Optimizing a Bayesian Hierarchical Adaptive Platform Trial Design for Stroke Patients

Guangyi Gao (ggao2@kumc.edu)
University of Kansas Medical Center https://orcid.org/0000-0003-3050-726X
Byron J Gajewski
University of Kansas Medical Center
Jo Wick
University of Kansas Medical Center
Jonathan Beall
Medical University of South Carolina
Jeffrey L Saver
University of California Los Angeles
Caitlyn Meinzer
Medical University of South Carolina

Research Article

Keywords: Platform trial design, Bayesian models, hierarchical models, response adaptive randomization, beta-binomial

Posted Date: June 25th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-584784/v1

Optimizing a Bayesian hierarchical adaptive platform trial design for stroke patients

Guangyi Gao[1,] Byron J. Gajewski[1], Jo Wick[1], Jonathan Beall[2], Jeffrey L. Saver[3], Caitlyn Meinzer,[2] the STEP-STONE Executive Committee

1. Department of Biostatistics & Data Science, University of Kansas Medical Center, Kansas City, KS, 66160

2. Department of Public Health Sciences, Medical University of South Carolina, Charleston, SC, 29425

3. Department of Neurology and Comprehensive Stroke Center, University of California, Los Angeles, Los Angeles, CA, 90095

In preparation for *Trials*, 06/01/2021

Abstract:

Background: Platform trials are well-known for their ability to investigate multiple arms on heterogeneous patient populations and their flexibility to add/drop treatment arms due to efficacy/lack of efficacy. Because of their complexity, it is important to develop highly optimized, transparent, and rigorous designs that are cost-efficient, offer high statistical power, maximize patient benefit and are robust to changes over time.

Methods: To address these needs, we present a Bayesian platform trial design based on a Beta-Binomial model for binary outcomes that uses three key strategies: 1) Hierarchical modelling of subgroups within treatment arms that allows for borrowing of information across subgroups, 2) utilization of response-adaptive randomization (RAR) schemes that seek a tradeoff between statistical power and patient benefit, and 3) adjustment for potential drift over time. Motivated by a proposed clinical trial that aims to find the appropriate treatment for different subgroup populations of ischemic stroke patients, extensive simulation studies were performed to validate the approach, compare different allocation rules and study the model operating characteristics.

Results & Conclusions: Our proposed approach achieved high statistical power, good patient benefit and was also robust against population drift over time. Our design provided a nice balance between the strengths of both the traditional RAR scheme and fixed 1:1 allocation and may be a promising choice for dichotomous outcomes trials investigating multiple subgroups.

Keywords: Platform trial design, Bayesian models; hierarchical models; response adaptive randomization; beta-binomial

1. Introduction

Master protocols, including umbrella, basket, and platform trials, are clinical trial designs which have received increased interest in the past few years. They simultaneously evaluate multiple drugs and/or multiple populations in multiple sub-studies and thus can accelerate the drug development process [1, 2]. Platform trials simultaneously investigate multiple treatments on multiple populations and are often referred to as "multi-arm, multi-stage" (MAMS) design trials [3-7]. This type of design allows for either a fixed number of treatments or an adaptive number of treatments by dropping and/or adding treatments during the process of the trial [8]. Compared to standalone designs, they are more efficient at identifying effective treatments for specific subpopulations and can require enrollment of fewer subjects for specific subpopulations [8]. While they may still result in a larger overall trial they can answer treatment questions for specific subpopulations. Basket trials and umbrella trials are subtypes of platform trials. Basket trials include a single investigational drug or device being tested on multiple diseases that share a specific biomarker or mutation [9-11]. They are often used in phase II studies with the goal to explore potential uses of a treatment or identify sub-populations in which a target treatment performs well or poorly [12]. Umbrella trials, on the other hand, compare multiple investigational drugs or devices in a single disease population [13-15]. They can identify treatments that perform well or poorly for a specific disease.

Recent studies [16-24] have shown endovascular thrombectomy (EVT) is a treatment of substantial benefit in select acute ischemic stroke patients and suggested EVT is a promising potential treatment in additional, not yet interrogated, sub-populations of acute ischemic stroke patients. Given the large difference in positive outcomes for subjects treated with EVT plus medical management versus standard medical management (MM) alone observed in these trials,

significant enthusiasm exists for expanding indication to additional subgroups not yet studied, as well as evaluating whether additional synergistic interventions exist. NIH-NINDS has published a notice of special interest (NOSI) in establishing a platform with a master protocol for multi-arm, multi-stage EVT trials [25]. To respond to the NOSI, the current authors, in collaboration with a team of clinical investigators, have developed the design proposed herein, with a focus on developing a first trial for performance on the platform that studies indication expansion to additional patient subgroups. If funded, the proposed "StrokeNet ThrombEctomy Platform - STarting with OptimizatioN of Eligibility" (STEP-STONE) trial, a companion trial to the platform, is a prospective, adaptive, registry-anchored trial that compares EVT plus MM medical to standard MM treatment alone, with the goal to identify patient subpopulations which can benefit from EVT treatment. In this trial, since we expect similarities in treatment differences across all subpopulations, a Bayesian hierarchical model was used to borrow information across different subgroups within a treatment arm thus improving the trial's efficiency. In addition, given the high efficacy observed in previous subgroups, clinicians do not have equipoise to randomize at a fixed equal allocation; instead response adaptive randomization is proposed to allocate patients to the more promising treatment as supportive evidence is acquired, facilitating investigator willingness to enroll.

Bayesian methods are attractive in adaptive trials, since they allow for continuous updating of posterior decision quantities as new information becomes available and thus they facilitate adapting to information obtained as a trial progresses [8, 26-30]. Motivated by previous work [31], the Bayesian hierarchical Beta-Binomial model used in the STEP-STONE trial included a tuning parameter in the prior distribution of response rates, that adjusts the "strength" of borrowing within treatment arms.

Response adaptive randomization (RAR) was utilized in the STEP-STONE trial to maximize patient benefit throughout the trial. While traditional clinical trials use fixed allocation and usually balance sample size equally in different treatment groups to eliminate bias, RAR is a patient allocation algorithm that has been commonly used in adaptive clinical trials to alter patient randomization probabilities based on interim results obtained from the trial. Updating the patient allocation ratio during the trial allows to randomize more patients to the more beneficial treatment, thus reduces the overall number of harmful events from the clinical trial and improves individual ethics [32-36].

There are many challenges accompanying the use of RAR in clinical trials, with one major challenge being patient population parameter drift [8, 37-39]. Drift occurs when the treatment response rates change over time. Without properly adjusting for drift effects, biased estimates could be obtained thus leading to wrong conclusions in the trial [37, 40, 41]. To alleviate this problem, Angus et al. used a first-order normal dynamic linear model (NDLM) to account for treatment response rates changing over time in the REMAP-CAP platform trial [42]. Motivated by their work, a drift parameter was also incorporated in the design of the STEP-STONE trial to capture the change in treatment response rates.

Another potential problem that arises in complex designs independently from RAR is the multiplicity issue. Multiplicity concerns arise when multiple comparison objectives are being evaluated in the same clinical trial [43] and failing to account for multiplicity results in inflation of type 1 errors. In the STEP-STONE trial, multiplicity occurred since multiple patient populations were included and multiple interims analysis were performed. In practice, controlling of familywise type 1 error in Bayesian designs often relies on simulation [8, 38, 44, 45]. In the STEP-

STONE study, thresholds of parameters were determined to ensure overall type 1 error being controlled at 0.05 level through extensive simulation studies.

The STEP-STONE trial is a two-arm, response adaptive platform trial. Previous research has shown [39, 46-48] that in the two-arm trial setting, compared with equal allocation, response adaptive allocation achieved lower statistical power due to unequal sample sizes [49]. To find a compromise between high statistical power in equal allocation and the high patient benefit obtained from response adaptive allocation, an innovative RAR scheme, "RARCOMP", is proposed.

In summary, to address the needs of the STEP-STONE trial, we proposed a two-arm adaptive platform trial design. Our approach has three distinct characteristics: (1) the use of a Bayesian hierarchical model that allows to gain efficiency by borrowing information between subgroups; (2) an innovative RAR allocation scheme (RARCOMP) that achieves a good balance between statistical power and patient benefit and (3) robustness to changes in the response over time. While covering all details and issues relating to platform trials is beyond the scope of this paper, the viability of our approach in two-arm trials with multiple sub-groups and binary primary endpoints is demonstrated via extensive simulation studies and the proposed RARCOMP scheme could be easily adapted to multi-arm settings.

2. Methods

2.1 Motivating Trial

In the STEP-STONE trial, the primary endpoint is binary and denotes if a favorable global disability level was observed at 90 days. Favorable outcome is assessed using prognosis-adjusted, sliding dichotomy analysis of the modified Rankin scale (mRS) [50-52]. Information obtained from both a prospective registry and previous related populations level I evidence trials,

is used to conduct patient allocation for each patient subpopulation. During the STEP-STONE trial, the patient allocation ratio is adaptively updated based on patients' treatment responses at each interim. Once a prespecified success criterion is identified for a subgroup, all future participants in that subgroup will be assigned to the superior treatment.

2.1.1. Patient Subgroups

The STEP-STONE trial will target three previously under-studied patient characteristics including 1) Individuals with Large Ischemic Cores, 2) Individuals with Mild Deficits, and 3) Individuals with Distal Vessel Occlusions. A **Large Ischemic Core** is defined as a substantial amount of already-injured brain tissue visualized using neuroimaging; a **Mild Deficit** is defined as few impairments in cognition, strength, vision, and other neurologic functions quantified using the National Institute of Health Stroke Scale; finally a **Distal Vessel Occlusion** refers to those strokes in which the causative clot(s) are located in intermediate (rather than large) diameter brain arteries. Of note, these characteristics are not mutually exclusive, though some combinations are clinically rare or highly unlikely, e.g. a large area of injury but only resulting in mild deficits (Large Ischemic Core + Mild Deficit).

Depending on whether these characteristics are present or not, patients are grouped into five mutually exclusive subgroups, which are Large Core Only, Mild Deficit Only, Distal Occlusion Only, Large core + Distal, and Mild Deficit + Distal. Figure 1 shows the Venn diagram of the five patient subpopulations and their respective expected population proportion.

Figure 1. Subgroup proportion summary.

2.2 Models

Three different models will be discussed in this section. We will first start with the simplest model, the Bayesian logistic independent model in section 2.2.1, as it is a commonly

used standard model for binary outcomes, and it is also the standard model in the Fixed and Adaptive Clinical Trials Simulator (FACTS) software. We then compare this model with two Bayesian hierarchical models. In section 2.2.2, we present a Bayesian hierarchical Beta-Binomial model that allows for borrowing information across patient subgroups within each treatment. A modified Bayesian hierarchical Beta-Binomial model which also accounts for patient response rates drift will be discussed in detail in section 2.2.3.

In all models, each of the five occurring combinations of the three patient characteristics (Mild Deficit, Distal Occlusion, Large Core) is treated as a unique subgroup without specific consideration for the base characteristics themselves. For each patient population subgroup $j = 1, \ldots, J$ and each treatment $k = 1, \ldots, K$, where in the STEP-STONE trial $J = 5$ and $K = 2$, the number of favorable outcomes Y_{jk} follows a binomial distribution, with parameters n_{jk} and P_{jk}. n_{jk}, where n_{jk} represents number of participants and P_{jk} is the probability of obtaining a favorable outcome. We will introduce the three Bayesian models accordingly.

2.2.1 Bayesian Logistic Independent Model

For the Bayesian logistic independent model, the response rates of subgroups are modeled separately each with its own prior in the independent model. In the STEP-STONE project, we assume the prior distribution for the log odds of the response rate in each subgroup and each treatment arm follows a simple normal distribution. The complete model can be written as:

$$\log\left(\frac{P_{jk}}{1 - P_{jk}}\right) = \beta_{jk}$$

Which means

$$P_{jk} = \Pr(favorable\ outcomes) = \frac{e^{\beta_{jk}}}{1 + e^{\beta_{jk}}}$$

Where the log odds β_{jk} uses the following prior distribution:

$$\beta_{jk} \sim N(0, 1.82^2)$$

This prior distribution is roughly equivalent to an uninformative $Beta\ (1,1)$ distribution on the response rate P_{jk}.

2.2.2 Bayesian Hierarchical Model

Motivated from previous work [31], in the **Bayesian hierarchical model**, the prior distribution of the response rate in each subgroup after receiving treatment k follows a Beta distribution with hyperparameters $(mP_k, m(1-P_k))$. This prior distribution allows response rates to borrow information among all subgroups within treatment k. Here, m is a constant that represents how strong our prior belief is, before the trial starts, that the treatment response rates P_{jk} are close to the average response rate P_k in treatment k. For example, if m is large, it means we are confident that P_{jk} is very close to P_k. In this paper, different m values were tested and compared.

The hyperparameter P_k is modeled through a Beta distribution with parameters (α_k, β_k). In this project, since we did not have strong prior knowledge about the average response rate of favorable outcomes, we considered an uninformative uniform prior, which is equivalent to setting α_k and β_k to 1. This uniform prior and can be interpreted as every possible probability of success from 0% to 100% being equally likely.

With this setup, the complete hierarchical model can be written as follows:

$$Y_{jk} \sim Binomial(n_{jk}, P_{jk})$$

$$P_{jk} \sim Beta(mP_k, m(1-P_k))$$

$$P_k \sim Beta(\alpha_k, \beta_k)$$

$$\alpha_k = \beta_k = 1$$

2.2.3 Bayesian Hierarchical Drift Model

Previous models assume the treatment response rates to not change over time. However, this is not always the case in real life clinical trials. If the response rates changed over time and were not adjusted for properly, severely biased estimates could be obtained thus leading to wrong decisions. Using the same approach described in the REMAP-CAP study [42], the previous Bayesian hierarchical model can be modified to include a drift parameter that accounts for treatment response rates changing over time.

In this model, we consider time points to correspond to interim analyses and the final analysis after completion of the trial. The time-indicating variable t is an integer ranging from 1 to T, with T representing the most recent time point. Each treatment response rate for the most recent time point P_{jkT} is modelled using the same structure as in the previous hierarchical model. For every previous time point, the response rate is modelled on the log odds scale as the sum of the response rate of the most recent time point and the time effect θ_t. The time effect parameters $\boldsymbol{\theta}_t$ are modelled with a first-order normal dynamic linear model (NDLM). The hyper prior of the drift parameter τ follows an inverse-gamma distribution. This model allows for borrowing among effects of adjacent time periods and can robustly handle different trends over time. The borrowing is controlled by the drift parameter τ. The full model can be summarized below, for the last time point T,

$$Y_{jkT} \sim Binomial(n_{jkT}, P_{jkT})$$

$$P_{jkT} \sim Beta(mP_k, m(1 - P_k))$$

For all $t < T$

$$Y_{jkt} \sim Binomial(n_{jkt}, P_{jkt})$$

$$logit(P_{j,k,t}) = logit(P_{jkT}) + \theta_t$$

$$\theta_{t-1} \sim N(\theta_t, \tau)$$

$$\tau \sim InvGamma(0.25, 0.1)$$

2.3 Bayesian Quantities of Interest

2.3.1 Posterior Probability of Treatment Difference

For each treatment, $k = EVT, MM$, the posterior probability of treatment difference $P(p_k - p_{k'} > 0)$ can be understood as the posterior probability that one treatment k is superior to another treatment k'.

After samples are drawn from each respective posterior distribution, the probabilities of treatment difference are calculated as the proportion of posterior samples where respectively either $p_{EVT} - p_{MM}$ or $p_{MM} - p_{EVT}$ is greater than 0.

2.3.2 Odds Ratio

For each subgroup, we calculate the posterior odds ratio of the probability of obtaining a favorable outcome response between two treatments as such:

$$OR_j = \frac{\frac{P_{jMM}}{1 - P_{jMM}}}{\frac{P_{jEVT}}{1 - P_{jEVT}}}$$

2.4 Study Design and Patient Accrual

The motivating study envisions a trial which recruits and follows subjects for four years with three interim analyses and one final analysis. The first interim is scheduled to occur after 2500 participants have enrolled into the trial. Subsequent interims will be conducted after every additional 2500 participants are enrolled and will continue until a total of 10,000 participants are enrolled. Since interim analyses are defined by participants enrolled, the timing of the interims, is random and will depend on the rate at which participants accrue to the trial. Overall, since we expect to enroll 10,000 participants in 4 years, an average of 52 participants have to be enrolled

per week. Here, we assume the patient accrual will follow a Poisson distribution with parameter 52.

$$Y \sim Poisson(52)$$

We considered three different study designs:

1) A fixed allocation design in which patients are always allocated to the two treatment arms in a 1:1 ratio. No interim analysis will be performed during the trial process.

2) A response adaptive randomization (RAR) design that updates allocation to favor the more promising treatment at each interim based on the Bayesian quantities of interest.

3) A modified RAR design that finds a compromise between the 1:1 and the pure RAR allocation ratios, named "RARCOMP".

For both RAR and RARCOMP designs, three interim analyses and one final analysis were performed as described above. Details about the adaptive randomization schemes will be explained in the next section.

2.5 Patient Allocation in Adaptive Designs

Adaptive randomization will begin right after the trial starts, using within subgroup prior information, and is performed at each interim, with the goal to allocate more subjects to the treatment that appears to be more promising. Bayesian quantities of interest discussed above were used to guide decisions. The patient allocation flowchart in Figure 2 briefly summarizes how patients were allocated in a single clinical trial. The posterior response rates for both treatment arms were compared for each patient subgroup. If a superiority criterion was satisfied for any subgroup, all future patients would be allocated to the superior arm for that patient subgroup. Equivalence of the two treatments were tested if the superiority criterion was not met. An establishment of equivalence would lead to all future patients being allocated to the MM

treatment for lower cost. If neither superiority nor equivalence were established, the patient allocation rates for the two treatments would be calculated using prespecified allocation schemes. Details about patient allocation will be provided in this section.

Figure 2. Adaptive Patient Allocation Flowchart.

2.5.1 Allocation for Expected Success

Patient randomization information may change at each interim analysis due to expected success and allocate all future participants to the superior treatment if the following criteria is satisfied, where k and k' represent different treatments:

$$P(P_{jk} > P_{jk'}) > \gamma$$

The value of γ was obtained based on simulation and controlled for overall type 1 error to be close to 0.05. It varies for different m values and for different randomization schemes (Section 3.1).

2.5.2 Allocation for Equivalence (effectively MM should be used)

During each interim, if the expected success condition is not met, the trial may change patient randomization due to equivalence and allocate all future participants to the MM treatment, since it is a less expensive treatment option. Equivalence is established if the following criterion is satisfied:

$$P\big(0.8 < OR_j < 1.2\big) > 0.7$$

The utilized boundaries for odds ratios in the above criterion have been traditionally used in bioequivalence studies and were selected based on this fact.

2.5.3 Allocation When No Success or Equivalence is Met

If neither superiority nor equivalence is identified, the patient allocation rates are calculated based on prespecified randomization schemes. They are 1) the common RAR allocation scheme and 2) RARCOMP - the modified RAR scheme.

For RAR, the probability V_{jk} of the next participant being allocated to treatment k in subgroup j, was calculated such that it satisfies the formula shown below:

$$V_{jkRAR} \propto \sqrt{P\left(P_{jk} - P_{jk'} > 0\right)\frac{Var\left(P_{jk}\right)}{n_{jk}+1}}$$

where $Var(P_{jk})$ are the posterior variances of the mean response rates, n_{jk} is the current number of participants in subgroup j assigned to treatment k, and k' being the treatment arm other than k. The randomization probabilities for treatments will be updated once at each interim.

For the fixed 1:1 allocation, these probabilities were both 0.5.

$$V_{jkfixed} = V'_{jkfixed} = \frac{1}{K} = 0.5$$

RARCOMP represents a trade-off between RAR and the fixed 1:1 allocation, where the allocation rate for treatment k is then the average of V_k and 0.5. The allocation rate for the new RAR compromise patient allocation scheme can thus be written as:

$$V_{jk_{compromise}} = \frac{V_{jkRAR} + V_{jkfixed}}{2}$$

2.5.4 Initial Allocation in Adaptive Designs

In the RAR and RARCOMP schemes, prior knowledge provided by the experts was used to inform patient allocation within subgroups at the start of the trial. This was done as follows: Let P_{jk}^0 be the current understanding of the rate of favorable outcome for treatment k in subgroup j. Let n_{jk}^0 be the prior sample size for the treatment k in subgroup j. Before the trial starts, create

a pseudo-dataset with $n_{jk} = n_{jk}^0 = 10$ observations in each subgroup-treatment combination and a response of $Y_{jk} = P_{jk}^0 \times n_{jk}^0$. Sample size for the pseudo-dataset was chosen to be 10, so that previous information was incorporated in the design but not too overpowering to bias the estimates. Based on this data calculate posterior quantities of interest and follow the allocation rules of the study protocol. Prior knowledge about P_{jk}^0 utilized in this trial is shown in Table 1.

Table 1. Prior information to calculate the patient allocation before trial starts

Subgroup	P_{MM}	P_{EVT}
Large Core Only	0.10	0.25
Mild Deficit Only	0.70	0.84
Distal Occlusion Only	0.35	0.55
Distal Occlusion + Large core	0.25	0.45
Distal Occlusion + Mild Deficit	0.75	0.85

2.6 Simulation Study

In this paper, we investigated and compared scenarios where the m value varies from 1 to 30 ($m \in \{1, 10, 20, 30\}$) for Bayesian hierarchical models.

2.6.1 Simulating Data without Drift Effect

We simulated 10000 clinical trial studies to investigate the model operating characteristics for each design. In order to study design performance, five simulation scenarios were considered: 1) one "**equal**" scenario in which the favorable outcome rates of MM and EVT are simulated to be the same (averaging across MM and EVT treatment for each subgroup), 2) an "**expected**" scenario where the favorable outcome response rates in EVT is simulated to be higher than in MM based on the previous knowledge, 3) a "**reverse**" scenario where the favorable outcome response rate in MM is simulated to be higher than in EVT, 4) an "**extreme**" case where the favorable outcome response rate in EVT is simulated to be much higher than in

MM and 5) a scenario in which "**single subgroup**" is better in EVT while the two treatments are the same for the rest of the subgroups. Details about the four scenarios are shown in Table 2.

Table 2. A summary of five simulation scenarios without drift effect.

Subgroup	**Equal**		**Expected**		**Reversed**		**Extreme EVT**		**Single Subgroup**	
	P_{MM}	P_{EVT}	P_{MM}	P_{EVT}	P_{MM}	P_{EVT}	P_{MM}	P_{EVT}	P_{MM}	P_{EVT}
Large Core Only	0.1	0.1	0.1	0.25	0.25	0.1	0.1	0.30	0.25	0.45
Mild Deficit Only	0.7	0.7	0.7	0.84	0.84	0.7	0.7	0.89	0.84	0.84
Distal Occlusion Only	0.35	0.35	0.35	0.55	0.55	0.35	0.35	0.60	0.55	0.55
Distal Occlusion + Large core	0.25	0.25	0.25	0.45	0.45	0.25	0.25	0.50	0.45	0.45
Distal Occlusion + Mild Deficit	0.75	0.75	0.75	0.85	0.85	0.75	0.75	0.90	0.85	0.85

2.6.2 Simulating Data with Drift Effect

Similarly, we also simulated all five scenarios when a drift effect was present in the data. To achieve this, the true response rates for the last time point P_{jkT} were chosen to be the same as the values in Table 2. However, the log odds of response rates for previous time points were set to decrease linearly over time. Under this simulation setup, response rates in earlier stages of the trial were higher than in the later stages. The simulated response rates for each time point are summarized in Table 3.

Table 3. Linear time effects for response rates used in simulation studies.

	$t = 1$	$t = 2$	$t = 3$	$t = 4$
θ_t	0.75	0.5	0.25	0
$logit(P_{jkt})$	$logit(P_{jkT}) + 0.75$	$logit(P_{jkT}) + 0.5$	$logit(P_{jkT}) + 0.25$	$logit(P_{jkT}) + 0$

t represents each interim analysis, θ_t represents time effects at different time points.

2.6.3 Model Operating Characteristics Evaluation

Bayesian hierarchical modelling was performed using the R (Version 3.5.3) package "Nimble" [53] (Version 0.9.0) (code provided in Appendix (c) in the supporting material). The results of the adaptive designs were then compared with two versions of fixed 1:1 allocation

designs: one using the Bayesian hierarchical model fit in Nimble, and another using the independent model fit in the Fixed and Adaptive Clinical Trial Simulator (FACTS) (Berry & Sanil, 2010) software [54] (Version 6.3), having no interims. The independent model fitted in FACTS is served as the standard design, however it is limited.

The type 1 error was obtained from the "Equal" scenario and was calibrated to the 0.05 level by adjusting γ in designs simulated in R and NIMBLE. Using these γ values statistical power was then evaluated in the remaining scenarios. Generally, the type 1 error was calculated as the proportion of simulations in which either EVT is superior, or MM is superior under the true scenario that EVT and MM have the same response rate; while power was calculated as the proportion of simulations which correctly exhibit superiority of either treatment under scenario "Expected", "Reversed" and "Extreme EVT".

3 Results

3.1 Type 1 Error Calibration

Overall type 1 error was successfully controlled at the 0.05 level in all simulated scenarios. The required γ thresholds tended to decrease with m when employing a fixed allocation scheme but remained stable at approximately 0.995 when employing the response adaptive designs. The simulation based overall type 1 errors for all scenarios as well as their corresponding γ values are provided in the Appendix (a) in the supporting material.

3.2 Bayesian Hierarchical Model on Data without Drift Effect

Statistical power for different randomization schemes was compared after calibrating the overall type 1 error at the 0.05 level. Figure 3 shows a comparison among three randomization schemes using Bayesian hierarchical model fit data that does not have a time drift effect under all alternative scenarios when m value is set to be 1. With y-axis being the difference in power

between the fixed design (with independent model) and the three adaptive designs with various randomization schemes respectively (with hierarchical models), for example, one of the y value could be $\text{Power}_{\text{fixed independent model}} - \text{Power}_{\text{adaptive Bayesian hierarchical}}$. Since an equal allocation of patients in the two-arm setting provides higher power, treating the independent fixed model as a reference, a smaller y value in Figure 3 indicates a higher statistical power. When $m = 1$, in all scenarios, among all randomization schemes, fixed allocation appeared to have the highest statistical power. This power difference was not strong for the first three subgroups as the sample sizes in those subgroups were large. However, the differences were extremely obvious for subgroups "Distal Occlusion + Large Core" and "Distal Occlusion + Mild Deficit" in scenarios "Expected", "Reversed" and "Extreme EVT" due to small sample sizes. RAR randomization scheme appeared to have the lowest statistical power as the y values for RAR tend to be the highest among the three schemes. RARCOMP scheme provided power higher than RAR, but lower than fixed 1:1 allocation rule.

Figure 3. Power difference among three adaptive design schemes relative to the fixed design when m=1.

Increasing the m value from 1 to 30, the statistical power increased for all randomization schemes in all scenarios for subgroup "Distal Occlusion + Mild Deficit" as the y values for that subgroup dropped to 0 for all schemes. However, the power was decreased for subgroup "Distal Occlusion + Large core" (Figure 4). In this subgroup, we can reach the same conclusion as before that the fixed 1:1 allocation obtained the highest power followed by RARCOMP allocation scheme. The power obtained from the RAR scheme was the lowest among all three schemes.

Figure 4. Power difference among three adaptive design schemes relative to the fixed design when m=30.

The inconsistent behavior in statistical power between the last two subgroups was caused by increased estimation bias when increasing the m value in the model. A brief demonstration of how power changes for scenario "Expected" can be found in Appendix (b) in the supporting material.

One of the benefits of using adaptive designs is to allocate more patients to the better performed treatment, thus improving the patient benefit within the trial. Figure 5 shows the patient benefit comparison among the three randomization schemes stratified by m value. The Y-axis represent the difference between the hypothetical subjects' proportion with good outcomes and the observed subjects proportion with good outcomes, with the former being the proportion of subjects that would experience a good outcome in a perfect world, where all subjects are always allocated to the treatment arm with the highest success rate, and the latter being the proportion of observed good outcomes in the simulated trials. In this way, a smaller y-value indicates higher patient benefits. In Figure 5, the RAR scheme obtained the highest patient benefit, which was closely followed by the RARCOMP scheme. The Fixed allocation scheme achieved the lowest patient benefit among all three schemes. Comparing m=1 to m=30 alone, although the differences were small, m=30 obtained higher patient benefit under all schemes for most of the scenarios.

Figure 5. Patient benefit comparison for three randomization schemes.

In summary, the RARCOMP randomization scheme has shown to improve statistical power compared to the regular RAR scheme, without compromising too much patient benefit. Also, increasing m led to a higher statistical power but also more biased estimates. Since m=30

provided the best performance (higher power, higher patient benefit with moderately biased estimates), in this paper we will focus on the model performance under m=30 setup.

3.3 Bayesian Hierarchical Model on Data with Drift Effect

The previous results compare the three randomization schemes when Bayesian hierarchical models were fit to the data without a time drift effect. When fitting the same model to data in which the response rates changed over time and using response adaptive randomization instead of a fixed 1:1 allocation scheme, a huge inflation in type 1 error was observed (Figure 6 (a)).

Figure 6. Type 1 error comparison for Bayesian hierarchical model and Bayesian hierarchical drift model.

3.4 Bayesian Hierarchical Drift Model on Data with Drift Effect

Fitting our Bayesian hierarchical drift model to data in which the response rates changed over time using response adaptive randomization, type 1 error was well controlled below 0.1 for both randomization schemes (Figure 6 (b)).

In addition to well-controlled type 1 error, the Bayesian hierarchical drift model also established a very high performance. High statistical power was observed for all alternative scenarios for both randomization schemes. RARCOMP appeared to have higher power than RAR in subgroup “Distal Occlusion + Large core”. Although the differences were small, RAR showed higher power for “Distal Occlusion + Mild Deficit” in scenarios “Expected” and “Reversed” compared with RARCOMP (Figure 7).

Figure 7. Power plots: fit Bayesian hierarchical drift model to linear time effect drift data.

Patient benefit was also compared between RAR and RARCOMP scheme. In Figure 8, a red circle was used to indicate the scenario when the drift model was fitted to the linear drift effect data. The y-axis represents the differences in patient benefit between the hypothetical proportion of patients obtaining good outcomes and the observed proportion of patients with good outcomes, a smaller y-axis value indicating a higher patient benefit. RAR and RARCOMP both obtained very high patient benefit, with the RAR scheme achieving slightly higher values.

Figure 8. Patient benefit when fitting Bayesian drift model.

3.5 Bayesian Hierarchical Drift Model on Data without Drift Effect

We have shown previously that the Bayesian hierarchical drift model handles response rates drift over time well when the drift effect is linear across time. However, to be a promising and robust model, the model still needs to perform well in situations where the time drift effect is absent. Figure 9 shows the power remained very high even when our drift model was fitted to a dataset that does not have a linear time effect. Comparing RAR and RARCOMP, the RARCOMP allocation scheme achieved higher statistical power especially in the subgroup "Distal Occlusion + Large Core".

Patient benefit was also evaluated for this setup. In figure 8, a blue box was used to indicate the scenario when the drift model was fitted to the data without a drift effect. Comparing the patient benefit when fitting the same model to both data with linear drift effect (red circle in Figure 8) and to data without linear drift effect (blue box in Figure 8), values were very similar, suggesting that this model was very robust against whether or not a linear time effect was present in the data.

Figure 9. Power plots: fit Bayesian hierarchical drift model to data without a linear effect.

4 Discussion

Our simulation studies have shown that the RARCOMP scheme can provide high statistical power while maintaining high patient benefit in all simulated scenarios. However, the use of RAR in two arm studies has been controversial[55-57]. Previous work has shown using RAR in two arm trials without careful planning and calibration could result in biased estimates and might even lead to wrong conclusions. In addition to its ability to balance statistical power and patient benefit, the RARCOMP scheme could help to mitigate this issue. The fact that it averages allocation ratios between the naïve RAR and the fixed 1:1 randomization, prevents the allocation process from creating highly unbalanced sample sizes between the two treatments, and makes it more robust to RAR bias.

As response adaptive designs are more susceptible to drift effects[57], in this paper, we also incorporated a drift parameter in the Bayesian model to account for response rates drift over time. Simulation results demonstrated our drift model can accurately estimate linear trend drift effects over time and account for these changes when comparing treatments. Moreover, even when time effects were absent in the data, our drift model still performed well and retained high statistical power. In combination with the fact that the NDLM component used to estimate time effects is able to flexibly model different shapes, our results suggest that this approach can be robustly applied in many clinical trial scenarios.

Current simulation results have confirmed our drift model works well on data with linear drift effect. More work needs to be done to confirm the drift model also maintains high performance in other situations. However, since we used noninformative informative priors on the drift effect τ, as long as the change in the response rates are not dramatic during a period of time, it is safe to guess our model could perform well even when the time effects are nonlinear.

In conclusion, with the ability to have high power, good patient benefit and to account for population drift, our design using the Bayesian hierarchical drift model with the RARCOMP scheme is a promising choice for adaptive trials. This article introduces the novel idea of combining the traditional RAR scheme and fixed 1:1 allocation to provide a nice balance between them. Our design is robust against both severely unbalanced allocation and drift over time.

5 List of abbreviations

RAR: Response adaptive randomization

MAMS: Multi-arm, multi-stage

EVT: Endovascular thrombectomy

MM: Medical management

NOSI: Notice of special interest

STEP-STONE: StrokeNet ThrombEctomy Platform - STarting with OptimizatioN of Eligibility

NDLM: Normal dynamic linear model

RARCOMP: An innovative RAR allocation scheme

mRS: Modified Rankin scale

FACTS: Fixed and Adaptive Clinical Trials Simulator

6 Declarations

6.1 Ethics approval and consent to participate

Not applicable.

6.2 Consent for publication

Not applicable.

6.3 Availability of data and materials

No real-world data was collected in this study. Nimble code to perform model fitting is provided in Appendix (c) in the supporting material.

6.4 Competing interests

The authors declare that they have no competing interests.

6.5 Funding

This work was partially supported by a CTSA grant from NCATS awarded to the University of Kansas for Frontiers: University of Kansas Clinical and Translational Science Institute (# UL1TR002366). The contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH or NCATS.

6.6 Authors' contributions

All authors conceptualized the study. GG and JB completed the data simulation and analyses. GG drafted the initial draft of the manuscript. All authors read, reviewed, and approved the final version of the manuscript.

6.7 Acknowledgements

Drs. Saver and Meinzer provided input on the development of this manuscript on behalf of the members of the STEP Executive committee; we would like to extend our gratitude to the other members of the STEP-STONE Executive Committee members, Dr. Colin Derdeyn, Dr. David Fiorella, Dr. Tudor Jovin, Dr. Pooja Khatri, Dr. Eva Mistry, Dr. J Mocco, Dr. Raul Nogueira, and Dr. Adnan Siddiqui for their scientific and clinical input leading to the conceptual development of this platform and associated STEP-STONE trial.

7 References

1. Park JJH, Siden E, Zoratti MJ, Dron L, Harari O, Singer J, et al. Systematic review of basket trials, umbrella trials, and platform trials: a landscape analysis of master protocols. Trials. 2019;20(1):572.

2. Woodcock J, LaVange LM. Master Protocols to Study Multiple Therapies, Multiple Diseases, or Both. N Engl J Med. 2017;377(1):62-70.
3. Parmar MK, Sydes MR, Cafferty FH, Choodari-Oskooei B, Langley RE, Brown L, et al. Testing many treatments within a single protocol over 10 years at MRC Clinical Trials Unit at UCL: Multi-arm, multi-stage platform, umbrella and basket protocols. Clin Trials. 2017;14(5):451-61.
4. Bateman RJ, Benzinger TL, Berry S, Clifford DB, Duggan C, Fagan AM, et al. The DIAN-TU Next Generation Alzheimer's prevention trial: Adaptive design and disease progression model. Alzheimers Dement. 2017;13(1):8-19.
5. Alexander BM, Ba S, Berger MS, Berry DA, Cavenee WK, Chang SM, et al. Adaptive Global Innovative Learning Environment for Glioblastoma: GBM AGILE. Clin Cancer Res. 2018;24(4):737-43.
6. Park JW, Liu MC, Yee D, Yau C, van 't Veer LJ, Symmans WF, et al. Adaptive Randomization of Neratinib in Early Breast Cancer. N Engl J Med. 2016;375(1):11-22.
7. Barker AD, Sigman CC, Kelloff GJ, Hylton NM, Berry DA, Esserman LJ. I-SPY 2: an adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy. Clin Pharmacol Ther. 2009;86(1):97-100.
8. Saville BR, Berry SM. Efficiencies of platform clinical trials: A vision of the future. Clin Trials. 2016;13(3):358-66.
9. Hainsworth JD, Meric-Bernstam F, Swanton C, Hurwitz H, Spigel DR, Sweeney C, et al. Targeted Therapy for Advanced Solid Tumors on the Basis of Molecular Profiles: Results From MyPathway, an Open-Label, Phase IIa Multiple Basket Study. J Clin Oncol. 2018;36(6):536-42.
10. Li BT, Shen R, Buonocore D, Olah ZT, Ni A, Ginsberg MS, et al. Ado-Trastuzumab Emtansine for Patients With HER2-Mutant Lung Cancers: Results From a Phase II Basket Trial. J Clin Oncol. 2018;36(24):2532-7.
11. Schoffski P, Sufliarsky J, Gelderblom H, Blay JY, Strauss SJ, Stacchiotti S, et al. Crizotinib in patients with advanced, inoperable inflammatory myofibroblastic tumours with and without anaplastic lymphoma kinase gene alterations (European Organisation for Research and Treatment of Cancer 90101 CREATE): a multicentre, single-drug, prospective, non-randomised phase 2 trial. Lancet Respir Med. 2018;6(6):431-41.
12. Simon R. Critical Review of Umbrella, Basket, and Platform Designs for Oncology Clinical Trials. Clin Pharmacol Ther. 2017;102(6):934-41.
13. Kim ES, Herbst RS, Wistuba, II, Lee JJ, Blumenschein GR, Jr., Tsao A, et al. The BATTLE trial: personalizing therapy for lung cancer. Cancer Discov. 2011;1(1):44-53.
14. Herbst RS, Gandara DR, Hirsch FR, Redman MW, LeBlanc M, Mack PC, et al. Lung Master Protocol (Lung-MAP)-A Biomarker-Driven Protocol for Accelerating Development of Therapies for Squamous Cell Lung Cancer: SWOG S1400. Clin Cancer Res. 2015;21(7):1514-24.
15. Papadimitrakopoulou V, Lee JJ, Wistuba, II, Tsao AS, Fossella FV, Kalhor N, et al. The BATTLE-2 Study: A Biomarker-Integrated Targeted Therapy Study in Previously Treated Patients With Advanced Non-Small-Cell Lung Cancer. J Clin Oncol. 2016;34(30):3638-47.
16. Albers GW, Marks MP, Kemp S, Christensen S, Tsai JP, Ortega-Gutierrez S, et al. Thrombectomy for stroke at 6 to 16 hours with selection by perfusion imaging. New England Journal of Medicine. 2018;378(8):708-18.
17. Berkhemer OA, Fransen PS, Beumer D, Van Den Berg LA, Lingsma HF, Yoo AJ, et al. A randomized trial of intraarterial treatment for acute ischemic stroke. n Engl J Med. 2015;372:11-20.
18. Campbell BC, Mitchell PJ, Kleinig TJ, Dewey HM, Churilov L, Yassi N, et al. Endovascular therapy for ischemic stroke with perfusion-imaging selection. New England Journal of Medicine. 2015;372(11):1009-18.
19. Goyal M, Demchuk AM, Menon BK, Eesa M, Rempel JL, Thornton J, et al. Randomized assessment of rapid endovascular treatment of ischemic stroke. New England Journal of Medicine. 2015;372(11):1019-30.

20. Jovin TG, Chamorro A, Cobo E, de Miquel MA, Molina CA, Rovira A, et al. Thrombectomy within 8 hours after symptom onset in ischemic stroke. New England Journal of Medicine. 2015;372(24):2296-306.
21. Mocco J, Zaidat OO, von Kummer R, Yoo AJ, Gupta R, Lopes D, et al. Aspiration thrombectomy after intravenous alteplase versus intravenous alteplase alone. Stroke. 2016;47(9):2331-8.
22. Nogueira RG, Jadhav AP, Haussen DC, Bonafe A, Budzik RF, Bhuva P, et al. Thrombectomy 6 to 24 hours after stroke with a mismatch between deficit and infarct. New England Journal of Medicine. 2018;378(1):11-21.
23. Saver JL, Goyal M, Bonafe A, Diener H-C, Levy EI, Pereira VM, et al. Stent-retriever thrombectomy after intravenous t-PA vs. t-PA alone in stroke. New England Journal of Medicine. 2015;372(24):2285-95.
24. Muir KW, Ford GA, Messow C-M, Ford I, Murray A, Clifton A, et al. Endovascular therapy for acute ischaemic stroke: the Pragmatic Ischaemic Stroke Thrombectomy Evaluation (PISTE) randomised, controlled trial. Journal of Neurology, Neurosurgery & Psychiatry. 2017;88(1):38-44.
25. (NINDS) NIoNDaS. Notice of Special Interest (NOSI): Platform Trials of Thrombectomy in Acute Stroke Treatment 2020 [Available from: https://grants.nih.gov/grants/guide/notice-files/NOT-NS-20-095.html.
26. Berry DA. Bayesian clinical trials. Nat Rev Drug Discov. 2006;5(1):27-36.
27. Berry DA. The Brave New World of clinical cancer research: Adaptive biomarker-driven trials integrating clinical practice with clinical research. Mol Oncol. 2015;9(5):951-9.
28. Berry SM, Connor JT, Lewis RJ. The platform trial: an efficient strategy for evaluating multiple treatments. JAMA. 2015;313(16):1619-20.
29. Hirakawa A, Asano J, Sato H, Teramukai S. Master protocol trials in oncology: Review and new trial designs. Contemp Clin Trials Commun. 2018;12:1-8.
30. Schnell PM, Tang Q, Offen WW, Carlin BP. A Bayesian credible subgroups approach to identifying patient subgroups with positive treatment effects. Biometrics. 2016;72(4):1026-36.
31. Zang Y, Guo B, Han Y, Cao S, Zhang C. A Bayesian adaptive marker-stratified design for molecularly targeted agents with customized hierarchical modeling. Stat Med. 2019;38(15):2883-96.
32. Park JJ, Thorlund K, Mills EJ. Critical concepts in adaptive clinical trials. Clin Epidemiol. 2018;10:343-51.
33. Lee JJ, Xuemin G, Suyu L. Bayesian adaptive randomization designs for targeted agent development. Clin Trials. 2010;7(5):584-96.
34. Cheng Y, Berry DA. Optimal adaptive randomized designs for clinical trials. Biometrika. 2007;94(3):673-89.
35. Thall PF, Wathen JK. Practical Bayesian adaptive randomisation in clinical trials. Eur J Cancer. 2007;43(5):859-66.
36. Hu F, Rosenberger WF. The theory of response-adaptive randomization in clinical trials: John Wiley & Sons; 2006.
37. Thall PF, Fox PS, Wathen JK. Some caveats for outcome adaptive randomization in clinical trials. Modern adaptive randomized clinical trials: statistical and practical aspects: Chapman and Hall/CRC; 2015. p. 287-305.
38. Connor JT, Elm JJ, Broglio KR, Esett, Investigators A-I. Bayesian adaptive trials offer advantages in comparative effectiveness trials: an example in status epilepticus. J Clin Epidemiol. 2013;66(8 Suppl):S130-7.
39. Thall P, Fox P, Wathen J. Statistical controversies in clinical research: scientific and ethical problems with adaptive randomization in comparative clinical trials. Annals of Oncology. 2015;26(8):1621-8.

40. Karrison TG, Huo D, Chappell R. A group sequential, response-adaptive design for randomized clinical trials. Control Clin Trials. 2003;24(5):506-22.
41. Cook JD. The effect of population drift on adaptively randomized trials. 2007.
42. Angus DC, Berry S, Lewis RJ, Al-Beidh F, Arabi Y, van Bentum-Puijk W, et al. The REMAP-CAP (Randomized Embedded Multifactorial Adaptive Platform for Community-acquired Pneumonia) Study. Rationale and Design. Ann Am Thorac Soc. 2020;17(7):879-91.
43. Dmitrienko A, D'Agostino RB, Sr. Multiplicity Considerations in Clinical Trials. N Engl J Med. 2018;378(22):2115-22.
44. Ryan EG, Brock K, Gates S, Slade D. Do we need to adjust for interim analyses in a Bayesian adaptive trial design? BMC Medical Research Methodology. 2020;20(1):1-9.
45. Broglio KR, Connor JT, Berry SM. Not too big, not too small: a goldilocks approach to sample size selection. J Biopharm Stat. 2014;24(3):685-705.
46. Korn EL, Freidlin B. Outcome-adaptive randomization: is it useful? Journal of Clinical Oncology. 2011;29(6):771.
47. Hey SP, Kimmelman J. Are outcome-adaptive allocation trials ethical? Clin Trials. 2015;12(2):102-6.
48. Viele K, Broglio K, McGlothlin A, Saville BR. Comparison of methods for control allocation in multiple arm studies using response adaptive randomization. Clinical Trials. 2020;17(1):52-60.
49. Korn EL, Freidlin B. Adaptive clinical trials: advantages and disadvantages of various adaptive design elements. JNCI: Journal of the National Cancer Institute. 2017;109(6):djx013.
50. Murray GD, Barer D, Choi S, Fernandes H, Gregson B, Lees KR, et al. Design and analysis of phase III trials with ordered outcome scales: the concept of the sliding dichotomy. J Neurotrauma. 2005;22(5):511-7.
51. Saver JL. Optimal end points for acute stroke therapy trials: best ways to measure treatment effects of drugs and devices. Stroke. 2011;42(8):2356-62.
52. Yeatts SD. Novel methodologic approaches to phase I, II, and III trials. Stroke. 2013;44(6 Suppl 1):S116-8.
53. de Valpine P, Turek D, Paciorek CJ, Anderson-Bergman C, Lang DT, Bodik R. Programming with models: writing statistical algorithms for general model structures with NIMBLE. Journal of Computational and Graphical Statistics. 2017;26(2):403-13.
54. Berry S, Sanil A. FACTSTM Dose finding: single endpoint engine specification. Tessela, Newton, MA. 2010.
55. Korn EL, Freidlin B. Outcome--adaptive randomization: is it useful? J Clin Oncol. 2011;29(6):771-6.
56. Thall P, Fox P, Wathen J. Statistical controversies in clinical research: scientific and ethical problems with adaptive randomization in comparative clinical trials. Ann Oncol. 2015;26(8):1621-8.
57. Berry DA. Adaptive clinical trials: the promise and the caution. J Clin Oncol. 2011;29(6):606-9.

Figures

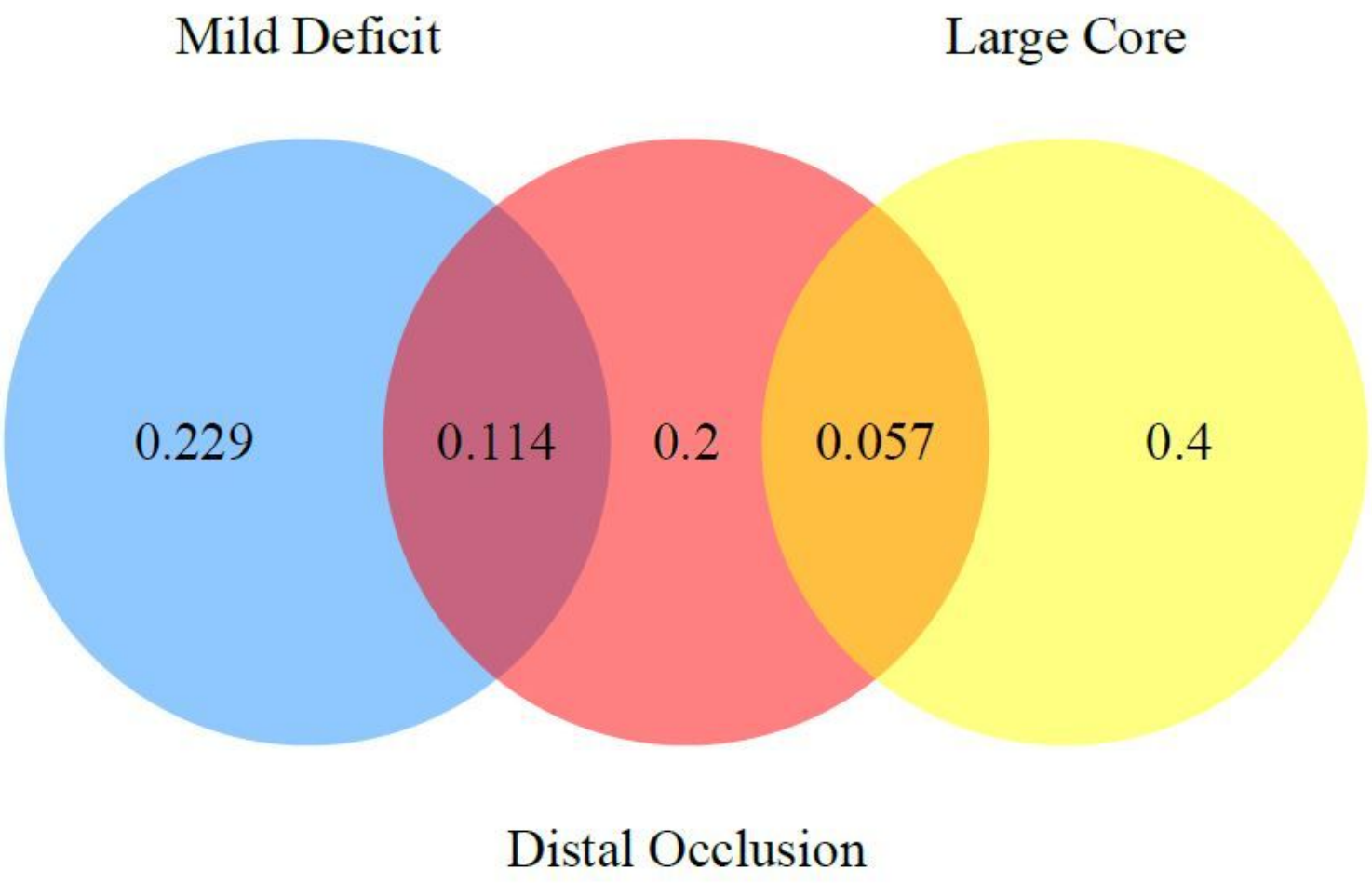

Figure 1

Subgroup proportion summary.

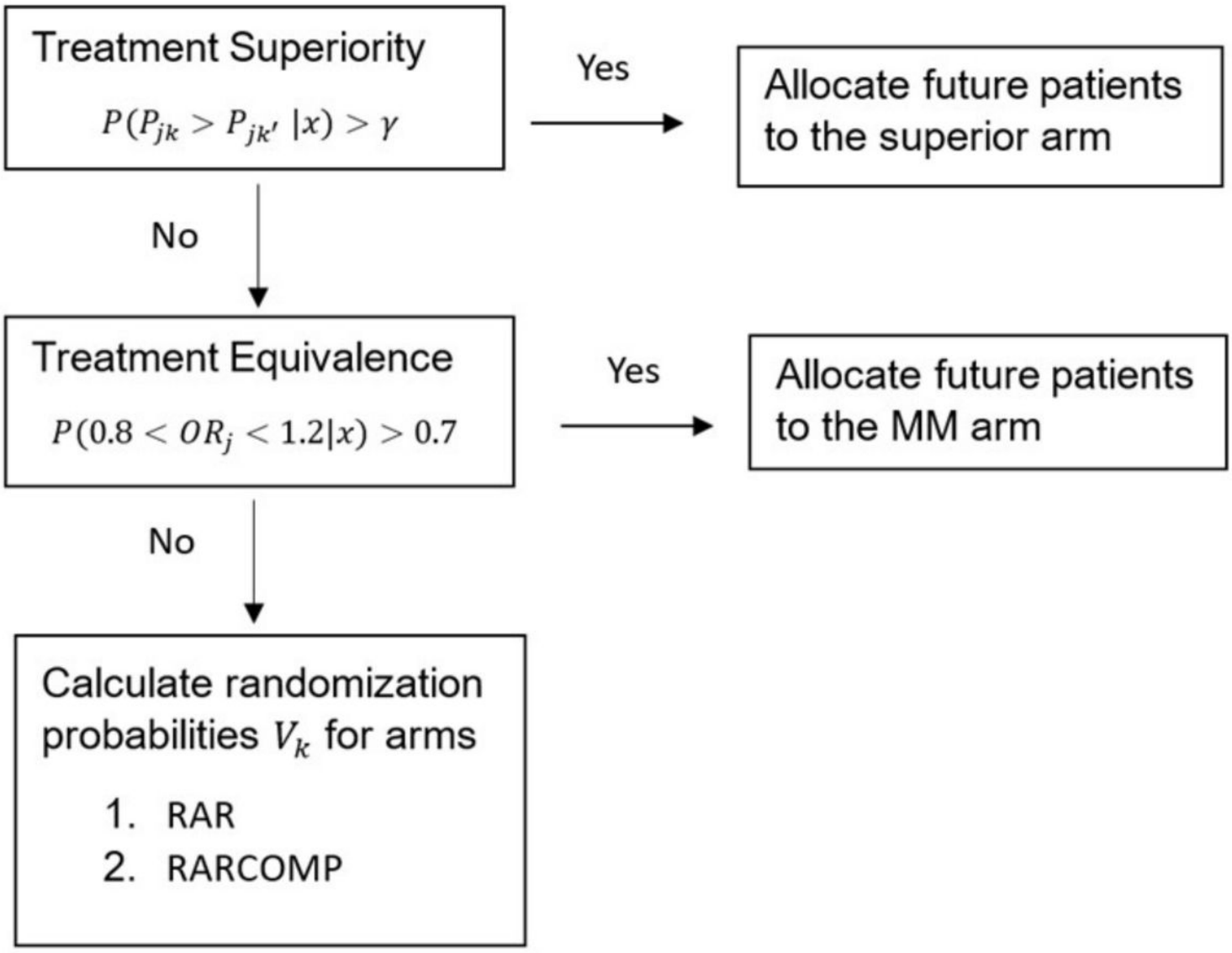

Figure 2

Adaptive Patient Allocation Flowchart.

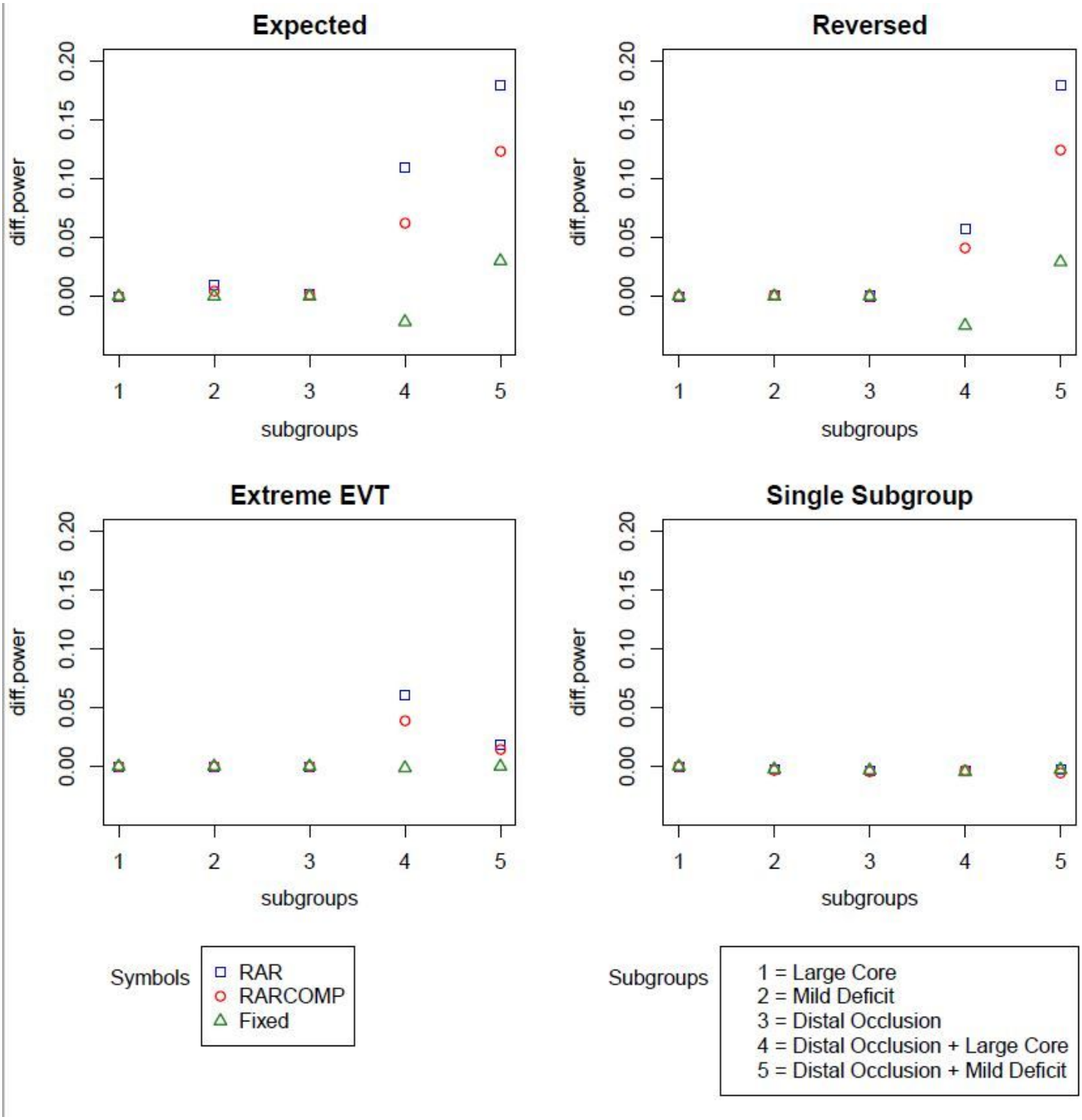

Figure 3

Power difference among three adaptive design schemes relative to the fixed design when m=1.

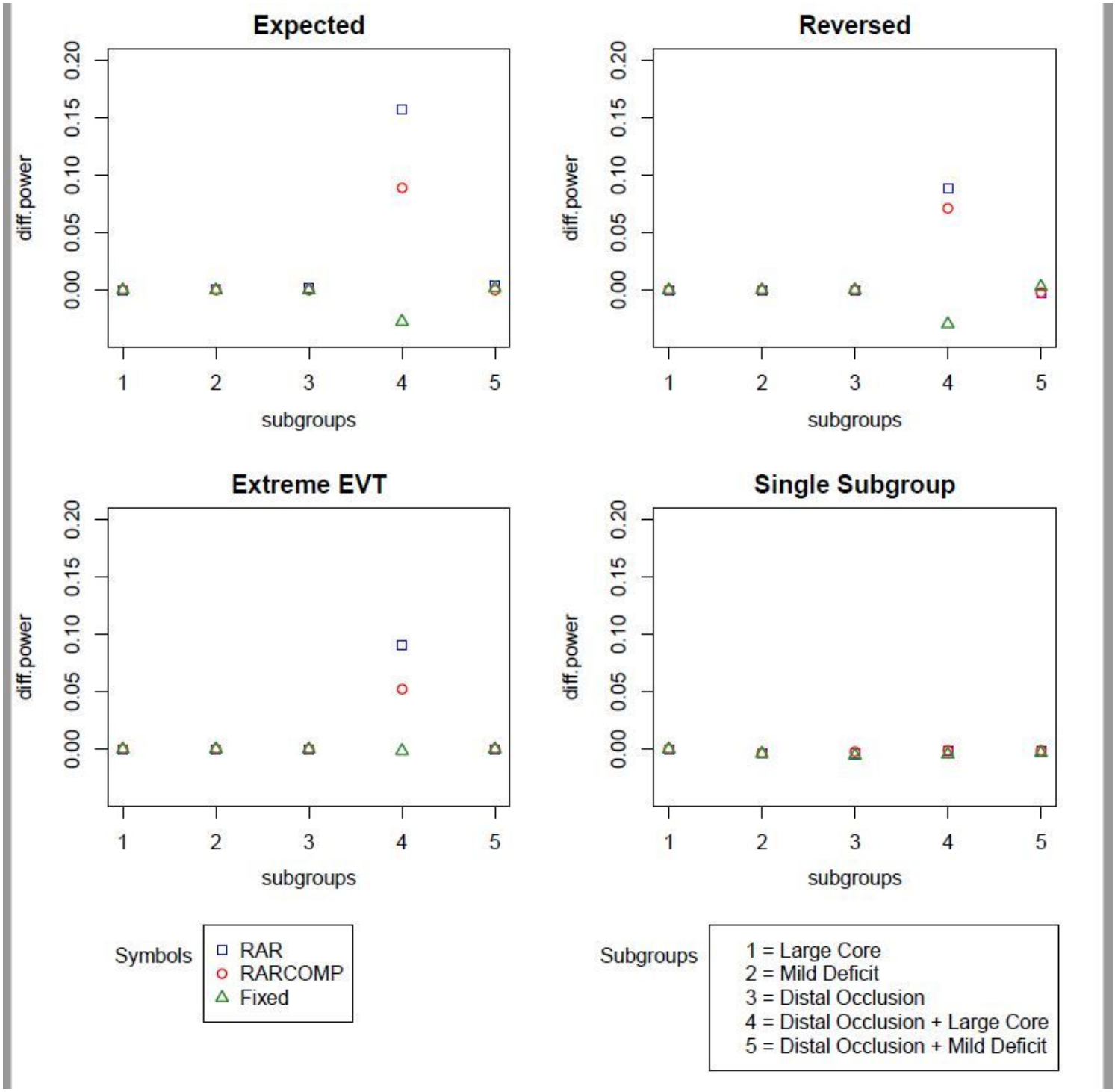

Figure 4

Power difference among three adaptive design schemes relative to the fixed design when m=30.

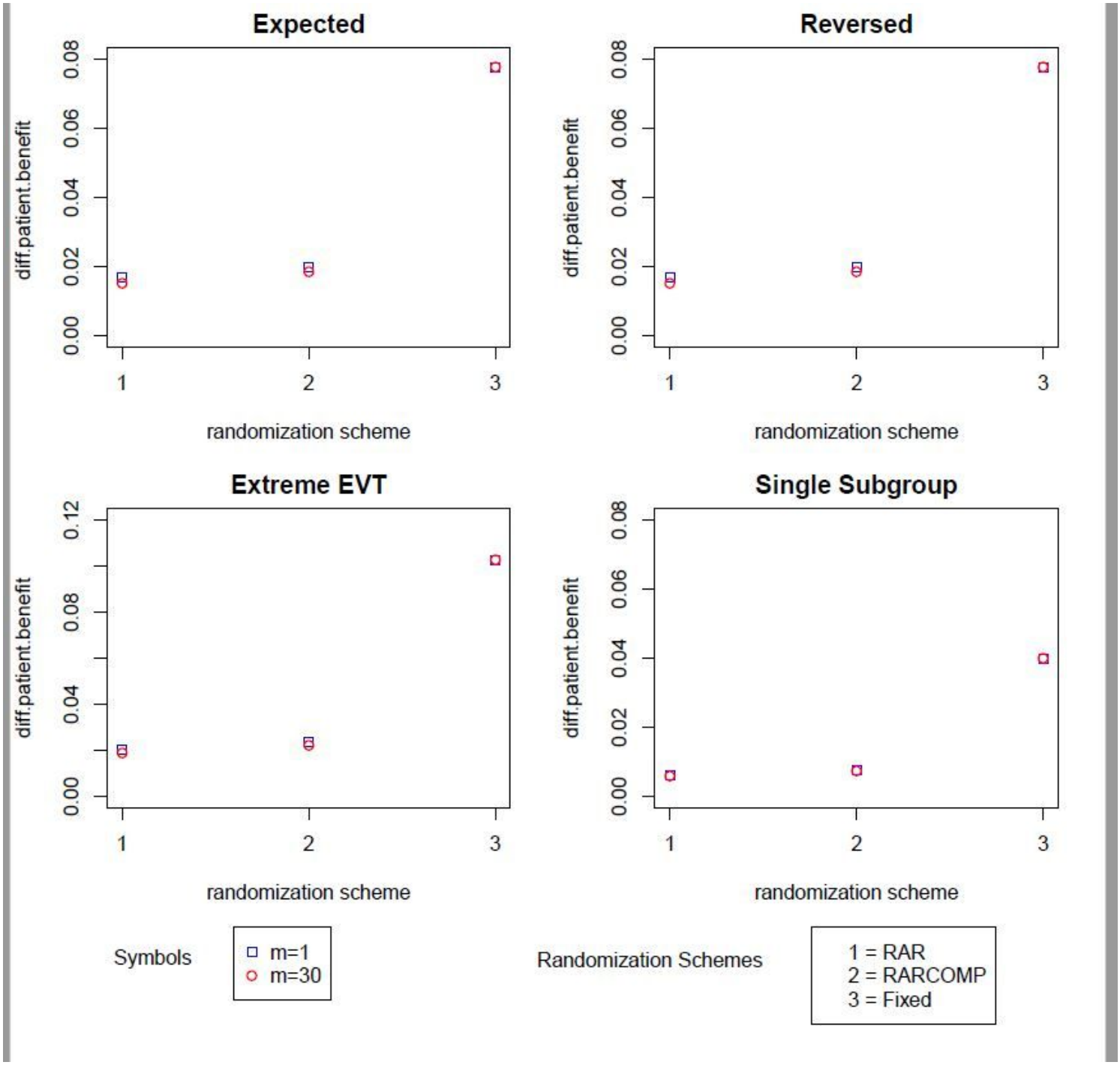

Figure 5

Patient benefit comparison for three randomization schemes.

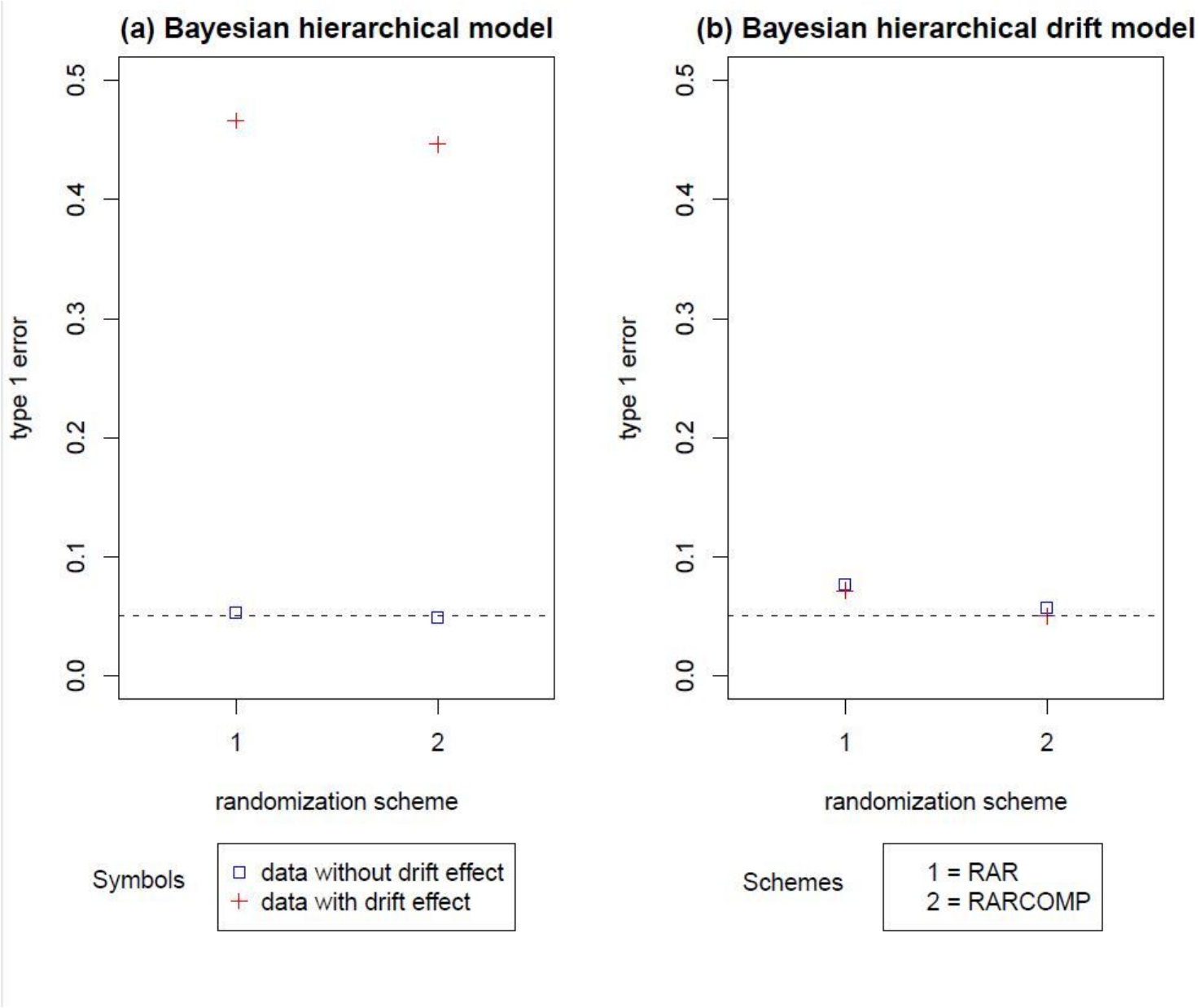

Figure 6

Type 1 error comparison for Bayesian hierarchical model and Bayesian hierarchical drift model.

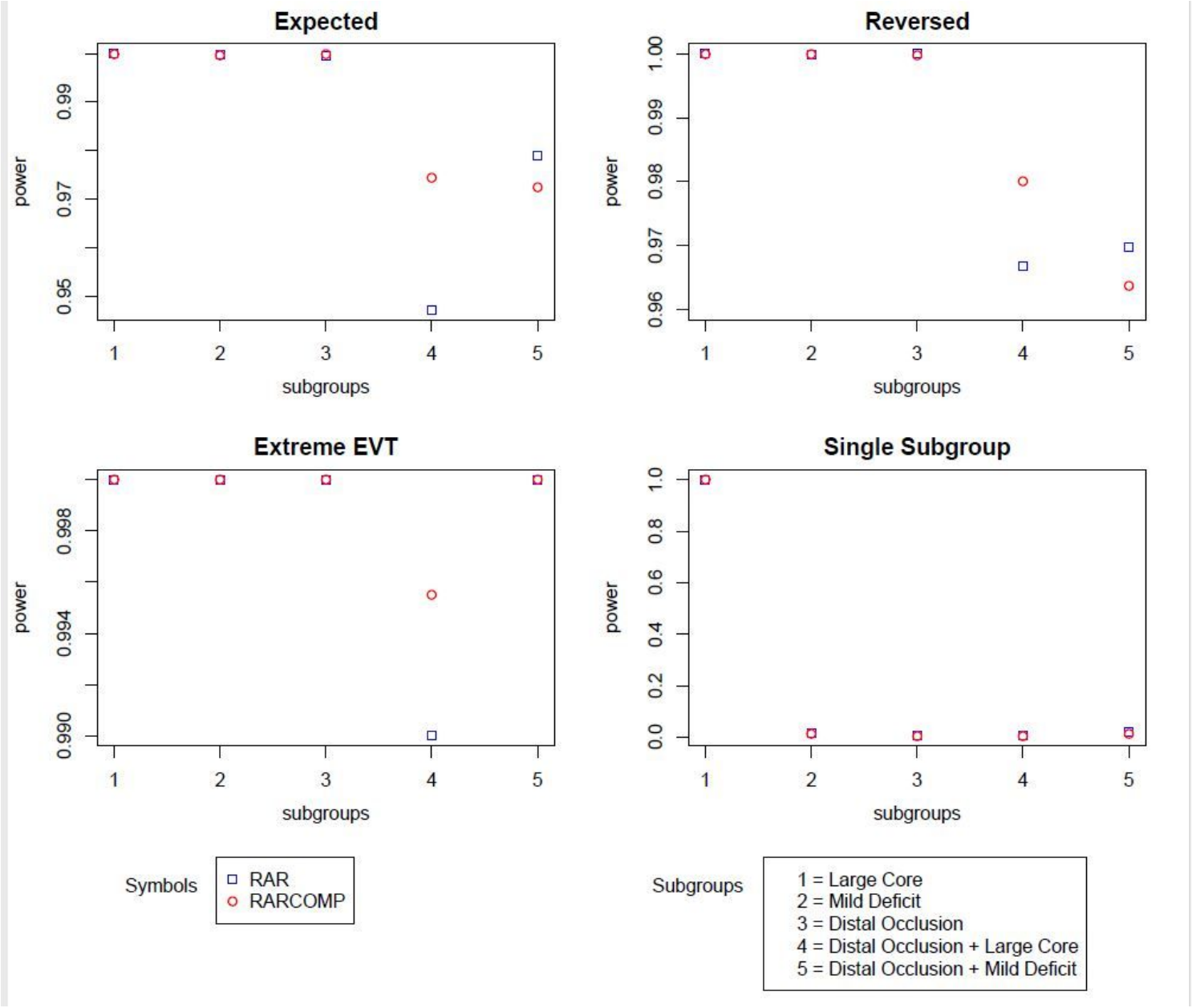

Figure 7

Power plots: fit Bayesian hierarchical drift model to linear time effect drift data.

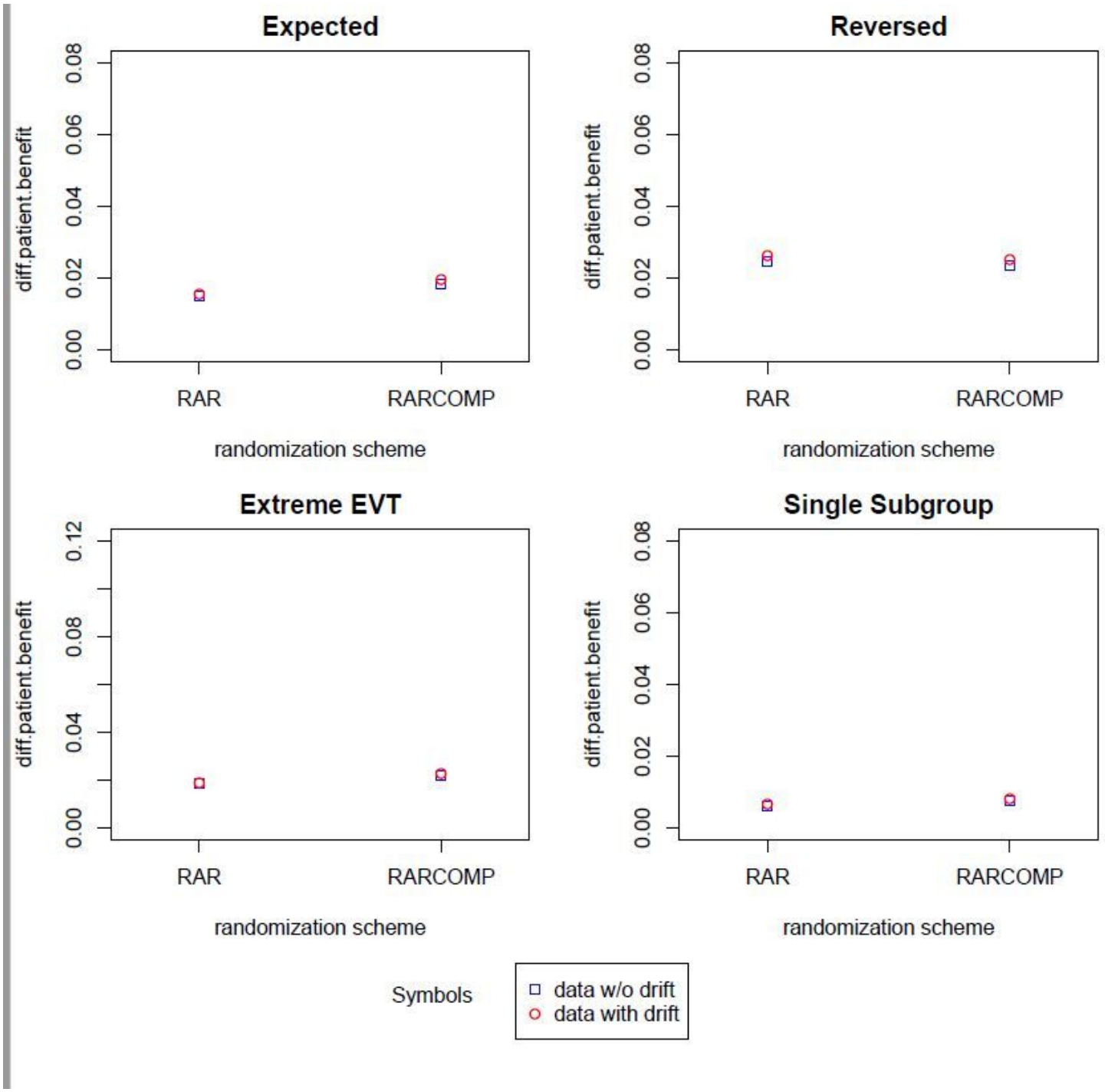

Figure 8

Patient benefit when fitting Bayesian drift model.

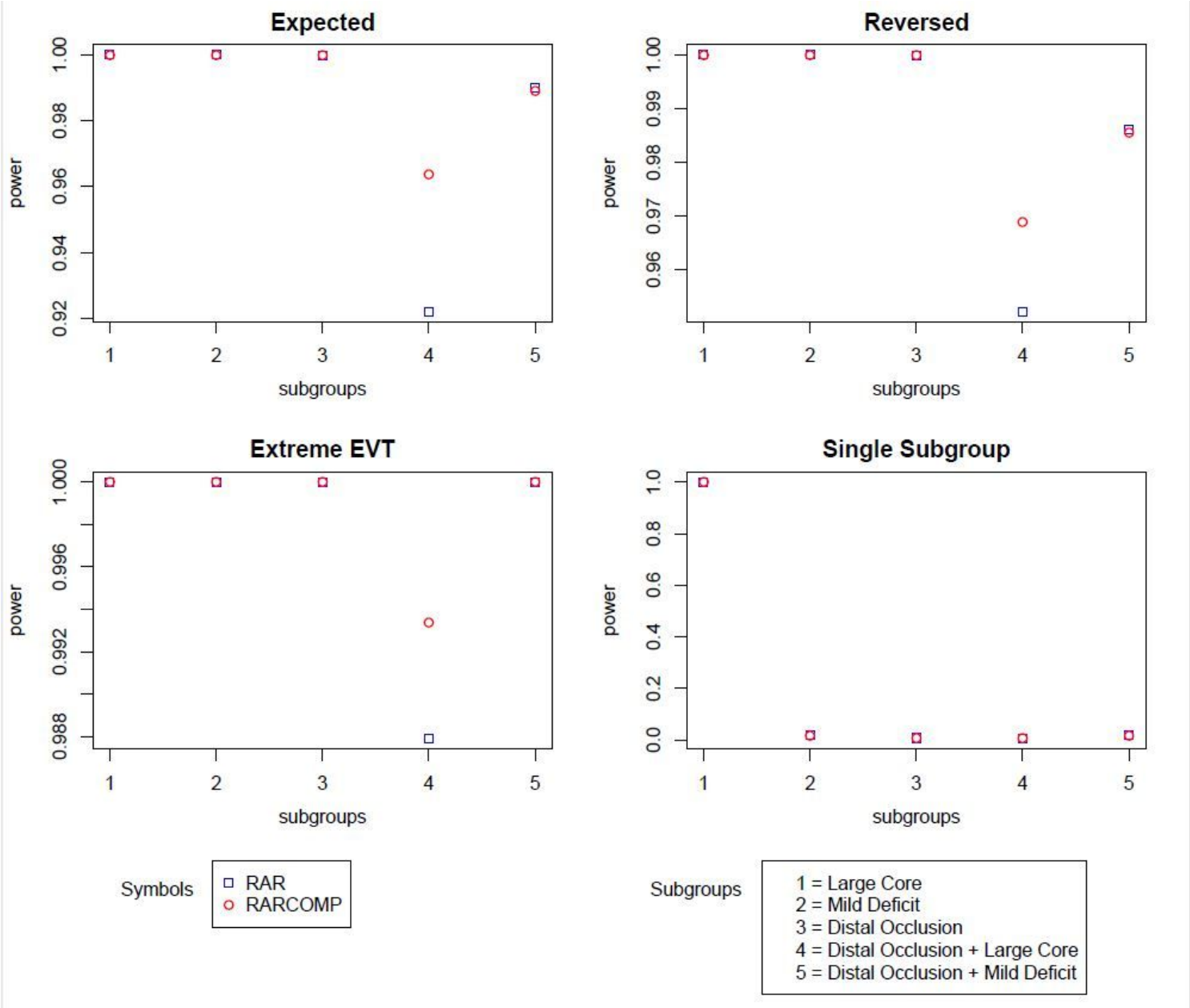

Figure 9

Power plots: fit Bayesian hierarchical drift model to data without a linear effect.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Appendix.docx