would be allocated at least 2 of the 16 non-control participants in expectation. If an arm allocation is set to 0, the remaining allocation probabilities for the active arms are re-normalized to sum to 2/3 (1/3 is always allocated to control).

As mentioned above, AD designs are implemented with two allocation ratio versions. Version A of each design uses randomization blocks of size 24 with 8 participants (i.e., 1/3) always allocated to the control arm, as in Viele et al. [2020b]. This results in a fixed number of control participants between each interim analysis, but the allocation ratios between each experimental arm and control may vary. Version B of each design fixes the allocation ratio for control and each experimental arm to 2:1, which mitigates the risk of non-comparability of treatment groups (cf. subsection 2.4), so the number of control participants between each interim may vary if arms are dropped.

Four additional group-sequential procedures are described and evaluated in Appendix B: MAMS (see Appendix A for details) with spending function with a parameter $\gamma = -0.5$, and more conservative futility bounds from a Hwang-Shih-DeCani spending function with a parameter $\gamma = -4$, which approximates O'Brien-Fleming bounds, for MAMS, AD PBO A, and AD PBO B.

For Peer Review Only

3.2 Analysis model

At each interim and the final analysis, we analyze all of the current data with a logistic regression model and estimate an odds ratio comparing the response rate for each experimental arm to the response rate of the control arm. We consider models with and without a covariate adjustment for time. We denote the response for participant i as y_i and model $y_i \sim \text{Bernoulli}(p_i)$ where p_i is the probability of a response. Without adjustment for time, the model specification is:

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \boldsymbol{x}_i'\boldsymbol{\theta} \qquad (2)$$

where $\boldsymbol{x}_i$ is a 4-dimensional vector of treatment indicators for each of the experimental arms and $\boldsymbol{\theta}$ is a 4-dimensional vector of coefficients for each of the experimental arms. The $\boldsymbol{\theta}$ coefficients are treatment effects for each experimental arm relative to control and can be interpreted as log odds ratios relative to control. Including adjustments for time, the model specification is:

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \boldsymbol{x}_i'\boldsymbol{\theta} + \boldsymbol{t}_i'\boldsymbol{\beta} \tag{3}$$

where $\boldsymbol{t}_i$ is a vector of indicators for each non-reference time period and $\boldsymbol{\beta}$ is the vector of time effects. The time periods are defined based on the timing of interim analyses so that allocation ratios are constant for all participants within a time period. Time period 1 is the reference period in the model and includes all participants randomized in the run-in period before the first interim analysis. At the first interim analysis, there is no time adjustment since all participants have been randomized in the same time period.

In the AD PBO designs, arm dropping is based on the p-value of the coefficient for each experimental arm in the frequentist logistic regression models given in (2) and (3). For designs based on Pr(max), the logistic regression model is fit within the Bayesian paradigm. Because the Bayesian analysis provides a joint posterior distribution for all experimental arm treatment effects, $\Pr(\max)_j$ is simple to compute for a given experimental arm as the proportion of posterior draws where that arm has the maximum coefficient across all experimental arms (see (1)). In the Bayesian analysis model, prior distributions are specified for each unknown model parameter. For α and each element of the vectors $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$, we specify independent, normally distributed prior distributions with mean 0 and standard deviation 2. For α, this prior on the log odds scale induces a prior that is approximately uniformly distributed on the probability scale. For log odds ratios $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$, this prior is weakly informative, centered on 0 (no effect), with a 95% interval ranging approximately from -4 to 4. The resulting posterior distribution is insensitive to other choices of weakly informative priors on the log-odds scale.

For Peer Review Only

3.3 Final analysis

The final analysis occurs at the maximum sample size of 240 participants (except for MAMS, see Appendix A). At the final analysis, we evaluate the following hypotheses of interest:

$$H_0: \quad \theta_1 = \theta_2 = \theta_3 = \theta_4 = 0 \tag{4}$$

$$H_1: \quad \max_{j \in \mathcal{J} \setminus \{0\}} \theta_j > 0 \tag{5}$$

where θ_j is the log odds ratio for each experimental arm compared to control. For all designs, the final analysis model is a Bayesian logistic regression model. We use j^* to denote the arm with the highest $\Pr(\max)_j$ among the active arms. We evaluate the hypotheses based on the posterior probability that arm j^* is better than control (i.e., a log odds ratio greater than 0). Specifically, the trial will claim superiority of the best arm over control if

$$Pr(\theta_{j^*} > 0 \mid \text{data}) > \delta \tag{6}$$

where the threshold δ is specific to each design. For each design, δ is selected by simulation to control the one-sided type I error rate at 2.5% (the standard for confirmatory trials, see U.S. Food and Drug Administration [2019]) under the global null hypothesis without a time trend. Note that only the active arms are eligible to be selected as the best arm at the final analysis. That is, if an arm is dropped in an arm dropping design, it cannot be selected as best. This restriction reflects common practice.

3.4 Scenarios of interest

Each of the designs described above is implemented using every combination of the seven treatment effect scenarios in Table 1 and five time trend scenarios in Table 2 as the underlying true response rate scenarios for trial simulation. The treatment effect scenarios are similar to those considered in Viele et al. [2020b] and Wason and Trippa [2014]. In all scenarios, the initial response rate of the control arm is 30% and we vary the experimental arm response rates in different treatment patterns. Table 1 presents each treatment effect scenario in terms of the response rate by arm (assuming no time trends are simulated) and the log odds ratio for each experimental arm relative to control.

Crossing the 7 treatment effect scenarios with 5 additive time trend scenarios and 6 designs results in 210 separate simulation scenarios. For data simulation, time periods are defined by the 48 participants run-in (time period 1) and every 24 participants thereafter

For Peer Review Only

Scenario	Response rate (no time trends)					Log odds ratio (relative to control)			
	Control	Arm 1	Arm 2	Arm 3	Arm 4	Arm 1	Arm 2	Arm 3	Arm 4
Null	0.30	0.30	0.30	0.30	0.30	0	0	0	0
Nugget	0.30	0.30	0.30	0.30	**0.53**	0	0	0	**0.97**
Two Low	0.30	0.30	0.30	**0.41**	**0.53**	0	0	**0.48**	**0.97**
Two High	0.30	0.30	0.30	**0.45**	**0.54**	0	0	**0.65**	**1.00**
Three Mixed	0.30	0.30	**0.40**	**0.45**	**0.50**	0	**0.44**	**0.65**	**0.85**
Least Favorable	0.30	**0.40**	**0.40**	**0.40**	**0.53**	**0.44**	**0.44**	**0.44**	**0.97**
Mixed	0.30	**0.35**	**0.41**	**0.47**	**0.53**	**0.23**	**0.48**	**0.73**	**0.97**

Table 1: Treatment effect scenarios presented in terms of the response rate and log odds ratios. These scenarios represent a variety of possible underlying true response rates, with varying numbers of effective arms and differing rates for effective but suboptimal arms.

For Peer Review Only

(time periods 2-9). A time scenario consists of shifts for all arms within time periods 2-9. We consider a flat scenario with no time trend, linear trends going up and down, a seasonal trend, and a distinct changepoint scenario.

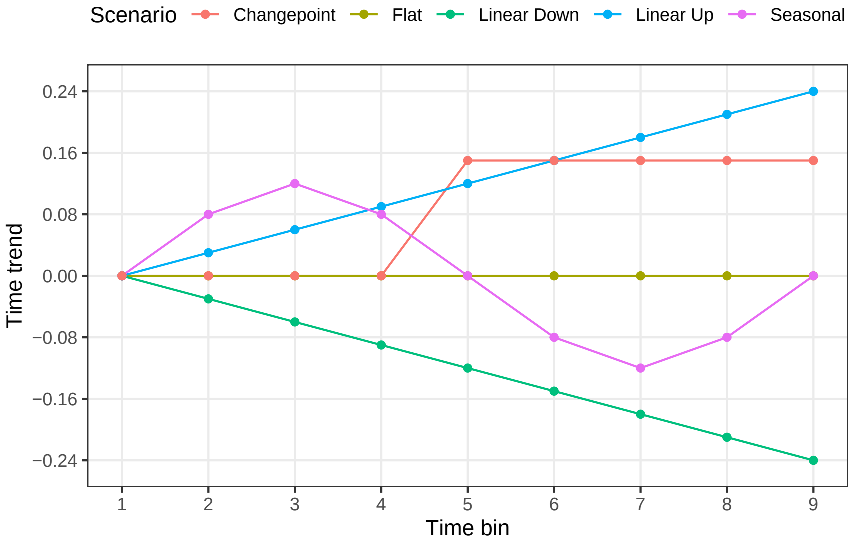
The time trends in Table 2 are expressed in terms of their response rate change for the control arm, but all shifts are on the log odds scale. Thus, in the linear up time trend scenario, in time period 3 the control response rate increases by 0.06 (from 0.30 during run-in to 0.36 in time period 3). This corresponds to a log odds increase of 0.2719. All the active arms in time period 3 are also shifted by 0.2719 on the log odds scale (thus a 35% response rate would become 41.4%).

3.5 Metrics for comparison

We compare the allocation methods above using a suite of metrics similar to Viele et al. [2020b] and Gajewski et al. [2019]. These metrics are computed from 10,000 simulated trials for each combination of efficacy and time scenarios. The metrics include:

1. **Statistical Power**: The probability that the trial declares statistical significance for at least one experimental arm compared to control.

2. **Estimation Accuracy**: Mean squared error (MSE) of the treatment effect estimate



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

Table 2: Time trend scenarios (change in response rate).

iew Only

(on the log odds scale) for the experimental arm selected as best. The MSE is the sum of the variance and squared bias of the treatment effect estimate.

3. **Regret / Responders Below Optimal**: The "cumulative regret" of a trial is the number of successes observed in the trial subtracted from the maximum number of successes that could have been achieved had the success rates for each arm been known in advance [Robbins, 1952]. We present the average cumulative regret. The best value of regret is 0 and positive values indicate a greater number of failures on average.

4. **Average Sample Size on Selected Arm**: The average number of participants randomized to the experimental arm selected as best.

5. **Best Arm Identification**: The probability that the true best arm is selected and demonstrates superiority to control.

6. **Ideal Design Percentage**: A metric that captures both power and best arm identification. At the end of the trial, we identify a best experimental arm and assess whether that arm is statistically significantly better than control. For a given trial, let p_{sel} be the true response rate for the selected arm, where "selected" means the arm was identified as the best arm and it is statistically significant. The control arm is "selected" if statistical significance is not achieved for an experimental arm. The "selected" arm is intended to represent the arm that might be applied to future external participants based on the conclusion of the trial. $E[p_{sel}]$ is computed by averaging over the power and arm selection across simulated trials.

   The ideal design percentage (IDP) linearly rescales $E[p_{sel}]$ to be between 0% and 100% based on the location of $E[p_{sel}]$ between the minimal and maximal true response rates in the experiment. An IDP of 100% indicates perfect power and arm selection (i.e., the correct best arm is always identified and declared significant). An IDP of 0% indicates the true worst arm is always chosen, a very unlikely result for any rational experiment, but if the control arm is the true worst arm in the experiment, an IDP of 0% also corresponds to not conducting the experiment at all, when future patients would continue to receive the control as standard of care. The IDP combines power

For Peer Review Only

and best arm identification and takes into account the magnitude of arm misselection. For example, if the two best arms in an experiment are similar, picking the true 2nd best arm is a minor mistake. This minor mistake would significantly reduce the probability the exact best arm is selected (previous metric) while IDP correctly notes the small magnitude of the error.

4 Results

All simulations described above were conducted with custom code in the R programming language [R Core Team, 2022]. The code is publicly available at [censored link]. The results presented below are based on 10,000 simulation replications per efficacy and time trend scenario for each design (resulting in a Monte Carlo standard deviation of 0.16% for simulated type I error rates). We divide the results into three subsections. The first accounts for situations where no time trend is present and no time trend is fit in the model, consistent with much of the prior work in this area. We then consider scenarios where a time trend is present, but time is not included in the model, where we might expect poor performance and biases resulting from poor model fit. Finally, we consider scenarios where a covariate for time trend is included in the model.

For Peer Review Only

4.1 Flat time scenario without covariate adjustment for time

Figure 1 shows the results for the flat time scenario with no time adjustment in the model. Three primary results are evident:

1. Across all metrics and scenarios, AD MAX A and RAR are among the best options, with very similar performance. AD MAX A has a slight advantage over RAR in terms of the sample size on the selected arm and the treatment effect MSE, while RAR has a slight advantage in terms of power, IDP, arm selection, and regret. In practice it seems unlikely the small differences between AD MAX A and RAR would exclude either as a top choice. The differences between AD MAX A and B are minimal in terms of power, IDP and effect MSE. In terms of the average sample size of the selected arm, AD MAX A has a clear advantage due to the restrictions on allocation

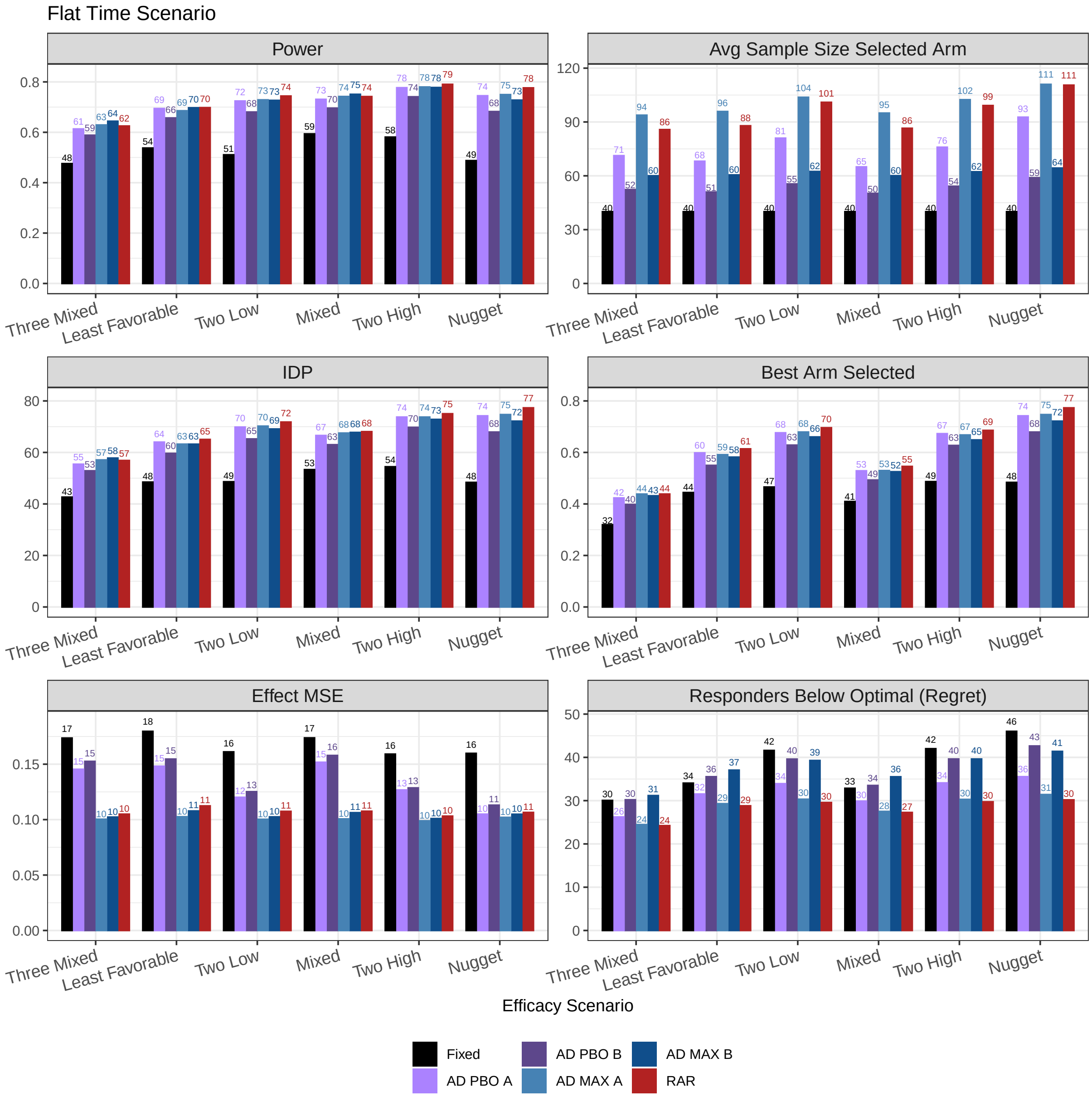


Figure 1: Simulation results in the flat time scenario with no modeled covariate adjustment for time. Each panel summarizes a performance metric on the y-axis. The six non-null efficacy scenarios are shown on the x-axis.

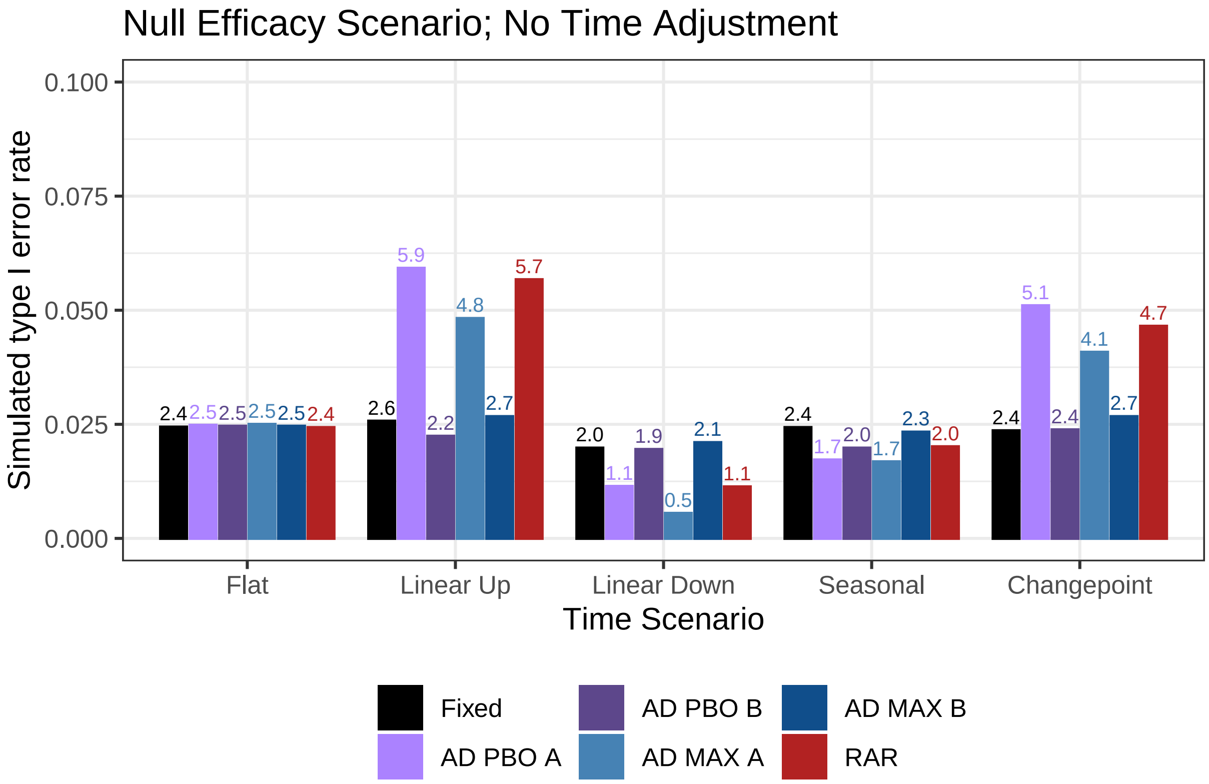


Figure 2: Simulated type I error rates in the null scenario without covariate adjustment for time. The five time trend scenarios are shown on the x-axis.

eview Only

that are imposed on AD MAX B. AD MAX A also has an advantage in selecting the best arm and regret compared to AD MAX B.

2. While AD PBO A is among the best options for power, IDP, and arm selection, AD PBO A noticeably under-performs on the sample size allocated to the selected arm, effect MSE, and regret. Note that AD PBO B performs worse than AD PBO A across the range of metrics. In this application where interest is centered on finding the best arm, AD MAX appears to be a meaningfully better arm dropping option than AD PBO.

3. The non-adaptive fixed design performs significantly worse on all metrics.

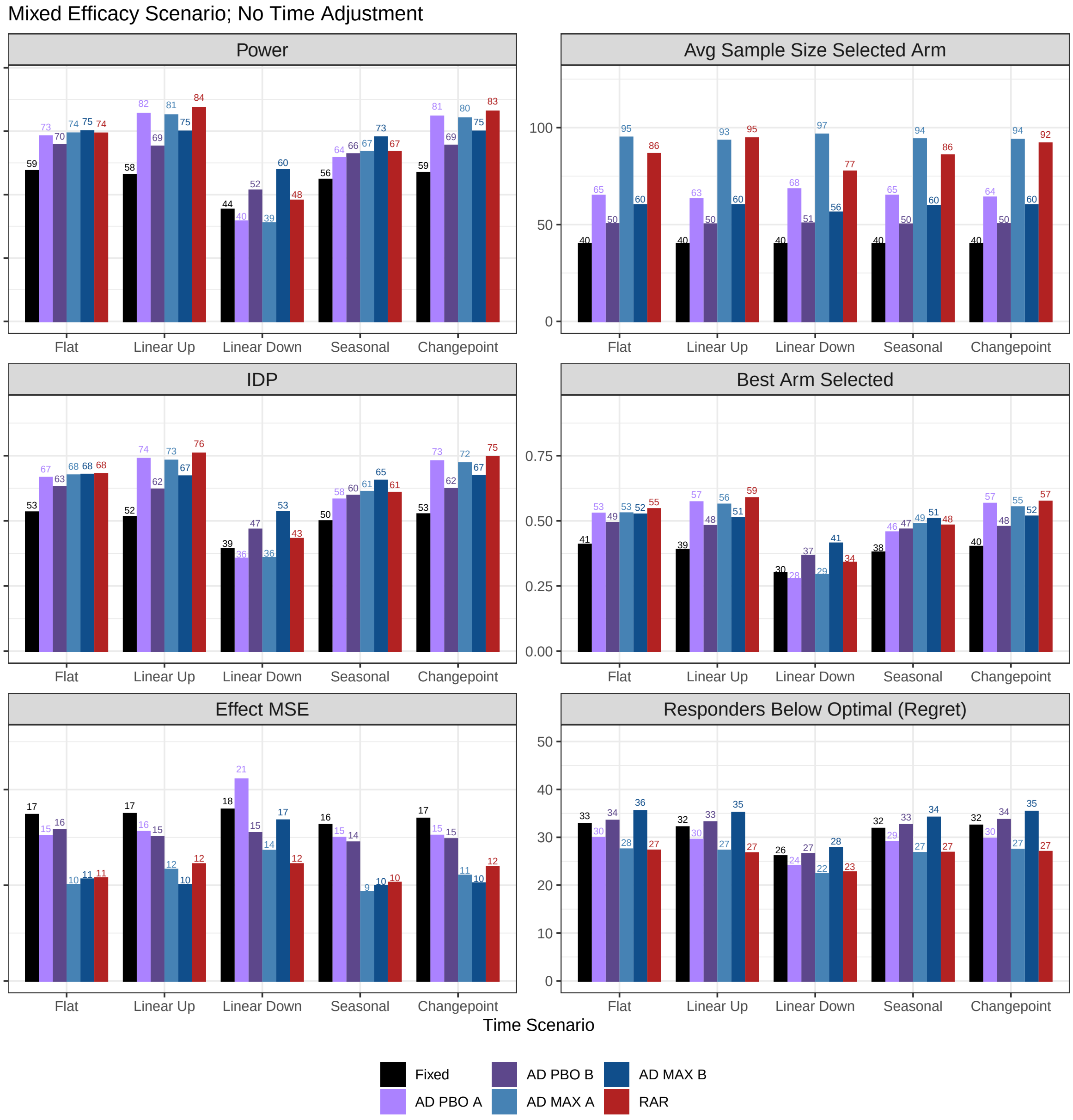


Figure 3: Simulation results in the Mixed efficacy scenario when time trends are simulated but no covariate adjustment for time is included in the analysis model. Each panel summarizes a separate metric on the y-axis. The x-axis shows the five time trend scenarios.

4.2 Time trend scenarios without covariate adjustment for time

Next, we present results from trials simulated with time trends, but without modeled covariate adjustment for time in the analysis models. Omitting this needed covariate significantly negatively impacts the performance of the adaptive designs. While we recommend including a time covariate, we include these results here as unadjusted models have been used in practice and it is important to document the risks associated with omitting time from the modeling. Figure 2 (null scenario) and Figure 3 (Mixed efficacy scenario) illustrate the type I error rates and performance on all metrics, respectively. In both figures, the x-axis in each panel is the simulated time trend scenario. The flat time scenario results are identical to those shown for the Mixed efficacy scenario in subsection 4.1. We observe the following:

For Peer Review Only

1. Most importantly, Figure 2 shows that type I error can be significantly inflated depending on the design and form of the time trend. The designs that maintain a 2:1 randomization ratio of placebo to active arms (Fixed, AD PBO B, AD MAX B) are generally protected from the effect of (additive) time trends, even if no explicit covariate is included, because they maintain strict balance between control and all active arms in every time period. However, for adaptive designs in which the relative randomization ratio of placebo to active arms can change between time periods (AD PBO A, AD MAX A, and RAR), for example, the "linear up" time scenario produces one-sided type I error rates of 5-6% (instead of the desired 2.5%). This inflation of type I error in time trend scenarios with increasing response rates throughout the trial (linear up and changepoint) presumably occurs due to a higher proportion of participants from later time periods with better response rates being assigned to active experimental arms than to control. In the linear down scenario, where response rates constantly decrease, type I error can be deflated since a higher proportion of patients from later time periods with worse response rates are allocated to active experimental arms than to control.

2. Comparing methods in Figure 3, the overall ordering of methods remains similar to Figure 1, although the overall picture is more complex. AD MAX A and RAR

together have the lowest regret and the highest average sample size on the selected arm across all time scenarios. However, these methods, along with AD PBO A, are less robust to the time effects in the metrics of power, IDP, and best arm selection than the methods that maintain a 2:1 randomization ratio. For example, in the best arm selection metric, AD PBO A, AD MAX A, and RAR tend to be lower than the other methods in the linear down time trend scenario, and higher than the others in the linear up scenario (see point below). As with the previous results, the fixed allocation design is generally among the worst performing designs on all metrics. Note that given the potential for inflated type 1 error for time unadjusted models, any inferential advantages should be weighed carefully against the increased risk of false claims of efficacy.

3. The robustness that fixed allocation ratios provides in the presence of unadjusted time trends is evident in these results. In the designs with constant 2:1 randomization ratios, underlying time trend changes impact the control and active arms with the same weight over the course of the trial. For example, if there was a sudden jump in the response rate for all participants enrolled in the last two randomization periods of the trial, each arm would observe that increased rate in exactly 20% of the randomized participants. To contrast, in the designs with fixed 8 participant control allocation per time period, the same jump in response rate may impact different percentages of participants in the control and active arms.

For Peer Review Only

4.3 Covariate adjustment for time

Finally we consider models that include a covariate adjustment for time. Figure 4 displays the simulated type I error rates in the null scenario for each time trend scenario and Figure 5 summarizes all of the performance metrics for the Mixed efficacy scenario across the five time trend scenarios. Additional efficacy scenarios are included in Appendix B. We make the following observations and comments:

1. The threshold for success was calibrated to control the type I error in the flat time trend scenario, but the resulting threshold reasonably controls type I error for all time

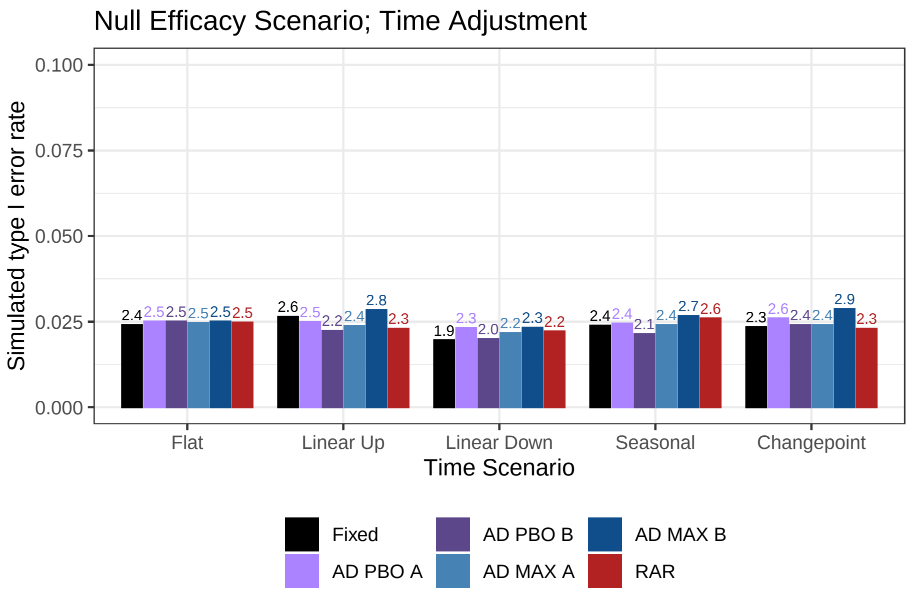


Figure 4: Simulated type I error rates in the null scenario with covariate adjustment for time. The five time trend scenarios are shown on the x-axis.

Peer Review Only

trend scenarios considered here when time is included as a covariate in the model (see Figure 4).

2. Comparing methods in Figure 5, the ordering of methods is very similar to the results under the flat time trend. AD MAX A, AD MAX B and RAR have similar performance in power, IDP, effect MSE, and best arm selected, and outperform the Fixed and AD PBO A and B for these metrics. AD MAX A and RAR clearly outperform the other designs in average sample size on the selected arm and regret. Patterns that emerged when time was not included as a covariate disappear with the inclusion of time in the models (for example, the poorer performance of RAR in average sample size of the selected arm and best arm selected for the linear down time trend scenario). Again, the fixed design performs the worst when comparing across all metrics.

3. Within allocation strategy (i.e. the A and B designs), AD MAX tends to outperform AD PBO. In particular, the AD MAX strategy has a lower effect MSE and higher average sample size on the selected arm. The result is intuitive in this setting. When

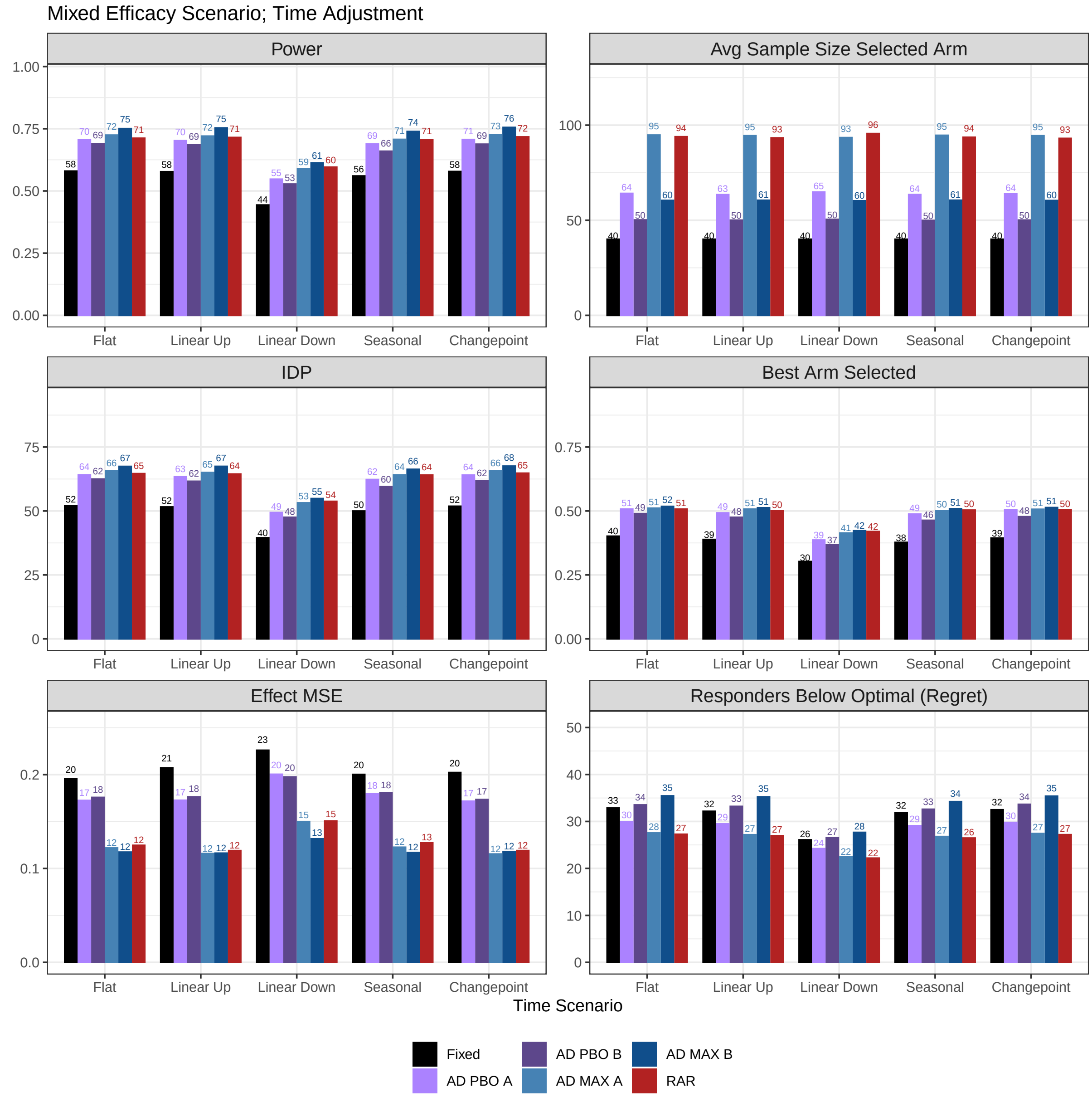


Figure 5: Simulation results in the Mixed efficacy scenario when time trends are simulated and a covariate adjustment for time is included in the analysis model. Each panel summarizes a separate metric on the y-axis. The x-axis shows the five time trend scenarios.

the goal is to find the best arm in a trial, it is natural to rely on the $\Pr(\max)_j$ as a direct comparison of the arms rather than each arm's comparison to the placebo control.

4. Linear down is the most difficult time trend for all of the designs in terms of the performance metrics provided here. While the ordering of the designs is consistent in the linear down scenario, there are simply fewer observed events, and thus we expect lower performance on all metrics (except regret, due to fewer expected events).

5. We examined the "cost" of adding a time covariate to the model in the flat time scenario. We compare the flat time scenario without a time covariate in the model to the flat time scenario with a time covariate in the model. In this case, time is an unneeded covariate and we would expect some cost to estimating these "extra" parameters.

   Figure 6 shows the difference of each metric in the flat time scenario, with and without a modeled time trend. Care must be taken not to over-interpret the ordering of the methods in this comparison. A method which is 30% better than a non-adaptive design on a metric can lose 10% of its value when adjusting for time, and still be better than a non-adaptive design that loses nothing by incorporating time. These graphs are informative in the setting where a time trend is unexpected but we wish to assess whether a time trend should be modeled for robustness. Generally speaking, incorporating the time trend when it is unneeded results in a 1-5% power loss (largest in the nugget scenario followed by two low) with similar impacts on arm selection and IDP. Effect MSE also tends to increase by 1-2% when including the "unneeded" time parameter. Interestingly, the only method affected in terms of average sample size on the best arm is RAR, adding 2-7 participants on the selected arm. Noting the y-axis in the regret graph, including time results in minimal change for all designs.

For Peer Review Only

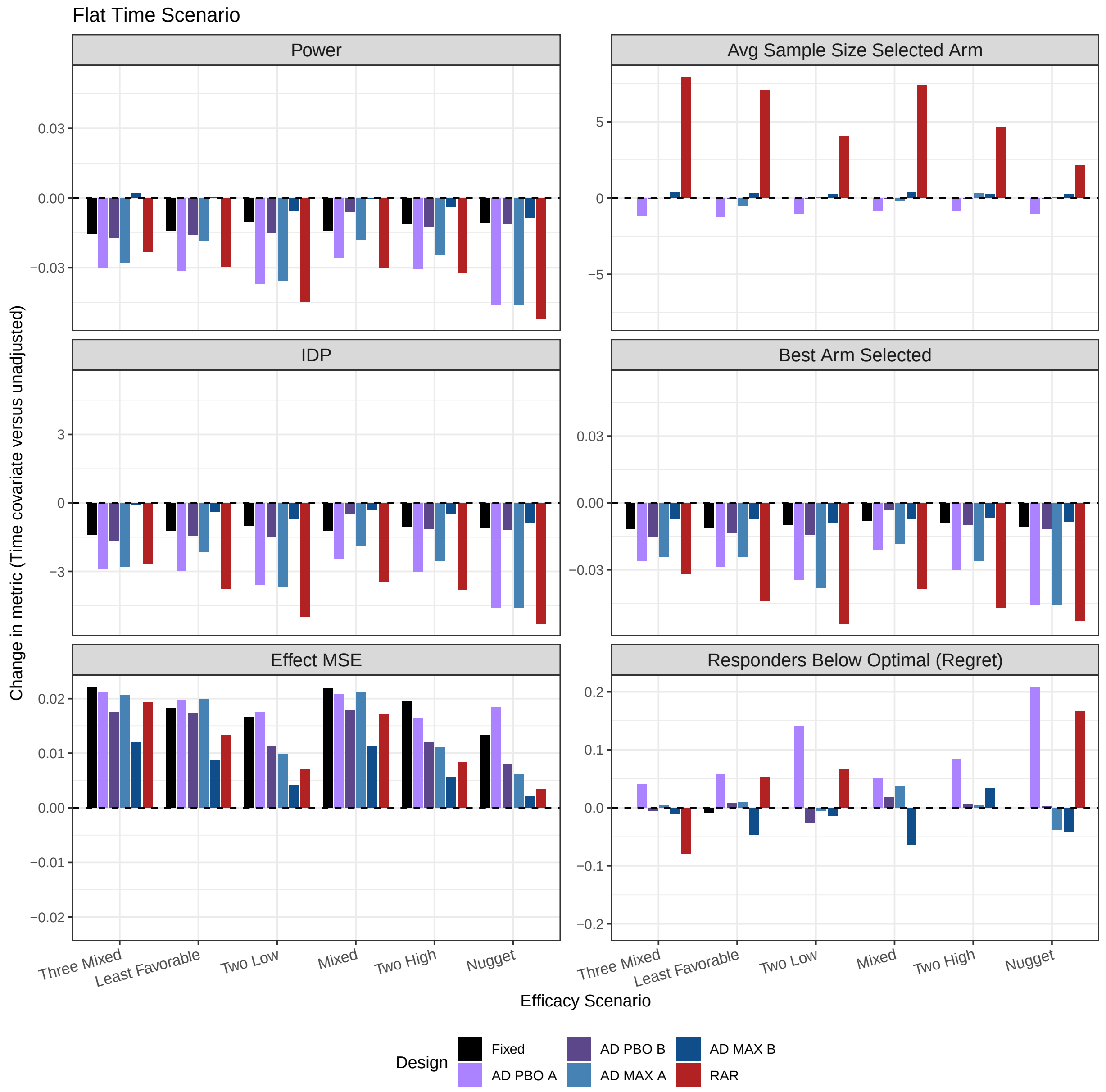


Figure 6: Absolute change in metric for the time covariate model versus the unadjusted model in the Flat time scenario. Each panel shows a separate metric on the y-axis.

5 Discussion

Within the context of multi-arm trials aimed at identifying a best arm and comparing that arm to control, our results align with previously published results showing that adaptive designs, including RAR and arm dropping, result in higher power and expected sample size on the true best arm when compared to fixed allocation in multi-arm trials [Ryan et al., 2020]. We have extended those results in terms of direct comparisons between different arm dropping designs and also demonstrated the results hold when time trends are present and accounted for in the model. We find that arm dropping based on the probability each arm is best is better than arm dropping based on p-values, and find no consequential differences between arm dropping based on the probability each arm is best and response-adaptive randomization. Adaptive designs may incur higher operational costs from interim analyses and/or the modification of randomization ratios. Sponsors should carefully consider these added costs relative to the benefits of adaptive designs. We recommend full simulation of trial designs to quantify the added value of adaptive features within the specific context of a trial including the indication, population, number of arms, and characteristics of the primary endpoint.

One important result found here is the difference in performance between AD PBO and AD MAX, with AD MAX being superior. As noted previously, this superiority is evident when interest centers only on the best arm, as opposed to all arms, but it indicates the importance of aligning adaptations to trial goals. This result for arm dropping is consistent with Viele et al. [2020b], who found a similar result comparing RAR driven by Pr(max) (a design very similar to design 4) to be superior to RAR driven by the probability of superiority to control.

As long as time trends are additive (as all time trends here were), then all adaptive methods considered here behave well after accounting for time in the model, extending similar results from Villar et al. [2018] for a bandit-based RAR procedure. Note that this result would not hold if time/treatment interactions were present, as opposed to additive time trends where all arms change by the same amounts over time. Fitting an additive model when an interaction is present may result in estimates that do not apply accurately to any fixed time period. This issue is discussed in more detail in Roig et al. [2022].

For Peer Review Only

If time trends are present, then incorporating time as a covariate is necessary to avoid type I error inflation in these adaptive allocation procedures. However, when time trends are unexpected it remains an open question whether a time covariate should be included. Maintaining a fixed allocation ratio (e.g. 2:1) of control to experimental arms in the trial can provide some protection against unanticipated time trends when using a method that does not include a covariate for time. In general, few previous trials have incorporated time effects in their modeling, but we expect the recent emphasis on time trends may change this practice. Our results indicate that including time as a covariate when time trends are not present involves a small loss on most metrics. This should be weighed against the costs of running into an unexpected time trend. Trial designers should engage with subject matter experts regarding their expectations for time trends. In some indications, time trends may be unexpected while others may anticipate differing populations over seasons (more severe, difficult to treat cases in winter, for example), variants of changing severity during an infectious disease epidemic, or other plausible mechanisms of time trends. Simulations presented to regulatory bodies should include scientifically plausible time trends but need not include every possibility as the FDA CDER/CBER adaptive guidance (section VI.A) states: "While it is impossible to simulate every scenario compatible with the null hypothesis, it may be possible to determine a limited set of scenarios that adequately represent the plausible range of potential false positives" [U.S. Food and Drug Administration, 2019]. Of course, if the eventual trial data indicate a time trend that was viewed as apriori unlikely, this may create a regulatory review issue. Given the relatively small cost of including a time trend relative to the efficiency gained by adaptive methods in general, we recommend including a time trend for robustness whenever feasible.

In practice, we also recommend including an appropriate futility rule that could stop a study for ethical and/or resource stewardship concerns. Futility rules have been omitted from these designs, which are all guaranteed the same sample size, to facilitate a direct comparison of inferential capabilities across allocation strategies. The most appropriate futility rule may vary across designs. A natural futility rule in the AD PBO design may stop the trial if all arms met the arm dropping criteria, while the RAR design may stop enrollment if the posterior probability that the best arm is superior to control is sufficiently

For Peer Review Only

low.

If standard frequentist model summaries are preferred over Bayesian posterior probabilities, then a nominal p-value could be reported at the final analysis with minimal impact on design performance. The simulations provided above would also justify type I error control for that analysis. However, it is unclear how far these results generalize in terms of other types of endpoints. For example, including covariates in logistic regression models has different impacts than in continuous settings, as described in [Robinson and Jewell, 1991]. They demonstrate that, in a logistic regression model, adding covariates can only increase the variance of treatment effect estimates. However, when there is no confounding between treatment assignment and time (which is true of fixed trials), omitting a "real" time covariate biases the treatment effect estimate towards 0. Robinson and Jewell [1991] go on to show that, for fixed trials, adjusting for covariates is more powerful than an unadjusted model. Thus, it remains important to explore this issue in a continuous setting in future work to see if any differences emerge.

For Peer Review Only

References

A D Barker, C C Sigman, G J Kelloff, N M Hylton, D A Berry, and L J Esserman. I-SPY 2: An adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy. *Clinical Pharmacology & Therapeutics*, 86(1):97–100, 2009. doi: 10.1038/clpt.2009.68.

P Bauer. Multiple testing in clinical trials. *Statistics in Medicine*, 10(6):871–889, 1991. doi: 10.1002/sim.4780100609.

Richard Bellman. A problem in the sequential design of experiments. *Sankhyā: The Indian Journal of Statistics*, 16(3/4):221–229, 1956.

Scott M Berry and Kert Viele. Comment: Response adaptive randomization in practice. *Statistical Science*, 38(2):229 – 232, 2023. doi: 10.1214/23-STS865F.

Swati Biswas, Diane D Liu, J Jack Lee, and Donald A Berry. Bayesian clinical trials at the University of Texas M. D. Anderson Cancer Center. *Clinical Trials*, 6(3):205–216, 2009. doi: 10.1177/1740774509104992.

R N Bradt, S M Johnson, and S Karlin. On sequential designs for maximizing the sum of n observations. *Annals of Mathematical Statistics*, 27(4):1060–1074, 1956. doi: 10.1214/aoms/1177728073.

Charles W Dunnett. A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association*, 50(272):1096–1121, 1955. doi: 10.2307/2281208.

Boris Freidlin and Edward L Korn. Adaptive randomization versus interim monitoring. *Journal of Clinical Oncology*, 31(7):969–970, 2013. doi: 10.1200/JCO.2012.45.0254.

Boris Freidlin, Edward L Korn, Robert Gray, and Alison Martin. Multi-arm clinical trials of new agents: Some design considerations. *Clinical Cancer Research*, 14(14):4368–4371, 07 2008. doi: 10.1158/1078-0432.CCR-08-0325.

Byron J Gajewski, Caitlyn Meinzer, Scott M Berry, Gaylan L Rockswold, William G Barsan, Frederick K Korley, and Renee' H Martin. Bayesian hierarchical EMAX model for dose-response in early phase efficacy clinical trials. *Statistics in Medicine*, 38(17):3123–3138, 2019. doi: 10.1002/sim.8167.

Andrew P Grieve. Response-adaptive clinical trials: case studies in the medical literature. *Pharmaceutical Statistics*, 16(1):64–86, 2017. doi: 10.1002/pst.1778.

Xuemin Gu, Nan Chen, Caimiao Wei, Suyu Liu, Vassiliki A Papadimitrakopoulou, Roy S Herbst, and J Jack Lee. Bayesian two-stage biomarker-based adaptive design for targeted therapy development. *Statistics in Biosciences*, 8(1):99–128, 2016. doi: 10.1007/s12561-014-9124-2.

Irving K Hwang, Weichung J Shih, and John S De Cani. Group sequential designs using a family of type I error probability spending functions. *Statistics in Medicine*, 9:1439–1445, 1990. doi: 10.1002/sim.4780091207.

Peter Jacko. The finite-horizon two-armed bandit problem with binary responses: A multidisciplinary survey of the history, state of the art, and myths. 2019. Management Science Working Paper 2019:3, Lancaster University Management School. arXiv:1906.10173.

For Peer Review Only

Patrick J Kelly, Nigel Stallard, and Susan Todd. An adaptive group sequential design for phase II/III clinical trials that select a single treatment from several. *Journal of Biopharmaceutical Statistics*, 15(4):641–658, jul 2005. doi: 10.1081/bip-200062857.

Edward S Kim, Roy S Herbst, Ignacio I Wistuba, J Jack Lee, George R Blumenschein, Anne Tsao, David J Stewart, Marshall E Hicks, Jeremy Erasmus, Sanjay Gupta, Christine M Alden, Suyu Liu, Ximing Tang, Fadlo R Khuri, Hai T Tran, Bruce E Johnson, John V Heymach, Li Mao, Frank Fossella, Merrill S Kies, Vassiliki Papadimitrakopoulou, Suzanne E Davis, Scott M Lippman, and Waun K Hong. The BATTLE trial: Personalizing therapy for lung cancer. *Cancer Discovery*, 1(1):44–53, 2011. doi: 10.1158/2159-8274.CD-10-0010.

Edward L Korn and Boris Freidlin. Time trends with response-adaptive randomization: The inevitability of inefficiency. *Clinical Trials*, 19(2):158–161, 04 2022. doi: 10.1177/17407745211065762.

J Jack Lee, Xuemin Gu, and Suyu Liu. Bayesian adaptive randomization designs for targeted agent development. *Clinical Trials*, 7(5):584–596, 2010. doi: 10.1177/1740774510373120.

Dominic Magirr, Thomas Jaki, and John Whitehead. A generalized Dunnett test for multiarm multi-stage clinical studies with treatment selection. *Biometrika*, 99(2):494–501, 03 2012. doi: 10.1093/biomet/ass002.

Vassiliki Papadimitrakopoulou, J Jack Lee, Ignacio I Wistuba, Anne S Tsao, Frank V Fossella, Neda Kalhor, Sanjay Gupta, Lauren Averett Byers, Julie G Izzo, Scott N Gettinger, Sarah B Goldberg, Ximing Tang, Vincent A Miller, Ferdinandos Skoulidis, Don L Gibbons, Li Shen, Caimiao Wei, Lixia Diao, S Andrew Peng, Jing Wang, Alda L Tam, Kevin R Coombes, Ja Seok Koo, David J Mauro, Eric H Rubin, John V Heymach, Waun Ki Hong, and Roy S Herbst. The BATTLE-2 study: A biomarker-integrated targeted therapy study in previously treated patients with advanced non-small-cell lung cancer. *Journal of Clinical Oncology*, 34(30):3638, 2016. doi: 10.1200/JCO.2015.66.0084.

For Peer Review Only

R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2022. URL https://www.R-project.org/.

Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527 – 535, 1952. doi: bams/1183517370.

David S Robertson, Kim May Lee, Boryana C López-Kolkovska, and Sofía S Villar. Response-adaptive randomization in clinical trials: From myths to practical considerations. *Statistical Science*, 38(2):185 – 208, 2023. doi: 10.1214/22-STS865.

Laurence D Robinson and Nicholas P Jewell. Some surprising results about covariate adjustment in logistic regression models. *International Statistical Review*, 59:227–240, 1991. doi: 10.2307/1403444.

Marta Bofill Roig, Pavla Krotka, Carl-Fredrik Burman, Ekkehard Glimm, Stefan M. Gold, Katharina Hees, Peter Jacko, Franz Koenig, Dominic Magirr, Peter Mesenbrink, Kert Viele, and Martin Posch. On model-based time trend adjustments in platform trials with non-concurrent controls. *BMC Medical Research Methodology*, 22(1):228, 2022. doi: 10.1186/s12874-022-01683-w.

Elizabeth G Ryan, Sarah E Lamb, Esther Williamson, and Simon Gates. Bayesian adaptive designs for multi-arm trials: an orthopaedic case study. *Trials*, 21(1):83, 2020. doi: 10.1186/s13063-019-4021-0.

Benjamin R Saville, Donald A Berry, Nicholas S Berry, Kert Viele, and Scott M Berry. The Bayesian time machine: Accounting for temporal drift in multi-arm platform trials. *Clinical Trials*, 2022. doi: 10.1177/17407745221112013.

Sama Shrestha and Sonia Jain. A Bayesian-bandit adaptive design for N-of-1 clinical trials. *Statistics in Medicine*, 2021. doi: 10.1002/sim.8873.

Nigel Stallard and Tim Friede. A group-sequential design for clinical trials with treatment selection. *Statistics in Medicine*, 27(29):6209–6227, 2008. doi: 10.1002/sim.3436.

For Peer Review Only

Nigel Stallard and Susan Todd. Sequential designs for phase II and phase III clinical trials incorporating treatment selection. *Statistics in medicine*, 22:689–703, 03 2003. doi: 10.1002/sim.1362.

Peter F Thall and J Kyle Wathen. Practical Bayesian adaptive randomisation in clinical trials. *European Journal of Cancer*, 43(5):859–866, 2007. doi: 10.1016/j.ejca.2007.01.006.

Peter F Thall, Richard Simon, and Susan S Ellenberg. Two-stage selection and testing designs for comparative clinical trials. *Biometrika*, 75(2):303–310, June 1988. doi: 10.1093/biomet/75.2.303.

Peter F Thall, Patricia S Fox, and J Kyle Wathen. Statistical controversies in clinical research: scientific and ethical problems with adaptive randomization in comparative clinical trials. *Annals of Oncology*, 26(8):1621–1628, 2015. doi: 10.1093/annonc/mdv238.

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933. doi: 10.2307/2332286.

Lorenzo Trippa, Eudocia Q Lee, Patrick Y Wen, Tracy T Batchelor, Timothy Cloughesy, Giovanni Parmigiani, and Brian M Alexander. Bayesian adaptive randomized trial design for patients with recurrent glioblastoma. *Journal of Clinical Oncology*, 30(26):3258, 2012. doi: 10.1200/JCO.2011.39.8420.

U.S. Food and Drug Administration. *Adaptive Designs for Clinical Trials of Drugs and Biologics: Guidance for Industry*. Center for Drug Evaluation and Research. Center for Biologics Evaluation and Research, 2019.

Kert Viele, Kristine Broglio, Anna McGlothlin, and Benjamin R Saville. Comparison of methods for control allocation in multiple arm studies using response adaptive randomization. *Clinical Trials*, 17(1):52–60, 2020a. doi: 10.1177/1740774519877836.

Kert Viele, Benjamin R Saville, Anna McGlothlin, and Kristine Broglio. Comparison of response adaptive randomization features in multiarm clinical trials with control. *Pharmaceutical Statistics*, 19(5):602–612, 2020b. doi: 10.1002/pst.2015.

For Peer Review Only

Sofía S Villar, Jack Bowden, and James Wason. Response-adaptive randomisation for multi-arm clinical trials using the forward looking Gittins index rule. *Biometrics*, 71(4): 969–978, 2015. doi: 10.1111/biom.12337.

Sofía S Villar, Jack Bowden, and James Wason. Response-adaptive designs for binary responses: How to offer patient benefit while being robust to time trends? *Pharm Stat*, 17(2):182–197, Mar 2018. doi: 10.1002/pst.1845.

Jixian Wang. Response-adaptive trial designs with accelerated Thompson sampling. *Pharmaceutical Statistics*, 20:645–656, 2021. doi: 10.1002/pst.2098.

James Wason, Dominic Magirr, Martin Law, and Thomas Jaki. Some recommendations for multi-arm multi-stage trials. *Statistical Methods in Medical Research*, 25(2):716–727, 2012. doi: 10.1177/0962280212465498.

James Wason, Nigel Stallard, Jack Bowden, and Christopher Jennison. A multi-stage drop-the-losers design for multi-arm clinical trials. *Statistical Methods in Medical Research*, 26(1):508–524, 2017. doi: 10.1177/0962280214550759.

James M S Wason and Lorenzo Trippa. A comparison of Bayesian adaptive randomization and multi-stage designs for multi-arm clinical trials. *Statistics in Medicine*, 33(13):2206–2221, 2014. doi: 10.1002/sim.6086.

James M S Wason, Jean E Abraham, Richard D Baird, Ioannis Gournaris, Anne-Laure Vallier, James D Brenton, Helena M Earl, and Adrian P Mander. A Bayesian adaptive design for biomarker trials with linked treatments. *British Journal of Cancer*, 113(5): 699–705, 2015. doi: 10.1038/bjc.2015.278.

S Faye Williamson. *Bayesian Bandit Models for the Design of Clinical Trials*. PhD thesis, University of Lancaster, February 2020.

S Faye Williamson, Peter Jacko, and Thomas Jaki. Generalisations of a Bayesian decision-theoretic randomisation procedure and the impact of delayed responses. *Comput Stat Data Anal*, 174:107407, 2022. doi: 10.1016/j.csda.2021.107407.

For Peer Review Only

A Additional designs

In Appendix B, we present the additional allocation procedures, including the multi-arm multi-stage (MAMS) design. The MAMS design differs from the other adaptive allocation procedures in terms of the timing of interim analyses. After the initial run-in period of 48 participants, subsequent analyses will be performed after a fixed number of control participants rather than a fixed number of total participants. Between each interim, 16 participants are randomized to control and 8 participants are randomized to each active experimental arm, so the number of total participants enrolled between interims may vary. If all 4 arms are active, 48 participants will be enrolled between interims. If a single arm is active, 24 participants will be enrolled between interims. The trial continues until all experimental arms are dropped or until the maximum sample size of 80 control participants is reached (corresponding with 240 total participants if no arms are dropped). Similar to the AD PBO design, experimental arms can be dropped based on a p-value comparison to control using the group sequential futility bounds from a Hwang-Shih-DeCani spending function with a parameter of $\gamma = -0.5$ [Hwang et al., 1990]. If the p-value for an active arm exceeds the futility bound, the arm is permanently dropped from the trial. A key difference between the MAMS and AD PBO design is that, if the p-values for all active arms exceed the futility bound, then the MAMS trial is halted for futility.

Results for the six performance metrics are shown for the MAMS design in Appendix B. When summarizing the average sample size and regret metrics, if a trial stops at a sample size below 240 participants, it is assumed that the remaining participants up to 240 total will receive the selected arm where "selected" means identified as the best arm and found to be statistically significant. If no arm is statistically significant, it is assumed that the remaining participants up to 240 receive the placebo control. This assumption is made to avoid penalizing the MAMS design for dropping ineffective arms and thus enrolling fewer than 240 participants. Note that the effect MSE metric does not incorporate this overrun up to 240, so the calculation may be based on an analysis including fewer than 240 participants.

For Peer Review Only

B Additional simulation results

Table 3 shows the success thresholds (δ) for each design as described in subsection 3.3.

Figures 7-11 reproduce the results from the main paper with the MAMS design (see Appendix A) added in. Evaluation of the MAMS design performance should consider that it is the only design that can stop enrollment early for futility, and in practice we recommend including an appropriate futility rule (see Discussion).

Figure 12 shows the average sample size and effect MSE metrics under the null scenario with no time adjustment. Four metrics (Power/IDP/regret/best arm selection) are not shown since they are not defined in the global null scenario.

Figures 13-17 show the results for the 5 remaining alternative efficacy scenarios when time trends are simulated and a time adjustment is included in the analysis model. The relative performance of the designs is similar to the Mixed scenario which is summarized in the main paper results.

Figures 18-19 compare the results for the AD PBO and MAMS designs for alternative choices of futility thresholds. The futility thresholds are based on the Hwang-Shih-DeCani (HSD) spending function. For the HSD spending function, a parameter value of -4 approximates O'Brien-Fleming futility bounds and a parameter of 1 approximates Pocock futility bounds. In the main paper, a parameter of -0.5 is utilized which falls in between O'Brien-Fleming and Pocock. We also present results for the HSD(-4) spending function. Figure 20 presents the futility bounds by interim for each spending function. Compared to the HSD(-0.5) designs, the designs with the approximate O'Brien-Fleming futility bounds have lower power, IDP, and arm selection performance and fewer participants are allocated to the selected arm. The MAMS design with more conservative futility bounds has lower effect MSE while the more conservative AD PBO designs have slightly higher effect MSE. There is not a noticeable difference in the regret metric across the futility bounds.

For Peer Review Only

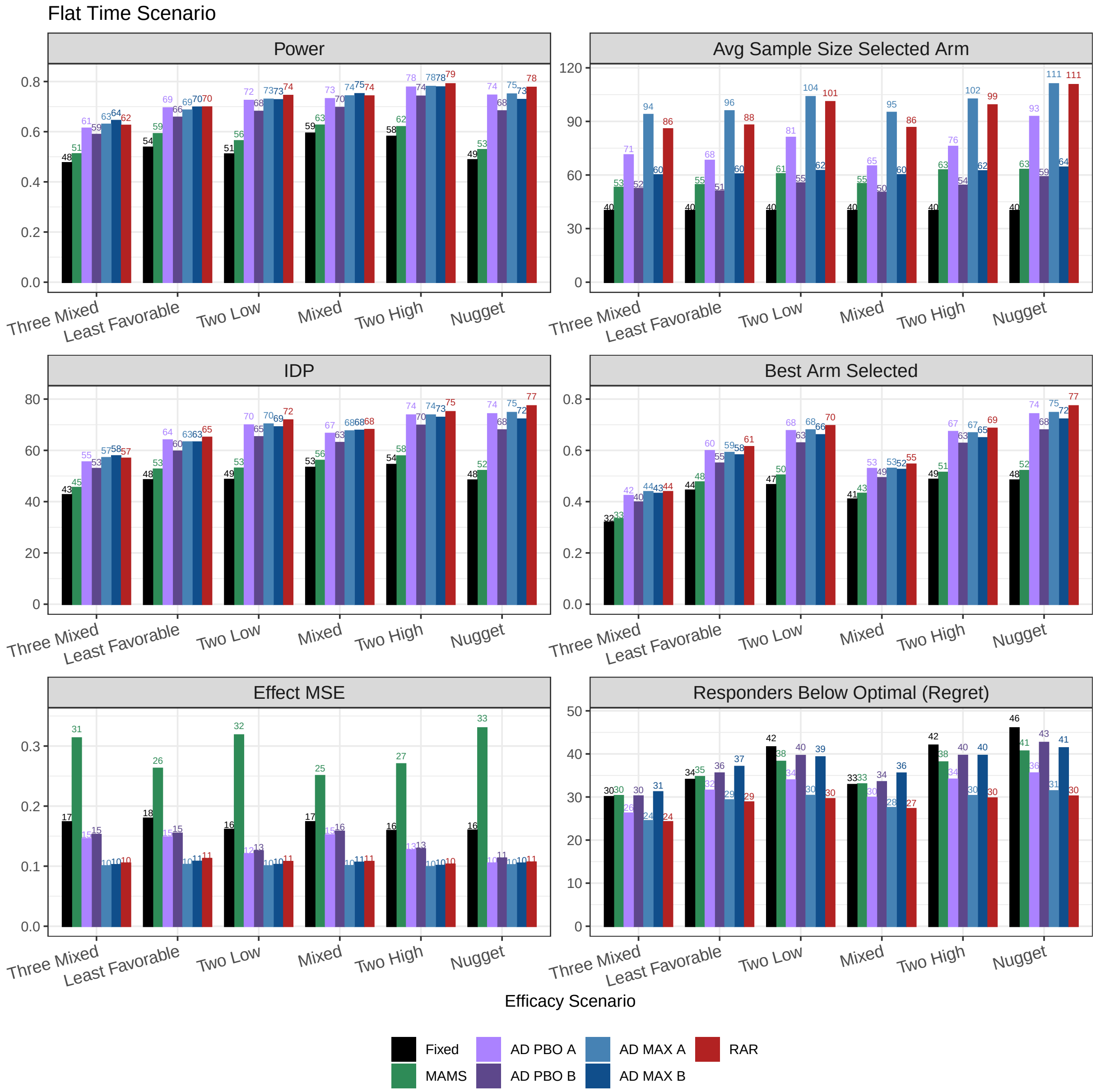


Figure 7: Simulation results in the flat time scenario with no modeled covariate adjustment for time. Each panel summarizes a performance metric on the y-axis. The six non-null efficacy scenarios are shown on the x-axis.

Design	Success Threshold δ	
	No Time Covariate	Yes Time Covariate
Fixed	0.993	0.994
AD PBO A	0.989	0.993
AD PBO B	0.993	0.994
AD MAX A	0.988	0.993
AD MAX B	0.988	0.990
RAR	0.988	0.992
MAMS	0.990	0.992

Table 3: Summary of success thresholds (δ) for each design as described in subsection 3.3.

For Pee

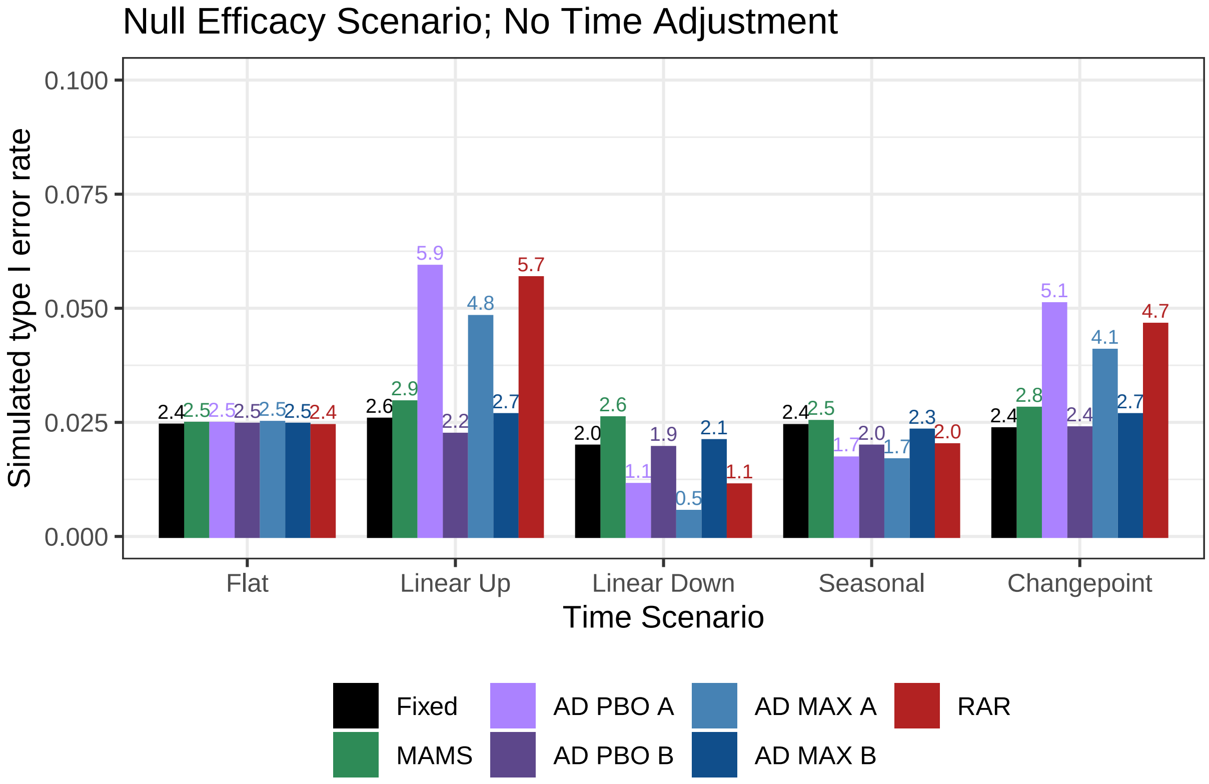


Figure 8: Simulated type I error rates in the null scenario without covariate adjustment for time. The five time trend scenarios are shown on the x-axis.

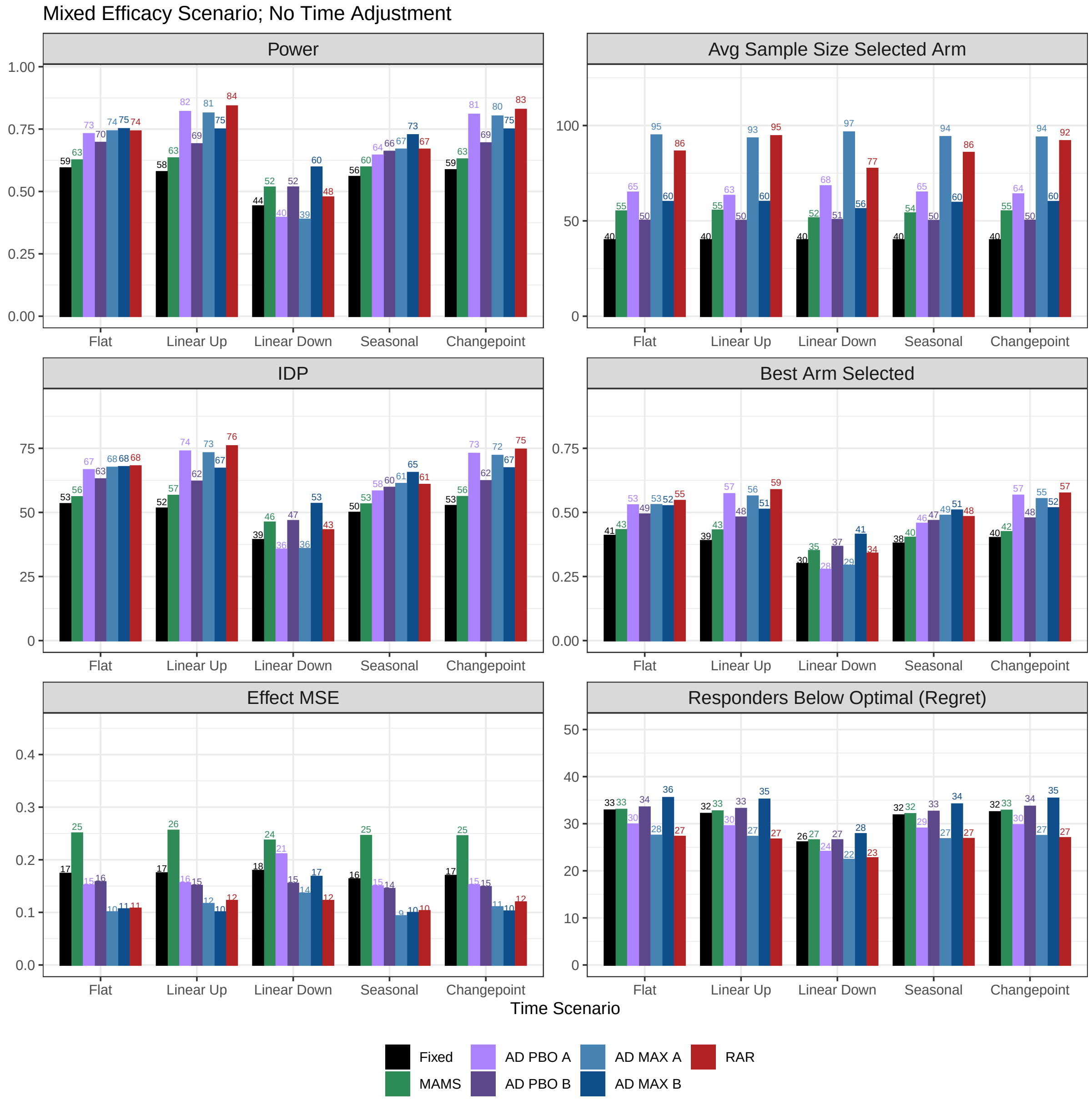


Figure 9: Simulation results in the Mixed efficacy scenario when time trends are simulated but no covariate adjustment for time is included in the analysis model. Each panel summarizes a separate metric on the y-axis. The x-axis shows the five time trend scenarios.

Fo Only

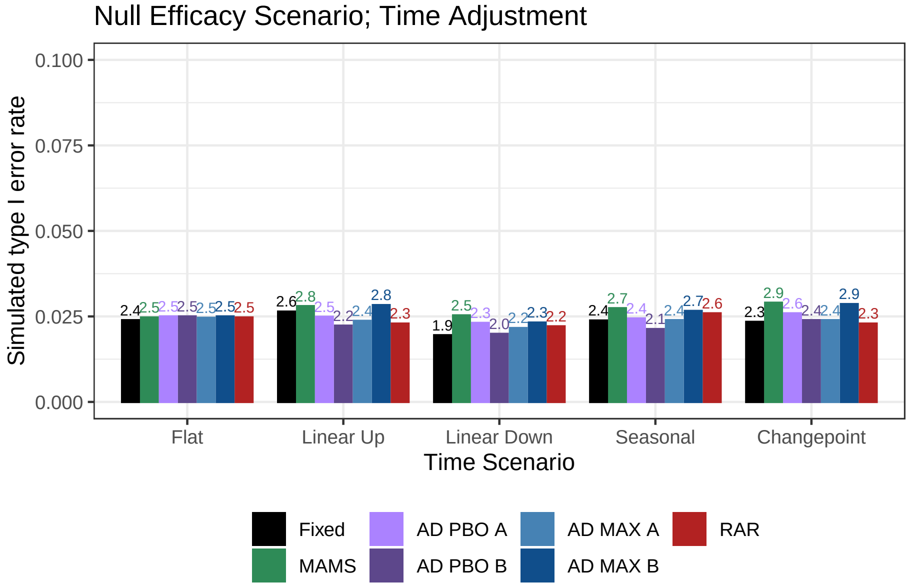


Figure 10: Simulated type I error rates in the null scenario with covariate adjustment for time. The five time trend scenarios are shown on the x-axis.

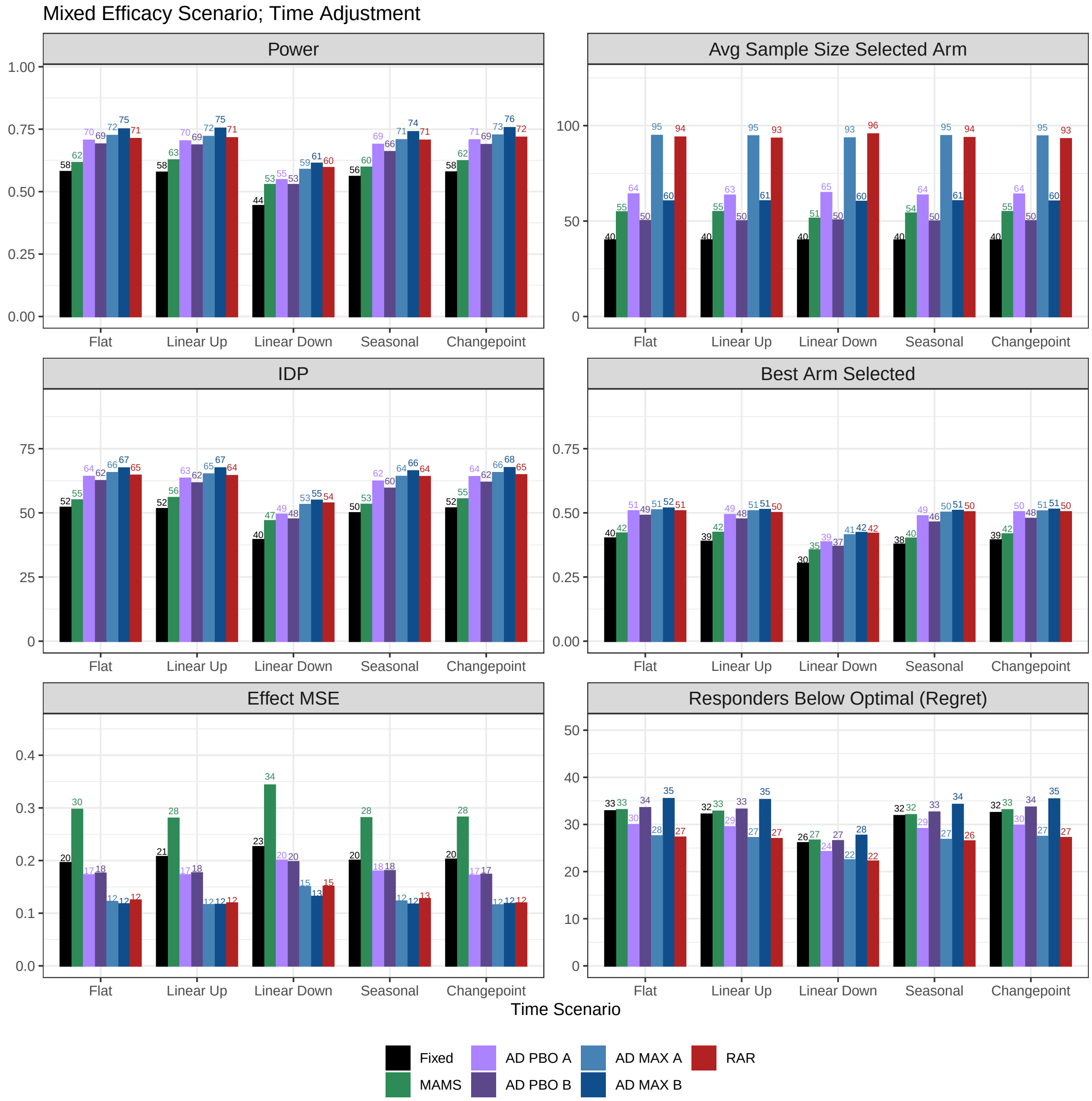


Figure 11: Simulation results in the Mixed efficacy scenario when time trends are simulated and a covariate adjustment for time is included in the analysis model. Each panel summarizes a separate metric on the y-axis. The x-axis shows the five time trend scenarios.

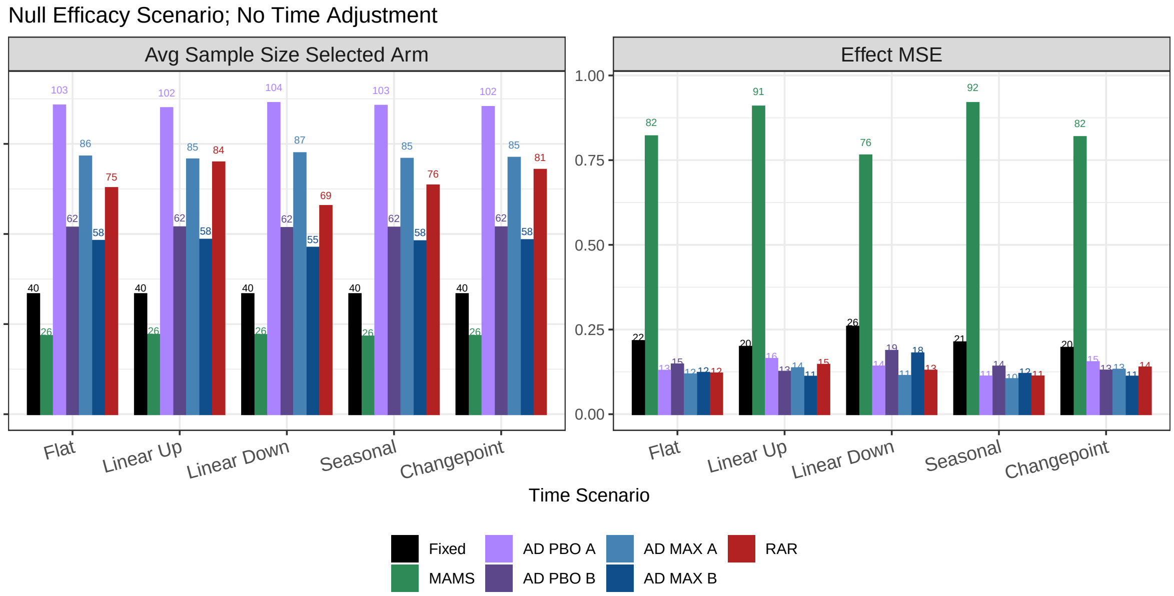


Figure 12: Simulation results in the Null efficacy scenario with no time adjustment. Four metrics (Power/IDP/Regret/Best Arm Selected) are not shown since they are not applicable in the global null scenario.

Only

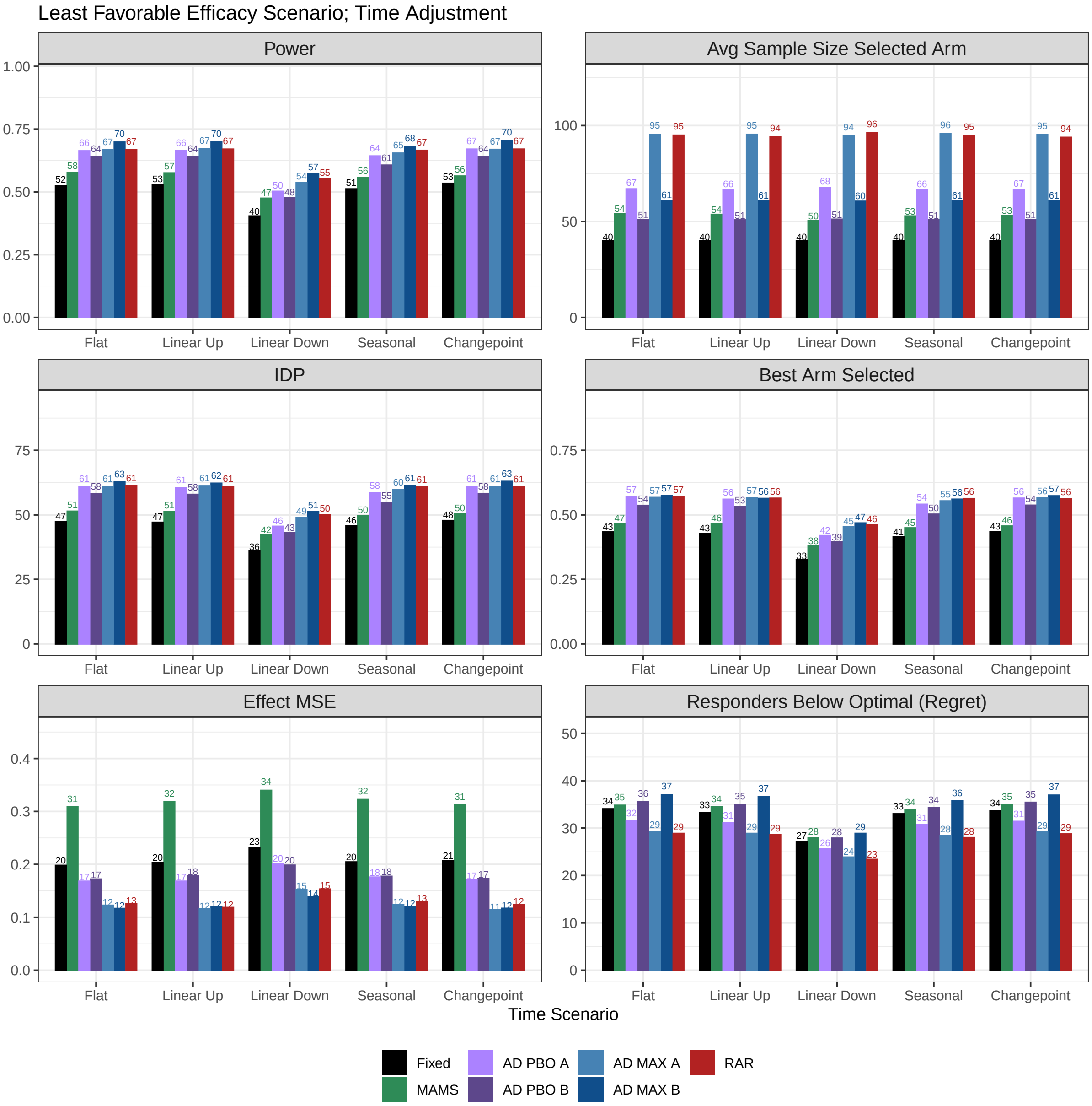


Figure 13: Simulation results in the Least Favorable efficacy scenario when time trends are simulated and a covariate adjustment for time is included in the analysis model. Each panel summarizes a separate metric on the y-axis. The x-axis shows the five time trend scenarios.

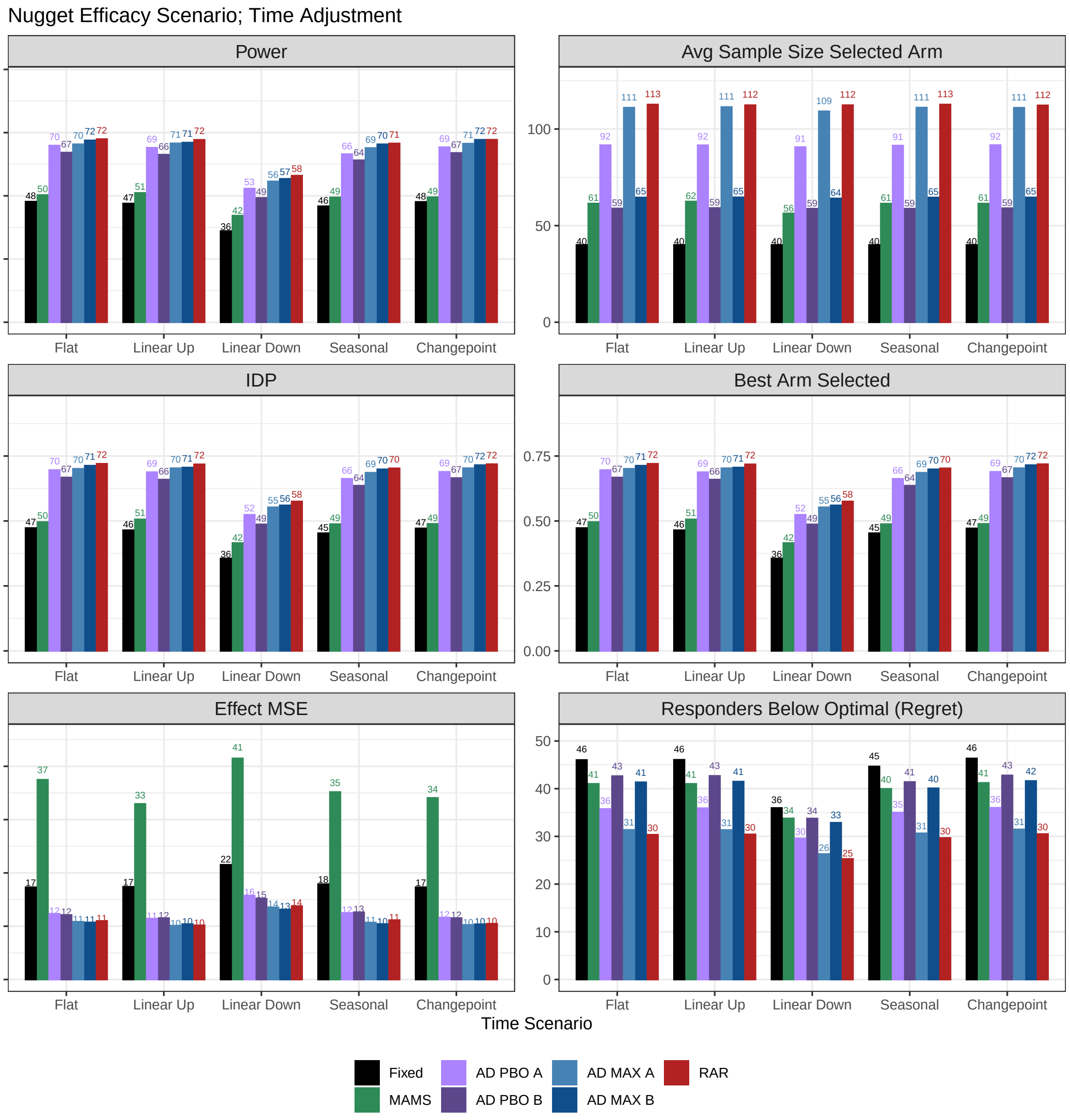


Figure 14: Simulation results in the Nugget efficacy scenario when time trends are simulated and a covariate adjustment for time is included in the analysis model. Each panel summarizes a separate metric on the y-axis. The x-axis shows the five time trend scenarios.

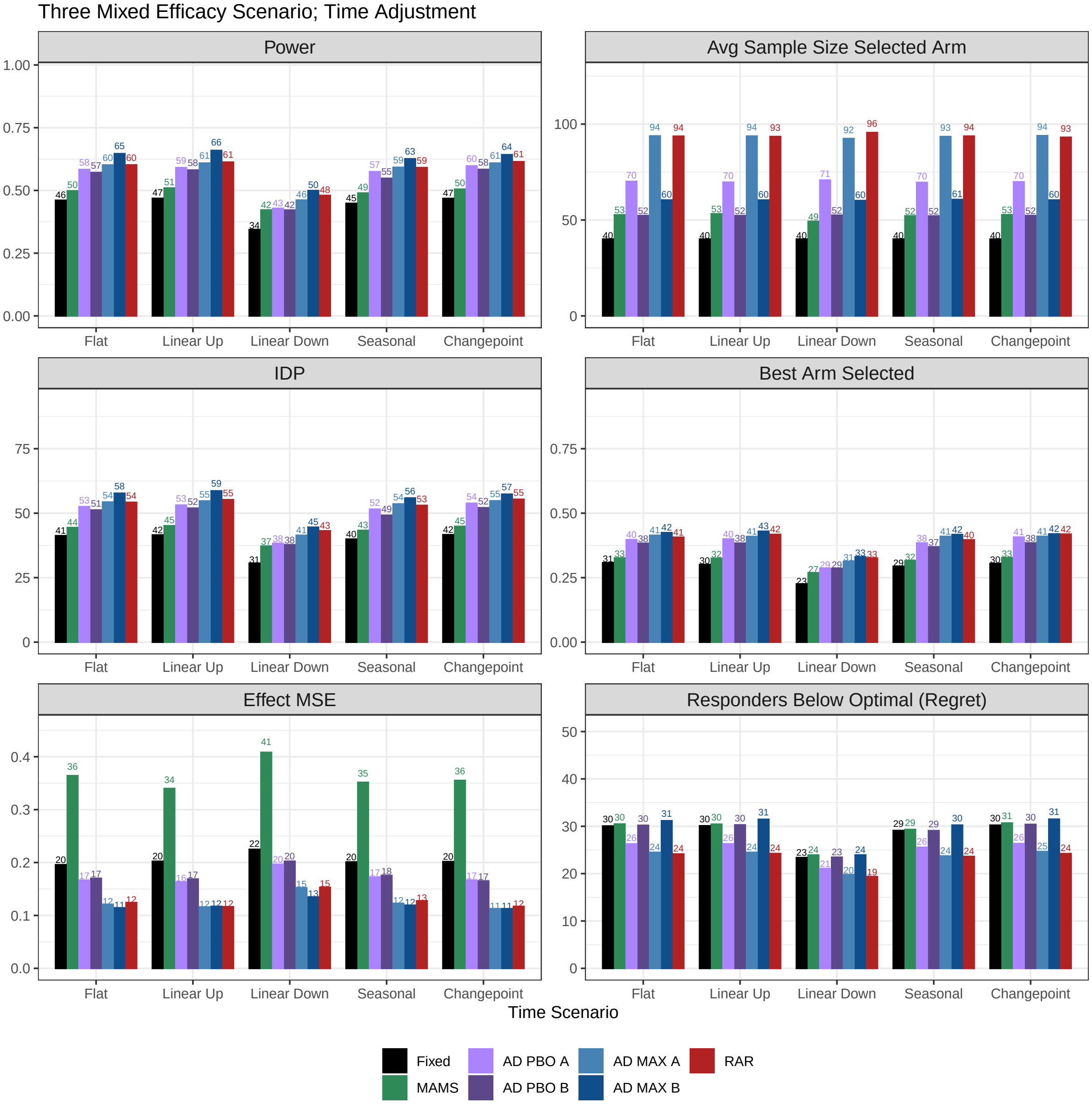


Figure 15: Simulation results in the Three Mixed efficacy scenario when time trends are simulated and a covariate adjustment for time is included in the analysis model. Each panel summarizes a separate metric on the y-axis. The x-axis shows the five time trend scenarios.

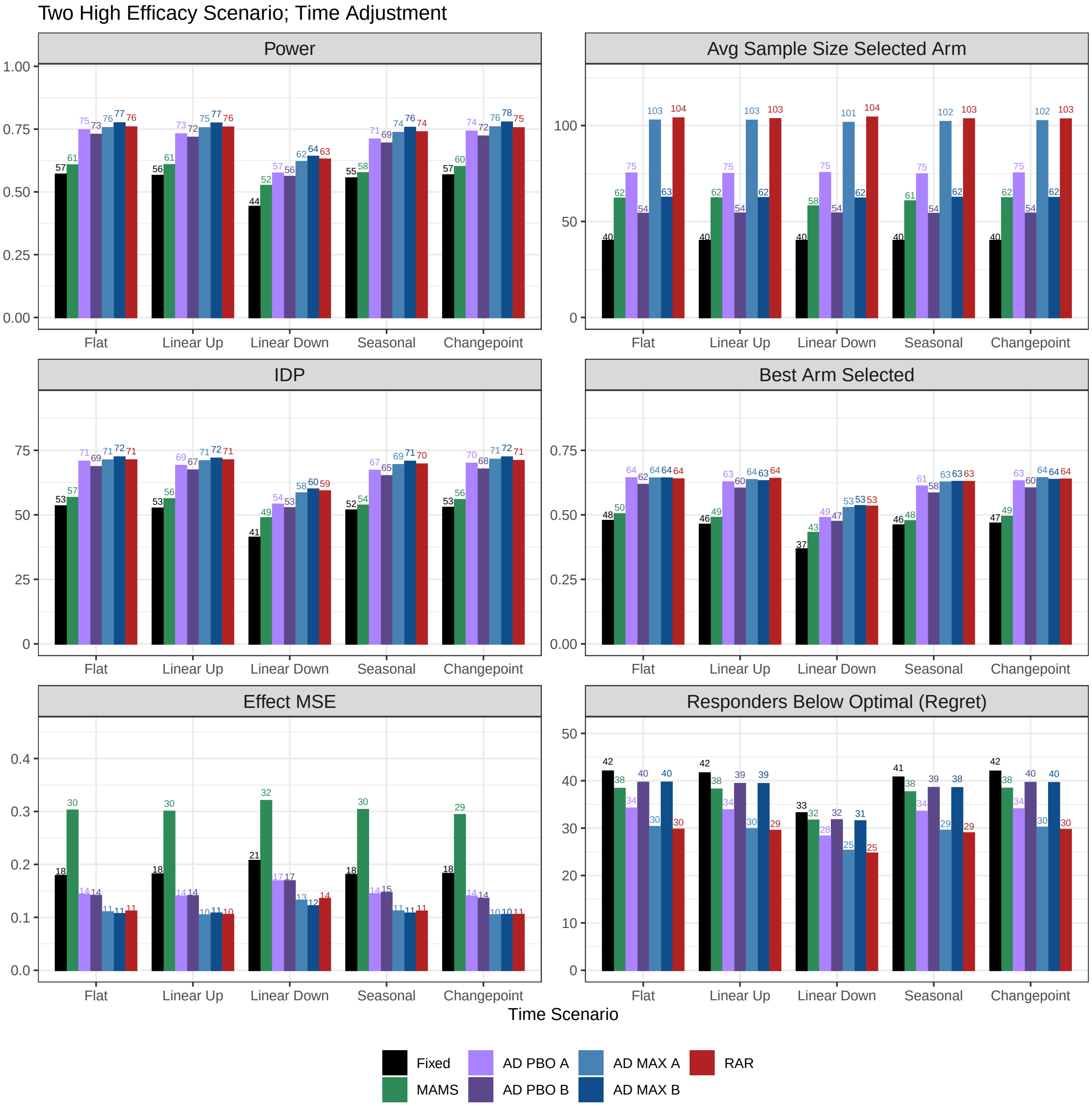


Figure 16: Simulation results in the Two High efficacy scenario when time trends are simulated and a covariate adjustment for time is included in the analysis model. Each panel summarizes a separate metric on the y-axis. The x-axis shows the five time trend scenarios.

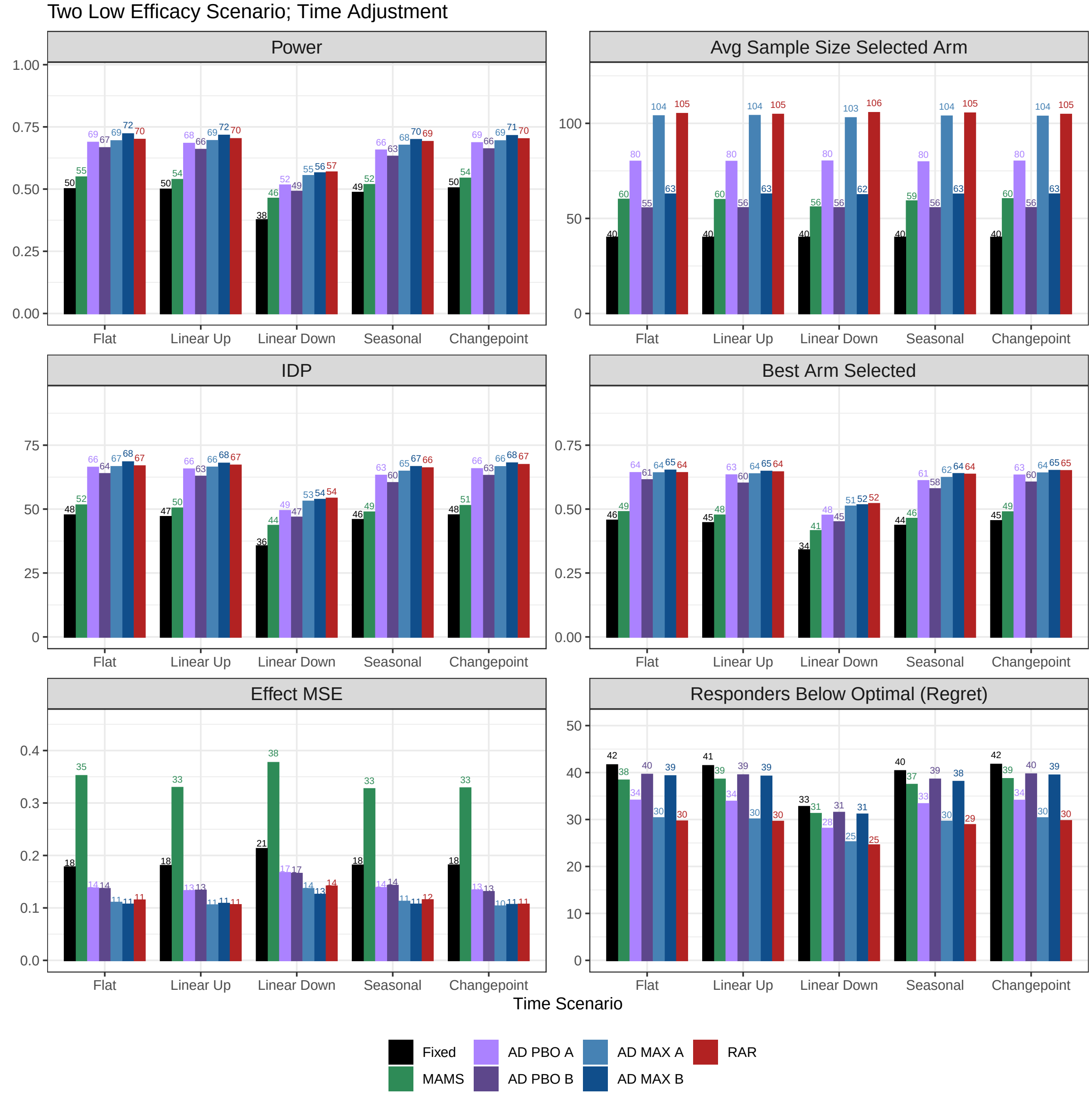


Figure 17: Simulation results in the Two Low efficacy scenario when time trends are simulated and a covariate adjustment for time is included in the analysis model. Each panel summarizes a separate metric on the y-axis. The x-axis shows the five time trend scenarios.

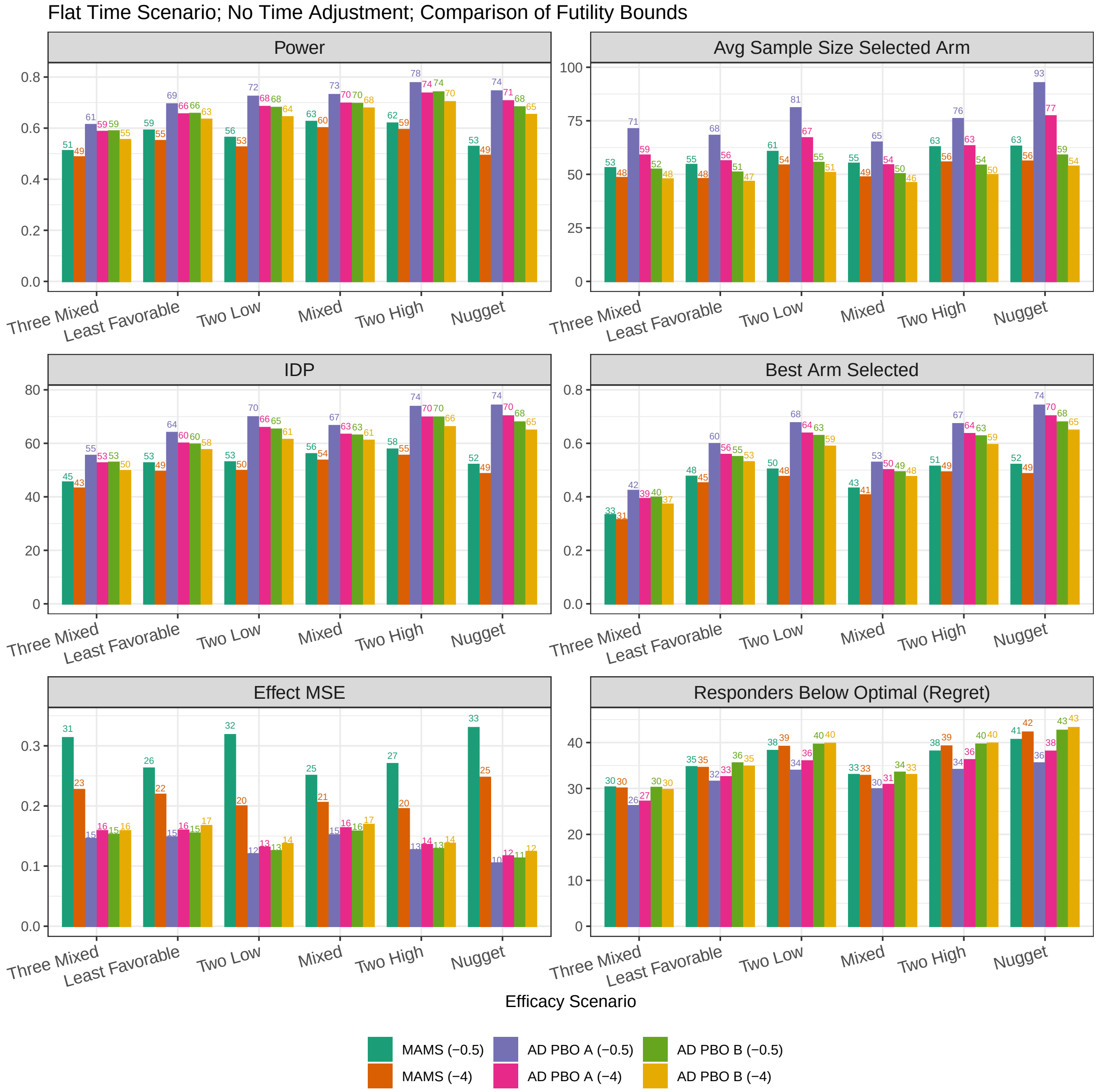


Figure 18: Comparison of metrics in the AD PBO and MAMS designs with different Hwang-Shih-DeCani spending function parameters (shown in parentheses in Design label). Simulation results in the flat time scenario with no modeled covariate adjustment for time. Each panel summarizes a performance metric on the y-axis. The six non-null efficacy scenarios are shown on the x-axis.

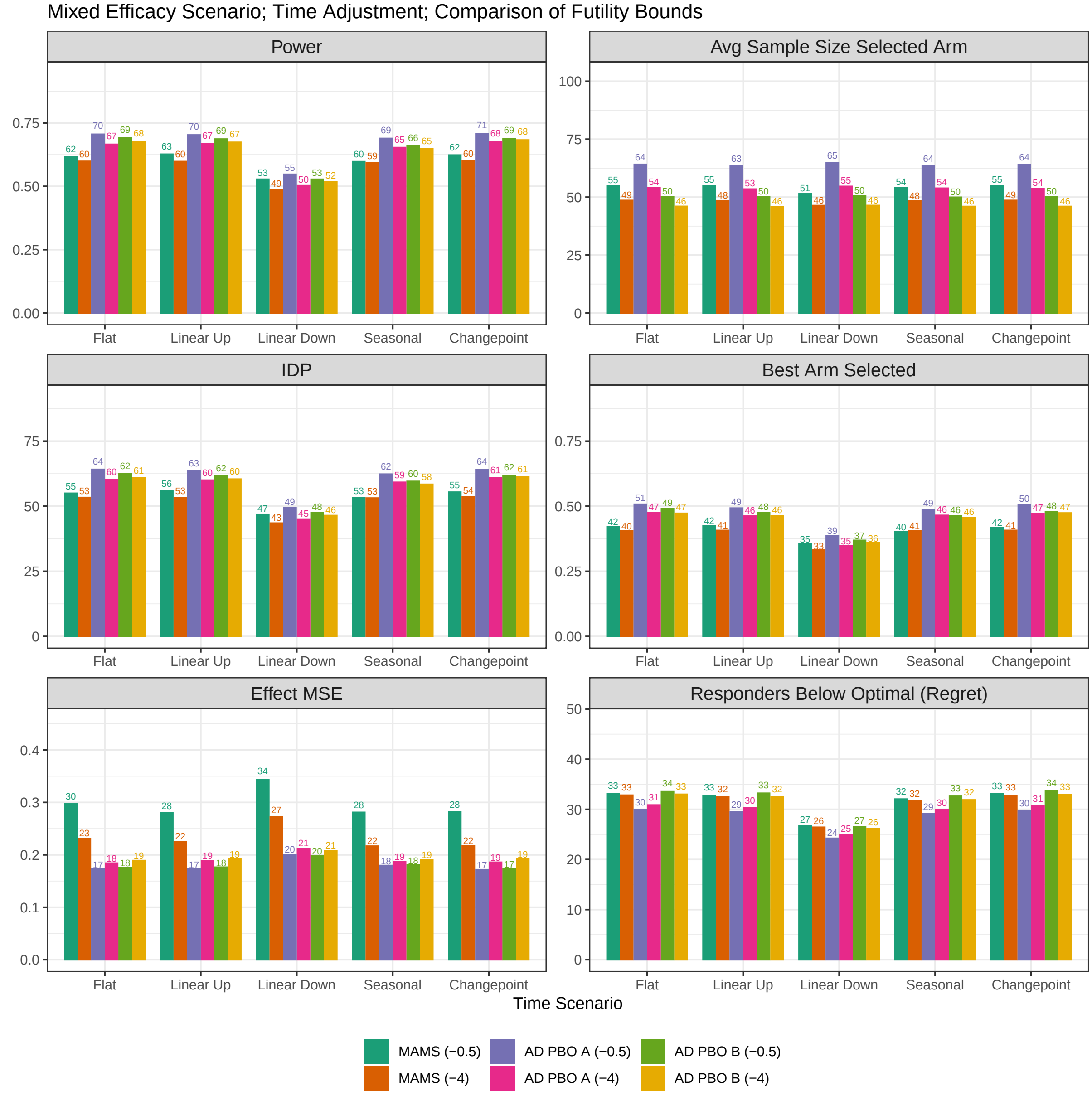
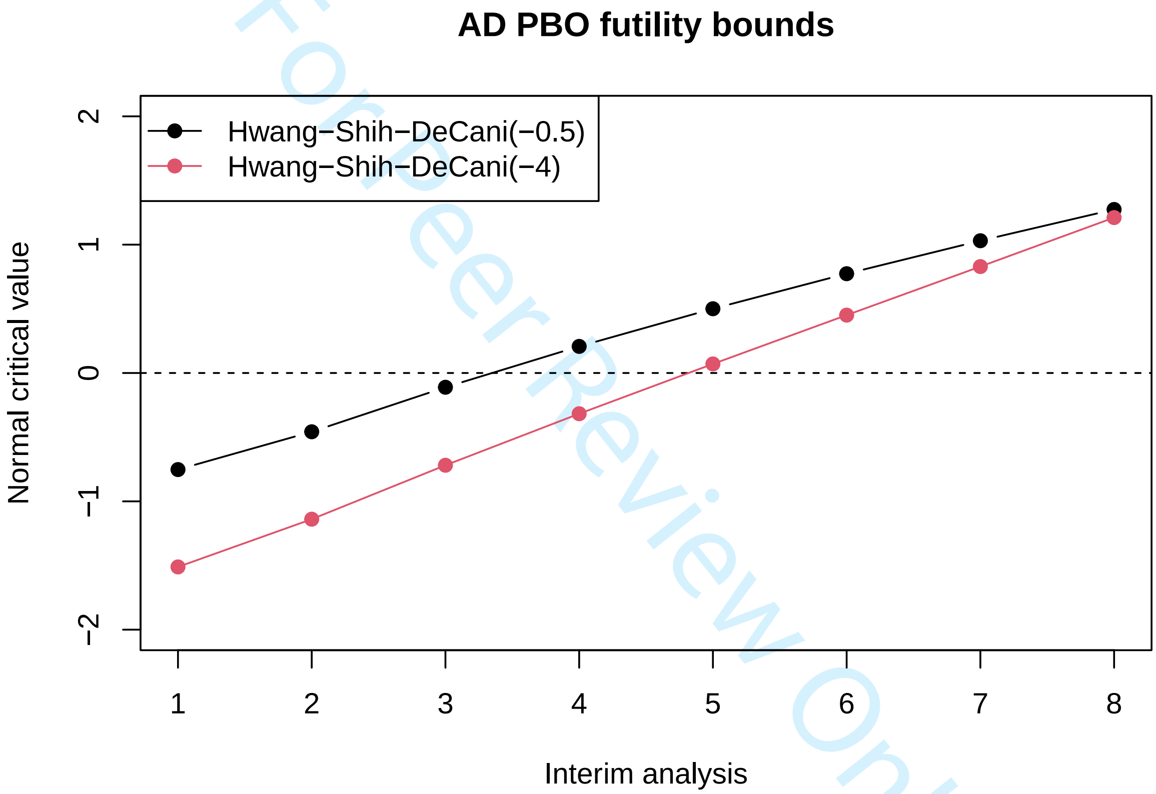


Figure 19: Comparison of metrics in the AD PBO and MAMS designs with different Hwang-Shih-DeCani spending function parameters (shown in parentheses in Design label). Simulation results in the Mixed efficacy scenario when time trends are simulated and a covariate adjustment for time is included in the analysis model. Each panel summarizes a separate metric on the y-axis. The x-axis shows the five time trend scenarios.



For Peer Review Only

Figure 20: AD PBO futility bounds (test statistic critical value) by interim.

Reply to Reviews
Effects of allocation method and time trends on identification of the best arm in multi-arm trials
SBR-22-091.R1

For Peer Review Only

November 10, 2023

1 Referee 2

The authors have addressed most of my comments and I am satisfied with their explanation. However, there are still some points in the present manuscript that needs improvement and addressing these points would enhance the manuscript as well as the applicability of the proposed methods in real life scenarios.

Thank you for your careful review and helpful recommendations. We have incorporated these suggestions as described in the point-by-point responses below.

1) Notation of $Pr(max)_j$ is not consistent. Section 3 (Page 10 for AD MAX A and AD MAX B) it is referred to as $Pr(max)_j$ as it is introduced in (1), however in the subsection of RAR in the same page it is referred to as Pr(max). There needs to be consistency in the notations. Pr(max) is also stated in page 12 instead of $Pr(max)_j$.

We have changed the notation to say Pr(max)$_j$ throughout the paper when we are referring to the specific quantity for an arm. There are a few instances where we left "Pr(Max)" to refer to the general quantity of the probability an arm is best.

2) In the Discussion section please detail clearly on how the benefits from the proposed methodologies would surpass the operational challenges in implementing these designs in practice.

We have added the following comment to the discussion:
"Adaptive designs may incur higher operational costs from interim analyses and/or the modification of randomization ratios. Sponsors should carefully consider these added costs relative to the benefits of adaptive designs. We recommend full simulation of trial designs to quantify the added value of adaptive features within the specific context of a trial including the indication, population, number of arms, and characteristics of the primary endpoint."

3) Please look carefully on the typographical errors.

Thank you, we have carefully reviewed the paper and corrected any typographical errors we identified.

4) In page 12 for selection of the priors for the model parameters, please include the sentence about other non-informative priors being considered but the effect of them on the posterior being minimal.

We have included the following sentence in the Methods section following the description of priors: The resulting posterior distribution is insensitive to other choices of weakly informative priors on the log-odds scale.

5) Since there are conflicting thoughts in the regulatory bodies in using Bayesian methods for primary analysis for pivotal trials, I think it would be useful if some explanation is included on if the conclusions differ with the final analysis being performed using a frequentist method (at same 2.5% type I error rate) with only important related results included in the appendix supporting this explanation. This would enhance the applicability of the proposed methods in real life scenarios.

We agree that it is important to point out how these results translate using a frequentist final analysis. We have added the following to the discussion section: "If standard frequentist model summaries are preferred over Bayesian posterior probabilities, then a nominal p-value could be reported at the final analysis with minimal impact on design performance. The simulations provided above would also justify type I error control for that analysis."

2 Referee 3

The authors are to be commended for providing a substantive amount of work to address the many reviewer questions! Some previous major concerns remain before I can give a more favorable review. In short they are:

1) Providing the reader (not just the reviewer) the explanations/justifications of feedback raised by previous review.

2) Giving credit to, and distinguishing from, similar work (Villar 2018) which addresses RAR in the presence of time trends. This work includes Thompson Sampling and explicit model adjustment for covariates. While this paper was been cited, the connection should, in my read, be much more explicit.

Thank you for the suggestions. We have responded to each detailed comment below, and noted our modifications to the paper to address each comment.

In more detail:

1. RE Explanation for why the designs departed from practice — for the sake of understanding

For Peer Review Only

the impact of allocation versus futility rules: This explanation was helpful and should be included in the text for the reader. While this should be emphasized, the limitation should be acknowledged that the designs under simulation deviate from practice and the importance of futility rules. . . . In the authors response, I don't yet buy the argument that the MAMS design which stopped early overall performs poorly. Under the null hypothesis (not presented), I expect early stopping to be more ethical than other designs.

We have added the following paragraph to the discussion to address the importance and impact of futility: "In practice, we also recommend including an appropriate futility rule that could stop a study for ethical and/or resource stewardship concerns. Futility rules have been omitted from these designs, which are all guaranteed the same sample size, to facilitate a direct comparison of inferential capabilities across allocation strategies. The most appropriate futility rule may vary across designs. For example, a natural futility rule in the AD PBO design may stop the trial if all arms met the arm dropping criteria, while a more natural futility rule in the RAR design could stop enrollment if the posterior probability that the best arm is superior to control is sufficiently low. "

We agree that the comparison to the MAMS design is complicated by the incorporation of a futility rule. We have replaced the original text in the Appendix about the MAMS design being outperformed with the following sentence addressing the difference in inclusion of futility rules: "Evaluation of the MAMS design performance should consider that it is the only design that can stop enrollment early for futility, and, in practice, we recommend including an appropriate futility rule (see Discussion)."

2. RE Major comment #1: The "Best" arm objective is addressed by pointing to Berry and Viele 2023. Thank you and please also state an overview of what were the objectives of the real-world applications cited in the paper.

We have modified this sentence to include the objective of the real world applications: "For a number of real world examples that have an objective of finding the best arm, see Berry and Viele [2023]."

3. RE Major comment #2: See later comment regarding Villar 2018. The novel contribution should still be clarified.

Please see our response to comment number 5 below.

4. RE Major comment #6: Thank you for the very thorough response! As a minor comment, the submitted manuscript still claims p-value arm dropping is the most common and still needs justification. Or the statement could be softened such as commenting on the author's experiences/observations.

We have removed the claim that p-value arm dropping is the most common approach in the introduction.

5. RE Major comment #7: This remains my biggest concern. First, please see Section 4 "Ad-

For Peer Review Only

justing the model for a time trend" of Villar 2018 where they do expressly include adjusting for a time covariate in modelling. "Parts (I) and (III) in Table 3 show the results for the estimation of the models' parameters using standard maximum likelihood estimation, when the (logistic) model is correctly specified. These results indicate, perhaps unsurprisingly, that for both designs, the treatment effect is found to be significant in less than 5% of the 5000 trials, which suggests that by including a correctly modelled time trend, type I error inflation is avoided." A more explicit connection to adjusting for covariates should be provided and similar findings should be reaffirmed in the discussion. Second, the paper does use Thompson Sampling (see 2.1 (a)). If using the RAR method as is in the paper (which seems reasonable for making the point), it would be good to acknowledge that other TS methods such Villar 2018 were not explored but may be anticipated to also be beneficial.

We apologize for overlooking aspects of the Villar 2018 paper in our previous response to this comment. We agree that Villar 2018 is the most relevant previous paper on this topic, and we have expanded upon our novel contributions. Most importantly, Villar et al [2018] focuses on one class of adaptive allocation procedures (RAR) versus Fixed randomization while we also consider Arm Dropping designs. Additionally, our comparison involves more time trend scenarios (Villar et al focus on a linear time trend scenario) and more efficacy scenarios (Villar et al focus on the "nugget" scenario where one arm is effective). There are also two key differences between the Thompson Sampling method in Villar et al and the method used in the current paper. First, we use simulation-based control of type I error when no time trends are present to achieve a type I error rate equivalent to the fixed design. The TS method in Villar et al has an unadjusted final analysis and thus results in inflated type I error even when no time trends are present (as seen in their Figure 2). Second, our RAR design maintains the control allocation throughout the trial to avoid reduction of power and poor estimation of the treatment effect as recommended in Viele et al. [2020a] and Villar et al. [2018] (and implemented in their CFLGI method).

We have added the following paragraph to Section 2.4:

Villar et al. [2018] discuss the issue of type I error inflation due to unknown time trends with several variants of RAR including a Thompson Sampling approach, and demonstrate that, when linear time trends are present, covariate adjustment for time trends can avoid type I error inflation in a two-arm trial using a bandit-based variant of RAR. The RAR design we evaluate has two key differences from the Villar et al. [2018] TS method: 1) simulation-based control of type I error when no time trends are present to allow for a direct comparison across allocation procedures and 2) maintenance of the control allocation throughout the trial as recommended in Villar et al. [2018] and Viele et al. [2020a] to avoid reduction of power and poor estimation of the treatment effect. We extend the Villar et al.'s exploration of the impact of covariate adjustment for time to the multi-arm context for several adaptive allocation procedures (including our variant of Thompson Sampling-based RAR) and a broad range of time trend and efficacy scenarios.

We have also added the following to the Discussion (changes in bold):

*As long as time trends are additive (as all time trends here were), then all adaptive methods considered here behave well after accounting for time in the model**, extending similar results from Villar et al. [2018] for a multi-armed-bandit-based RAR procedure.***

Minor comments:

For Peer Review Only

1. Include the number of simulation replicates. For figure of Type I error, are the differences within monte-carlo sampling error? Consider including error bands.

We have included this sentence at the beginning of Section 4: "The results presented below are based on 10,000 simulation replicates per efficacy and time trend scenario (resulting in a Monte Carlo standard deviation of 0.16% for simulated type I error rates)."

Independent sampling of 10,000 simulations would result in a Monte Carlo standard deviation of 0.16%. For example, the simulated values of 2.8%-2.9% in Figure 4 appear to be slightly above what would be expected if the true value was 2.5%. The height of the text labels on the figure is about 2 standard deviations (0.3%), so adding error bands would likely not be readable.

2. Change the Y-Axis in figures to only include the range of outcomes. Example, Figure 1 could roughly be from 0.45 to 0.80.

After careful consideration of this recommendation, we prefer the figures without modifying the range of the y-axes. Zero is a meaningful minimum value for all of the metrics, and starting the y-axes at zero allows the reader to interpret the magnitude of differences across the scale without overemphasis on small differences. This approach is consistent with the area principle and data visualization guidelines produced by the Royal Statistical Society (see publication).

3 Associate Editor

Thank you for submitting the revised manuscript. It has been much improved. There are some outstanding items to be addressed - please see the two reviewers' comments for details.

For Peer Review Only