A comparison of Bayesian adaptive randomization and multi-stage designs for multi-arm clinical trials

James M. S. Wason\textsuperscript{a,*} and Lorenzo Trippa\textsuperscript{b,c}

When several experimental treatments are available for testing, multi-arm trials provide gains in efficiency over separate trials. Including interim analyses allows the investigator to effectively use the data gathered during the trial. Bayesian adaptive randomization (AR) and multi-arm multi-stage (MAMS) designs are two distinct methods that use patient outcomes to improve the efficiency and ethics of the trial. AR allocates a greater proportion of future patients to treatments that have performed well; MAMS designs use pre-specified stopping boundaries to determine whether experimental treatments should be dropped. There is little consensus on which method is more suitable for clinical trials, and so in this paper, we compare the two under several simulation scenarios and in the context of a real multi-arm phase II breast cancer trial. We compare the methods in terms of their efficiency and ethical properties. We also consider the practical problem of a delay between recruitment of patients and assessment of their treatment response. Both methods are more efficient and ethical than a multi-arm trial without interim analyses. Delay between recruitment and response assessment attenuates this efficiency gain. We also consider futility stopping rules for response adaptive trials that add efficiency when all treatments are ineffective. Our comparisons show that AR is more efficient than MAMS designs when there is an effective experimental treatment, whereas if none of the experimental treatments is effective, then MAMS designs slightly outperform AR. © 2014 The Authors Statistics in Medicine Published by John Wiley & Sons, Ltd.

Keywords: adaptive randomization; group-sequential designs; multi-arm trials; multiple testing

1. Introduction

In many disease indications, there are multiple drugs in the same phase of clinical development. Specifically, in oncology, there are over 1500 cancer therapeutics in the clinical pipeline [1]. For several malignancies, the number of putative treatments has rapidly increased because of development of molecularly targeted agents [2]. In such a context, efficient alternatives to traditional single-arm and two-arm trials are needed to maximize the number of treatments tested considering the limited number of patients and resources that are available for trials [3].

A multi-arm trial, in which several new drugs are compared with a control treatment within a single trial, is a class of designs that have potential to greatly increase the efficiency of the drug development process. There is a growing literature on multi-arm trials, with new trial designs investigated [3–5], compared with traditional designs [6, 7] and applied in practice [8–11]. The main advantage of a multi-arm trial is that the shared control group means fewer patients are required compared with conducting separate two-arm trials of the different experimental drugs. There are also other advantages. First, patient recruitment to clinical trials can often be increased if there is a greater chance of being allocated a new treatment [12]. Second, the overall administrative burden in setting up and running a multi-arm trial can
be lower than running several separate trials [13]. Third, in early phase cancer trials, most treatments fail to show superiority [14], and it is attractive to concentrate patient allocation to the most promising arms.

As patients are randomized and treated during a multi-arm trial, a considerable amount of information will be gathered on the effect of different treatments. Several designs have been proposed to use this information in order to focus resources on the most promising arms. Two of the most commonly proposed approaches to achieve this goal are response-adaptive randomized designs and multi-arm multi-stage (MAMS) designs.

Adaptive randomization (AR) allows changes to be made to the randomization probabilities to treatments during the trial. In most cases, the aim of the procedure is to allocate a greater proportion of patients to treatments that have so far demonstrated evidence of a better performance than other arms. A number of papers have discussed AR (see, for example, [15–19]). Here, we will consider Bayesian AR procedures. Such procedures are characterized by time varying randomization probabilities that are defined and recursively updated through a Bayesian model for patients’ outcomes. Hu and Rosenberger [20] have also developed and investigated alternative non-Bayesian adaptive procedures for multi-arm trials, with similar goals.

In the case of a single experimental treatment and a control arm, AR can be less efficient than a balanced randomized trial when the design power is considered but can also result in patients on the trial having a better average treatment response. A paper by Coad and Rosenberger [21] compared a form of AR to group-sequential designs in this situation and found the latter to be more efficient. AR is thus often argued by its proponents to be more ethical and by its detractors to be inefficient. This trade-off only applies to the two-arm setting. In the case of multi-arm trials, it has been shown that applying AR to the experimental treatments, but protecting the allocation to the control group, is more efficient than balanced randomization in many situations [7].

Multi-arm multi-stage designs, like AR, use interim data. Instead of changing the allocation proportions, stopping thresholds are set on test statistics, with arms that show insufficient evidence of effectiveness dropped. The allocation to each remaining arm is fixed in MAMS trials, although assigning more patients to the control group than to the experimental arms can provide small gains in efficiency over balanced randomization [22], and has been used in practice [13].

Although both methods have been proposed to improve efficiency compared with a multi-arm trial without interim analyses, no systematic comparison of the two methods has been undertaken. These two approaches, and the investigation of several anticancer treatments within single studies, have been identified in the literature as relevant strategies for streamlining the development of new drugs. The major aim of this paper is to rigorously compare the two methods and to make recommendations about their relative advantages and disadvantages.

We performed the study so that the overall complexity level of the trial is identical under the two approaches; in particular, we compare MAMS and AR designs with identical numbers of interim analyses. The literature reports several trade-off relationships between the complexity of specific designs, for example, the number of interim analyses, and the resulting efficiency and operating characteristics. High complexity levels of the design are, in most cases, associated with higher costs, such as the infrastructures for updating randomization probabilities and coordinating patients assignment in multi-center trials. We therefore keep the trial complexity at identical levels when comparing MAMS and AR trials.

We focus on a binary outcome with a short delay between recruitment and outcome observation. This is a scenario motivated by adjuvant phase II breast cancer trials, where the primary outcome is pathological complete response (pCR), a binary outcome observed after a chemotherapy regimen, but before surgery. We discuss simulations of AR and MAMS trials under scenarios that mimic a recent multi-arm trial in breast cancer described in Gianni et al. [9]. We also consider the effect of the recruitment rate when there is a delay between recruitment and observation of the treatment outcome. An advantage in using this example for illustrating the relative merits of the two competing approaches, AR and MAMS designs, is the possibility of constructing realistic scenarios based on historical data.

The literature on Bayesian clinical trials includes relevant discussions on whether the final decisions at completion of the trial should be based on Bayesian or frequentist arguments [23]. Berry [24] has also considered some intermediate solutions. Here, we do not discuss these alternatives. Instead, we assume that the final recommendations, for each of the tested agents, has to be constrained by standard frequentist criteria. That is, either the type I error probability of each null hypothesis or the overall familywise error rate (FWER) are controlled at a pre-specified α level. Our choice should not necessarily be interpreted as a preference for the frequentist approach, which is often incompatible with the foundational likelihood principle [25] in multi-stage and adaptive experiments. This choice has two main motivations.
First, anchoring a comparative study to a single approach for the final analyses, at completion of the trial, is necessary to obtain easy to interpret results. Second, in most clinical trials, some of the stakeholders, including in several cases regulators such as the Food and Drug Administration, require final decisions based on frequentist criteria.

2. Designs

2.1. Multi-arm multi-stage designs for binary outcomes

Methodology for MAMS trials has mainly focused on situations where the primary endpoint is assumed normally distributed [5, 26] or time-to-event [27]. In this section, we describe the procedure we used to design MAMS trials with binary outcomes.

We consider a $J$-stage trial with $K$ experimental arms and a control arm. The response probability of the $k$-th arm is $p_k$. We labeled the control arm and the $K$ experimental arms as $k = 0$ and $k = 1, \ldots, K$, respectively. We designed the trial to test $K$ null hypotheses $H_0^{(1)}, H_0^{(2)}, \ldots, H_0^{(K)}$, where the $k$-th null hypothesis is

$$H_0^{(k)} : p_k \leq p_0$$

At each stage, we recruited $n$ patients to the control arm, and $n$ patients are recruited to each experimental arm that has not been previously dropped. The value of $n$ is called the group size. We define the number of responses and the number of patients recruited to arm $k$ at the completion of the $j$-th stage, respectively, as $Y_{jk}$ and $n_{jk}$. As suggested in Jennison and Turnbull [28], a suitable test statistic for contrasting arm $k \neq 0$ with the control is

$$Z_{jk} = \left( \frac{Y_{jk}}{n_{jk}} - \frac{Y_{j0}}{n_{j0}} \right) \left( \left( \frac{1}{n_{j0}} + \frac{1}{n_{jk}} \right) \tilde{p}_{jk} (1 - \tilde{p}_{jk}) \right)^{-1/2}$$

where $\tilde{p}_{jk} = (Y_{jk} + Y_{j0})/(n_{jk} + n_{j0})$. Asymptotically, under the hypothesis of no treatment effect, the test statistic in (1) is distributed as a standard Normal random variable.

We specify futility boundaries for each stage, $f = (f_1, \ldots, f_J)$. At stage $j < J$, if $Z_{jk} \leq f_j$, then arm $k$ is dropped for futility. Once experimental arm $k$ is dropped for futility, no further patients are recruited to it, and the null hypothesis $H_0^{(k)}$ is not rejected. If all experimental arms are dropped for futility, the trial terminates. Otherwise the remaining arms continue to stage $j + 1$. At stage $J$, if the test statistic $Z_{Jk}$ is above $f_J$, then $H_0^{(k)}$ is rejected.

In the case of a single experimental treatment and a control arm, that is, $K = 1$, the work of Jennison and Turnbull [28] provided analytical formulae and asymptotic results for the probability of rejecting the null hypothesis. The formulae involves recursive summation, over all possible pairs of the response frequencies on the two arms, at each interim analysis, taking into account the futility boundaries. If the type I error rate and power of each null hypothesis is of interest, then this method can be directly used in the multi-arm case. In our study, we also consider the control of multiple-comparison operating characteristics such as the FWER. For calculation of such quantities, because of the shared control group, recursive summation over all possible $(K + 1)$-tuples of the number of responses on each arm is required. This quickly becomes computationally demanding as $K$ increases.

For this reason, we use a Monte Carlo procedure, similar to the one used in Wason and Jaki [5], to find a design that has pre-specified operating characteristics. This procedure involves simulating a large number of independent arrays of random variables. Unless otherwise stated, we use 250,000 arrays. Each one is a $J \times (K + 1)$ matrix of independent uniform variables, $U_{jk} \sim U(0, 1)$. For any fixed value of $(p_0, p_1, \ldots, p_K)$, the random variable $U_{jk}$ can be transformed to the number of responses from the set of patients recruited to the $k$-th treatment at the $j$-th stage by using the quantile function of the binomial distribution. This simulates the number of responses that would be observed at each stage if the treatment is not dropped. For each simulation replicate, given a futility boundary $f$, one can determine which (if any) null hypotheses are rejected.

Throughout the paper, we define the power (unless otherwise stated) under the least favorable configuration (LFC) [29]. The LFC is the configuration where one experimental arm, without loss of generality treatment 1, has response probability equal to a clinically relevant value $p_1^*$, and all other arms have some uninteresting treatment effect, which we set equal to $p_0^* < p_1^*$, on the basis of historical data or investigators expectations.
We now describe how to select futility boundaries that control the FWER, which is the probability of rejecting any true null hypothesis. The FWER is strongly controlled at level $\alpha$ if the maximum probability of rejecting a true null hypothesis, over $(p_0, p_1, \ldots, p_K) \in [0, 1]^{K+1}$, is less than or equal to $\alpha$. Magirr, Jaki, and Whitehead [26] showed that the maximum probability of rejecting a true null hypothesis, in the case of normally distributed outcomes, is attained when all of the experimental treatments have the same effect as the control treatment. The proof also applies when the outcomes are binary. When $p_0 = p_1 = \ldots = p_K$, the probability of rejecting a null hypothesis $H_0^{(k)}$, on the basis of pre-selected thresholds $f$ and the summary statistics in (1), depends on the common response probability. For small samples and fixed boundaries $f$, the type I error rate can vary considerably with the value of the identical response probabilities. Jennison and Turnbull [28], in a similar context, suggested that the type I error rate should be evaluated at a number of plausible values of $p_0$ and to use the maximum to bound the type I error rate. We adopt this approach for finding MAMS designs that control the FWER.

The procedure for choosing the boundaries $(f_1, \ldots, f_J)$ is based on the triangular test (refer to [30] and [31]). This approach has been compared with several other types of boundaries and found to perform well for group-sequential designs and MAMS designs [5, 32] in terms of the expected number of patients recruited.

For a MAMS design, the boundaries can be adjusted so that the FWER is controlled. Given any approach for early stopping boundaries in the two arms setting, say the triangular method or alternative methods based on power prediction [33], one can tune the parameters of the procedure so that the desired MAMS operating characteristic (in our case, the FWER) is matched. These procedures define two arm designs with futility boundaries $f$ on the basis of (i) a desired significance level $\alpha$ and (ii) a targeted power $\beta$ at a reference scenario $(p_0^*, p_1^*)$. With several experimental arms, we consider a continuum of boundaries obtained varying only the significance level, then we select the boundary with maximum FWER equal to the pre-specified target, where the maximum is with respect to a grid of values for $p_0$. That is, we plug in the largest significance level $\alpha$ that allows us to control the FWER at the desired level.

In some cases, it is not of interest to control the FWER or, more generally, to correct for multiple testing. The view as to whether the FWER is relevant in multi-arm trials has been discussed in the literature (refer to [6] and [22]). In our comparative study, we also consider settings in which the investigator only controls, for each hypothesis, the type I error at some $\alpha$ level, without constraints on the FWER.

### 2.2. Adaptive randomization for binary outcomes

We use a Bayesian adaptive design with overall sample size $N$. The hypotheses to be tested, $H_0^{(1)}, \ldots, H_0^{(K)}$, are the same as in the previous section. The design includes $J - 1$ interim analyses. At each interim analysis, we update the Bayesian model used for setting the randomization probabilities. The $j$-th interim analysis takes place after $(j/J)N$ patients have been randomized. If $(j/J)N$ is non-integer, it is rounded to the nearest integer.

We use a similar AR procedure to the one used in Trippa et al. [7]; the approach is also strictly related with previous proposals discussed in Thall and Wathen [15]. The procedure requires specification of a prior distribution on the response probabilities $(p_0, \ldots, p_K)$. At each interim analysis, we compute the conditional probabilities $P(p_k > p_0|\text{data})$, that is, the posterior probability of arm $k$’s response rate being higher than the control arm’s response rate given the data so far. Using notation from the previous section, we found that this posterior probability is $P(p_k > p_0|Y_{jk}, Y_{j0}, n_{jk}, n_{j0})$. We used these conditional probabilities to obtain dynamic randomization probabilities.

We use independent beta $(a_1, a_2)$ prior distributions for the response probabilities $p_0, \ldots, p_K$. The parameters of the prior are chosen to be $a_1 = p_0^*$ and $a_2 = 1 - p_0^*$ so that the the prior is centered around an initial guess $p_0^*$.

We set the randomization probabilities $(\pi_0, \ldots, \pi_K)$, during the $(j+1)$-th stage of the trial, given the current arm-specific sample sizes $n_0', \ldots, n_K'$, with $j \times N/J \leq n' = \sum_{i=0}^{K} n_i' \leq (j + 1) \times N/J$, to

$$
\pi_k \propto \begin{cases} 
\frac{\mathbb{P} (p_k > p_0|Y_{jk}, Y_{j0}, n_{jk}, n_{j0})^\gamma(j/J)}{\sum_{i=1}^{K} \mathbb{P} (p_i > p_0|Y_{ji}, Y_{j0}, n_{ji}, n_{j0})^\gamma(j/J)} & \text{if } k = 1, \ldots, K; \\
\frac{1}{K} \exp \left( \max(n_1', n_2', \ldots, n_K') - n_0' \right)^{n(a'/N)} & \text{if } k = 0
\end{cases}
$$

(2)
The vector of randomization probabilities is therefore a function of the posterior probabilities contrasting each experimental arm with the control. Recall that \( n_{jk} \) is the number of patients assigned to treatment \( k \) before the \( j \)-th interim analysis, that is, before computation of the posterior probabilities; this integer-valued variable differs from \( n'_k \) in (2), which denotes the current number of patients that have been randomized to arm \( k \).

Although we do not consider the possibility further in this paper, the posterior probabilities for each experimental arm could vary across patients with different covariates. This would mean that patients with different covariate values would have different allocation probabilities. Recent examples that account for covariates in the computation of the adaptive probabilities are the ISPY2 trial [11] and the BATTLE trial [8].

If we consider a scenario with a single effective experimental treatment, and a total sample size that diverges, then a suitable \( \eta \) function would result in approximately 50% of patients assigned to the effective treatment and approximately 50% assigned to the control arm. Also, when the overall sample size diverges, if two or more experimental treatments are effective, with identical treatment effects, then approximately 100% of patients assigned to an experimental treatment will receive one of the best treatments. These facts directly follow from standard results in sequential analysis; see, for example, [34, 35]. Note that, assuming \( \eta > 0 \), the dominance of the exponential term \( \exp(\cdot) \) in (2) implies that, as the overall sample size diverges, the proportions of patients allocated to the control and to the experimental arm with the highest final sample size become identical.

The functions \( \eta \) and \( \gamma \) govern to what extent the allocation probabilities deviate from equal allocation. For example, if \( \gamma(j/J) = 0 \), each experimental arm has identical allocation probability. In contrast, as \( \gamma(j/J) \) increases to infinity, only randomization to the control and the most promising experimental treatment is allowed.

A good choice of the tuning function \( \gamma \) balances the so-called exploration versus exploitation trade-off [36]. A positive \( \eta \) function allows one to approximately match the number of patients \( \max(n_1, n_2, \ldots, n_K) \) receiving the most frequently assigned experimental treatment and the current sample size \( n'_0 \) of the control arm. This characteristic of the adaptive design we consider is the only major difference with respect to AR procedures previously discussed in the literature. In our simulations, we observed negligible sensitivity of the operating characteristics illustrated in this paper to the specific choice of \( \eta \). All presented results are based on a linear \( \eta \) function with final value \( \eta(1) = 0.25 \). In contrast, we have to carefully select the \( \gamma \) function.

We explored parametric \( \gamma \) functions such as \( \gamma(n/N) = a \left( \frac{n}{N} \right)^b \) and \( \gamma(n/N) = a \times 1 \left( \frac{n}{N} > b \right) \). To simplify computations, we have considered only functions with two parameters. We obtained the tuning of \( \gamma \) with Monte Carlo simulations. We iteratively simulated trials at each value of the \( \gamma \) parameters over a grid. We considered relevant scenarios for selecting \( \gamma \). For cancer studies, where there is a paucity of effective new treatments, we used scenarios with only one effective experimental arm, say arm 1. To choose the parameters, we focused on the final evidence in favor of a treatment effect. In particular, at each simulation, we computed the Bayes factor in favor of the hypothesis of a positive treatment effect for arm 1. This produces a rank of the \( \gamma \) functions by computing the proportion of simulations with Bayes factor above a threshold, say 2, under each \( \gamma \) function. This optimization procedure used scenarios with \( p_0 \) uniformly distributed in \((0.2, 0.8)\) and \( p_1 = p_0 + 0.15 \). Additional simulations suggested negligible sensitivity to the distribution of \( p_0 \) and \( p_1 \) across iterations.

At completion of the trial, we obtained the control of the type I error rate or FWER by computing a test statistic \( Z_K \) for each experimental arm and rejecting the associated null hypotheses for those arms whose test statistic exceeds a threshold \( c \). We obtained the threshold using a simulation algorithm aimed at approximating for each possible value of \((p_0, \ldots, p_K) \in [0, 1]^{K+1} \) the corresponding joint distribution of \( Z_1, \ldots, Z_K \). Once these approximations are obtained, we select the minimum \( c \) value that controls under all \((p_0, \ldots, p_K) \in [0, 1]^{K+1} \) combinations the type I error rate or the FWER below the pre-specified target.

The algorithm is a direct application of the importance sampling method. It involves two steps. We first iteratively simulate clinical trials varying at each iteration \((p_0, \ldots, p_K) \). We sampled the vector of response probabilities at each simulation from independent beta distributions. That is, each trial simulation \( T \) is based on random response probabilities \((p_0, \ldots, p_K) \sim g \), where \( g \) is a conveniently selected distribution. Let \( L(t; p_0, \ldots, p_K) \) be the likelihood of a trial under the adaptive scheme; this is the probability of a specific sequence of outcomes and treatment assignments at a fixed value of the vector
(p₀, . . . , p_K). We chose the distribution g so that for each generated trial T, the importance weights

$$w(T; \tilde{p}_0, . . . , \tilde{p}_K) = \frac{L(T; \tilde{p}_0, . . . , \tilde{p}_K)}{\int L(T; p_0, . . . , p_K) d g} \quad (3)$$

can be analytically obtained at any fixed value of \(\tilde{p}_0, . . . , \tilde{p}_K\). The computation of the ratio in expression (3) does not involve considerations on the adaptive scheme. We only need to compute the ratio between likelihoods under binomial and beta-binomial models.

The importance weights, one for each simulated trial, can be used to estimate the distribution of \(Z_1, . . . , Z_K\) at any combination \((\tilde{p}_0, . . . , \tilde{p}_K)\). Weighted estimates are obtained exploiting the equalities

$$E(w(T; \tilde{p}_0, . . . , \tilde{p}_K) \times 1(T = t)) = L(t; \tilde{p}_0, . . . , \tilde{p}_K)$$

for any possible value \(t\) of \(T\). The aforementioned expression implies that for any subset \(B \subset \mathbb{R}^K\), the equality

$$E(w(T; \tilde{p}_0, . . . , \tilde{p}_K) \times 1(Z(T) \in B)) = \mathbb{P}((Z_1, . . . , Z_K) \in B | \tilde{p}_0, . . . , \tilde{p}_K) \quad (4)$$

where \(Z(T)\) denotes the \(K\) test statistics associated to \(T\) and the right hand of the equation is the probability of the event \((Z_1, . . . , Z_K) \in B\) at a fixed value of response probabilities \(\tilde{p}_0, . . . , \tilde{p}_K\).

Standard Monte Carlo arguments allow one to approximate the probabilities of interest on the right hand of (4) by iteratively simulating the trial \(T\), which in turn produces a simulation of the random variable \(w(T; \tilde{p}_0, . . . , \tilde{p}_K) \times 1(Z(T) \in B)\). Each simulated trial \(T\) defines on \([0, 1]^{K+1}\) the function \((\tilde{p}_0, . . . , \tilde{p}_K) \rightarrow w(T; \tilde{p}_0, . . . , \tilde{p}_K) \times 1(Z(T) \in B)\) and therefore contributes to the estimates of all the joint \(Z_1, . . . , Z_K\) distributions, which vary across the \((\tilde{p}_0, . . . , \tilde{p}_K)\) combinations.

The second step consists in minimizing \(c\) with the constraints given by the estimates of the joint distributions of \(Z_1, . . . , Z_K\) over a grid of values for \(p_0, . . . , p_K\). In other words, we select \(c\) such that the maximum FWER or type I error rate estimate, across possible \(p_0, . . . , p_K\), is bounded by a pre-specified value. To speed up computations, one can consider an importance distribution \(g\) that generates \(p_0 = p_1\) with probability 1 and a grid of \(p_0, . . . , p_K\) with \(p_0\) always equal to \(p_1\), without compromising the described procedure for bounding the type I error probabilities or the FWER.

Most of the computing time to implement the importance sampling is dedicated to the simulation of adaptive trials with different scenarios \(\theta = (p_0, . . . , p_K)\) varying across the parameter space \(\Theta = [0, 1]^{K+1}\). In the subsequent optimization stage of the algorithm, with a moderate number of arms, say up to \(K = 5\), we used a regular grid that partitioned each dimension into 35 equally spaced subintervals. When we considered designs with a larger number of arms (say 10), we first optimize over a coarse grid with up to \(10^7\) points. Then, after the optimum \(\theta^*\) over this grid is obtained, the grid is rescaled and centered at \(\theta^*\).

The computational efficiency of the algorithm can be directly compared with a Monte Carlo procedure that estimates \(Z\) percentiles at fixed \(\theta\)s on a grid. We considered, for instance, a trial with \(K = 2\) and an overall sample size equal to 150. The grid is two dimensional. We fixed the overall number of simulated trials (10^6) to estimate \(Z\) percentiles at each \(\theta\) value on the grid. Then, by iterating the computational procedures, we obtained estimates for the mean squared errors. As expected, the mean squared errors of the Monte Carlo procedure depend on the grid granularity, because each estimate is based on a number of simulated trials that decreases with the number of points on the grid. If the grid has 100 points, then each estimate is based on 10^4 simulated trials. In contrast, this relation between granularity and the estimation accuracy is absent under the importance sampling approach. In the example given earlier, with up to 16 points on the grid, the efficiency of the Monte Carlo procedure is comparable with our approach, but with as few as 100 points, the higher efficiency of the importance sampling becomes apparent.

3. Case study

We consider the NeoSphere trial, a multi-arm trial in women with human epidermal growth factor receptor 2 breast cancer [9]. Human epidermal growth factor receptor 2 is a biomarker, which defines a subgroup with notably shortened survival [37]. In NeoSphere, the four arms considered were combination therapies, trastuzumab plus docetaxel (arm A), pertuzumab plus docetaxel (arm B), pertuzumab with trastuzumab (arm C), and pertuzumab plus docetaxel plus trastuzumab (arm D). The trial was powered to compare arm A with arms B and C, and arm B against arm D. The primary endpoint was pCR, which is defined as the absence of invasive neoplastic cells at surgery.
In total, 407 patients were recruited, with an equal allocation probability to each arm. Although not explicitly mentioned in this study, we will assume that pCR was assessed 15 weeks after randomization, which allows 3 weeks between the last day of treatment and the response assessment. Recruitment lasted from December 2007 to December 2009. This corresponds to a recruitment rate of approximately four patients per week. We will investigate the effects of a faster or slower recruitment in Section 6.

The final endpoint is observed with delay from the individual enrollment. We incorporated this characteristic into our comparison of AR and MAMS designs in Section 6. Each patient is randomly assigned a recruitment time. The time (in weeks) between recruitment of two consecutive patients is assumed to be an Exponential\((1/m)\) random variable, where \(m\) is the mean recruitment rate per week.

### 4. Comparison of adaptive randomization and multi-arm multi-stage trials

We compare MAMS and AR designs using simulated trials under a variety of scenarios when (i) the FWER or (ii) the type I error rate of each comparison is controlled. The baseline scenario is a trial with five interim analyses, four experimental arms, a target FWER of 0.2, or target type I error rate of 0.05, depending on which quantity is being controlled. Although a FWER of 0.2 may seem high, we consider it to be adequate for typical multi-arm phase II trials. Also, Wason, Jaki, and Stallard recently gave a detailed motivation for controlling the FWER at levels of at least 0.2, on the basis of optimality criteria, in [38]. We used a target power of 0.8 when \(p^*_0 = 0.2\) and \(p^*_1 = 0.4\). We then varied each of these parameters and obtained the list of scenarios in Table I. The first six columns illustrate these variations; the acronyms T-FWER, T-Power, and T-TOER are used to indicate the targeted FWER, power, and type I error rate, respectively. Table I provides a description of scenarios, the resulting cutoff value \(c\) for the AR design, used at completion of the trial for reporting significant treatment effects, and relevant operating characteristics. The table has two main components: the top half shows designs when the FWER is controlled, and the bottom half shows designs when the type I error rate is controlled.

For the MAMS design, the sample size is random, so the expected sample size (ESS) under the global null hypothesis \(H_G\), that all treatment effects are null, and under the LFC, are shown. The expected values are indicated in Table I as ESS-\(H_G\) and ESS-LFC. For the AR designs, we report the overall (fixed) sample size \(N\). We also report, for the MAMS trial, the group size \(n\), which is the number of patients per arm at each stage.

Table I suggests that the procedures described in Section 2 control the FWER at approximately the targeted value. We estimated the reported control of the FWER under \(H_G\) and the power under the LFC for AR and MAMS designs in Table I from independent simulations performed after the designs had been selected. These estimates are based on 100,000 simulations. Only in a few scenarios is, the FWER/type I error rate slightly above the target level. For example, the scenario in which \(p^*_0 = 0.5\) appears to result in a small inflation in FWER and type I error rate using the MAMS design. We also note that the MAMS design generally seems more sensitive to the value of \(p^*_0\), as the observed FWER is often noticeably lower than the target level. This is not true to the same extent with the AR design.

In all scenarios, there is a positive difference between the maximum sample size under the MAMS design and the sample size of the AR design. For instance, under the first scenario we considered, the maximum MAMS sample size is \(J \times (K + 1) \times n = 325\), whereas the AR design sample size is \(N = 190\). These differences are similar when we consider the control of the FWER and of the type I error rate.

In most scenarios, the expected sample size under \(H_G\) for the MAMS design is smaller than the sample size required by the AR design. This suggests that the MAMS design can be more efficient when all treatments are ineffective. In contrast, when the LFC is true, that is, there is a single effective experimental treatment, the AR design sample size \(N\) is always lower than the MAMS expected sample size.

The distribution of the MAMS design sample size across simulations is summarized in Figure 1 through box plots. The panels consider MAMS and AR trials designed to control the type I error rate of each comparison at 0.05, with a power of 0.8 at \((p^*_0, p^*_1) = (0.2, 0.4)\). This figure emphasizes the variability of the sample size used by the MAMS design, with at least a threefold difference between the minimum and maximum observed sample sizes across simulations. These boxplots suggest that the choice of implementing a MAMS design or an AR study should not depend exclusively on the comparison of the average sample size under relevant scenarios and that sample size variability should be included among the most relevant operating characteristics. The figure also displays, under six scenarios, the AR sample size \(N\), which always remains below the third quartile of the sample size distribution of the MAMS design.
### Table I. Design parameters and operating characteristics of MAMS designs and AR designs for different scenarios.

| Design parameters MAMS | Adaptive randomisation |
|------------------------|------------------------|
| JK | $\rho_0^*$ | $\rho_1^*$ | T-FWER | T-Power | n | TOER | FWER | Power | ESS-$H_G$ | ESS-LFC | ENF | N' | c | TOER | FWER | Power | ENF |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Target familywise error rate required** |
| 5 | 4 | 0.2 | 0.4 | 0.2 | 0.8 | 13 | 0.068 | 0.186 | 0.809 | 171.9 | 215.3 | 160.4 | 190 | 1.40 | 0.071 | 0.197 | 0.815 | 140.5 |
| 3 | 4 | 0.2 | 0.4 | 0.2 | 0.8 | 21 | 0.078 | 0.205 | 0.815 | 193.4 | 227.9 | 170.4 | 200 | 1.40 | 0.072 | 0.199 | 0.809 | 148.5 |
| 10 | 4 | 0.2 | 0.4 | 0.2 | 0.8 | 7 | 0.066 | 0.180 | 0.810 | 165.3 | 217.7 | 161.6 | 180 | 1.38 | 0.072 | 0.198 | 0.813 | 132.4 |
| 5 | 2 | 0.2 | 0.4 | 0.2 | 0.8 | 12 | 0.120 | 0.197 | 0.813 | 93.0 | 136.1 | 98.2 | 110 | 1.14 | 0.116 | 0.197 | 0.813 | 79.2 |
| 6 | 4 | 0.2 | 0.4 | 0.2 | 0.8 | 14 | 0.055 | 0.196 | 0.797 | 259.6 | 303.3 | 229.8 | 250 | 1.51 | 0.057 | 0.203 | 0.807 | 187.3 |
| 5 | 4 | 0.2 | 0.4 | 0.05 | 0.8 | 21 | 0.015 | 0.047 | 0.818 | 260.5 | 345.1 | 256.3 | 300 | 2.13 | 0.016 | 0.051 | 0.805 | 221.3 |
| 5 | 4 | 0.2 | 0.4 | 0.05 | 0.95 | 26 | 0.072 | 0.194 | 0.955 | 445.6 | 331.2 | 340 | 340 | 1.39 | 0.071 | 0.203 | 0.958 | 250.7 |
| 5 | 4 | 0.5 | 0.7 | 0.2 | 0.8 | 16 | 0.078 | 0.212 | 0.824 | 211.7 | 265.3 | 184.6 | 230 | 1.62 | 0.048 | 0.140 | 0.804 | 171.0 |
| 5 | 4 | 0.2 | 0.5 | 0.2 | 0.8 | 7 | 0.070 | 0.188 | 0.837 | 91.1 | 115.5 | 82.7 | 90 | 1.40 | 0.072 | 0.191 | 0.810 | 63.3 |
| 5 | 4 | 0.3 | 0.2 | 0.2 | 0.8 | 50 | 0.0701| 0.195 | 0.818 | 658.2 | 826.1 | 638.1 | 700 | 1.39 | 0.069 | 0.197 | 0.803 | 540.1 |
| **Target type I error rate per comparison required** |
| 5 | 4 | 0.2 | 0.4 | 0.05 | 0.8 | 15 | 0.048 | 0.136 | 0.808 | 197.3 | 249.5 | 185.8 | 220 | 1.61 | 0.047 | 0.136 | 0.808 | 162.6 |
| 3 | 4 | 0.2 | 0.4 | 0.05 | 0.8 | 24 | 0.048 | 0.139 | 0.801 | 207.2 | 251.3 | 184.6 | 230 | 1.62 | 0.047 | 0.139 | 0.801 | 171.0 |
| 10 | 4 | 0.2 | 0.4 | 0.05 | 0.8 | 8 | 0.047 | 0.136 | 0.812 | 189.5 | 251.3 | 186.5 | 200 | 1.59 | 0.048 | 0.144 | 0.799 | 147.1 |
| 5 | 2 | 0.2 | 0.4 | 0.05 | 0.8 | 15 | 0.048 | 0.084 | 0.809 | 117.2 | 175.3 | 126.4 | 170 | 1.63 | 0.048 | 0.086 | 0.824 | 123.2 |
| 6 | 0.2 | 0.4 | 0.05 | 0.8 | 15 | 0.049 | 0.176 | 0.809 | 274.2 | 323.4 | 244.9 | 270 | 1.59 | 0.048 | 0.179 | 0.814 | 202.3 |
| 5 | 4 | 0.2 | 0.4 | 0.01 | 0.8 | 22 | 0.009 | 0.031 | 0.800 | 270.1 | 345.1 | 256.3 | 340 | 2.30 | 0.010 | 0.035 | 0.803 | 221.3 |
| 5 | 4 | 0.2 | 0.4 | 0.05 | 0.95 | 27 | 0.049 | 0.140 | 0.950 | 350.8 | 460.3 | 341.9 | 360 | 1.61 | 0.048 | 0.141 | 0.949 | 265.4 |
| 5 | 4 | 0.5 | 0.7 | 0.05 | 0.8 | 17 | 0.054 | 0.154 | 0.810 | 223.1 | 281.5 | 244.9 | 250 | 1.61 | 0.048 | 0.144 | 0.812 | 190.7 |
| 5 | 4 | 0.2 | 0.5 | 0.05 | 0.8 | 7 | 0.049 | 0.136 | 0.803 | 91.3 | 116.1 | 83.2 | 110 | 1.61 | 0.048 | 0.135 | 0.831 | 73.3 |
| 5 | 4 | 0.2 | 0.3 | 0.05 | 0.8 | 53 | 0.049 | 0.190 | 0.798 | 684.1 | 868.8 | 670.8 | 800 | 1.61 | 0.047 | 0.145 | 0.807 | 617.2 |

AR, adaptive randomization; MAMS, multi-arm multi-stage; $J$, number of interim analyses; $K$, number of experimental arms; $\rho_0^*$, response probability for the control arm; $\rho_1^*$, response probability for the effective treatment under the least favorable configuration; T-FWER, target familywise error rate; T-TOER, target type I error rate per comparison; T-Power, target power under the least favorable configuration; $n$, group size in MAMS design; TOER, type I error rate under the $H_G$ hypothesis; FWER, familywise error rate under the $H_G$ hypothesis; Power, power under the least favorable configuration; ESS-$H_G$, expected sample size of the trial under the global null hypothesis $H_G$; ESS-LFC, expected sample size of the trial under the least favorable configuration; ENF, expected number of treatment failures; $N'$, sample size for the AR design; $c$, final critical value used for hypothesis testing.
Figure 1. Sample size distribution of multi-arm multi-stage (boxplots) and adaptive randomization (dashed lines) trials, for four experimental arms as the number of interim analyses \((J)\) varies. The panels consider multi-arm multi-stage and adaptive randomization trials designed to control the type I error of each comparison at 0.05, with a power of 0.8 at \((p_0^*, p_1^*) = (0.2, 0.4)\). Sample sizes are shown under the global null hypothesis \((H_G)\) and under the least favorable configuration (LFC).

Figure 2. Boxplots showing the distribution, from 250,000 replicates, of the proportion of patients allocated to each arm in a five-arm adaptive randomization design with five interim analyses. The design parameters are given in the first row of Table I.

In Figure 2, we showed the distribution (over 250,000 simulation replicates) of the proportion of patients allocated to each arm when there is one effective experimental treatment and when there are two equally effective experimental treatments. When there is one effective experimental treatment, an average of 0.31 of the patients receive the best available treatment. A lower proportion of patients is allocated to the three ineffective experimental arms. When there are two effective experimental arms, the average allocation proportion of the control group is the highest across the \(K = 5\) arms. This is because, in each simulation replicate, one of the effective treatments is assigned more frequently than the other by chance, whereas the final sample size of the control arm approximately matches the maximum of the sample sizes across experimental arms. Similar simulations for the MAMS design showed that when there was one effective experimental treatment, the median allocation to the control treatment was 0.281 (interquartile range (IQR) 0.25–0.313), the median allocation to the effective experimental treatment was 0.278 (IQR 0.238–0.313), and the median allocation to each ineffective experimental treatment was 0.150 (IQR 0.100–0.200). For two effective experimental treatments, the median allocation to the control treatment was 0.250 (IQR 0.235–0.278), the median allocation to each effective experimental
We defined random variables under scenarios with more than one effective treatment. We selected these designs to control the FWER. With both designs, we observed efficiency gains, in terms of sample size requirements, when the number of interim analyses increases. This fact becomes apparent by comparing designs with three and five interim analyses. As one may expect, the marginal gain of each additional interim analysis diminishes with the number of stages $J$. In particular, this trend is suggested by the trial characteristics when 3, 5, and 10 interim analyses are considered.

The expected number of treatment failures (ENF) can be used to assess trial designs when possible ethical issues are considered. We report in Table I the ENF under the LFC, while under the $H_G$ hypothesis can be directly obtained by multiplying $(1 − p_{0i})$ and the expected sample size. In all scenarios, the ENF is lower using AR than using a MAMS design. This is partly expected as the sample size used by the AR design under the LFC is lower.

Table I shows scenarios with different number of experimental arms. As the number of experimental arms increases, the relative efficiency of the two designs shows slight changes in favor of AR. For instance, with two experimental arms, the expected sample size of the MAMS design under $H_G$ is 15% lower than the sample size of the AR design. Conversely, it is 4% higher when there are six experimental arms.

We attempted to check if the major differences between MAMS and AR designs described in the previous paragraphs are related or robust with respect to a few particular aspects involved in the comparison. First, we checked whether the results of the comparative study are confirmed when a different test statistic is used at completion of the simulated AR trials. Second, we checked whether the described differences persist when an alternative approach is used to define the futility boundaries $(f_1, . . . , f_J)$.

We substituted the test statistic used throughout the simulation study (1) with the $p$-values produced by the Fisher exact procedure for 2×2 tables. The adjective exact could be misleading in our adaptive context—these $p$-values are only used as $Z_k$ summary statistics. Under the considered scenarios, we observed negligible variations of the operating characteristics. We implemented, for instance, the AR procedure with $N = 220$, $K = 4$, and $J = 5$, using Fisher’s $p$-values to control the type I error at 0.05. The resulting cutoff was $c = 0.064$, and the power at the LFC considered in the first row of Table I was 0.801.

We implemented, for instance, the AR procedure with $N = 220$, $K = 4$, and $J = 5$, using Fisher’s $p$-values to control the type I error at 0.05. The resulting cutoff was $c = 0.064$, and the power at the LFC considered in the first row of Table I was 0.801. The procedure described in section 2 with $N = 220$ selected $\gamma (j/J) = 13.5 \times (j/J)^{2.75}$. Under the LFC, the resulting average proportion, across simulations, of randomizations to the effective arm, $p^*_E = 0.4$, and the control, $p^*_C = 0.2$, were 33.1% and 33.3% respectively, with the remaining average of 33.6% assigned to the three experimental arms with no treatment effect.

The main differences between AR and MAMS designs were also confirmed when we substituted the triangular method with the predictive power approach [33] for selecting the MAMS boundaries $(f_1, . . . , f_J)$. When we considered, for instance, the first scenario in Table I with the control of the type I error at 0.05 and redefined the futility boundaries the resulting average sample size under the LFC and $H_G$ across simulations become 250.1 and 200.6, respectively. Also in this case, the variability of the sample size across simulations is relevant, and the IQR is equal to 120 patients under the $H_G$ hypothesis and 65 under the LFC.

Next, we explored the operating characteristics of the AR and MAMS designs in row 1 of Table I under scenarios with more than one effective treatment. We selected these designs to control the FWER. We defined random variables $E$ and $I$ as the number of $H_0^{(k)}$ hypotheses that are correctly and erroneously rejected in a given trial. We estimated the probabilities that $E \geq 1$ and $I \geq 1$, the latter being the FWER, together with the expected value of $E$ and the expected number of treatment failures. We show the results in Table II.

Table II shows that the differences between the two designs are small for all criteria considered, with the exception of the ENF. However, MAMS designs do perform marginally better in terms of recommending effective treatments and avoiding recommending ineffective treatments. They tend to have a higher expected number of treatment failures, although as the number of effective treatments increases, this comparison is mainly driven by higher sample sizes under the MAMS designs. The fourth scenario, in which all experimental treatments are inferior to control, is the only one in which the MAMS design performs clearly better. The differences in Table II have an intuitive explanation—AR is conceived to sequentially allocate increasing resources to the comparison of the most promising experiential arm versus the control. But when two arms have similar treatment effects, there is a trade-off between the number of patients randomized to each of them. In contrast, under the MAMS design, the distribution of the number of patients randomized to an experimental arm, say arm 1, is not related to the treatment effects of the other $(K−1)$ experimental treatments.
Table II. Properties of five-arm five-stage multi-arm multi-stage and adaptive randomization trials for a range of response probability configurations.

| p                  | AR | MAMS | AR | MAMS | AR | MAMS | AR | MAMS | AR | MAMS |
|--------------------|----|------|----|------|----|------|----|------|----|------|
| (0.2,0.4,0.4,0.4,0.4,0.2) | 0.914 | 0.917 | 1.482 | 1.601 | 0.121 | 0.109 | 133.6 | 172.4 |
| (0.2,0.4,0.4,0.4,0.4,0.2) | 0.942 | 0.954 | 2.106 | 2.404 | 0.072 | 0.064 | 128.1 | 183.5 |
| (0.2,0.4,0.4,0.4,0.4,0.4,0.4) | 0.954 | 0.970 | 2.710 | 3.205 | 0 | 0 | 123.2 | 194.0 |
| (0.3,0.2,0.2,0.2,0.2,0.2) | 0 | 0 | 0 | 0 | 0.015 | 0.011 | 146.3 | 91.8 |

In the last scenario, the control group response probability is 0.2 and the response probability of each experimental treatment is independently simulated from a uniform U(0.1,0.4) distribution.

\( \mathbb{P}(E \geq 1) \), probability of a trial ending with recommendation of one or more effective treatments; \( \mathbb{E}(E) \), expected number of effective treatments recommended in a trial; \( \mathbb{P}(I \geq 1) \), probability of a trial ending with recommendation of one or more ineffective treatments; ENF, expected number of failures; AR, adaptive randomization; MAMS, multi-arm multi-stage.

5. Futility stopping rules for adaptive randomization designs

The results in Section 4 show the main weakness of the AR procedure; it performs worse than MAMS designs when all experimental treatments are ineffective. Also, ethical concerns arise when scenarios with all experimental arms having detrimental effects are realistic and need to be considered. This is because there is no futility rule that stops the trial if all experimental arms perform poorly in comparison with the control treatment.

We consider a stopping rule that terminates the trial earlier when none of the experimental arms has sufficient evidence to suggest a positive treatment effect. The stopping rule exploits the posterior probabilities of the events \( \{p_0 < p_k\} \) computed at each interim analysis. If all these posterior probabilities fall below a pre-specified threshold \( \ell_j \) at the \( j \)-th interim analysis, then the trial is stopped. We use thresholds defined by the equalities \( \ell_j = \lambda (j/J)^v \) for a suitable choice of \( \lambda \) and \( v \). Note that these additional thresholds do not add computational costs to the procedure in Section 2 for controlling either the type I error rate or the FWER. That is, the described stopping rule can be directly included in the algorithm for selecting the threshold \( c \) that is used at completion of the trial for reporting positive treatment effects.

Table III lists the operating characteristics (estimated from 10,000 replicates) that result from a variety of potential stopping rules. Some values of \( (\lambda, v) \), for example, \((0.6, 2)\), result in minimal loss in power and a reduction in the expected sample size under \( H_G \) of approximately 10 patients. In no cases does the stopping rule lead to an inflation in the type I error rate or FWER. We then considered again scenarios with all experimental arms having a detrimental effect and observed across simulations substantial
reduction of the average sample sizes. For instance, under the fourth scenario listed in Table II, by setting \((\lambda, v)\) equal to \((0.6, 2)\), the trial is stopped at the third and fourth interim analyses with probabilities equal to 0.25 and 0.52, respectively.

6. Comparisons based on a case study

The case study that we use to compare AR and MAMS designs is the NeoSphere trial reported by Gianni et al. [9]. We consider planning a multi-arm trial similar to Neosphere, which compares three experimental treatments to a control arm with a planned sample size of approximately 400 patients. Using the methods in Section 2, we first found AR and MAMS designs that controlled the FWER at 0.1, for \(J \in \{2, 3, 5, 10\}\) and \(K = 3\), assuming that there was no delay between patient recruitment and response evaluation. We chose the FWER of 0.1 as it is close to the FWER of the actual trial design, which controlled the type I error rate of each comparison at the 5% level (one sided). We chose the AR design so that its sample size \(N\) was 400, whereas the MAMS design was chosen so that its expected sample size under the LFC was as close to 400 patients as possible. The AR designs included a futility stopping rule, as described in Section 5, with parameters \(\lambda = 0.6\) and \(v = 2\). We then evaluated the operating characteristics of the designs assuming that the endpoint of pCR was observed 15 weeks after recruitment. We considered recruitment rates, in terms of mean number of patients recruited per week, equal to 2, 4, and 8.

We first investigated MAMS designs in which interim analyses took place once a pre-planned number of patients per arm had been recruited, as described in Section 2. Using the futility boundaries found in the no-delay case, the MAMS designs, under all scenarios with delayed responses, had both FWER and power considerably lower than the targeted FWER and power. For example, with \(J = 10\), \(p_0^* = 0.2\), \(p_1^* = 0.4\) and a mean of four patients per week, the estimated FWER and power were 0.059 and 0.436, respectively, compared with planned values of 0.1 and 0.65. This indicates that ignoring the delay between recruitment and assessment leads to relevant deviations from the planned operating characteristics of MAMS trials. The decrease in power is because the futility boundaries are found assuming that the interim analyses are equally spaced in terms of number of patients assessed. Under all scenarios with delayed responses, interim analyses take place when outcome data are available only for a subset of enrolled patients. We therefore chose to focus on MAMS designs with interim analyses shifted in time by approximately 15 weeks, exactly the time interval between each patient enrollment and the subsequent response assessment. The operating characteristics of the AR design were less sensitive to the delay between patient accrual and response evaluation. Therefore, the interim analyses of the AR design were not shifted, and at each interim analysis, the necessary posterior probabilities of the events \((p_0 < p_k)\) are updated using the available outcome data.

Table IV shows relevant operating characteristics of the designs under different recruitment rates. This simulation study suggested that the FWER was not affected by the described delays and, in all considered scenarios, was controlled at the 10% level as planned. The recruitment rate impacts the designs in different ways. The AR design loses some power as the recruitment rate increases, with a loss in power (compared with the no-delay case) of approximately 6–8% when a high recruitment rate is considered. If the recruitment rate is slow, that is, two patients per week, then only around 1–2% power is lost. The power of the MAMS design is not affected by the recruitment rate, but its expected sample size is affected considerably. Even with a slow recruitment rate, the expected sample size under \(H_G\) increases by approximately 15%.

Table 2 of [9] showed the actual results of the trial. The estimated pCR rates were 0.29, 0.458, 0.168, and 0.240 for arms A–D, respectively. With a FWER of 0.1 and a sample size of 400 patients, the power to conclude that arm B is significantly better than arm A is 0.776 if no interim analyses are used, assuming that these pCR estimates are the truth. We simulated data under this configuration of treatment effects with the designs used in Table IV and assuming a mean recruitment rate of four patients per week. Table 2 in the Supporting information shows the results. In this case, the power to recommend arm B is higher using AR (0.881 for AR, 0.839 for MAMS with \(J = 5\)), with the difference increasing in the number of interim analyses. On the other hand, the expected sample size of the MAMS design also decreases as \(J\) increases. Both methods lead to a notable decrease in the expected number of failures in the trial (284.4 if no interim analyses are used, 269.7 and 266.3 for AR and MAMS respectively with \(J = 5\)).

This case study illustrates that AR and MAMS are generally more efficient than a multi-arm trial without interim analyses. This gain in efficiency depends on the recruitment rate of the trial. Under AR, there is a gain in power, and on average, more patients are assigned to arm B. Although using a MAMS
Table IV. Operating characteristics of adaptive randomization and multi-arm multi-stage designs for case-study scenarios as the number of interim analysis and the mean number of patients recruited per week varies.

| $J$ | Recruitment rate | Power | ENF | ESS-$H_G$ | ESS-LFC |
|-----|------------------|-------|-----|-----------|---------|
|     |                  | AR    | MAMS | AR        | MAMS    | AR      | MAMS    |
| 2   | $0^+$            | 0.773 | 0.760| 278.0     | 281.2   | 393.6   | 334.9   | 399.6   | 398.6   |
| 2   | 2                | 0.759 | 0.757| 279.8     | 290.3   | 396.6   | 355.7   | 399.7   | 410.9   |
| 2   | 4                | 0.751 | 0.759| 281.6     | 299.6   | 398.2   | 375.3   | 399.8   | 423.4   |
| 2   | 8                | 0.734 | 0.758| 284.2     | 318.3   | 399.7   | 415.6   | 400.0   | 448.7   |
| 3   | $0^+$            | 0.789 | 0.790| 281.9     | 285.2   | 386.1   | 305.8   | 399.2   | 393.8   |
| 3   | 2                | 0.775 | 0.791| 281.4     | 303.9   | 395.1   | 379.9   | 399.6   | 438.2   |
| 3   | 4                | 0.755 | 0.791| 281.7     | 309.3   | 391.5   | 351.1   | 399.5   | 421.9   |
| 3   | 8                | 0.723 | 0.790| 282.8     | 332.3   | 399.7   | 438.2   | 400.0   | 469.4   |
| 5   | $0^+$            | 0.820 | 0.783| 279.8     | 275.6   | 377.8   | 290.3   | 398.9   | 393.1   |
| 5   | 2                | 0.811 | 0.782| 279.6     | 297.3   | 391.5   | 351.1   | 399.5   | 411.9   |
| 5   | 4                | 0.789 | 0.782| 280.4     | 303.7   | 390.6   | 379.9   | 399.1   | 431.2   |
| 5   | 8                | 0.754 | 0.786| 282.4     | 323.2   | 396.7   | 416.5   | 399.7   | 464.9   |
| 10  | $0^+$            | 0.842 | 0.774| 278.5     | 259.6   | 366.7   | 259.7   | 398.3   | 371.2   |
| 10  | 2                | 0.823 | 0.772| 278.8     | 274.9   | 377.7   | 297.3   | 398.6   | 392.1   |
| 10  | 4                | 0.807 | 0.773| 279.3     | 290.3   | 385.0   | 333.9   | 398.9   | 413.0   |
| 10  | 8                | 0.768 | 0.781| 281.9     | 318.2   | 394.3   | 399.5   | 399.4   | 451.2   |

Scenario parameters for the LFC were $p_0^* = 0.25$, $p_1^* = 0.4$ and target FWER equal to 0.1. Recruitment rate of $0^+$ refers to the case where there is no delay between a patient being recruited and assessed for treatment response. FWER, familywise error rate under the $H_G$ hypothesis; ENF, expected number of failures; ESS-$H_G$, expected sample size of the trial under the global null hypothesis $H_G$; ESS-LFC, expected sample size of the trial under the least favorable configuration; AR, adaptive randomization; MAMS, multi-arm multi-stage.

Design led to a loss in power, compared with AR, there was also a slight drop in the expected number of patients recruited to the trial. We also considered setting the MAMS design maximum sample size to be as close as possible to 400. Table 1 in the Supporting information shows these results. In this case, the MAMS design has considerably lower power and expected sample size compared with AR.

At completion of group sequential or adaptive trials, the investigator reports the arm specific response probabilities and the treatment effects differences between the control and the investigated agents. In most cases, maximum likelihood estimation is used, or equivalently, when the primary outcome is binary, the number of responses and failures are summarized in a table. It is known that the maximum likelihood estimates (MLEs) following group sequential or adaptive experiments can be biased; that is, the expected values of the estimates can be substantially different from the true parameters. Several correction strategies have been proposed but are rarely used in the medical literature. We assessed these discrepancies with 100,000 Monte Carlo iterations under the scenario tailored to the NeoSphere trial after having selected AR and MAMS designs with a power of 80% at $(p_0^* = 0.2, p_1^* = 0.4)$ and FWER controlled at 0.2. The bias of the empirical estimates produced at completion of the MAMS trial results larger than the bias under the AR design. The differences $(\hat{p}_0 - p_j) - E(\hat{p}_0 - p_j)$, with $j = 1, 2, 3$, between the true treatment effects and the estimates under the MAMS (AR) design are equal to $-0.036(-0.001)$, $-0.037(-0.011)$, and $-0.041(-0.016)$, respectively. These results suggests that AR can be considerably less impacted by this problem of bias. AR designs have been shown to result in asymptotically consistent and normally distributed treatment effect MLEs under mild conditions [39]. The main results in [39] can be applied to our AR scheme. For small sample sizes, however, the bias of MLEs is still noticeable for non-effective arms.

### 7. Discussion

When several new agents or combinations are at the same stage of clinical development, a multi-arm trial provides efficiency gains over separate trials of each experimental agent against a control treatment.
Two major classes of designs, which can add further efficiency, are AR and MAMS designs. In this paper, we discussed designing AR and MAMS trials and compared their properties. Our study compared the two approaches using easily interpretable operating characteristics such as power, overall sample size, sample size variability, and expected number of treatment failures.

Both AR and MAMS designs provide efficiency advantages over multi-arm trials without interim analyses. Efficiency can be in terms of extra power or a reduced sample size required for a fixed power. Although we did not explicitly consider the trial duration, a reduction in expected sample size will also lead to a reduction of the expected time to complete the study. The relationship depends on the accrual rate and, in some cases, is assumed linear. In practice, the recruitment rate can vary during the trial, and the proportion of time saved by AR or MAMS may slightly deviate from the expected sample size reduction. The relative advantages of both methods depend on the accrual rate and the delay between individual recruitment and outcome assessment. Parmar et al. [40] motivated the use of MAMS designs in oncology but stated that they should only be used if the final endpoint, or some suitable intermediate endpoint, is observed relatively quickly. The same is true for AR designs, and a phase II trial where the final endpoint is observed quickly is the ideal setting for using the designs discussed in this paper.

In several settings, and in particular in confirmatory studies, the investigator has to preform frequentist hypothesis testing with a stringent control of the type I error rate. For group sequential studies, there are several well-established methods to do this, and the performances and robustness of these procedures are well documented in the literature. Under adaptive designs, the control of the type I error rate, in general, is more challenging. The use of standard Monte Carlo simulations under a few representative scenarios is unsatisfactory. An attempt to control the type I error rate with simulations can be insufficient for regulatory agencies. Our application of the importance sampling method has the goal of controlling the probability of a type I error under the complete set of possible arm-specific response probabilities. We do not claim that this is a general solution for any adaptive trial. In our setting, with multi-arm trials and binary outcomes, we assessed the algorithm. This evaluation included independent Monte Carlo simulations. Overall, the proposed computational procedure matches the nominal $\alpha$ level and the actual control of the type I error rate. More generally, the study of computational strategies applicable to the control of the type I error in adaptive trials can facilitate the use of adaptive designs.

The results in this paper show that AR designs are often more efficient than MAMS designs when there is a single effective experimental treatment. With multiple effective experimental treatments, the differences between the two approaches become less evident. Also, when there are no effective experimental treatments in the trial, the MAMS design tends to have a lower expected sample size, but including a futility rule in the AR design narrows this gap. An additional advantage of AR designs is a lower sensitivity of the operating characteristics when the number of assessed outcomes at each interim analysis is different from the originally planned number.

Adaptive randomization designs allow combining a fixed overall sample size with easy to interpret rules for dropping experimental arms; this may be convenient for planning the trial. The MAMS approach could be modified to have a fixed sample size by fixing the number of patients recruited at each stage (as opposed to a fixed number of patients recruited per arm per stage). Similar to the AR design, the sample size is fixed, unless all experimental arms are dropped at some interim analysis. The stopping boundaries should be adequately chosen, possibly with different values depending on the number of treatments remaining in the trial after each interim analysis, but despite this complication, it may be a useful design to consider in future work.

Adaptive randomization has previously been criticized as being less efficient than a fixed randomization approach (see for example, Korn and Freidlin [41]), and indeed, this is true when a control is compared with a single experimental treatment. However, the results in this paper emphasize that controlling the allocation to the control group and applying AR to the experimental groups, both the power and the rate of positive outcomes across the enrolled patients, can be improved compared with more standard designs with fixed randomization probabilities.

We have assumed that the different arms represent separate treatments as opposed to different doses or different combinations of treatments (with overlap of treatments between arms). If the arms are different doses or combinations, then the approach we have used is valid, but not necessarily the most efficient. Regression models can be used to improve AR and MAMS designs. The AR design could, for instance, integrate a dose-response model for estimating sequentially the effectiveness of different doses. This could potentially improve the allocation procedure. Similar considerations apply when the arms are combinations of agents (with treatments in common): prior models that include the possibility of synergistic effects and agent-specific effects can be considered. Analogous remarks hold for MAMS.
designs; additionally, more complex decision rules on dropping arms can be used, such as only dropping a higher dose for futility if the lower dose has been dropped.

When we considered MAMS designs, we were forced to use interim analyses taking place after a pre-specified number of outcomes had been assessed on each arm. This is not always a practical solution, for example, if there is relevant variability in the time intervals between individual accrual and outcome assessment. In order to obtain the same flexibility of the AR designs, one would need MAMS designs with non-deterministic futility boundaries defined by empirical spending functions (see, for example, [42]). This would allow adaption of the futility boundary according to the number of available outcomes. To our knowledge, this approach has not been applied to trials with multiple experimental arms. Overall, we observed that, with delay between recruitment and response, an increased recruitment rate leads to AR designs having lower power and MAMS designs having higher expected sample size.

We focused on the statistical properties of MAMS and AR designs, but there are other considerations when choosing a design. Both AR and MAMS designs will cause operational issues [43]. Examples of such issues are managing the randomization and changing the allocation ratios after an interim analysis. Both designs will cause problems with the supply of drugs/treatments (as it is not known in advance how many patients will be on each arm); AR could cause greater problems as there is scope for greater deviation from equal randomization. Blinding treatment assignment is another important aspect that needs careful consideration. When blinding is to be maintained, it is important to have properly maintained firewalls between the statistician conducting the interim analyses and other investigators.

A more complex setting in which AR designs have been proposed consists of trials evaluating the effects of several treatments on patients from multiple biomarker subgroups. Examples include the BATTLE [8] and iSPY2 [11] trials. In addition, MAMS approaches have been used to design the FOCUS-4 trial, which will test several treatments in biomarker subgroups. A comparison of AR and MAMS designs in this context would be useful and is an area for future research.

Acknowledgements

This work is sponsored by MRC (grant numbers G0800860 and MR/J004979/1). We thank the reviewers for useful comments that helped improve the paper.

References

1. Reed J. Toward a new era in cancer treatment: message from the new editor-in-chief. *Molecular Cancer Therapeutics* 2012; 11:1621–1622.
2. Schilsky R. Personalized medicine in oncology: the future is now. *Nature Reviews Drug Discovery* 2010; 9(5):363–366.
3. Berry DA. Adaptive clinical trials in oncology. *Nature Reviews Clinical Oncology* 2011; 9:199–207.
4. Barthel F, Parmar M, Royston P. How do multi-stage, multi-arm trials compare to the traditional two-arm parallel group design—a reanalysis of 4 trials. *Trials* 2009; 20(21).
5. Wason J, Jaki T. Optimal design of multi-arm multi-stage trials. *Statistics in Medicine* 2012; 31:4269–4279.
6. Freidlin B, Korn EL, Gray R, Martin A. Multi-arm clinical trials of new agents: some design considerations. *Clinical Cancer Research* 2008; 14(14):4368–71.
7. Trippa L, Lee EK, 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.
8. Kim ES, Herbst RS, Wistuba II, Lee JJ, Blumenschein GR, Tsao AS, David J, Hicks ME, Erasmus J, Gupta S, Alden CM, Liu S, Tang X, Khuri FR, Tran HT, Johnson BE, Heymach JV, Mao L, Fossella F, Kies MS, Papadimitrakopoulou V, Davis SE, Lippman SM, Hong WK. The BATTLE trial: personalizing therapy for lung cancer. *Cancer Discovery* 2011; 1(1):44–53.
9. Gianni L, Pienkowski T, Im YH, Roman L, Tseng LM, Liu MC, Lluch A, Staroslawska E, de la Haba-Rodriguez J, Im SA, Pedrini JL, Poirier B, Morandi P, Semiglazov V, Srimuninnimit V, Bianchi G, Szado T, Ratnayake J, Ross G, Valagussa P. Efficacy and safety of neoadjuvant pertuzumab and trastuzumab in women with locally advanced, inflammatory, or early HER2-positive breast cancer (neosphere): a randomised multicentre, open-label, phase 2 trial. *The Lancet Oncology* 2011; 13(1):44–53.
10. Giles FJ, Kantarjian HM, Cortes JE, Garcia-Manero G, Verstovsek S, Faderl S, Thomas DA, Ferrajoli A, O’Brien S, Wathen JK, Xiao LC, Berry DA, Estey EH. Adaptive randomized study of idarubicin and cytarabine versus troxatilabine and cytarabine versus troxatilabine and idarubicin in untreated patients 50 years or older with adverse karyotype acute myeloid leukemia. *Journal of Clinical Oncology* 2003; 21(9):1722–1727.
11. Barker A, Sigman K, Kellogg G, Hylton N, Berry D, Esserman L. i-spy 2: an adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy. *Clinical Pharmacology & Therapeutics* 2009; 86(1):97–100.
12. Dumville J, Hahn S, Miles J, Togerson D. The use of unequal randomisation ratios in clinical trials: a review. *Contemporary Clinical Trials* 2006; 27:1–12.
13. Sydes MR, Parmar MK, James ND, Clarke NW, Dearmaile DP, Mason MD, Morgan RC, Sanders K, Royston P. Issues in applying multi-arm multi-stage methodology to a clinical trial in prostate cancer: the MRC STAMPEDE trial. Trials 2009; 10. DOI: 10.1186/1745-6215-10-39.

14. Ledford H. Translational research: 4 ways to fix the clinical trial. Nature 2011; 477(7366):526–528.

15. Thall P, Wathen J. Practical Bayesian adaptive randomisation in clinical trials. European Journal of Cancer 2007; 43(5):859–866.

16. Lee J, Gu X, Liu S. Bayesian adaptive randomization designs for targeted agent development. Clinical Trials 2010; 7(5):584–596.

17. Berry S, Carlin B, Lee J, Mülter P. Bayesian Adaptive Methods for Clinical Trials, Vol. 38. CRC Press Llc: Boca Raton, Florida, 2010.

18. Yuan Y, Yin G. On the usefulness of outcome-adaptive randomization. Journal of Clinical Oncology 2011; 29:e390–e392.

19. Alexander B, Wen P, Trippa L, Rearson D, Yung W, Parmigiani G, Berry D. Interim analysis in clinical trials: the role of the likelihood principle. Statistics in Medicine 2013; 32:2239–2256. DOI: 10.1002/sim.1430.

20. Hu F, Rosenberger W. The Theory of Response-Adaptive Randomization in Clinical Trials, Vol. 525. Wiley-Interscience: Hoboken NJ, 2006.

21. Coad D, Rosenberger W. A comparison of the randomized play-the-winner rule and the triangular test for clinical trials with binary responses. Statistics in Medicine 1999; 18:761–769.

22. Wason J, Magirr D, Law M, Jaki T. Some recommendations for multi-arm multi-stage trials. Accepted by Statistical Methods in Medical Research 2013.

23. Lee JJ. Demystify statistical significance: time to move on from the p value to bayesian analysis. Journal of the National Cancer Institute 2013; 105(1):2–3.

24. Berry DA. A case for Bayesianism in clinical trials. Journal of the National Cancer Institute 2003; 95(2):111–122.

25. Magirr D, Jaki T, Whitehead J. A generalized Dunnett test for multiarm-multistage clinical studies with treatment selection. Biometrika 2012; 99:494–501. DOI: 10.1093/biomet/ass002.

26. Whitehead J, Stratton I. Group sequential clinical trials with triangular continuation regions. Biometrics 1983; 39:227–236.

27. Whitehead J, Matsushita T. Stopping clinical trials because of treatment ineffectiveness: a comparison of a futility design with a method of stochastic curtailment. Statistics in Medicine 2003; 22(5):677–687.

28. Wason J, Mander A, Thompson S. Optimal multi-stage designs for randomised clinical trials with continuous outcomes. Statistics in Medicine 2012; 31:301–312. DOI: 10.1002/sim.4421.

29. Dunnett C. Selection of the best treatment in comparison to a control with an application to a medical trial. In Design of Experiments: Ranking and Selection, Santner T, Tamhane A (eds). Marcel Dekker: New York, 1984; 47–66.

30. Whitehead J, Stratton I. Group sequential clinical trials with triangular continuation regions. Biometrics 1983; 39:227–236.

31. Whitehead J, Matsushita T. Stopping clinical trials because of treatment ineffectiveness: a comparison of a futility design with a method of stochastic curtailment. Statistics in Medicine 2003; 22(5):677–687.

32. Wason J, Mander A, Thompson S. Optimal multi-stage designs for randomised clinical trials with continuous outcomes. Statistics in Medicine 2012; 31:301–312. DOI: 10.1002/sim.4421.

33. Betensky RA. Alternative derivations of a rule for early stopping in favor of h 0. The American Statistician 2000; 54(1):35–39.

34. Flournoy N. Adaptive Designs. Encyclopedia of Environmetrics. Wiley: New York NY, 2006.

35. Muliere P, Maria Paganoni A, Secchi P. A randomly reinforced urn. Journal of Statistical Planning and Inference 2006; 136(6):1853–1874.

36. Cohen J, McClure S, Angela J. Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. Philosophical Transactions of the Royal Society B: Biological Sciences 2007; 362:933–942.

37. Ross J, Slodkowska E, Symmans W, Pusztai L, Ravdin P, Hortobagyi G. The HER-2 receptor and breast cancer: ten years of targeted anti-HER-2 therapy and personalized medicine. Oncologist 2009; 14:320–368.

38. Wason J, Magirr D, Law M, Jaki T. Some recommendations for multi-arm multi-stage trials. Accepted by Statistical Methods in Medical Research 2013.

39. Melfi V, Page C. Estimation after adaptive allocation. Journal of Statistical Planning and Inference 2000; 87:353–363.

40. Parmar MKB, Barthel FMS, Sydes M, Langley R, Kaplan R, Eisenhauer E, Brady M, James N, Bookman MA, Swart AM, Qian W, Royston P. Speeding up the evaluation of new agents in cancer. Journal of the National Cancer Institute 2008; 100:1204–1214.

41. Korn E, Freidlin B. Outcome-adaptive randomization: is it useful? Journal of Clinical Oncology 2011; 29:771–776.

42. Pampallona S, Tsiatis A, Kim K. Interim monitoring of group sequential trials using spending functions for the type I and type II error probabilities. Therapeutic Innovation & Regulatory Science 2001; 35:1113–1121.

43. Gallo P. Operational challenges in adaptive design implementation. Pharmaceutical Statistics 2006; 5:119–124.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher’s web site.