Response adaptive designs for binary responses: how to offer patient benefit while being robust to time trends?

Sofía S. Villar[1,*], **Jack Bowden**[2], and **James Wason**[1]

[1]MRC Biostatistics Unit, Cambridge Institute of Public Health, Forvie Site, Robinson Way, Cambridge Biomedical Campus, Cambridge CB2 0SR United Kingdom

[2]MRC Integrative Epidemiology Unit, University of Bristol Oakfield House Bristol BS8 2BN

Abstract

Response-adaptive randomisation (RAR) can considerably improve the chances of a successful treatment outcome for patients in a clinical trial by skewing the allocation probability towards better performing treatments as data accumulates. There is considerable interest in using RAR designs in drug development for rare diseases, where traditional designs are not either feasible or ethically questionable. In this paper we discuss and address a major criticism levelled at RAR: namely, type I error inflation due to an unknown time trend over the course of the trial. The most common cause of this phenomenon is changes in the characteristics of recruited patients - referred to as *patient drift*. This is a realistic concern for clinical trials in rare diseases due to their lengthily accrual rate. We compute the type I error inflation as a function of the time trend magnitude in order to determine in which contexts the problem is most exacerbated. We then assess the ability of different correction methods to preserve type I error in these contexts and their performance in terms of other operating characteristics, including patient benefit and power. We make recommendations as to which correction methods are most suitable in the rare disease context for several RAR rules, differentiating between the two-armed and the multi-armed case. We further propose a RAR design for multi-armed clinical trials, which is computationally efficient and robust to several time trends considered.

Keywords

Response-adaptive randomisation; Type I error; randomisation test; Power; clinical trials

1 Introduction

Randomised controlled trials (RCTs) are considered the gold standard approach to learn about the relative efficacy of competing treatment options for evidence based patient care. The information provided by a RCT can subsequently be used to better treat future populations. Traditionally, patients are allocated with a fixed and equal probability to either an experimental treatment or standard therapy arm. We will refer to RCTs implemented in this way as incorporating complete randomization (CR). However, there is generally a

---
*Corresponding author: sofia.villar@mrc-bsu.cam.ac.uk, Phone: +44 (0)1223 330385, Fax: +44 (0)1223 330365.

conflict between the individual benefit of patients in the trial and the collective benefit of future patients. CR, by definition, does not provide the flexibility to alter the allocation ratios to each arm, even if information emerges that breaks the initial trial equipoise. This conflict becomes more acute in the context of rare life-threatening conditions, because the trial participants generally make up a sizeable proportion of the total patient population.

Response adaptive randomisation (RAR) offers a way of simultaneously learning about treatment efficacy while also benefiting patients inside the trial. It achieves this by skewing allocation to a better performing treatment, if it exists, as data is accrued. When RAR rules are used in a multi-armed trial, they also increase the probability of finding a successful treatment and speed up the process of doing so (Meurer et al, 2012, Wason and Trippa, 2014).

However, RAR is still infrequently used in practice. One of the most prominent recent arguments against its use is the concern that the false positive error rate (or Type I error rate) may not be controlled at the nominal level (Thall et al, 2015). This can easily occur if the distribution of patient outcomes changes over time, and the traditional methods of analysis are used (Simon and Simon, 2011). One such example is when the underlying prognosis of patients recruited in the early stages of a trial differs from those recruited in the latter stages. This is often referred to as 'patient drift'. (Karrison et al, 2003) investigate the type I error inflation induced by various RAR rules implemented within a two-armed group sequential design with a binary outcome in which, depending on the observed value of the corresponding z-statistics, the next group of patients is allocated in one of four possible fixed ratios $R(z)$. They show that if all success rates increase by 0.12 over the course of a study with three interim analysis, the type I error rate achieved by a group sequential design is 'unacceptably high', with the inflation being worst for the most aggressive RAR rules

Time trends are more likely to occur in studies that have a long duration. Consider for example, the Lung Cancer Elimination (BATTLE)-1 phase II trial which recruited patients for 3 years (2006-2009). It was found that more smokers and patients who had previously received the control treatment enrolled in the latter part of the study compared to the beginning of the study (Liu and Lee, 2015). Trials that last more than 3 years will often be required for rare diseases because of the recruitment challenge. It is also exactly for this case where the use of RAR can be most desirable as the trial patients represent a higher proportion of the total patient population and the suboptimality gap of traditional RCTs -in terms of overall expected patient benefit- increases as the prevalence of the disease decreases (Cheng and Berry, 2007).

There has been little work in the literature considering the impact of time trends on different RAR rules. In Coad (1991a) sequential tests for some RAR rules that allow for time trends are constructed while estimation and the issue of bias within this context are addressed in Coad (1991b). In Rosenberger, Vidyashankar and Agarwal (2001), a covariate-adjusted response adaptive mechanism for a two-armed trial that can take a specific time trend as a covariate is introduced. In the trial context investigated by Karrison et al (2003) an analysis stratified by trial stage eliminates the type I error inflation induced by a simple upward trend of all the success rates. A similar stratified analysis is used in Coad (1992). More recently,

Simon and Simon (2011) considered broad RAR rules for the two-armed case and proposed a randomisation test to correct for type I error inflation caused by unknown time trends of any type. There have been several recent papers comparing different classes of RAR rules under various perspectives (see e.g. Biswas and Bhattacharya, 2016, Gu and Lee, 2010, Flournoy et al, 2013, Ivanova and Rosenberger, 2000) yet most of these do not examine the effects of time trends as done in Coad (1992). An exception to this is Thall et al (2015) in which type I error inflation under time trends is pointed out as an important criticism of RAR. However, their paper only investigates a special class of RAR (based on regular updates of posterior probabilities). In this paper we identify and address a number of unanswered questions which we describe below, including the study of the multi-armed case.

If one is considering designing a clinical trial using a particular RAR rule then a fundamental question to consider is how large the temporal change in the trial data has to be to materially affect the results. In section 2 we address this question for a representative selection of RAR procedures.

If the possibility of a large drift occurring during the trial is a concern and a RAR scheme is being considered for designing such a trial, then subsequent and related questions are: Do any 'robust' hypothesis testing procedures exist that naturally preserve type I error in the presence of an unknown time trend? Should these procedures be different for two-armed and for multi-armed trials? Should they differ depending on the RAR rule in use? And, finally, what is their effect on statistical power? section 2 and section 3 address these questions for different RAR procedures. In section 4 we consider whether time trends can be effectively detected and adjusted for in the analysis, and how extended modelling approaches for modelling a time trend compare to model-free approaches in order to control for type I error. In section 5 some conclusions and recommendations for addressing this specific concern are given.

2   RAR rules, time trends and type I error rates

In this section we assess the impact of different time trend assumptions on the type I error rate of distinct RAR procedures. We assume that patients are enrolled in the trial sequentially, in groups of equal size b over J stages. We do not consider monitoring the trial for early stopping and therefore the trial size is fixed and equal to $T = b \times J$. We have omitted it the possibility of early stopping in this paper to isolate the effects of an unaccounted for time trend in a trial design using RAR. Patients are initially allocated with an equal probability to each treatment arm. After the first interim analysis allocation probabilities are updated based on data and according to different RAR rules. In a real trial, this initial CR start-up phase could be replaced by a restricted randomisation phase (e.g. a permuted block design) to minimise sample imbalances and improve the subsequent probabilities updates (Haines and Hassan, 2015). For simplicity of presentation we consider a binary outcome variable $Y_{i,j,k}$ for patient i allocated to treatment k at stage j, (with $Y_{i,j,k} = 1$ representing a success and $Y_{i,j,k} = 0$ a failure) that is observed relatively quickly after the allocation. An example might be whether a surgery is considered to have been successful or not.

We consider a trial with K 1 experimental arms and a control arm and assume that every patient in the trial can only receive one treatment. We will also assume that for every $j < J$ before making the treatment decisions for the $(j + 1)^{th}$ block of patients the outcome information of the jth block of patients is fully available. Patients in block j are randomised to treatment k with probability $\pi_{j,k}$ (for $j = 1, \ldots, J$ and $k = 0, 1, \ldots, K$). For example, a traditional CR design will have $\pi_{j,k} = 1/(K + 1) \ \forall_{j,k}$. Patient treatment allocations are recorded by binary variables $a_{i,j,k}$ that take the value 1 when patient i in block j is allocated to treatment k and 0 otherwise. Because we assume that patients can only receive one treatment, we impose that $\sum_{k=0}^{K} a_{i,j,k} = 1$ for all i, j. We denote the control treatment by $k = 0$. Updating the allocation probabilities after blocks of patients rather than after every patient makes the application of RAR rules more practical in real trials (Rosenberger and Lachin, 1993).

An appropriate test statistic is used to test the hypotheses that the outcome probability in each experimental treatment is equal to that of the control. That is, if we let $Pr(Y_{i,j,k} = 1/a_{i,j,k} = 1) = p_k$, then we consider the global null to be $H_{0,k} : p_0 = p_k$ for $k = 1, \ldots, K$. Generally, any sensible test statistic will produce reliable inferences if the outcome probability in each arm conditional on treatment remains constant over the course of the trial. If this is not the case, then the analysis may be subject to bias. To illustrate this we shall assume the following model for the outcome variable Y

$$\text{Logit} \left[Pr(Y_{i,j,k} = 1 | Z_{i,j}, a_{i,j,k} = 1) \right] = \begin{cases} \beta_0 + \beta_t t_j + \beta_z Z_{i,j} & k = 0 \\ \beta_0 + \beta_t t_j + \beta_z Z_{i,j} + \beta_k & k \geq 1 \end{cases} \quad (1)$$

where $t_j = (j - 1)$, $Z_{i,j}$ is a patient-level covariate (e.g. a binary indicator variable representing whether a patient characteristic is present or absent) and therefore β_t is a time trend effect, β_z is the patient covariate effect and β_k is treatment's k main effect. We shall assume that $Z_{i,j} \sim Bern(q_j)$ and define $\text{Expit}(u) = \frac{exp(u)}{1 + exp(u)}$. Furthermore, we shall assume that the global null hypothesis is true, meaning $H_{0,k}$ holds for $k = 1, \ldots, K$, or equivalently $\beta_1 = \ldots = \beta_k = 0$. Patients with $Z_{i\cdot} = 1$ will have success rate when allocated to arm k equal to $\text{Expit}(\beta_0 + \beta_t t_j + \beta_z)$ while patients with a negative value $Z_{i\cdot} = 0$ will have a success rate of $\text{Expit}(\beta_0 + \beta_t t_j)$.

If the covariate variable Z is unobservable then when analysing the data, response rates will in effect be marginalised over Z as follows:

$$Pr(Y_{i,j,k} = 1 | a_{i,j,k} = 1) = \sum_{f=0}^{1} Pr(Y_{i,j,k} = 1 | Z_{i,j} = f, a_{i,j,k} = 1) Pr(Z_{i,j} = f) \quad (2)$$
$$= \text{Expit}(\beta_0 + \beta_t t_j)(1 - q_j) + \text{Expit}(\beta_0 + \beta_t t_j + \beta_z) q_j$$

Assuming that equal numbers of patients are recruited at each of J stages then the mean response rate in arm k will be

$$Pr(Y_{i,.,k} = 1 | a_{i,.,k} = 1) = \frac{1}{J} \sum_{j=1}^{J} Pr(Y_{i,j,k} = 1 \big| a_{i,j,k} = 1) \quad (3)$$

The inclusion of t_j and $Z_{i,j}$ allow us not only to introduce time trends of different magnitude but also to describe two distinct scenarios that are likely to be a concern in modern clinical trials: *changes in the standard of care* (Scenario (i)) - or changes in the effectiveness of the control treatment, and *patient drift* (Scenario (ii)) - or changes in the baseline characteristics of patients. Under model (1) we shall consider that a case of Scenario (i) occurs if $\beta_t \neq 0$ while $\beta_z = 0$ whereas an instance of Scenario (ii) happens if $\beta_z \neq 0$ while $\beta_t = 0$ and q_j evolves over *j*.

In this section we consider the global null hypothesis by setting $\beta_k = 0$ for all *k* for both scenarios. In subsection 3.3 and section 4 we consider extensions of these scenarios where $\beta_k > 0$ for some $k \geq 1$. Specifically, we consider alternative hypotheses of the form $H_{1,k} : p_k - p_0 = \Delta_p > 0$ for some $k \geq 1$ with the treatment effect Δ_p defined as $\Delta_p = Pr(Y_{i,.,k} = 1 | a_{i,.,k} = 1) - Pr(Y_{i,.,k} = 1 | a_{i,.,0} = 1)$.

2.1 RAR procedures considered

Many variants of RAR have been proposed in the literature. However, different RAR procedures often perform similarly, because they obey the same fundamental principle. *Myopic* procedures determine the 'best' allocation probabilities for the next patient (or block of patients) according to some criteria based on the accumulated data (on both responses and allocations) up to the last treated patient. *Non-myopic* procedures consider not only current data but also all possible future allocations and responses to determine the allocation probability of every patient (or block of patients) in the trial (see chapter 1 in Hu and Rosenberger, 2006). Furthermore, RAR procedures can be considered to be *patient benefit-oriented* if they are defined with the goal of maximising the exposure to a best arm (when it exists). Additionally, RAR procedures can also be defined with the goal of attaining a certain level of statistical power to detect a relevant treatment effect, thus being *power-oriented*. RAR rules that score highly in terms of patient benefit generally have lower power.

Thus, in order to illustrate these four types of rules we focus on the following RAR rules: 'Thompson Sampling' *(Myopic-Patient benfit oriented)*, 'Minimise failures given power' *(Myopic-Power oriented)*, the 'Forward Looking Gittins Index rule' *(Non-Myopic-patient benefit oriented)*, and its controlled version, the 'Controlled Forward Looking Gittins Index rule' *(Non-Myopic-Power oriented)*. A short summary of these approaches is now given, for a more detailed description see (Villar, et al., 2015).

(a) *'Thompson Sampling'* (TS): (Thompson, 1933) was the first to recommend allocating patients to treatment arms based on their posterior probability of having the largest response rate.

We shall compute the TS allocation probabilities using a simple Monte-Carlo approximation. Moreover, we shall introduce a tuning parameter *c* defined as

$\frac{(j-1) \times b}{2T}$ where $(j-1) \times b$ and T are the current and maximum sample size respectively. This parameter tunes the *aggressiveness* of TS allocation rule based in the accumulated data so that the allocation probabilities become more skewed towards the current best arm only as more and more data accumulates. Notice that TS is essentially the only class of RAR considered in Thall et al (2015).

(b) '*Minimise failures given power*' (RSIHR): (Rosenberger et al, 2001) proposed and studied an optimal allocation ratio for two-armed trials in which the allocation probability to the experimental arm is defined as:

$$\pi_{j,1} = \frac{\sqrt{p_1}}{\sqrt{p_0} + \sqrt{p_1}}, \quad (4)$$

This allocation procedure is optimal in the sense that it minimises the expected number of failures for a fixed variance of the estimator under the alternative hypothesis that there is a positive treatment effect $p = p_1 - p_0 > 0$. The optimal allocation ratios that extend equation (4) for the general case in which $K > 1$ do not admit a closed form, however numerical solutions can be implemented as in (Tymofyeyev et al, 2007).

In practice the allocation probabilities may be computed by plugging in a suitable estimate for the p_k's using the data up to stage $j-1$. In our simulations we implemented the optimal allocation ratio for RSIHR using the doubly-adaptive biased coin design. Specifically, we used Hu and Zhang's randomization procedure with allocation probability function given by equation (2) in Tymofyeyev et al (2007). and $\gamma = 2$. Notice that we estimated the success rate parameters from the mean of its prior distribution for the first block of patients (when no data were available) and the posterior mean thereafter.

(c) '*Forward Looking Gittins Index* rule' (FLGI): in (Villar, et al., 2015), we introduced a block randomised implementation of the optimal deterministic solution to the *classic* multi-armed bandit problem, first derived in (Gittins, 1979, Gittins and Jones, 1974). The FLGI probabilities are designed to mimic what a rule based on the Gittins Index (GI) would do. See Section 3 and Figure 1 in (Villar, et al., 2015) for a more detailed explanation of how these probabilities are defined and approximately computed via Monte-Carlo. The near optimality attained by this rule differs from the one targeted in procedure (b) in the sense that average patient outcome is nearly maximised with no constraint on the power levels that should be attained. Notice that before the introduction of this procedure based on the Gittins index an important limitation to the practical implementation of non-myopic RAR rules such as those in Cheng and Berry (2007), Williamson, et al. (2017) was computational, particularly in a multi-armed scenario.

(d) '*Controlled FLGI*' (CFLGI): In addition to the rule described in (c), for the multi-armed case (i.e. $K > 1$) we consider a group allocation rule which,

similarly to the procedure proposed in (Trippa et al., 2012), protects the allocation to the control treatment so it never goes below $1/(K+1)$ (i.e. its fixed equal allocation probability) during the trial.

2.2   Simulation results

In this section we present the results of various simulation studies which show, for instances of scenarios (i) and (ii) (described in detail below), the degree to which the type I error rate can be inflated for different RAR rules relative to a CR design. As is the usual case when comparing RAR procedures we consider measures of efficiency (or variability) and ethical performance, assessing which ones of them (if any) provide a better compromise between these two goals (Flournoy et al, 2013). We therefore also compute the expected number of patients assigned to the best treatment (p^*) and expected number of patient successes (ENS). However, under the global null considered in this section ENS and p^* are identical for all designs and therefore we do not report them here. In Sections where we consider scenarios under various alternative hypotheses we report patient benefit measures as well as power. Specifically, we report p^* and the increment in the expected patient benefit that the RAR rule considered attains over a CR design, i.e.   ENS = $\text{ENS}_{RAR} - \text{ENS}_{CR}$.

For each scenario a total of 5000 trials were simulated under the global null and the same global null was tested. We used z-statistics for testing with RAR rules (a), (b) and (d) (when asymptotic normality can be assumed) and, given that bandit-based procedures can result in very small sample sizes for some arms, an adjusted Fisher's exact test for procedure (c). The adjustment for the bandit rules chooses the cutoff value to achieve a 5% type-I error rate (as in Villar et al (2015)). For multi-armed trials, we use the Bonferroni correction method to account for multiple testing and therefore ensure that the family-wise error rate is less than or equal to 5%. In all simulations and for all RAR rules we assumed uniform priors on all arms' success rates before treating the first block of patients.

2.2.1   Scenario (i): Changes in the standard of care—The first case we consider is that of a linear upward trend in the outcome probability of the control arm. This could be the case of a novel surgery technique that has recently become the standard of care but it requires a prolonged initial training period for the majority of surgeons to become proficient in these complex procedures until "failure" is eliminated or reduced to a minimum constant rate. In terms of the model described in equation (1) this corresponds to varying β_t with all else fixed.

Specifically, we let β_t take a value such that the overall time trend within the trial

$$D = Pr(Y_{i,j,.} = 1 \mid t_j = J - 1) - Pr(Y_{i,j,.} = 1 \mid t_j = 0),$$

varies in $D = \{0, 0.01, 0.02, 0.04, 0.08, 0.16, 0.24\}$. Figure 1 (left) shows the corresponding evolution of the per block success rate of every arm over time across the scenario in which $J = 5$ and for the cases of: no drift ($D = 0$, dark blue) and the strongest drift considered ($D = 0.24$, dark red).

Figure 2 summarises the simulation results. The top row of plots show the results for the two-armed trials (i.e. $K = 1$) and the bottom row plots show the results for $K = 2$. In both cases the trial size was $T = 100$. The value of the sample size T might be interpreted as the maximum possible sample size (i.e. including a very large proportion of the patient population) in the context of a rare disease setting. The plots in the left column assume a block size of 10, and the plots in the right column assume a block size of 20. The initial success rate was assumed to be equal to 0.3 (which corresponds to $\beta_0 \approx -0.8473$) for all the arms considered.

Under the assumption of no time trend (i.e. $\beta_t = D = 0$) the test statistics used preserve the type I error rate for all designs in all cases considered, except for Thompson Sampling for which the false positive rate is somewhat inflated (as pointed out in Thall et al (2015)). For CR the type I error rate is preserved even when a time trend is present and regardless of the block size, number of arms and the trend's magnitude.

The error rates for some of the RAR rules (FLGI and TS) are substantial when overall time trends are of 0.08 and more. This is because these rules are *patient benefit oriented*, i.e. they skew allocation towards an arm based on data more considerably and/or earlier on in the trial. On the other hand, the RSIHR procedure, being a *power-oriented* rule, remains practically unaffected by temporal trends in terms of type I error inflation. This very important difference in performance amongst RAR procedures has not been noted previously.

Multi-arm allocation rules that protect allocation to the control arm, like the CFLGI, are also unaffected by type I error inflation, even for large drifts. Generally, the type I error inflation suffered by the other RAR rules seems to be slightly larger for the three-armed case than for the two-armed case.

2.2.2 Scenario (ii): Patient drift—For this case we imagine a simplistic instance in which patients are classified into two groups according to their prognosis. This occurs if, for example, $Z_{i,j}$ in model (1) represents the presence or absence of a biomarker in patient i at stage j, where $Z_{i,j} = 1$ denotes a biomarker positive patient and $Z_{i,j} = 0$ denotes biomarker negative patient. Alternatively, $Z_{i,j}$ can capture any other patient feature. It could, for example, represent if a patient is a smoker and previously received the control arm, which would be a relevant covariate in the BATTLE-1 trial. Moreover, we let the recruitment rates of these two types of patients, i.e. q_j, vary as the trial progresses to induce the desired drift in the mix of patients over time. We will model this situation by letting $\beta_z > 0$ in (1) whilst holding all else fixed. We start by assuming that Z is unobserved. In subsection 4.1 we explore the case where Z is measured and can be adjusted for.

The middle and right-hand side plots in Figure 1 show the evolution of $Pr(Y_{i,.,k} = 1|a_{i,.,k} = 1)$ under differing patterns of *patient drift* over the course of the trial. The middle plot describes the case in which there is a linear trend in the average success rates of all arms created by the *patient drift* whereas the plot to the right considers the case of a more complex temporal evolution with the average success rates going up and then down or vice-versa. In both cases a trial of size $T = 200$ with $J = 10$ and therefore $b = 20$ was considered. The success rates for

all arms for the biomarker negative patients was $Pr[\,Y_{i,j.} = 1|Z_{i,j} = 0\,] = 0.3$ (so that $\beta_0 \approx$ −0.8473). For the biomarker positive group $Pr[\,Y_{i,j.} = 1|Z_{i,j} = 1\,] = 0.6$ (such that $\beta_z \approx$ 1.2528).

Figure 3 summarises the simulation results for the cases depicted in Figure 1 (middle and right). The results show that the RAR procedures most affected by type I error inflation are those that are *patient-benefit* oriented (FLGI and TS). Type I error inflation is high only for the moderately large recruitment rate evolution assumed. As before, the *power-oriented* rules (RSIHR and CFLGI) have type I error rates that do not significantly differ from those obtained by a CR design. This further supports the argument that not all RAR procedures are equally affected by the presence of the same temporal trend.

3   Testing procedures and RAR designs robust to patient drift

In this section we describe a hypothesis testing procedure for RAR rules in a two-armed trial context and a RAR design for multi-armed trials that preserves type I error rates in the presence of an unknown time trend.

3.1   Two-armed trials: randomization test and the FLGI

The type I error inflation shown in scenarios (i) and (ii) for some of the RAR rules is caused by the fact that the test statistics used assume every possible sequence of treatment allocations (i.e., every possible trial realisation) is equally likely. For instance, this is the case for the adjusted Fisher's exact test used for the FLGI in the previous sections. This assumption is not true in general as certain allocation sequences will be highly unlikely or even impossible for some RAR procedures. This is particularly well illustrated in the case of the FLGI rule where it is possible for one of the arms to be effectively 'selected' within the trial, since the probability of assigning a patient to any other arm from that point onwards is zero.

In this section we show the results of developing and computing a test statistic, introduced in (Simon and Simon, 2011), based on the distribution of the assignments induced by the FLGI under the null hypothesis. In their paper, the authors show that using a cut-off value from the distribution of the test statistic generated by the RAR rule under the null hypothesis, and conditional on the vector of observed outcomes, ensures the control of the type I error rate (see Theorem 1 in Simon and Simon (2011)). Their result applies to any RAR rule and any time trend in a two-armed trial, most importantly, its implementation does not require any knowledge or explicit modelling of the trend. In this paper we have chosen to implement it for the FLGI rule as this is the most recently proposed RAR procedure of the ones considered. Notice that Type I error rate preservation under time trends by means of randomisation based inference for restricted randomised procedures is established in (Rosenberger and Lachin, 2016, Section 6.10).

However, computation of the null distribution can be challenging under realistic trial scenarios as it requires the complete enumeration of all trial histories and it is infeasible for response adaptive rules that are deterministic as e.g. the GI rule is. Therefore, there is a need to find ways of computing such a randomization test efficiently for the sake of its practical

implementation as well as evaluating its effect on power, which might differ across different rules.

We implement a randomisation test for the FLGI rule that is based on a Monte-Carlo approximation of the exact randomisation test. More precisely, our approach does the following: for a given trial history $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_j)$, where $\mathbf{y}_j$ a vector of the b observed outcomes at stage j, we simulate M trials under the FLGI allocation rule. The FLGI allocation ratios are updated after each block using the allocation variables $a_{i,j,k}$ randomly generated under the FLGI rule by Monte-Carlo and the observed outcome data up to that point (i.e. $(\mathbf{y}_1, \ldots, \mathbf{y}_j)$). For each of these M simulated trials we compute the value of the test statistic to assemble an empirical distribution of the test statistic under the null. We can then compare the test statistic observed in the original trial to the empirical distribution, rejecting the null hypothesis at level α if it is more extreme than its α percentile for a one-sided test (or than its $\alpha/2$ or $1 - \alpha/2$ percentile for a two-sided test). Finally, we repeat this procedure for another Nr trial history replicates and report the average type I error rate achieved as well as the averages of the other ethical performance measures considered.

Table 1 shows the results from $Nr = 5000$ replicates using the approximate randomisation test for the case of scenario (i) displayed in Figure 2 (top-right). For each trial replicate the approximate randomisation-based test was computed using $M = 500$ simulated trials -or *resamplings*- to construct the empirical distribution function. The values of Nr and M are the same as those used by the simulations in Simon and Simon (2011). From Table 1, we see that the type I error rate is preserved at its 5% level even when the patient drift is severe. We also report p^* and ENS, as defined in Section 2.

3.2 Multi-armed trials: protecting allocation to control

As shown in the simulation results reported in section 2 the RAR rules that include a protection of the allocation to the control treatment (specifically, the CFLGI) preserve the type I error rate. Matching the number of patients allocated to control to that allocated to the best performing arm has also been found to produce designs that result in power levels higher than that of a CR design (See e.g. Trippa et al., 2012, Villar, et al., 2015). In the next section, we report on results that indicate that this *matching* feature not only preserves the type I error rates but also ensures high power levels when using the standard analysis methods as an approximation inference method and under the presence of time trends.

Therefore, if the design of the multi-arm trial incorporates protection of the control allocation there appears to be no need to implement a testing procedure that is specifically designed to be robust to type I error inflation. This is another important and novel finding.

3.3 Protecting against time trends and its effect on Power

Preserving the type I error rate is an important requirement for a clinical trial design. However, the learning goal of a trial also requires that, if a best experimental treatment exists, then the design should also have a high power to detect it. In this section, we therefore assess the power of the approximate randomisation test (for the FLGI) and the standard test (for the CFLGI).

We first explore an extension of an instance of scenario (i) in which we assume there is a treatment effect of 0.4 (where $p_0 = 0.3$ and $p_1 = 0.7$, therefore $\beta_1 \approx 1.6946$) which is maintained even in the cases where we also assume a positive time trend in the standard of care. The trial is of size $T = 150$ with $J = 5$ stages (so that $b = 30$). Under this design the assumed treatment effect is detected with approximately 80% power by the FLGI rule if there is no time trend and the adjusted Fisher's test is used. If a traditional CR design is used then the power attained is 99%, but the proportion of patients allocated to each arm is fixed at 1/2.

Table 2 shows the power to reject the null hypothesis as the overall time trend increases from 0 to 0.24 (i.e. for $\beta_t \in \{0, \ldots, 0.24\}$ while $\beta_z = 0$) for a treatment effect of 0.4 (i.e. for $\beta_1 \approx 1.6946$). We denote by $(1 - \beta_F)$ the power level attained by the adjusted Fisher's test and by $(1 - \beta_{RT})$ the power level when using the approximate randomisation test.

These results show that the power of the randomisation test is considerably reduced compared to that obtained using Fisher's exact test. However, the patient benefit properties of the FLGI over the CR design are preserved in all the scenarios. The improvement in patient response of the FLGI design over CR is around 30% regardless of the drift assumption.

Next, we consider the multi-arm case by assessing the effect on power on the RAR rules that protect allocation to the control arm. In order to do so under different trend assumptions we extend a case of scenario (i). We assume then that there is a treatment arm that has an additional benefit over the other two arms of 0.275 (where $p_0 = p_2 = 0.300$ and $p_1 = 0.575$) which is maintained even in the cases we assume a positive time trend in the success rate of the standard of care. Regardless of the trend assumption, a traditional CR design has a mean p^* value of 1/3 by design and detects a treatment effect of such a magnitude with approximately 80% power.

Figure 4 shows the ENS-power levels for the designs considered. The CR design performs as predicted in terms of power and ENS. The power of the CFLGI is unaffected except for a small increase when the trend is very high. This approach attains an improvement over CR on ENS for every trend magnitude assumption considered (the improvement in ENS goes from 15.81% to 13.44% in the case of the largest assumed trend).

Figure 4 also includes the results for the FLGI for comparison. Power levels are increased when there is a positive time trend, as assumed in this case, compared to the case of no time trend. This is due to the temporal upward trend which for the FLGI rule causes an overestimation of the treatment effect. Notice that the probability in these simulations of the allocation imbalance observed in the FLGI being in the wrong direction (i.e. towards inferior arms) only occurred in less than 4% of all replicates.

4 Adjusting the model for a time trend

In this section we illustrate the extent to which adjusting for covariates can help to reduce type I error inflation and affect power. This section also discusses the problems that can be encountered when doing this after having used a RAR procedure and how to address them.

4.1   Two-armed trials

We first study covariate-adjustment under instances of scenario (i). We consider a two-armed trial of size $T = 100$ with $J = 5$ and $b = 20$. We shall focus on the most extreme case considered in Figure 1 in which the overall time trend was $D = 0.24$ (or $\beta_t \approx 0.2719$). The initial success rates of both arms were set to 0.3 (i.e. $\beta_0 \approx -0.8473$).

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. However, we note that there is a strong deflation in the type I error rate of the FLGI design. This occurs because the testing procedure used in this section does not include an adjustment similar to the one used with Fisher's exact test in section 2 and section 3. When we look at the mean estimated coefficient for the time trend we note that CR only slightly underestimates it, having a 40% power to detect it as significantly different from 0. The FLGI design results in a larger underestimation of the time trend coefficient. This underestimation is consistent with that observed in Villar et al (2015) for reasons clarified in Bowden and Trippa (2015). The power to detect a significant time trend for the FLGI is more than halved compared to CR. Since its estimate is negatively correlated with that of the time trend coefficient, the baseline effect β_0 is also overestimated in both designs, but more severely for the FLGI.

Another consequence of the under-estimation of the time trend is that complete or quasi-complete separation is more likely to occur (See Albert and Anderson, 2002). This happens for the FLGI for example when all the observations of one of the arms are failures (and few in number) and this arm is therefore dropped early from the trial (i.e. its allocation probability goes to 0 and never goes above 0 again within the trial).

When this problem occurs in a trial realization, the maximum likelihood estimates are highly unstable and will not be well defined. This can be observed in the MSE value for $\hat{\beta}_0$ in Table 3 (III) for the FLGI. In order to address this, we applied Firth's penalized likelihood approach (Firth, 1993) which is a method for dealing with issues of separability, small sample sizes, and bias of the parameter estimates (using the R package "logistf"). The use of Firth's approach mitigates the bias due to the separability issue but it will not correct for the bias caused by the RAR procedure, which is addressed in Bowden and Trippa (2015) or Coad and Ivanova (2001). To the best of our knowledge, methods that simultaneously adjusts for both sources of bias do not exist. These results suggest that developing bias-correction methods specially designed for the FLGI within this context could offer improved estimation results than those obtained here. In Table 3 parts (II) and (IV) results of deploying the Firth correction are displayed. These results show an improvement in the estimation of the baseline effect when using the FLGI design: the MSE value is significantly reduced and the average estimate of β_0 is closer to its true value (though it is still overestimated). For the CR design there is also an improvement. Also, note that the type I error deflation has also been almost fully corrected by the Firth's adjustment in the FLGI design.

To assess the effect on statistical power in Table 4 we replicate the estimation procedure for the case studied in the 3rd row of Table 2 in which we let the treatment effect of arm 1 be positive (i.e. $\beta_1 \approx 1.6946$) while the overall drift assumed corresponds with $D = 0.16$ (or $\beta_t \approx 0.1840$ and $\beta_z = 0$). The initial success rate in the control arm was equal to 0.3 (i.e. $\beta_0 \approx -0.8473$). Because complete (or quasi-complete) separation affected the FLGI rule in all the scenarios considered here, Table 4 and the following tables only display the results using Firth's correction.

As expected the power of a CR design displayed in Table 4 coincides with the value reported in subsection 3.3, which using both procedures (i.e. adjusting for covariates or hypothesis testing) yields an average value of 99%. The power value of the FLGI design when fitting the GLM model is lower than the 80% value reported in subsection 3.3 for the case of no time trend (i.e. $\approx 74\%$). The difference in power levels is explained by the adjustment in Fisher's exact test done in that section which raises power by correcting for the deflation of the type I error rate of the standard Fisher's test.

Our results suggest that correctly modelling a time trend and adjusting for separation via Firth correction can safeguard the validity of trial analyses using RAR. That is, by maintaining correct type I error rates and delivering a level of statistical power similar to that obtainable when no trend is present.

4.2 Multi-armed trials

In this section we consider the case of multi-armed trials and an instance of Scenario (ii) or *patient drift*. Also, we shall remove the assumption that the patient covariate or biomarker is unobservable, and allow for the availability of this information before analysing and estimating the corresponding model in (1).

First, we study the case of scenario (ii) in which the proportion of biomarker positive patients evolves as $q_j = 0.5 + (j-1) \times 0.05$ for $j = 1, \ldots, 10$ (See Figure 1, middle). We simulated 5000 three-armed trials of size $T = 200$ with $J = 10$ and $b = 20$. The differential effect of being biomarker positive was assumed to be of 0.3, which corresponds with $\beta_z \approx 1.2528$. The initial success rates of all the arms for the biomarker negatives was equal to 0.3 (i.e. $\beta_0 \approx -0.8473$).

Table 5 displays the results of the CR, FLGI and CFLGI designs under the null hypothesis. These results suggest that all designs attain the same power to detect the biomarker effect (as the adaptation is not done using this information, all designs have similar numbers of patients with a positive and a negative biomarker status). More importantly, incorporating patient covariate data into the explicative model dramatically reduces type I error inflation for the FLGI. Although the observed rates are close to 6% and thus, above the 5% target, they are well below the levels observed without this adjustment (approx. 17%).

In Table 6 we examine the effect on power by replicating the previously described scenario but allowing for the experimental arm 1 to have an effect for all patients of 0.2 (i.e. $\beta_1 = 0.8473$). These results show how the power to detect the treatment effect in arm 1 with a FLGI design is almost halved compared to that attained by a traditional CR design. Yet, the

CFLGI improves on the power level of the CR design by approximately 15%. Also note that the type I error rate for arm 2 appears to be deflated for the FLGI and CFLGI designs.

These results suggest that fitting a model that includes a time trend after having used a RAR rule can protect against the type I error inflation caused by *patient drift* as long as the patient-covariate information is observable and available to adjust for. However, the power level attained by covariate adjustment is considerably less than that attained by a design that protects the allocation to the control arm.

Furthermore, these results fail to illustrate the learning-earning trade-off that characterises the choice between a CR and a RAR procedure and the reasons why the FLGI could be desirable to use from a patient benefit perspective (despite the power loss and the type I error inflation potential). The traditional CR design, which maximises learning about all arms, yields an average number of successfully treated patients (or ENS) of 116.83 when $p*$ remains fixed by design at 1/3. The FLGI design, on the other hand, is almost optimal from a patient benefit perspective, achieving an ENS value of 135.21, 15.73% higher than with CR, and it achieves this by skewing $p*$ to 0.7783. Finally, the CFLGI is a compromise between the two opposing goals that improves on the power levels attained by a CR design and also on its corresponding ENS value (though is below the value that could be attained with the unconstrained FLGI rule) by attaining an ENS vale of 126.32, 8.12% more than with a CR design, and a $p*$ of 0.5618.

5 Discussion

Over the past 65 years, RCTs have become the gold standard approach for evaluating treatments in human populations. Their inherent ability to protect against sources of bias is undoubtedly one of their most attractive features, and is also the reason that many are unwilling to recommend the use of RAR rules, feeling that this would be a "step in the wrong direction" (Simon, 1977). Recently, Thall et al (2015) have suggested that a severe type I error inflation could occur if RAR is used under the presence of an unaccounted for time trend. The temporal heterogeneity of the study population is a reasonable concern for trials in rare diseases as they tend to last several years. However, there remains a strong interest in the medical community to use RAR procedures in this very setting (Abrahamyan et al, 2016). The considerable patient benefit advantages offered by non-myopic RAR procedures can increase trial acceptability among both patients and physicians and enhance patient enrollment. Additionally, in the multi-armed case, the modified non-myopic procedures offer also increased statistical efficiency when compared with a traditional RCT.

We have assessed the level of type I error inflation of several RAR procedures by creating scenarios that are likely to be a concern in modern clinical trials that have a long duration. Our results suggest that the magnitude of the temporal trend necessary to seriously inflate the type I error of the *patient benefit oriented* RAR rules needs to be of an important magnitude (i.e. change larger than a 25% in its outcome probability) to be a source of concern. This supports the conclusion of (Karrison et al, 2003) in a group sequential design context. However, we also conclude that certain RAR rules are effectively immune to time trends. Specifically, those that are power oriented such as the CFLGI rule (Villar, et al.,

2015) or the '*Minimise failures given power*' rule (Rosenberger et al, 2001). This suggests that when criticising the use of RAR in real trials one must be careful not to include all RAR rules in the same class, as they have markedly different performances in the same situation.

In addition, we have recommended two different procedures that can be used in an RAR design to protect for type I error inflation. For two-armed trials, the use of a randomisation test (instead of traditional tests) preserves type I error under any type of unknown temporal trend. The use of randomisation as a basis for inference provides robust assumption-free testing procedures which depend explicitly on the randomisation procedure used, be it a RAR procedure or not. The cost of this robustness may be a computational burden and an reduction in statistical power (although most patients are still allocated to a superior arm when it exists if using the FLGI procedures). This particular feature highlights the need to develop computationally feasible testing procedures that are specifically tailored to the behaviour of a given RAR rule. For example, as pointed out by Villar et al (2015), bandit based rules such as the FLGI are extremely successful at identifying the truly best treatment but, as a direct result, often cannot subsequently declare its effect 'significant' using standard testing methods.

For multi-armed clinical trials protecting allocation to the control group (the recommended procedure) preserves the type I error while yielding a power increase with respect to a traditional CR design. However, despite rules such as the CFLGI being more robust to a time trend effect, they also offer a reduced patient benefit in the case there is a superior treatment, when compared to the *patient benefit oriented* RAR rules such as the FLGI.

Finally, we also assessed adjustment for a time trend both as an alternative protection procedure against type I error inflation and to highlight estimation problems that can be encountered when an RAR rule is implemented. Our conclusion is that adjustment can alleviate the type I error inflation of RAR rules (if the trend is correctly specified and the associated covariates are measured and available). However, for the multi-armed case this strategy attains a lower power than simply protecting the allocation to the control arm. Furthermore, the technical problem of separation also complicates estimation after the *patient benefit oriented* RAR rules have been implemented and severely impacts the power to detect a trend compared to an CR design.

Large unobserved time trends can result in an important type I error inflation, even if using restricted randomisation algorithms that seek to balance the number of patients in each treatment group (Section 6.10 Rosenberger and Lachin, 2016, Tamm et al., 2014). Therefore, an important recommendation of this work is that in the design stage of a clinical trial, when a RAR procedure is being considered, a similar detailed evaluation of type I error inflation and bias as done in this paper is performed in order to choose a suitable RAR procedure for the trial at hand.

Further research is also needed to assess the potential size of time trends through careful re-analysis of previous trial data (as (Karrison et al, 2003) do with data from (Kalish and Begg, 1987)). Our results suggest that even in the case of a large time trend being a realistic concern there are still some RAR rules that, both for the two-armed and the multi-armed

case, remain largely unaffected in all the cases we have considered. Of course, these rules offer increased patient benefit properties when compared to a traditional CR design but reduced when compared to the *patient benefit oriented* RAR rules.

Another area of future work is to explore the combination of randomisation tests with stopping rules in a group sequential context, specifically for the FLGI. Additionally, techniques for the efficient computation of approximate randomisation tests for the FLGI could be studied, similar to those explored in (Plamadeala and Rosenberger, 2012).

Acknowledgements

References

Abrahamyan, L., Feldman, BM., Tomlinson, G., et al. American Journal of Medical Genetics Part C: Seminars in Medical Genetics. Vol. 172. Wiley Online Library; 2016. Alternative designs for clinical trials in rare diseases; p. 313-331.

Albert A, Anderson JA. On the existence of maximum likelihood estimates in logistic regression models. Biometrika. 1984; 71(1):1–10.

Berry, Donald A., Eick, Stephen G. Statistics in medicine. Vol. 14. Wiley Online Library; 1995. Adaptive assignment versus balanced randomization in clinical trials: a decision analysis; p. 231-246.

Biswas, Atanu, Bhattacharya, Rahul. Response-adaptive designs for continuous treatment responses in phase III clinical trials: A review. Statistical methods in medical research. 2016; 25(1):81–100. [PubMed: 22425657]

Bowden, Jack, Trippa, Lorenzo. Statistical methods in medical research. SAGE Publications; 2015. Unbiased estimation for response adaptive clinical trials. To appear

Cheng Y, Berry DA. Optimal adaptive randomized designs for clinical trials. Biometrika. 2007; 94:673–689.

Coad, DS. Biometrika. Vol. 78. JSTOR; 1991a. Sequential tests for an unstable response variable; p. 113-121.

Coad, DS. Sequential analysis. Vol. 10. Taylor & Francis; 1991b. Sequential estimation with data-dependent allocation and time trends; p. 91-97.

Coad, DS. Journal of Statistical Computation and Simulation. Vol. 40. Taylor & Francis; 1992. A comparative study of some data-dependent allocation rules for Bernoulli data; p. 219-231.

Coad, DS., Ivanova, A. Sequential Analysis. Vol. 20. Taylor & Francis; 2001. Bias calculations for adaptive urn designs; p. 91-116.

Gittins JC. Bandit processes and dynamic allocation indices. J Roy Statist Soc Ser B. 1979; 41(2):148–177. with discussion.

Gittins, JC., Jones, DM. In: Gani, J.Sarkadi, K., Vincze, I., editors. A dynamic allocation index for the sequential design of experiments; Progress in Statistics (European Meeting of Statisticians, Budapest, 1972); North-Holland, Amsterdam, The Netherlands. 1974. p. 241-266.

Gu, Xuemin, Lee, J Jack. A simulation study for comparing testing statistics in response-adaptive randomization. BMC medical research methodology. 2010; 10(48)

Firth, David. Bias reduction of maximum likelihood estimates. Biometrika. 1993; 80(1):27–38.

Flournoy, Nancy, Haines, Linda M., Rosenberger, William F. A graphical comparison of response-adaptive randomization procedures. Statistics in Biopharmaceutical Research. 2013; 5(2):126–141.

Haines, Linda M., Sadiq, Hassan. Start-up designs for response-adaptive randomization procedures with sequential estimation. Statistics in medicine. 2015; 34(21):2958–2970. [PubMed: 25944171]

Hu, F., Rosenberger, WF. The theory of response-adaptive randomization in clinical trials. Vol. 525. John Wiley & Sons; 2006.

Ivanova, Anastasia, Rosenberger, William F. A comparison of urn designs for randomized clinical trials of $K > 2$ treatments. Journal of biopharmaceutical statistics. 2000; 10(1):93–107. [PubMed: 10709804]

Liu, Suyu, Lee, J Jack. An overview of the design and conduct of the BATTLE trials. Chinese Clinical Oncology. 2015; 4(3):1–13. [PubMed: 25841708]

Kalish, Leslie A., Begg, Colin B. Controlled Clinical Trials. Vol. 8. Elsevier; 1987. The impact of treatment allocation procedures on nominal significance levels and bias; p. 121-135.

Karrison, Theodore G., Huo, Dezheng, Chappell, Rick. Controlled Clinical Trials. Vol. 24. Elsevier; 2003. A group sequential, response-adaptive design for randomized clinical trials; p. 506-522.

Meurer, William J., Lewis, Roger J., Berry, Donald A. JAMA. Vol. 307. American Medical Association; 2012. Adaptive clinical trials: a partial remedy for the therapeutic misconception?; p. 2377-2378.

Plamadeala V, Rosenberger WF. Sequential monitoring with conditional randomization tests. The Annals of Statistics. 2012; 40(1):3044.

Rosenberger W, Lachin J. The use of response-adaptive designs in clinical trials. Controlled Clinical Trials. 1993; 14(6):471–484. [PubMed: 8119063]

Rosenberger, W., Lachin, J. Randomization in clinical trials: theory and practice. John Wiley & Sons; 2016.

Rosenberger W, Stallard N, Ivanova A, Harper C, Ricks M. Optimal adaptive designs for binary response trials. Biometrics. 2001; 57(3):0909–913.

Rosenberger WF, Vidyashankar AN, Agarwal DK. Covariate-adjusted response-adaptive designs for binary response. Journal of biopharmaceutical statistics. 2001; 11(4):227–236. [PubMed: 12018777]

Simon R. Adaptive treatment assignment methods and clinical trials. Biometrics. 1977; 33:743–749. [PubMed: 338044]

Simon, R., Simon, NR. Statistics & Probability Letters. Vol. 81. Elsevier; 2011. Using randomization tests to preserve type I error with response adaptive and covariate adaptive randomization; p. 767-772.

Tamm, M., Hilgers, RD., et al. Methods of information in medicine. Vol. 53. Schattauer Publishers; 2014. Chronological bias in randomized clinical trials arising from different types of unobserved time trends; p. 501-510.

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; 26:1621–1628. [PubMed: 25979922]

Thompson W. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Biometrika. 1933; 25(3/4):285–294.

Trippa L, Lee EQ, Wen PY, Batchelor TT, Cloughesy T, Parmigiani G, Alexander BM. Bayesian Adaptive Randomized Trial Design for Patients With Recurrent Glioblastoma. Journal of Clinical Oncology. 2012; 30:3258–3263. [PubMed: 22649140]

Tymofyeyev Y, Rosenberger WF, Hu F. Implementing optimal allocation in sequential binary response experiments. Journal of the American Statistical Association. 2007; 102(477)

Villar S, Bowden J, Wason J. Multi-armed Bandit Models for the Optimal Design of Clinical Trials: Benefits and Challenges. Statistical Science. 2015; 30(2):199–215. [PubMed: 27158186]

Villar S, Wason J, Bowden J. The forward looking Gittins index: a novel bandit approach to adaptive randomization in multi-arm clinical trials. Biometrics. 2015; 71:969–978. [PubMed: 26098023]

Wason, J., Trippa, L. Statistics in medicine. Vol. 33. Wiley Online Library; 2014. A comparison of bayesian adaptive randomization and multi-stage designs for multi-arm clinical trials; p. 2206-2221.

Williamson F, Jacko P, Villar S, Jaki T. A Bayesian adaptive design for clinical trials in rare diseases. Comp Statist Data Anal. 2017; 113C:136–153.
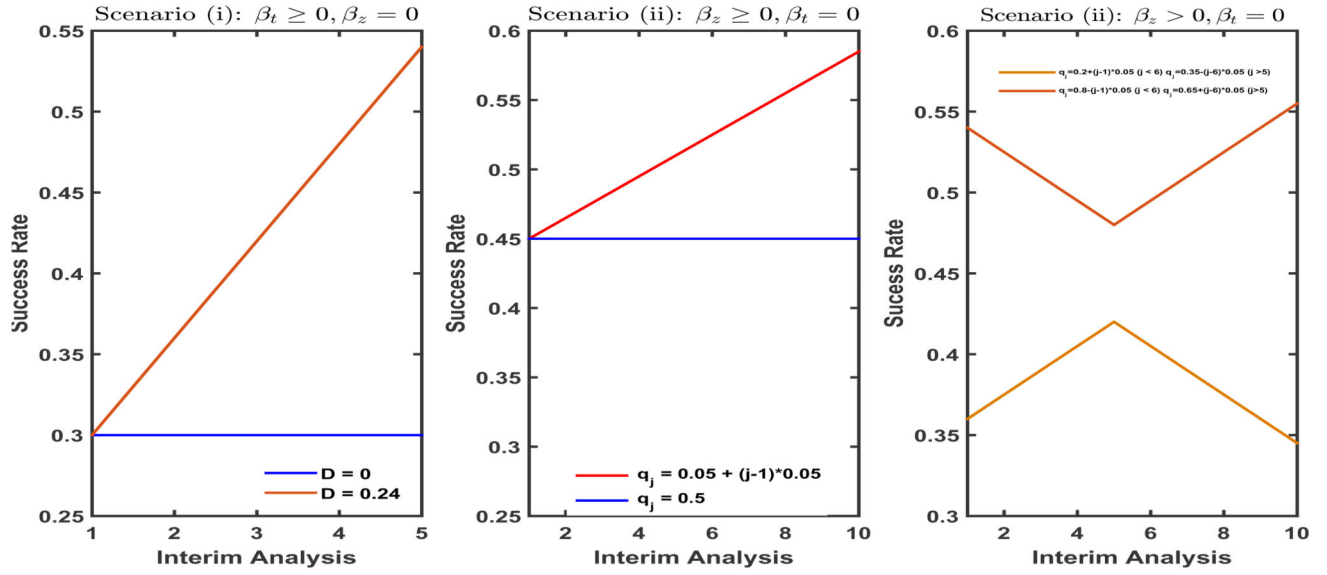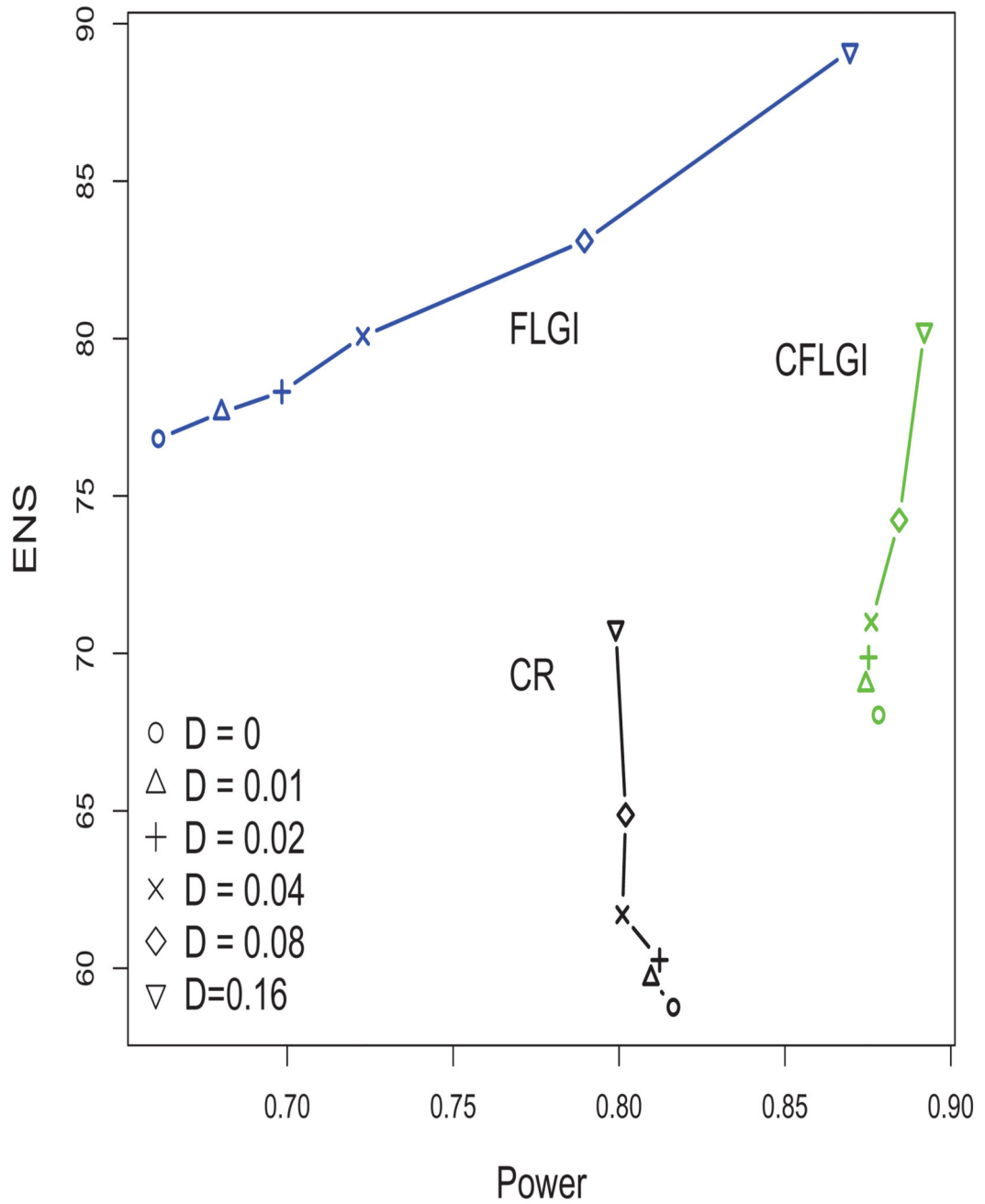
Figure 1.
The per block success rate under different time trend assumptions plotted over time. Left plot corresponds to scenario (i) (Changes in standard of care) and middle and right plot correspond to different cases of scenario (ii) (Patient drift).
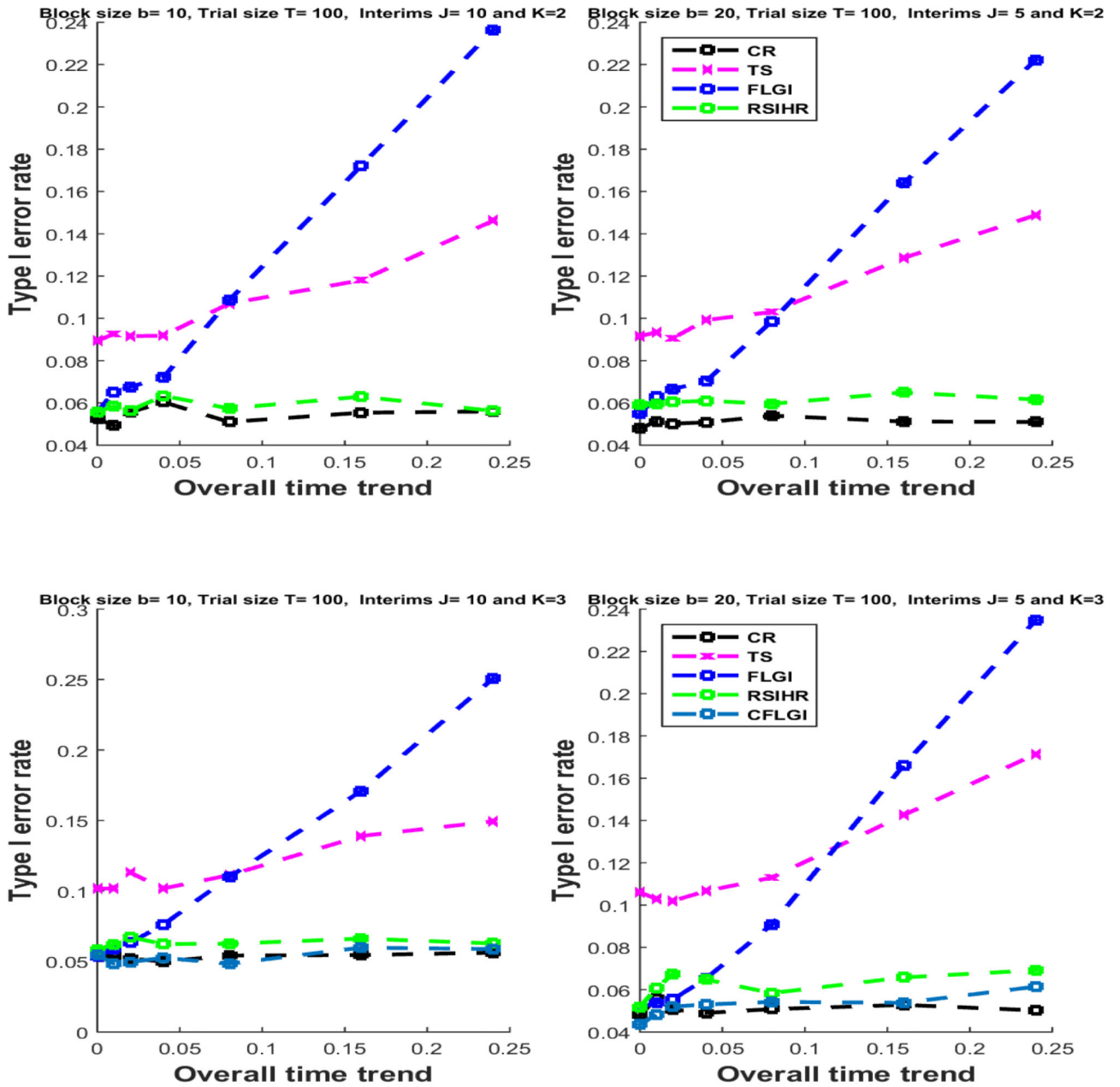
Figure 2.
The type I error rate for Scenario (i) (changes in the standard of care) under different linear time trends assumptions and different RAR rules.
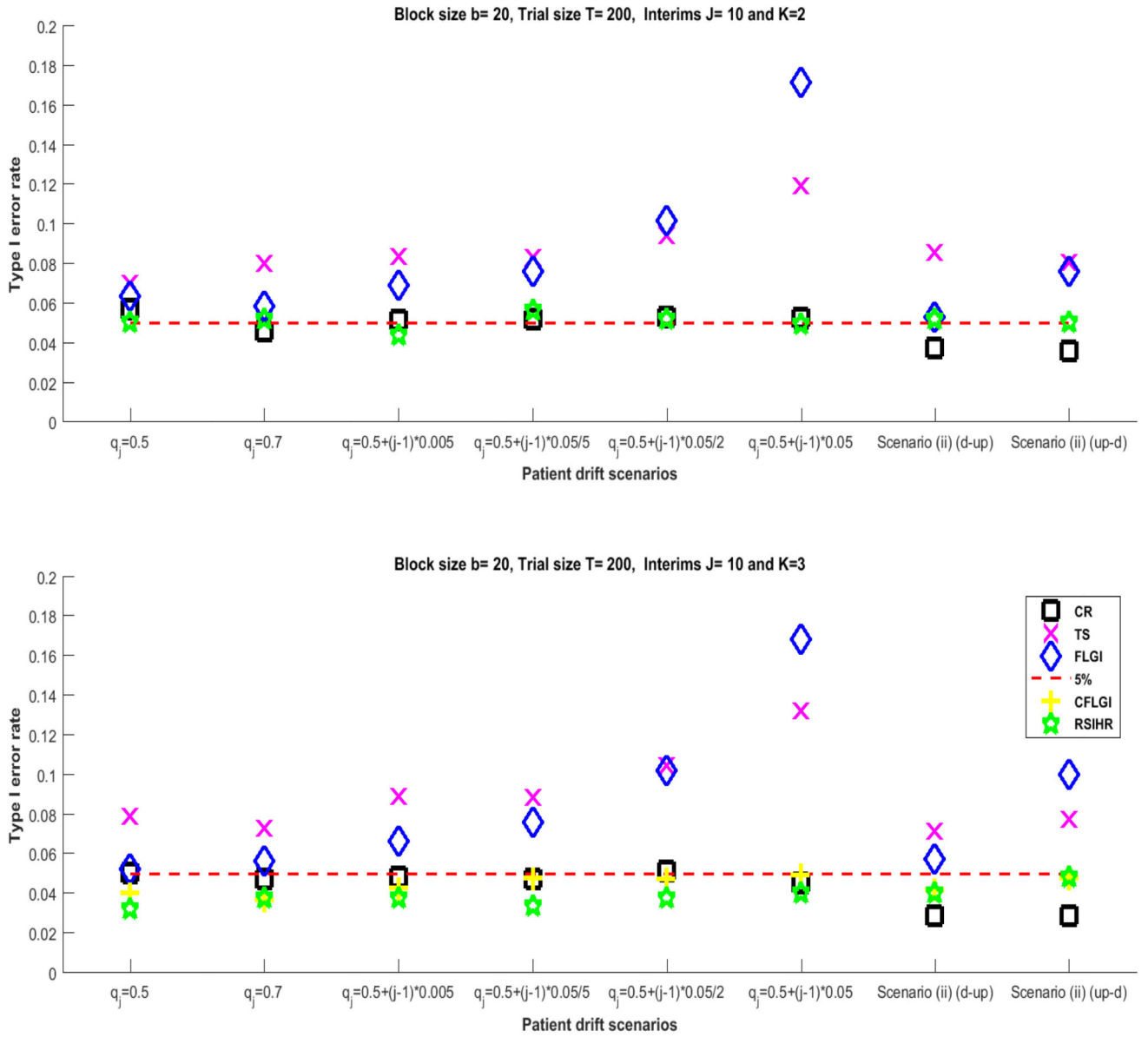
Figure 3.
The type I error rate for different group recruitment rates assumptions under scenario (ii) with $\beta_Z \approx 1.2528$

Figure 4.
ENS-Power trade-off of CR, CFLGI and FLGI in 5000 replicates of a 3-arm trial of size $T =$ 100 with block size $b = 20$ ($J = 5$) under a case of scenario (i) with a treatment effect of 0.275 for arm 1

Table 1

The type I error rate for the approximate randomisation test from 5000 replicates of a 2-arm trial of size $T = 100$ using a FLGI with block size $b = 20$ ($J = 5$) and under the case of Scenario (i) depicted in Figure 2 (top-right plot).

$\boldsymbol{a}$ (s.e.)	p^* (s.e.)	ENS	D
0.0445 (0.21)	0.501 (0.21)	0.19	0
0.0480 (0.21)	0.506 (0.22)	−0.17	0.08
0.0449 (0.20)	0.494 (0.23)	0.02	0.16
0.0445 (0.21)	0.499 (0.24)	0.23	0.24

Table 2

Power for the approximate randomisation test from 5000 replicates of a 2-arm trial of size $T = 150$ using a FLGI with block size $b = 30$ ($J = 5$) under a case of Scenario (i) with a treatment effect of 0.40.

$(1 - \beta_F)$(s.e.)	$(1 - \beta_{RT})$(s.e.)	p^*(s.e.)	ENS	D
0.8086 (0.39)	0.6057 (0.48)	0.871 (0.09)	22.04	0
0.8972 (0.30)	0.6080 (0.49)	0.881 (0.04)	24.01	0.08
0.9524 (0.21)	0.6021 (0.50)	0.878 (0.05)	23.99	0.16
0.9802 (0.14)	0.5851 (0.48)	0.882 (0.03)	23.73	0.24

Table 3

GLM estimated through MLE with and without Firth correction for $T = 100$, $J = 5$, $b = 20$ in a case of scenario (i) with $D = 0.24$. Results for 5000 trials. True values were assumed to be $\beta_0 \approx -0.8473$, $\beta_t \approx 0.2719$ and $\beta_z = \beta_1 = 0$.

(I) GLM fitting without correction for CR		
E $(\hat{\beta_i})$	E (MSE)	E $(p_{value} < 0.05)$
$\hat{\beta_0}$ -0.8684	0.1992	0.5174
$\hat{\beta_t}$ 0.2610	0.0243	0.4018
$\hat{\beta_1}$ 0.0070	0.1900	0.0544
(II) GLM fitting with correction for CR		
$\hat{\beta_0}$ -0.8370	0.1838	0.5224
$\hat{\beta_t}$ 0.2509	0.0227	0.4012
$\hat{\beta_1}$ 0.0067	0.1775	0.0534
(III) GLM fitting without correction for FLGI		
$\hat{\beta_0}$ -1.4465	8.9957	0.4110
$\hat{\beta_t}$ 0.1898	0.0307	0.1844
$\hat{\beta_1}$ 0.0038	18.2440	0.0142
(IV) GLM fitting with correction for FLGI		
$\hat{\beta_0}$ -0.9307	0.3947	0.4670
$\hat{\beta_t}$ 0.1825	0.0301	0.1858
$\hat{\beta_1}$ 0.0048	0.7993	0.0456

Table 4

GLM estimated through MLE with Firth correction for $T = 150$, $J = 5$, $b = 30$ in a case of Scenario (i) with $D = 0.16$ and $\beta_1 = 1.6946$. Results for 5000 trials. True values were assumed to be $\beta_0 \approx -0.8473$, $\beta_t \approx 0.1840$, $\beta_1 = 1.6946$ and $\beta_z = 0$

(II) GLM fitting with correction for CR		
E $(\hat{\beta}_i)$	E (MSE)	E $(p_{value} < 0.05)$
$\hat{\beta}_0$ -0.8951	0.1413	0.7194
$\hat{\beta}_t$ 0.1985	0.0175	0.3262
$\hat{\beta}_1$ 1.7831	0.1488	0.9994
(IV) GLM fitting with correction for FLGI		
$\hat{\beta}_0$ -0.8832	0.3192	0.3364
$\hat{\beta}_t$ 0.2408	0.0291	0.3062
$\hat{\beta}_1$ 1.6917	0.4313	0.7394

Table 5

GLM estimated through MLE with Firth correction for $T = 200$, $J = 10$, $b = 20$, $K = 3$ in a case of scenario (ii) in which $qj = [0.5 : 0.05 : 0.95]$. Results for 5000 trials. True values were assumed to be $\beta_0 \approx -0.8473$, $\beta_z \approx 1.2528$ and $\beta_t = \beta_1 = \beta_2 = 0$

	E $(\hat{\beta}_i)$	E (MSE)	E $(p_{value} < 0.05)$
(II) GLM fitting with correction for CR			
$\hat{\beta}_0$	-0.8527	0.1307	0.6778
$\hat{\beta}_z$	1.2597	0.1142	0.9758
$\hat{\beta}_1$	-0.0084	0.1305	0.0458
$\hat{\beta}_2$	-0.0029	0.1304	0.0486
(IV) GLM fitting with correction for FLGI			
$\hat{\beta}_0$	-0.8771	0.1724	0.6740
$\hat{\beta}_z$	1.2471	0.1169	0.9702
$\hat{\beta}_1$	0.0114	0.2228	0.0598
$\hat{\beta}_2$	-0.0097	0.2246	0.0620
(VI) GLM fitting with correction for CFLGI			
$\hat{\beta}_0$	-0.8455	0.1338	0.6632
$\hat{\beta}_z$	1.2505	0.1200	0.9686
$\hat{\beta}_1$	-0.0226	0.1471	0.0558
$\hat{\beta}_2$	-0.0196	0.1477	0.0492

Table 6

GLM estimated through MLE with Firth correction for $T = 200$, $J = 10$, $b = 20$, $K = 3$ in a case of scenario (ii) in which $q_j = [0.5 : 0.05 : 0.95]$. Results for 5000 trials. True values were assumed to be $\beta_0 \approx -0.8473$, $\beta_z \approx 1.2528$ and $\beta_1 = 0.8473$, $\beta_t = \beta_2 = 0$

(II) GLM fitting with correction for CR		
$E(\hat{\beta}_i)$	$E(MSE)$	$E(p_{value} < 0.05)$
$\hat{\beta}_0$ -0.8816	0.1324	0.7078
$\hat{\beta}_z$ 1.3006	0.1239	0.9726
$\hat{\beta}_1$ 0.9355	0.1549	0.6954
$\hat{\beta}_2$ -0.0032	0.1356	0.0516
(IV) GLM fitting with correction for FLGI		
$\hat{\beta}_0$ -1.1635	0.5391	0.3994
$\hat{\beta}_z$ 1.3492	0.1300	0.9762
$\hat{\beta}_1$ 1.1300	0.5845	0.3672
$\hat{\beta}_2$ 0.0041	0.7740	0.0246
(VI) GLM fitting with correction for CFLGI		
$\hat{\beta}_0$ -0.8861	0.1378	0.6966
$\hat{\beta}_z$ 1.3127	0.1243	0.9718
$\hat{\beta}_1$ 0.8862	0.2077	0.7718
$\hat{\beta}_2$ -0.2487	0.5027	0.0288