Received: 9 September 2022 | Revised: 16 January 2024 | Accepted: 15 April 2024

DOI: 10.1002/sim.10090

RESEARCH ARTICLE

Statistics in Medicine WILEY

Adding experimental treatment arms to multi-arm multi-stage platform trials in progress

Thomas Burnett[1] | **Franz König**[2] | **Thomas Jaki**[3,4]

[1]Department of Mathematical Sciences, University of Bath, Bath, UK

[2]Center for Medical Data Science, Medical University of Vienna, Vienna, Austria

[3]MRC Biostatistics Unit, University of Cambridge, Cambridge, UK

[4]Faculty of Computer Science and Data Science, University of Regensburg, Regensburg, Germany

Correspondence
Franz König, Center for Medical Data Science, Medical University of Vienna, Spitalgasse 23, Vienna 1090, Austria.
Email: franz.koenig@meduniwien.ac.at

Funding information
Innovative Medicines Initiative 2 Joint Undertaking, Grant/Award Number: 853966; National Institute for Health and Care Research, Grant/Award Number: NIHR-SRF-2015-08-001; NIHR Biomedical Research Centre, Grant/Award Number: BRC-1215-20014; Medical Research Council, Grant/Award Numbers: MC_UU_00002/14, MC_UU_00040/03, MR/V038419/1

Multi-arm multi-stage (MAMS) platform trials efficiently compare several treatments with a common control arm. Crucially MAMS designs allow for adjustment for multiplicity if required. If for example, the active treatment arms in a clinical trial relate to different dose levels or different routes of administration of a drug, the strict control of the family-wise error rate (FWER) is paramount. Suppose a further treatment becomes available, it is desirable to add this to the trial already in progress; to access both the practical and statistical benefits of the MAMS design. In any setting where control of the error rate is required, we must add corresponding hypotheses without compromising the validity of the testing procedure.To strongly control the FWER, MAMS designs use pre-planned decision rules that determine the recruitment of the next stage of the trial based on the available data. The addition of a treatment arm presents an unplanned change to the design that we must account for in the testing procedure. We demonstrate the use of the conditional error approach to add hypotheses to any testing procedure that strongly controls the FWER. We use this framework to add treatments to a MAMS trial in progress. Simulations illustrate the possible characteristics of such procedures.

KEYWORDS
adaptive designs, conditional error, design modification, multi-arm multi-stage

1 | INTRODUCTION

It is common to have several competing treatments during the clinical development process. These may be different doses of the same drug or entirely different treatment regimes. Jaki and Hampson[1] note that, given the high failure rate and cost of Phase III trials, it is key to give careful consideration to which treatments we investigate. Multi-arm multi-stage trials (MAMS)[2-4] compare several experimental treatments with a common control allowing for the efficient selection of appropriate treatments.[5]

MAMS trials reduce the expected number of patients by dropping treatments that accruing data suggest are ineffective or stopping the trial if the data demonstrate efficacy. Given the multiple hypotheses and highly adaptive nature of the design, MAMS studies require specialist testing methodology to control the error rate of the trial.[6] Magirr et al[7] introduce the generalised Dunnett family of tests. They define group sequential testing boundaries to account for the multiple

analyses and the correlation structure introduced by the comparison of several experimental arms to a common control;[8] Urach and Posch[9] extend this, directly defining all elements of the closed testing procedure.[10] Alternatively, fully flexible testing methods[11-15] allow decisions about which arms should remain in the study to function separately from the hypothesis testing. Both group sequential and fully flexible methods require the pre-definition of all study hypotheses, constructing the overall testing procedure to give strong control of the family-wise error rate (FWER).[16] Especially regulatory guidance documents stress that strict control of the FWER is required for confirmatory (pivotal) trials,[17-19] for example, for adaptive (seamless) designs with data-dependent selection and addition of treatments in (un-)planned interim analyses.[20-22]

It is possible that not all experimental treatments are available at the start of the trial, for example, see the STAMPEDE trial.[23] STAMPEDE started with five comparisons and subsequently added several more to the protocol. For further examples of trials comparing multiple experimental treatments see sect. 5 of the work of Bauer et al[24] or the discussion in the adaptive designs CONSORT extension, Dimairo et al.[25] Adding further treatments to the trial in progress maintains the benefits of a MAMS design, reducing logistical and administrative effort, speeding up the overall development process,[26] adding efficiency through the multiple comparisons and allowing direct comparisons of the treatments within the same trial. Though there seem some consensus arising that strong control of FWER is not required in such platform trials, particularly when the statistical claims are independent.[27] The question whether a strict control of the FWER is needed, depends on the consequences of erroneous rejections.[28] Especially in multi-armed trials where the treatment arms correspond to different dose levels, the hypotheses to be tested are not independent and therefore appropriate multiplicity adjustments such as (group-sequential) Dunnett tests[7,8,29] are needed. When adding new dose levels in an already ongoing trials, the multiple testing strategy and hypotheses to be tested needs to be adapted accordingly to control the FWER.

When adding treatments to a trial that strongly controls the FWER we must also add the corresponding hypotheses. If no data have been observed (including a requirement that there have been no interim analyses) we may adjust the pre-planned testing structure to incorporate the additional hypotheses. Bennett and Mander[30] demonstrate how to suitably adjust the sample size for each treatment arm for such additions. However, what if we wish to add treatments (and corresponding hypotheses) after an interim analysis? Our methods allow modification of the trial design in the presence of observed data. Furthermore, this modification of the trial design may be done without specifying when and how to add new treatments. In principle, any internal and external data could be utilised for this decision-making. The only restriction is that the trial has not been stopped before the addition of further experimental treatments, for example, when all treatments have to be stopped due to binding futility rules.

The conditional error approach[31] allows for design modifications during a trial, where these modifications have not been pre-planned. These modifications may be accounted for in the setting of treatment selection[14,32] however, as noted by Hommel[33] adding hypotheses to a testing framework requires further restrictions. We build a general framework using these principles, which allows the inclusion of additional hypotheses to any testing procedure that strongly controls the FWER and the use of existing trial information in this testing procedure. Though depending on the conditional error for some of the intersection procedures, adding a new treatment arm might not be the most efficient strategy compared to opening a new trial for the new treatment arm. We discuss which type of power can be of interest in this type of designs. We show how to apply the conditional error principle in the setting of MAMS designs, demonstrating how to construct an appropriate hypothesis testing structure for the updated trial.

2 | ALTERING A TRIAL IN PROGRESS

2.1 | A two arm trial

Suppose we plan a two-arm trial with a continuous outcome to compare a new treatment, T_1, and a control, T_0. Let μ_1 and μ_0 be the expected responses for patients on treatments T_1 and T_0 respectively, and define the treatment effect to be $\theta_1 = \mu_1 - \mu_0$. We investigate the null hypothesis $H_{01} : \theta_1 \leq 0$ versus the one sided alternative $H_{11} : \theta_1 > 0$.

We plan a trial recruiting a total of n patients. These patients are randomised equally between treatment and control. We collect observations $X_{i,k} \sim N(\mu_k, \sigma^2)$ for $i = 1, \ldots, n/2$ and $k = 0, 1$. Let $\hat{\theta}_1$ denote the estimate of the treatment effect

corresponding to these observations. Defining $\xi_1 = \frac{\theta_1\sqrt{n}}{2\sigma}$ this has corresponding Z-statistic,

$$Z_1 = \frac{\hat{\theta}_1\sqrt{n}}{2\sigma} \sim N(\xi_1, 1).$$

We reject H_{01} at level α when $Z_1 > \Phi^{-1}(1-\alpha)$, where Φ is the standard normal CDF.

2.2 | Adding a treatment

Suppose for some $\tau \in (0,1)$ after τn observations we wish to add a second experimental treatment, T_2, to the trial. Let μ_2 be the expected response for patients receiving this new treatment and define the corresponding treatment effect to be $\theta_2 = \mu_2 - \mu_0$. We add the corresponding null hypothesis $H_{02} : \theta_2 \leq 0$ versus the corresponding one-sided alternative $H_{12} : \theta_2 > 0$.

While not strictly necessary for what follows, for illustration we maintain the pre-planned elements of the trial concerning treatments T_1 and T_0, such as the same sample size per treatment. It is useful to consider the trial in two stages, stage 1 and stage 2 consisting of the patients recruited before and after the treatment is added. From the stage 1 data we find the Z-statistic

$$Z_1^{(1)} \sim N(\xi_1\sqrt{\tau}, 1)$$

and from the stage 2 data we find the Z-statistic

$$Z_1^{(2)} \sim N(\xi_1\sqrt{1-\tau}, 1).$$

The overall Z-statistic (that would be found from the pooled date) may be reconstructed from the stage-wise Z-statistics

$$Z_1 = \sqrt{\tau}Z_1^{(1)} + \sqrt{1-\tau}Z_1^{(2)}.$$

We may choose to recruit to T_2 as we please (indeed the choice of how to recruit the remainder of the trial is something that should be carefully considered in practice). Suppose again for illustration we maintain equal randomisation to each treatment, recruiting $(1-\tau)n/2$ patients to T_2 in stage 2. Since T_2 is added to the trial for stage 2 the estimate of the treatment effect $\hat{\theta}_2$ must be based only on the stage 2 data (in particular note that this means only control patients recruited in stage 2 are used). Thus for $\xi_2 = \frac{\theta_2\sqrt{n}}{2\sigma}$ we find the corresponding Z-statistics

$$Z_2 \sim N(\xi_2\sqrt{1-\tau}, 1).$$

Due to the common control and equal randomisation $Z_1^{(2)}$ and Z_2 have a known correlation $1/2$. We note that if we were to adjust the recruitment to each treatment group this correlation would change but still remain known due to the common control patients.

2.3 | Hypothesis testing

In Section 2.1, we construct the hypothesis test to ensure a pre-specified type I error rate. When adding H_{02} it is natural to extend this principle of error control and require strong control of the family-wise error rate (FWER). Let R be the event that that we reject one or more true null hypothesis then we achieve strong control of the FWER at level α when

$$P_{\boldsymbol{\theta}}(R) \leq \alpha \text{ for all } \boldsymbol{\theta} = (\theta_1, \theta_2). \tag{1}$$

Suppose we add T_2 at $\tau = 0.5$ and we test each null hypotheses at a nominal level $\alpha = 0.05$, the FWER in this case is 0.09 (Figure S1 in the supplementary material shows a similar inflation for all values of τ). Sugitani et al[34] propose methods

that account for the introduction of the additional hypothesis, testing any introduced hypothesis based strictly on the data collected after their introduction at level α.[33] We build on this approach, adjusting for multiplicity when adding hypotheses and incorporating existing information where possible.

We construct a closed testing procedure[10] accounting for the adaptive nature of the trial within each test.[14] We require level α tests of H_{01}, H_{02} and $H_{0,12} = H_{01} \cap H_{02} : \theta_1 \leq 0 \ \& \theta_2 \leq 0$ (we shall refer to these as local tests). We reject H_{01} globally when the local level α tests of H_{01} and $H_{0,12}$ are rejected. Similarly we reject H_{02} globally when the local level α tests of H_{02} and $H_{0,12}$ are rejected. Such a procedure ensures strong control of the FWER at leve α.

We have not considered any change regarding the local test of H_{01}, so we reject H_{01} locally when $Z_1 > \Phi^{-1}(1-\alpha)$. However, it is useful to discuss constructing this test using the conditional error principle.[31] Given stage 1 observation $z_1^{(1)}$ we define the conditional error rate,

$$A(z_1^{(1)}) = P_{\theta_1=0}(\text{Reject } H_{01} | Z_1^{(1)} = z_1^{(1)}).$$

The probability of subsequently rejecting H_{01} must not exceed $A(z_1^{(1)})$. Thus in stage 2 we locally reject H_{01} when $Z_1^{(2)} > \Phi^{-1}(1 - A(z_1^{(1)}))$. Let $f(z_1^{(1)})$ be the probability density function of $z_1^{(1)}$, under H_{01} we have that

$$P_{\theta_1=0}(\text{Reject } H_{01}) = \int_{z_1^{(1)}} f(z_1^{(1)}) A(z_1^{(1)}) \mathrm{d}z_1^{(1)} = \alpha. \tag{2}$$

Thus the local test of H_{01} is constructed at the pre-specified α as required.

There is no existing for H_{02} and so this test must be constructed at level α and based only on stage 2 observations. Thus we locally reject the test for H_{02} when $Z_2 > \Phi^{-1}(1-\alpha)$.

While there was no planned test for $H_{0,12}$, there is pre-existing information for H_{01} in the form of $Z_1^{(1)}$. Hommel[33] notes such stage 1 data may be used in the test by considering some initially excluded hypotheses. We apply this concept to the test of $H_{0,12}$. Clearly $H_{0,12}$ implies H_{01}. Under H_{01} we compute the conditional error rate $A(z_1^{(1)})$ as described previously, where under under H_{01} $z_1^{(1)}$ is distributed such that Equation (2) holds as before. Thus we may construct the local test of $H_{0,12}$ at level $A(z_1^{(1)})$ allowing for the incorporation of the stage one data given by $Z_1^{(1)}$.

To test $H_{0,12}$ we use a Dunnett test.[8] Let

$$Z_D = max(Z_1^{(2)}, Z_2)$$

and define the distribution

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1 \end{pmatrix} \right).$$

We construct the p-value,

$$P_D = P(X > Z_D \cup Y > Z_D)$$

and locally reject $H_{0,12}$ when $P_D < A(Z_1^{(1)})$.

The number of patients assigned to each treatment is free to vary provided the local tests of H_{01} and $H_{0,12}$ are constructed based on the stage 2 data and tested at $A(z_1^{(1)})$. Alternatively one could base the test on a combination across stages, such as the weighted inverse normal.[35-37] Critically if the ratio of patients allocated to each treatment differs between stages 1 and 2 the test may not be based on the pooled data.

2.4 | Simulation study

We assess the operating characteristics of the Dunnett-based closed testing procedure proposed in Section 2.3 via simulation. We compare this with a closed testing procedure with a local test for $H_{0,12}$ based only on evidence for H_{01}, that is we reject both H_{01} and $H_{0,12}$ locally when $Z_1 > \Phi^{-1}(1-\alpha)$. This alternative test is a gate keeping procedure[38] written in

TABLE 1 Rejection criteria for each of the local tests under the Dunnett-type and gate keeping testing procedures.

	Rejection criteria for local hypotheses		
Testing procedure	H_{01}	H_{02}	$H_{0,12}$
Dunnett type	$Z_1 > \Phi^{-1}(1-\alpha)$	$Z_2 > \Phi^{-1}(1-\alpha)$	$\max\left(Z_1^{(2)}, Z_2\right) > \Phi^{-1}\left(1 - A\left(Z_1^{(1)}\right)\right)$
Gate keeping	$Z_1 > \Phi^{-1}(1-\alpha)$	$Z_2 > \Phi^{-1}(1-\alpha)$	$Z_1 > \Phi^{-1}(1-\alpha)$

TABLE 2 Rejection probabilities of local and global hypothesis tests under each testing procedure.

		Rejection of local tests			Rejection of global tests			
ξ_1	ξ_2	$P(L_{01})$	$P(L_{02})$	$P(L_{0,12})$	$P(G_{01})$	$P(G_{02})$	$P(G_B)$	$P(G_A)$
Dunnett procedure for testing the intersection hypothesis								
0	0	**0.05**	**0.05**	**0.05**	**0.03**	**0.01**	**0.01**	**0.05**
δ	0	0.90	**0.05**	0.86	0.81	**0.00**	**0.05**	0.86
0	δ	**0.05**	0.66	0.36	**0.00**	0.29	**0.04**	0.33
δ	δ	0.90	0.66	0.92	0.26	0.03	0.62	0.91
Gate keeping procedure for testing the intersection hypothesis								
0	0	**0.05**	**0.05**	**0.05**	**0.04**	NA	**0.01**	**0.05**
δ	0	0.90	**0.05**	0.86	0.85	NA	**0.05**	0.90
0	δ	**0.05**	0.66	0.36	**0.01**	NA	**0.04**	0.05
δ	δ	0.90	0.66	0.90	0.28	NA	0.62	0.90

Note: Error rates highlighted in bold, $\delta = \Phi^{-1}(0.95) + \Phi^{-1}(0.9)$ such that we have power of 0.9 when testing H_{01} in the original trial. The events L_{01}, L_{02} and $L_{0,12}$ are the events that we locally reject H_{01}, H_{02} and $H_{0,12}$ respectively. Similarly the events G_{01}, G_{02}, G_B and G_A are the events that we globally reject only H_{01}, only H_{02}, both or any null hypothesis respectively. Note for the gate keeping procedure it is not possible to test only H_{02}, this is reflected in the table by the NA for $P(G_{02})$, though one may interpret this probability as zero.

the form of a closed testing procedure, the test for $H_{0,12}$ has the correct error rate of α since $H_{0,12}$ implies H_{01}. Further to this both procedures use the existing data $Z_1^{(1)}$ to test $H_{0,12}$ also by the argument that $H_{0,12}$ implies H_{01}. Table 1 shows the rejection criteria for the local test under each testing procedure, this shows that the difference in the procedures is entirely driven by the test of the intersection hypothesis.

For combinations (ξ_1, ξ_2), with $\sigma/\sqrt{n} = 1$, $\delta = \Phi^{-1}(0.95) + \Phi^{-1}(0.9)$ and $\tau = 0.5$ and strong control of the FWER at $\alpha = 0.05$ we simulate 1,000,000 realisations of Z_1 and Z_2 assuming equal sample size in each treatment at each stage in R.[39]

Table 2 shows the local and global rejection probabilities for the null hypotheses for each testing method. As required both procedures control error rate of all local tests at 0.05 and thus both strongly control the FWER.

We see that both procedures perform similarly when H_{01} is false. The most notable difference is the scenario where H_{01} is true and H_{02} is false. Since, by construction, the gate keeping procedure cannot produce a scenario where only H_{02} is rejected, we note that all of these values are not shown in Table 2. Thus the gate keeping procedure cannot reject H_{02} without making an error by rejecting H_{01} locally and the Dunnett-based procedure increases the probability of rejecting H_{02} by 0.29 (by being able to reject H_{02} only globally. This motivates our use of Dunnett-type procedures in Section 4.1.

In Figure 1, we explore the probabilities of rejecting the intersection hypothesis $H_{0,12}$ for all combinations of H_{01} and H_{02} true and false. When H_{01} is false the conditional error is likely to be higher than the pre-planned α, giving a high chance of rejecting $H_{0,12}$; when H_{01} is true and H_{02} is false there is a small reduction in the probability of rejecting $H_{0,12}$, this explains why the gate keeping procedure performs slight better when $\xi_1 = \delta$ and $\xi_2 = 0$. Conversely when H_{01} is true the conditional error is likely to be quite low: when both null hypotheses are true this corresponds to a low probability of rejecting $H_{0,12}$ however, when H_{02} is false we recover some possibility of rejecting $H_{0,12}$ allowing us to reject H_{02} globally.

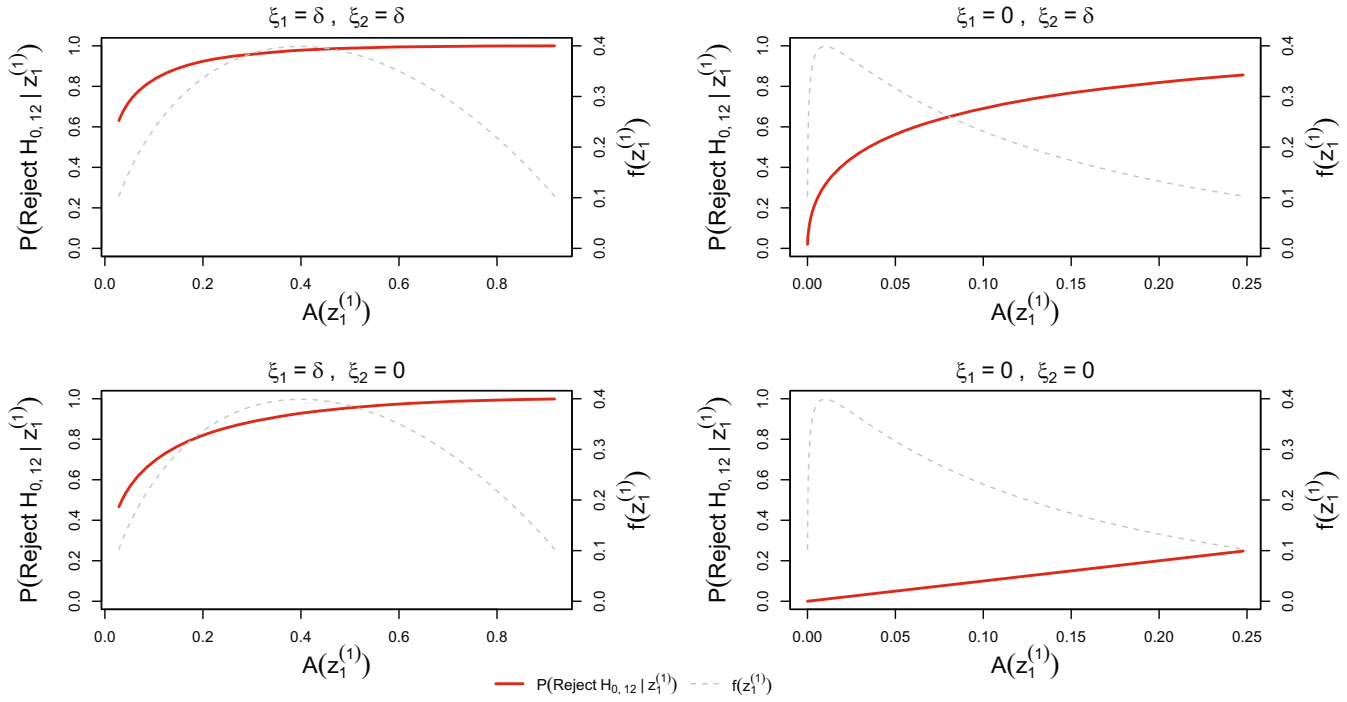

FIGURE 1 Conditional error rate, $A(z_1(1))$, against probability of rejecting the intersection hypothesis $P(\text{Reject } H_{0,12}|z_1^{(1)})$ and corresponding density of conditional error $f(z_1^{(1)})$, $\delta = \Phi^{-1}(0.95) + \Phi^{-1}(0.9)$ such that we have power of 0.9 when testing H_{01} in the original trial.

3 | GENERAL RULE FOR ADDING HYPOTHESES

Consider a trial that aims to test $m \in N$ null hypotheses of the form $H_{0i} : \theta_i \in \Theta_{0,i}$ versus alternative hypotheses of the form $H_{1i} : \theta_i \in \Theta_{1,i}$ where $\Theta_{1,i} = \Theta \setminus \Theta_{0,i}$ for all $i = 1, \ldots, n$. Suppose one wishes to construct a hypothesis testing procedure to strongly control the FWER at level α. Generalising Equation (1) if R is the event that we reject one or more true null hypotheses we require

$$P_\theta(R) \leq \alpha \text{ for all } \boldsymbol{\theta} = (\theta_1, \ldots, \theta_m).$$

To ensure this is achieved we require a closed testing procedure.[10] Let E be the set of these hypotheses $E = (H_{01}, \ldots, H_{0m})$. For any $e \subseteq E$ we may construct a level α test of the intersection hypothesis given by

$$H_e = \bigcap_{i \in e} H_{0i}. \tag{3}$$

To reject a null hypothesis H_{0i} globally (ie, to say, the hypotheses may be rejected in total by the testing procedure) in this testing procedure one must reject level α tests of all intersections in which it is involved. That is to globally reject H_{0i} for any $i = (1, \ldots, m)$ for all $e \subseteq E$ one must reject all H_e for which $H_{0i} \in e$.

Suppose we wish to add m' further hypotheses whilst this trial is in progress having observed some trial data (if we have not we still have the freedom to redefine the testing procedure and incorporate these hypotheses in a test of the original form). Let N be the set of additional null hypotheses $N = (H_{0m+1}, \ldots, H_{0m+m'})$. To update our closed testing procedure we now require that for any $e \subseteq E \cup N$ we can construct a level α test of the intersection hypothesis given by H_e defined as in Equation (3). As in the original procedure, we may globally reject H_{0i} for any $i = (1, \ldots, m)$ if we reject all H_e for which $H_{0i} \in e$ having considered all $e \subseteq E \cup N$. An updated closed testing procedure incorporating all hypotheses in the set $E \cup N$ has three forms of null hypothesis we must consider when constructing these tests.

H_e: For any $e \subseteq E$ let α'_e be the conditional error rate under the original design for the test of H_e given the data available before we modify the trial (ie, before the N hypotheses are added). The local test of H_e must be constructed at level α'_e.

H_n: For any $n \subseteq N$ the local test of H_n must be constructed at level α and based only on data collected after the N hypotheses are added.

$H_e \cap H_n$: For any $e \subseteq E$ and $n \subseteq N$ $H_e \cap H_n$ implies H_e and hence the data already available for H_e is distributed such that computing the corresponding conditional error α'_e will ensure the equivalent of Equation (2) holds. Thus the local test of $H_e \cap H_n$ must be constructed at level α'_e with any contribution from H_n being based only on data collected after the N hypotheses are added.

It is necessary to construct tests of $H_e \cap H_n$ in this way since we may choose what is added to the trial (and thus which hypotheses will be tested) and how (eg, choosing how many patients to recruit to each added treatment arm) based on the already accumulated data. In the case where H_n is tested then the test for $H_e \cap H_n$ may be based on the data relating to both H_e and H_n and is tested at α'_e; while if H_n is not added the test for $H_e \cap H_n$ is implicitly that of H_e also tested at α'_e.

The procedure is universally applicable to all testing procedures that strongly control the FWER, since any procedure that gives strong control of the FWER is a closed testing procedure Burnett and Jennison.[40] Thus we may add hypotheses to any procedure that ensures strong control of the FWER while maintaining the statistical integrity of the trial. The penalty for doing so compared is the test of hypotheses of the form $H_e \cap H_n$, though testing such intersection hypotheses is always a penalty when adjusting for multiplicity.

4 | ALTERATION OF A MULTI-ARM MULTI-STAGE TRIAL IN PROGRESS

4.1 | Multi-arm multi-stage trials

Returning to our motivation set out in Section 1 we now discuss how to add additional treatments (and the corresponding hypotheses) to a multi-arm multi-stage (MAMS) platform trials[5,41] in progress. MAMS allow us to compare the treatments in the same trial, while incorporating pre-planned interim analyses to facilitate early stopping (both for futility and efficacy). This early stopping can be done while formally testing null hypotheses and controlling the FWER (Equation 1), through the use of generalised Dunnett testing procedures[7] though the generalised Dunnett procedure is implicitly a closed testing procedure we follow the proposal of Urach and Posch,[9] directly defining all elements of the closed testing procedure.

Suppose we have K experimental treatments, $T_1, \ldots, T_K$ to compare against a common control. We define the null hypotheses $H_{0i} : \theta_i \leq 0$ and corresponding alternatives $H_{1i} : \theta_i > 0$ for all $i = 1, \ldots, K$. We consider a MAMS design that will simultaneously test these K hypotheses over J analyses.

Let n be the number of patients to be recruited to the control arm in the first stage of the trial. At analysis $j = 1, \ldots, J$ the trial will have recruited $r_k^{(j)} n$ patients to treatment $k = 0, 1, \ldots, K$ ($r_0^{(1)} = 1$ by construction). Suppose treatment $k \in 1, \ldots, K$ is dropped for futility at analysis $j * \in 1, \ldots, J$ we have $r_k^{(j)} = r_k^{(j*)}$ for all $j > j*$. If all $T_1, \ldots, T_K$ are dropped for futility the trial stops recruiting. Alternatively, if treatment/s are selected at an interim analysis (such as when the trial is stopped due to a treatment-control comparison yielding statistical significance[9]) the trial is concluded.

From the observations up to stage $j = 1, \ldots, J$ and treatment $k = 1, \ldots, K$ we construct estimates $\hat{\theta}_k^{(j)}$. Then defining

$$\mathcal{I}_k^{(j)} = \frac{r_k^{(j)} r_0^{(j)} n}{\sigma^2 (r_k^{(j)} + r_0^{(j)})},$$

we find the corresponding Z-statistics

$$Z_k^{(j)} = \hat{\theta}_k^{(j)} \sqrt{\mathcal{I}_k^{(j)}}.$$

In the testing procedures that follow we require that the ratio of patients assigned to each treatment remains fixed for the duration of the trial,[14] that is for all $k = 1, \ldots, K$ and $j, l = 1, \ldots, J$

$$\frac{r_0^{(j)}}{r_k^{(j)}} = \frac{r_0^{(l)}}{r_k^{(l)}}. \tag{4}$$

4.2 | Group sequential closed testing

Recall from Section 2.3 that when testing multiple hypotheses we desire strong control of the FWER. If R is the event that we reject one or more true null hypothesis then extending Equation (1) to K null hypothesis strong control requires that

$$P_{\theta}(R) \leq \alpha \text{ for all } \boldsymbol{\theta} = (\theta_1, \ldots, \theta_K).$$

Let $\mathcal{K}$ be the set such for any $I \subseteq (1, \ldots, K)$ we have that $\cap_{i \in I} H_{0i} \in \mathcal{K}$. To create a closed testing procedure we require local tests for each $H_{0m} \in \mathcal{K}$ at level α. We reject H_{0k} globally when all tests including H_{0k} are rejected at level α for $k = 1, \ldots, K$.

The generalised Dunnett method[7] simultaneously tests all null hypotheses, defining group sequential testing boundaries that account for the correlation structure of comparing multiple treatments to control to achieve the desired FWER. This is equivalent to the test of the global null hypothesis

$$H_G = \bigcap_{k=1}^{K} H_{0k}.$$

We follow this construction for each local null hypothesis.

For each $H_{0m} \in \mathcal{K}$ we choose efficacy boundaries $\boldsymbol{u}_m = (u_{1,m}, \ldots, u_{J,m})$. Where $H_{0,m}$ is rejected at analysis j if $Z_k^{(j)} > u_{j,m}$. The boundaries $\boldsymbol{u}_m$ must be chosen such that the local test of each H_{0m} is constructed at level α. We may also define futility boundaries $\boldsymbol{l} = (l_1, \ldots, l_J)$ where if $Z_k^{(j)} < l_j$ the corresponding treatment is dropped for futility (note we define these universally across all treatments for computational convenience but it is theoretically possible to have these differ). If such futility boundaries are to be binding then we may incorporate this into the choice of each $\boldsymbol{u}_m$, noting that futility is applied globally.

4.3 | Adding experimental treatment arms

Suppose at the J'^{th} ($J' \in 1, \ldots, J$) interim analysis of a MAMS trial in progress we wish to add $T \geq 1$ new treatments. We now have up to $K' = K + 1 + T$ treatments in total (in the case that all $K + 1$ original treatment arms are still in the trial). We had planned recruitment $r_k^{(j)} n$ for treatment $k = 1, \ldots, K + T$ at stage $j = 1, \ldots, J$ where $r_k^{(j)} = 0$ for all $k > K$. We modify the recruitment for each remaining stage of the trial $j = J', \ldots, J$ recruiting $r_k'^{(j)} n$ patients for each treatment $k = 1, \ldots, K + T$ (we could also use this opportunity to modify the number of stages). Note we fix the planned recruitment for the remainder of the trial at the point of modifying the trial, subject to the same restriction as Equation (4).

As in Section 2.2 it is computationally useful to split the trial according to patients recruited before and after the J'^{th} analysis. For $j = J' + 1, \ldots, J$ and $k = 0, 1, \ldots, K$ we would recruit $r_k^{\star(j)} n$ further patients, where $r_k^{\star(j)} = r_k^{(j)} - r_k^{(J')}$, from which we compute Z-statistics $Z_k^{\star(j)}$. For each $k = 1, \ldots, K$ and $j = J' + 1, \ldots, J$ we define weights,

$$w_{1,k}^{(j)} = \sqrt{\frac{r_k^{(J')} + r_0^{(J')}}{r_k^{(j)} + r_0^{(j)}}},$$
$$w_{2,k}^{(j)} = \sqrt{1 - w_{1,k}^{(j)} w_{1,k}^{(j)}}$$

and re-construct the Z-statistics for the remainder of the trial as

$$Z_k^{(j)} = w_{1,k}^{(j)} Z_k^{(J')} + w_{2,k}^{(j)} Z_k^{\star(j)}.$$

4.4 | Incorporating additional hypotheses

Adding hypotheses for the T new treatments, we now have null hypotheses $H_{0i} : \theta_i \leq 0$ and alternatives $H_{1i} : \theta_i > 0$ for all $i = 1, \ldots, K + T$. Suppose we still desire strong control of the FWER across all $K + T$ tests, we construct a closed

testing procedure following the principles of Section 3. Let $\mathcal{K}$ be the set of existing null hypotheses $H_{01}, \ldots, H_{0K}$ and all intersections, $\mathcal{T}$ be the set of added null hypotheses $H_{0K+1}, \ldots, H_{0K+T}$ and all intersections, and $\mathcal{KT}$ the set of all intersections between existing and added null hypotheses.

The conditional error rate of each test for $H_{0m} \in \mathcal{K}$ is maximised under the global null.[42] Given the existing estimates, $\hat{\boldsymbol{\theta}}^{(J')} = (\hat{\theta}_1^{(J')}, \ldots, \hat{\theta}_k^{(J')})$ and under the original trial described in Sections 4.1 and 4.2 we write the conditional error for each $H_{0m} \in \mathcal{K}$ under the global null as

$$B_m(\hat{\boldsymbol{\theta}}^{(J')}) = P_{\mathbf{0}}(\text{Reject } H_{0m} | \hat{\boldsymbol{\theta}}^{(J')}) \leq \alpha.$$

It is useful to re-write the testing boundaries for each $H_{0m} \in \mathcal{K}$ in terms of only the data collected after stage J'. That is for $j = J' + 1, \ldots, J$ and $k = 1, \ldots, K$

$$u_{j,k,m} = \frac{u_{j,m} - w_{1,k}^{(j)} Z_k^{(J')}}{w_{2,k}^{(j)}}$$

H_{0m} is rejected locally at stage j of the trial if $Z_k^{\star(j)} > u_{j,k,m}$. Similarly, if futility stopping is being used we write

$$l_{k,j} = \frac{l_j - w_{1,k}^{(j)} Z_k^{(J')}}{w_{2,k}^{(j)}}$$

where if $Z_k^{\star(j)} < l_{k,j}$ T_k is dropped for futility. This allows computation of the conditional error rate based on the distributions of the data that will be collected subsequent to the addition of the new treatments, $Z_k^{\star(j)}$ for $j = J' + 1, \ldots, J$ and $k = 0, 1, \ldots, K$.

For each $H_{0m} \in \mathcal{K}$ the hypothesis test must be constructed at level $B_m(\hat{\boldsymbol{\theta}}^{(J')})$. For each $H_{0m} \in \mathcal{T}$ the hypothesis test must be constructed at level α. For each $H_{0m} \in \mathcal{KT}$ the hypothesis test must be constructed at level $B_m(\hat{\boldsymbol{\theta}}^{(J')})$. Applying the closed testing procedure we reject H_{0m} globally when all corresponding local hypothesis tests are rejected at the appropriate level. This ensures strong control of the FWER at level α. Further to allowing the addition of the additional hypotheses to the testing procedure this testing structure would also allow other modifications to the design (such as the allocation ratios to each treatment, provided they remain fixed for the remainder of the trial).

For each hypothesis $H_{0m} \in K \cup T \cup KT$ we define the testing boundaries for the modified trial at the required error rate $\boldsymbol{u}'_m = (u'_{J'+1,m}, \ldots, u'_{J,m})$ and $\boldsymbol{l}'_m = (l'_{J'+1,m}, \ldots, l'_{J,m})$. At stage $j = J' + 1, \ldots, J$ for treatment $k = 0, 1, \ldots, K$ the recruitment is governed by $r_k^{\star\prime(j)} = r_k'^{(j)} - r_k^{(J')}$, with corresponding Z-statistics $Z_k^{\star\prime(j)}$. For each experimental treatment from the first stage of the trial $k = 1, \ldots, K$ we define weights for data before and after stage J', for $j = J' + 1, \ldots, J$ and $k = 1, \ldots, K$

$$w_{1,k}'^{(j)} = \sqrt{\frac{r_k^{(J')} + r_0^{(J')}}{r_k'^{(j)} + r_0'^{(j)}}},$$
$$w_{2,k}'^{(j)} = \sqrt{1 - w_{1,k}'^{(j)} w_{1,k}'^{(j)}}$$

and construct the Z-statistics for the hypothesis tests as

$$Z_k'^{(j)} = w_{1,k}'^{(j)} Z_k^{(J')} + w_{2,k}'^{(j)} Z_k^{\star\prime(j)}$$

allowing us to write the testing boundaries for each $H_{0m} \in \mathcal{K}$ in terms of only the data collected after stage J'. That is for $j = J' + 1, \ldots, J$ and $k = 1, \ldots, K$

$$u'_{j,k,m} = \frac{u'_{j,m} - w_{1,k}'^{(j)} Z_k^{(J')}}{w_{2,k}'^{(j)}}$$

rejecting $H_{0m} \in \mathcal{K}$ at analysis $j = J' + 1, \ldots, J$ if $Z_k^{\star\prime(j)} > u_{j,k,m}$ and

$$l'_{k,j} = \frac{l'_j - w_{1,k}^{\prime(j)} Z_k^{(J')}}{w_{2,k}^{\prime(j)}}$$

where if $Z_k^{\star\prime(j)} < l_{k,j}$ T_k dropped for futility (note for $k > K$ $u'_{j,m}$ and l'_j). With this in place $\boldsymbol{u}'_m$ and $\boldsymbol{l}'_m$ may be computed as per the generalised Dunnett test.

Note that we have not discussed how to decide whether the treatment (and hence corresponding hypotheses) should be added to the trial. As discussed in Section 3 the testing procedure is constructed such that strong control of the FWER is guaranteed whether the hypotheses are added to the procedure or not. We shall not go into details of how to make such decisions in this manuscript however the testing procedure will remain valid should the decision be based upon trial data or based on factors external to the trial.

5 | EXAMPLE

5.1 | Problem setup

Consider a MAMS trial comparing two experimental treatments with a common control over three stages (ie, with two interim analyses). Suppose $n = 10$ patients are recruited to each treatment at each stage of the trial. So for this trial we have $J = 3$, $K = 2$ and $\boldsymbol{r}_k = (1, 2, 3)$ for $k = 0, 1, 2$. Note the sample size here does not reflect the typical sample size one might expect for a MAMS trial, however, it will provide a useful scale for comparison. The maximum sample size for this trial is 90 patients while the expected sample sizes remain on a useful scale for ease of interpretation.

Under this design we test the null hypotheses $H_{01} : \theta_1 \leq 0$ and $H_{02} : \theta_2 \leq 0$. The testing boundaries are constructed for a FWER of $\alpha = 0.05$, let $\delta = \Phi^{-1}(0.75)\sqrt{2}$ and $\sigma = 1$. At a configuration of $\boldsymbol{\theta} = (\delta, 0)$ we have a target power of $1 - \beta = 0.9$. Defining the triangular testing boundaries[43] we first compute the testing boundary for $H_{01} \cap H_{02}$ using the `mams()` function of the MAMS package in R.[44] This sets the futility boundary for all tests with the upper boundaries computed for testing both H_{01} and H_{02} separately.

Suppose after the first analysis $J' = 1$ two further experimental treatments become available that we wish to add to the trial in progress, that is we have $T = 2$ further arms that we may wish to consider. Making the same comparison to the common control we shall further test the null hypotheses $H_{03} : \theta_3 \leq 0$ and $H_{04} : \theta_4 \leq 0$ (in addition to H_{01} and H_{02}). To illustrate how this may operate we compare our proposed method with two alternatives. Under each option, we allow the addition of a further 10 patients per treatment per stage for the added experimental treatments. This may not be the optimal choice for each design but ensures we have some common ground between the methods we are comparing, we shall discuss possible optimisation of such design choices further in Section 5.4.

5.2 | Designs for comparison

Design 1, adding to the existing trial: Using the methods outlined in Section 4 one may add the two new treatments to the design and analysis of the trial already in progress. That is up until the first analysis 10 patients are recruited to the control arm and experimental treatments 1 and 2. This first stage is conducted using the testing boundaries for the original three-arm three-stage design. That is the closed testing procedure is built to incorporate H_{01} and H_{02}.

After the first analysis (if the trial is still ongoing) treatments 3 and 4 are added to the trial with 10 patients being recruited to each treatment remaining in the trial for the remainder of the trial. In addition, the testing boundaries are updated following the methods outlined in Section 4. Thus the closed testing procedure now includes H_{01}, H_{02}, H_{03} and H_{04}, the testing boundaries for any tests involving H_{01} or H_{02} have been modified according to the data observed in the first stage of the trial following the methods described in Section 4.4.

Design 2, two separate testing procedures: Suppose one wished to incorporate all treatments into the same trial without modifying the testing procedure of the original design. In this case one may separate the testing procedures for the original hypotheses (H_{01} and H_{02}) and the new hypotheses (H_{03} and H_{04}), strongly controlling the FWER across each pair. This testing structure will not strongly control the FWER across the set of all four hypotheses. We make some

attempt to adjust for this by not sharing control patients between the two pairs of tests. This is something that would require careful consideration in practice as it does not align with the original intent for the trial to strongly control the FWER.

Thus for H_{01} and H_{02} the trial continues beyond the first interim analysis as originally planned, following exactly the original design. To test H_{03} and H_{04} we recruit 10 further patients across the control and treatments 3 and 4 for the remaining two stages of the trial. The testing structure for this is constructed as for a two-stage three-arm MAMS design, again using triangular testing boundaries.

Note that functionally this choice is equivalent to allowing the original trial to run to its' conclusion while simultaneously conducting a separate trial for treatments 3 and 4. Incorporating all recruitment into a single protocol would however avoid problems such as competing recruitment.

Design 3, a new testing procedure: Supposing alternatively that one wished to ensure strong control of the FWER without using the methods proposed in this work. One cannot modify the ongoing testing procedure once data have been seen, thus to conduct a trial that tests all four hypotheses H_{01}, H_{02}, H_{03} and H_{04} one must discard the existing data. To allow for a comparison with the other two designs we thus consider a two-stage trial recruiting 10 patients to the control and each of the four experimental treatments, using triangular boundaries as we had in the original design.

5.3 | Simulation results

We evaluate the performance of these three designs via simulation. For design 1 we estimate the operating characteristics using 10,000 simulations of the whole trial. This is a lower number of simulations than would be ideal, due to the computationally intensive nature of the simulation. For each iteration, we must simulate the first stage of the trial then compute the updated testing boundaries allowing us to simulate the remainder of the trial. In practice, we do not expect this to be used as a pre-planned scheme and hence only one set of updated testing boundaries needs computing, making longer simulations more viable. An illustration of adding treatments based on interim observations is given in the supplementary material Section 2. We use 1,000,000 simulations to estimate the operating characteristics for both designs 2 and 3 given the relative computational simplicity.

For the original trial (before any design modification): when $\boldsymbol{\theta} = (0, 0)$ the probability of the trial continuing past the first interim analysis is 0.64, with a probability of 0.34 of stopping for futility; when $\boldsymbol{\theta} = (\delta, 0)$ the probability of the trial continuing past the first interim analysis is still 0.64, with a probability of 0.35 of stopping for efficacy; and when $\boldsymbol{\theta} = (\delta, \delta)$ the probability of the trial continuing past the first interim analysis drops to 0.46, with a probability of 0.53 of stopping for efficacy. Table 3 shows the properties design 1 should the trial continue beyond the first interim analysis. We

TABLE 3 Design 1, probabilities of rejecting null hypotheses and expected sample size under the corresponding configuration of $\boldsymbol{\theta}$ when the trial continues beyond the first interim analysis.

$\boldsymbol{\theta}$	$P_\theta(R_1)$	$P_\theta(R_2)$	$P_\theta(R_3)$	$P_\theta(R_4)$	$E_\theta(N)$	$P_\theta(M_0)$	$P_\theta(M_1)$	$P_\theta(M_2)$	$P_\theta(M_3)$	$P_\theta(M_4)$
$(0, 0, 0, 0)$	0.03	0.02	0.01	0.01	78	0.94	0.05	0.01	0.00	0.00
$(\delta, 0, 0, 0)$	0.91	0.00	0.01	0.01	71	0.08	0.89	0.02	0.00	0.00
$(\delta, \delta, 0, 0)$	0.79	0.84	0.02	0.02	77	0.04	0.30	0.62	0.04	0.00
$(0, 0, \delta, 0)$	0.01	0.02	0.71	0.01	78	0.29	0.68	0.03	0.00	0.00
$(\delta, 0, \delta, 0)$	0.86	0.01	0.64	0.03	71	0.05	0.40	0.52	0.03	0.00
$(\delta, \delta, \delta, 0)$	0.82	0.82	0.61	0.01	75	0.01	0.19	0.35	0.44	0.01
$(0, 0, \delta, \delta)$	0.02	0.02	0.63	0.63	79	0.12	0.48	0.37	0.02	0.00
$(\delta, 0, \delta, \delta)$	0.82	0.01	0.60	0.60	70	0.03	0.27	0.35	0.33	0.01
$(\delta, \delta, \delta, \delta)$	0.77	0.77	0.61	0.61	76	0.02	0.15	0.21	0.30	0.32

Note: Where R_i is the event that we reject H_{oi}, N is the total sample size (including the 30 patients included in stage one) and M_j is the event that j null hypotheses are rejected.

TABLE 4 Design 2, probabilities of rejecting null hypotheses and expected sample size under the corresponding configuration of $\boldsymbol{\theta}$ when the trial continues beyond the first interim analysis.

θ	$P_\theta(R_1)$	$P_\theta(R_2)$	$E_\theta(N_1)$	$P_\theta(R_3)$	$P_\theta(R_4)$	$E_\theta(N_2)$
$(0,0,0,0)$	0.03	0.03	49	0.03	0.03	38
$(\delta,0,\delta,0)$	0.93	0.02	47	0.82	0.04	39
$(\delta,\delta,\delta,\delta)$	0.81	0.81	45	0.77	0.77	39
	Original analysis (H_{01} and H_{02})			**Additional analysis (H_{01} and H_{02})**		
θ	$P_\theta(M_0)$	$P_\theta(M_1)$	$P_\theta(M_2)$	$P_\theta(M_0)$	$P_\theta(M_1)$	$P_\theta(M_2)$
$(0,0,0,0)$	0.95	0.04	0.01	0.95	0.04	0.01
$(\delta,0,\delta,0)$	0.07	0.90	0.02	0.18	0.78	0.04
$(\delta,\delta,\delta,\delta)$	0.02	0.34	0.64	0.07	0.31	0.62

Note: Where R_i is the event that we reject H_{oi}, N_1 is the total sample size in the original trial, N_2 is the total sample size in the additional trial and M_j is the event that j null hypotheses are rejected.

observe lower probabilities of rejecting H_{03} or H_{04} than rejecting H_{01} or H_{02}, this is due to the already promising results for treatments 1 and 2 given the trial has proceeded beyond the first interim analysis.

Comparing design 1 with design 2 shown in Table 4, separating the analyses for the pairs of hypotheses produces similar probabilities of rejecting H_{01} or H_{02}, this is encouraging that the probability of success for the existing hypotheses will not be overly influenced by the addition of further treatments. Design 1 is shown to be more sensitive to the values of θ_3 and θ_4 due to their ability to also conclude the trial early. Design 2 increases the probabilities of rejecting H_{03} or H_{04}; this is partially due to the disconnect between the analyses, if one concludes early the other may continue and reject and null hypothesis. Given this and that patients are recruited to the control in both trials we see that design 1 significantly reduces the expected sample size, with 70–80 patients including the first stage of the trial for trials that continue beyond the first stage (for the trial as a whole this expected sample size drops to 50–60 over the scenarios we have examined) whereas design 2 requires 90–95 patients. Design 2 is of course flawed when considering the original goals of the trial, while this method incorporates all existing data for H_{01} and H_{02} there is no multiplicity adjustment between the existing and added hypotheses. Furthermore, should this be conducted as two separate trials of two pairs of null hypotheses each with a FWER of α, if we wish to select some subset of treatments for further study, there is no guarantee of direct comparability between each analysis.

Comparing the operating characteristics of design 3 in Table 5 with design 1 in Table 3, we see that, when starting a new analysis, the probabilities of rejecting H_{01} or H_{02} are lower while the probabilities of rejecting H_{03} or H_{04} are similar. This leads to a reduction in the probability of rejecting multiple hypotheses in design 3 when compared with design 1. For example, when $\theta_1 = \theta_2 = \theta_3 = \theta_4 = \delta$ the probabilities of rejecting H_{01}, H_{02}, H_{03} and H_{04} are 0.77, 0.77, 0.61 and 0.61 respectively under design 1, while they are all 0.64 under design 3 and the probability of rejecting two or more hypotheses falls by 0.12 compared to design 1. The expected sample size of the trial conducted under design 3 is reduced versus design 1 by 8–15 patients. However, this does not account for the fact that 30 patients have been recruited who do not contribute to the result, these patients lead to a significant improvement in the expected sample size for design 1.

To allow for a direct comparison, Table 6 collates the expected sample size under each design. We see here that using our proposed method design 1 offers an improvement directly in the expected sample size of the trial. Given the construction of the example to ensure the same sample for each treatment within each stage of the trial this is the most direct comparison available and shows the key potential benefit of our proposed method.

5.4 | Design choices

The choices we have made in this example have allowed us to explore and illustrate how our proposed method may operate while adding treatments to a trial. There are many elements of the design we encourage practitioners to consider when implementing these methods in practice. Indeed we expect one would need to consider an in-depth simulation

TABLE 5 Design 3, probabilities of rejecting null hypotheses and expected sample size under the corresponding configuration of θ when the trial continues beyond the first interim analysis.

θ	$P_\theta(R_1)$	$P_\theta(R_2)$	$P_\theta(R_3)$	$P_\theta(R_4)$	$E_\theta(N)$	$P_\theta(M_0)$	$P_\theta(M_1)$	$P_\theta(M_2)$	$P_\theta(M_3)$	$P_\theta(M_4)$
$(0,0,0,0)$	0.02	0.02	0.02	0.02	62 (+30)	0.95	0.04	0.01	0.00	0.00
$(\delta,0,0,0)$	0.75	0.01	0.01	0.01	63 (+30)	0.24	0.73	0.02	0.00	0.00
$(\delta,\delta,0,0)$	0.67	0.67	0.02	0.02	62 (+30)	0.11	0.43	0.43	0.02	0.01
$(0,0,\delta,0)$	0.01	0.01	0.75	0.01	63 (+30)	0.24	0.73	0.02	0.00	0.00
$(\delta,0,\delta,0)$	0.67	0.02	0.67	0.02	62 (+30)	0.11	0.43	0.43	0.02	0.00
$(\delta,\delta,\delta,0)$	0.64	0.64	0.64	0.03	62 (+30)	0.07	0.31	0.28	0.33	0.03
$(0,0,\delta,\delta)$	0.02	0.02	0.67	0.67	62 (+30)	0.11	0.43	0.43	0.02	0.00
$(\delta,0,\delta,\delta)$	0.64	0.03	0.64	0.64	62 (+30)	0.07	0.31	0.27	0.33	0.03
$(\delta,\delta,\delta,\delta)$	0.64	0.64	0.64	0.64	63 (+30)	0.05	0.24	0.19	0.18	0.35

Note: Where R_i is the event that we reject H_{oi} and N is the total sample size (note 30 additional patients are recruited but not used in the analysis).

TABLE 6 Comparing the expected sample size for each design under varying configurations of $\boldsymbol{\theta}$, note for design 3 the additional 30 patients are recruited but not used in the analysis while they are used in the analysis for both designs 1 and 2.

	Expected sample size		
θ	**Design 1**	**Design 2**	**Design 3**
$(0,0,0,0)$	78	87	62 (+30)
$(\delta,0,0,0)$	71	85	63 (+30)
$(\delta,\delta,0,0)$	77	83	62 (+30)
$(0,0,\delta,0)$	78	88	63 (+30)
$(\delta,0,\delta,0)$	71	86	62 (+30)
$(\delta,\delta,\delta,0)$	75	84	62 (+30)
$(0,0,\delta,\delta)$	79	88	62 (+30)
$(\delta,0,\delta,\delta)$	70	86	62 (+30)
$(\delta,\delta,\delta,\delta)$	76	84	63 (+30)

study taking into account the trials' specific objectives, there is a link to a github repository in Section 7 that contains the code used to generate the examples and may help with further exploration.

The first and most obvious choice beyond whether to add further experimental treatments is how many patients should be recruited to each treatment arm of the trial. One may wish to attempt to retain power for the original treatments in the trial and increase allocation to these arms or one may be concerned about the probability of rejecting at least one null hypothesis under a given configuration of the treatment effects which would drive an alternative choice. Similarly one may wish to use the opportunity presented by altering the trial to more carefully consider futility stopping for the treatments already in the trial, doing so in this way would present a non-binding futility decision which while having a potentially small impact on power (the impact is expected to be small since one would unlikely stop a well-performing treatment in such a way) will not inflate the error rate.

In the designs we have presented we have stopped the trial simultaneously for all treatments should one treatment be found to be effective. This is in keeping with how we envisage the intent of the original MAMS design that we proposed, where using the group sequential stopping rule is intended to end the trial early and reduce the expected sample size. However, this is not strictly necessary, with all elements of the closed test well-defined one could continue the trial to its'

conclusion (stopping recruitment to arms either when the corresponding null hypothesis is globally rejected or when the treatment is removed for futility).

6 | DISCUSSION

In this paper, we do not discuss whether strict control of the FWER is required or not in multi-armed clinical trials per-se such as in platform trials. Here we refer to several recent discussions on this topic.[27,28,45,46] For example, in COVID platform trials[47] such as RECOVERY or REMAP-CAP only the per-comparison error rate is controlled and no strict FWER control across treatment arms is required. We note, however, that at the start and in the early protocols of RECOVERY, for example, strong control of the FWER was required. In this work, we provide a framework for confirmatory multi-armed trials, where strict control of the FWER is required, for example, for regulatory purposes[17-22] and it is desirable to incorporate a new treatment in an already ongoing trial while preserving the integrity of the design and avoiding delays to the overall development process. Our proposed general framework for adding new experimental treatments (and corresponding null hypotheses) to a trial in progress builds upon the work of Hommel,[33] allowing the addition of null hypotheses to any trial that strongly controls the FWER. This testing structure allows other design alterations while utilising all available data in inference and decision-making.

We have applied this general framework in the setting of MAMS platform trials.[48,49] The methods proposed will achieve strong FWER control for the family of hypotheses of interest. This may include all hypotheses related to the MAMS trial concerned, but the methods are also applicable only for a subset of hypotheses in a larger platform trial with several substudies and treatment arms. For example, a platform where in a substudy several doses of a certain treatment shall be tested and for this family strict FWER control is needed for regulatory purposes. If additional doses shall be added later in a data-dependent way, for example, due to lack of efficacy of lower doses or safety issues if higher ones, our methods will ensure the confirmatory characteristics despite adding additional doses in an ongoing study. The examples in Section 5 demonstrate that this does indeed strongly control the FWER as expected.

The examples in Sections 2.4 and 5 show there is a penalty for adding treatments in this way. In the simulation scenarios investigated, there is a small reduction in the probability of rejecting the null hypotheses already in the trial. There is an impact in these examples from the reduced sample size and probability of early stopping on the probabilities of globally rejecting the hypotheses for the added treatments, which stay relatively low. In addition, the combination of utilising the existing data and the efficient use of control patients across the trial yields a reduction in the expected sample size when compared to alternatives that do not make such use of the existing data. Of course, there is the possibility that some existing treatments have a small or negative effect then investigators might be incentivised to open a new independent study as the new treatments are penalised by the existing data. At this point, the investigator must consider the trade-off between using a shared infrastructure (and potentially starting earlier) compared to starting from scratch (but having the full α for the investigation). However, if such scenarios are of sufficient concern this may be accounted for by the futility boundaries of the original MAMS trial. By allowing for further adaptations such as increasing the sample sizes for the newly added treatment and adapting the allocation ratios one could compensate for such lower conditional errors for some intersection hypotheses.

Our examples aim to illustrate how the method might be applied, the exact circumstances of adding arms will vary widely in practice. The statistical operating characteristics (specifically with regard to the probabilities of rejecting each hypothesis globally) were not our primary motivation however, this could be a more important consideration. In principle, one might consider more carefully the desired operating characteristics of the design for the remainder of the trial, optimising the proportion of recruitment to each treatment. We have discussed such options in more detail in Section 5.4. While making such decisions one would need to think more deeply about what the goals for the modified trial would be. In addition, other practical considerations should be thought about. For example, if sufficiently many patients must be added to the trial will this extend the timelines of the trial such that population shift becomes a concern? Thus investigators are encouraged to think carefully about such modifications on a case-by-case basis.

The general framework for adding hypotheses to a trial in progress has a broader application than MAMS designs, applying to any testing procedure that gives strong control of the FWER. The addition of hypotheses in this way allows the incorporation of all available data into decisions about how to plan the remainder of the trial. Thus, through consideration of the existing closed testing procedure and the application of the conditional error principle, this gives a versatile tool for ad hoc design modification when required. Although, in practice, one must carefully consider the appropriateness of its use.

7 | SOFTWARE

Code relating to the examples presented here available at https://github.com/Thomas-Burnett/Adding-treatments-to-clinical-trials-in-progress.git.

AUTHOR CONTRIBUTIONS

Franz König and Thomas Jaki share the last authorship equally.

ACKNOWLEDGEMENTS

This research was supported by the NIHR Biomedical Research Centre (BRC-1215-20014). This report is independent research supported by the National Institute for Health Research (Prof Jaki's Senior Research Fellowship, NIHR-SRF-2015-08-001). T Jaki received funding from the UK Medical Research Council (MC_UU_00002/14, MC_UU_00040/03 and MR/V038419/1). Franz König is a member of the EU Patient-Centric Clinical Trial Platform (EU-PEARL) which has received funding from the Innovative Medicines Initiative 2 Joint Undertaking, grant no 853966. This Joint Undertaking receives support from the EU Horizon 2020 Research and Innovation Programme, EFPIA, Children's Tumor Foundation, Global Alliance for TB Drug Development, and SpringWorks Therapeutics. The views expressed in this publication are those of the authors. They are not necessarily those of the NHS, the National Institute for Health Research or the Department of Health and Social Care (DHSC). The funders and associated partners are not responsible for any use that may be made of the information contained herein.

DATA AVAILABILITY STATEMENT

Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

ORCID

Thomas Burnett https://orcid.org/0000-0001-8912-2554
Franz König https://orcid.org/0000-0002-6893-3304
Thomas Jaki https://orcid.org/0000-0002-1096-188X

REFERENCES

1. Jaki T, Hampson LV. Designing multi-arm multi-stage clinical trials using a risk–benefit criterion for treatment selection. *Stat Med*. 2016;35(4):522-533.
2. Royston P, Parmar MK, Qian W. Novel designs for multi-arm clinical trials with survival outcomes with an application in ovarian cancer. *Stat Med*. 2003;22(14):2239-2256.
3. Jaki T, Magirr D. Considerations on covariates and endpoints in multi-arm multi-stage clinical trials selecting all promising treatments. *Stat Med*. 2013;32(7):1150-1163.
4. Wason JM, Jaki T. Optimal design of multi-arm multi-stage trials. *Stat Med*. 2012;31(30):4269-4279.
5. Jaki T. Multi-arm clinical trials with treatment selection: what can be gained and at what price? *Clin Investigat*. 2015;5(4):393-399.
6. Stallard N, Todd S. Sequential designs for phase III clinical trials incorporating treatment selection. *Stat Med*. 2003;22(5):689-703.
7. Magirr D, Jaki T, Whitehead J. A generalized Dunnett test for multi-arm multi-stage clinical studies with treatment selection. *Biometrika*. 2012;99(2):494-501.
8. Dunnett CW. A multiple comparison procedure for comparing several treatments with a control. *J Am Stat Assoc*. 1955;50(272):1096-1121.
9. Urach S, Posch M. Multi-arm group sequential designs with a simultaneous stopping rule. *Stat Med*. 2016;35(30):5536-5550.
10. Marcus R, Eric P, Gabriel KR. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*. 1976;63(3):655-660.
11. Bretz F, Schmidli H, König F, Racine A, Maurer W. Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: general concepts. *Biometr J: J Math Methods Biosci*. 2006;48(4):623-634.
12. Schmidli H, Bretz F, Racine A, Maurer W. Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: applications and practical considerations. *Biom J*. 2006;48(4):635-643.
13. Posch M, Koenig F, Branson M, Brannath W, Dunger-Baldauf C, Bauer P. Testing and estimation in flexible group sequential designs with adaptive treatment selection. *Stat Med*. 2005;24(24):3697-3714.
14. Koenig F, Brannath W, Bretz F, Posch M. Adaptive Dunnett tests for treatment selection. *Stat Med*. 2008;27(10):1612-1625.
15. Bauer P, Kieser M. Combining different phases in the development of medical treatments within a single trial. *Stat Med*. 1999;18(14):1833-1848.
16. Dmitrienko A, Tamhane AC, Bretz F. *Multiple Testing Problems in Pharmaceutical Statistics*. Boca Raton, FL: CRC Press; 2009.
17. European Medicines Agency. ICH Topic E9: Statistical Principles for Clinical Trials. 1998.

18. US Food and Drug Administration and others. Guidance for industry on multiple endpoints in clinical trials. 2017.
19. European Medicines Agency. Guideline on multiplicity issues in clinical trials. 2017.
20. European Medicines Agency. Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design. 2007.
21. US Food and Drug Administration and others. Draft guidance for industry: adaptive design clinical trials for drugs and biologics. 2010.
22. US Food and Drug Administration. Adaptive Designs for Clinical Trials of Drugs and Biologics Guidance for Industry. 2019.
23. Sydes MR, Parmar MK, James ND, et al. Issues in applying multi-arm multi-stage methodology to a clinical trial in prostate cancer: the MRC STAMPEDE trial. *Trials*. 2009;10(1):39.
24. Bauer P, Bretz F, Dragalin V, König F, Wassmer G. Twenty-five years of confirmatory adaptive designs: opportunities and pitfalls. *Stat Med*. 2016;35(3):325-347. doi:10.1002/sim.6472
25. Dimairo M, Pallmann P, Wason J, et al. The adaptive designs CONSORT extension (ACE) statement: a checklist with explanation and elaboration guideline for reporting randomised trials that use an adaptive design. *Bmj*. 2020;369:m115. doi:10.1136/bmj.m115
26. Parmar MK, Barthel FMS, Sydes M, et al. Speeding up the evaluation of new agents in cancer. *J Natl Cancer Inst*. 2008;100(17):1204-1214.
27. Collignon O, Gartner C, Haidich AB, et al. Current statistical considerations and regulatory perspectives on the planning of confirmatory basket, umbrella, and platform trials. *Clin Pharmacol Therapeut*. 2020;107(5):1059-1067.
28. Bretz F, König F. Commentary on parker and weir. *Clin Trials*. 2020;17(5):567-569.
29. Hlavin G, Hampson LV, Koenig F. Many-to-one comparisons after safety selection in multi-arm clinical trials. *PLoS One*. 2017;12(6):e0180131.
30. Bennett M, Mander AP. Designs for adding a treatment arm to an ongoing clinical trial. *Trials*. 2020;21(1):1-12.
31. Proschan MA, Hunsberger SA. Designed extension of studies based on conditional power. *Biometrics*. 1995;51(4):1315-1324.
32. Magirr D, Stallard N, Jaki T. Flexible sequential designs for multi-arm clinical trials. *Stat Med*. 2014;33(19):3269-3279.
33. Hommel G. Adaptive modifications of hypotheses after an interim analysis. *Biometr J: J Math Methods Biosci*. 2001;43(5):581-589.
34. Sugitani T, Posch M, Bretz F, Koenig F. Flexible alpha allocation strategies for confirmatory adaptive enrichment clinical trials with a prespecified subgroup. *Stat Med*. 2018;37(24):3387-3402.
35. Bauer P, Köhne K. Evaluation of experiments with adaptive interim analyses. *Biometrics*. 1994;50:1029-1041.
36. Lehmacher W, Wassmer G. Adaptive sample size calculations in group sequential trials. *Biometrics*. 1999;55:1286-1290.
37. Hartung J. A note on combining dependent tests of significance. *Biom J*. 1999;41:849-855.
38. Dmitrienko A, Tamhane AC. Gatekeeping procedures with clinical trial applications. *Pharmaceut Stat: J Appl Stat Pharmaceut Ind*. 2007;6(3):171-180.
39. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2019.
40. Burnett T, Jennison C. Adaptive enrichment trials: what are the benefits? *Stat Med*. 2021;40(3):690-711.
41. Wason J, Magirr D, Law M, Jaki T. Some recommendations for multi-arm multi-stage trials. *Stat Methods Med Res*. 2016;25(2):716-727.
42. Stallard N, Kunz CU, Todd S, Parsons N, Friede T. Flexible selection of a single treatment incorporating short-term endpoint information in a phase II/III clinical trial. *Stat Med*. 2015;34(23):3104-3115.
43. Whitehead J. *The Design and Analysis of Sequential Clinical Trials*. Chichester: John Wiley & Sons; 1997.
44. Jaki TF, Pallmann PS, Magirr D. The R package MAMS for designing multi-arm multi-stage clinical trials. *J Stat Softw*. 2019;88(4):1-25.
45. Stallard N, Todd S, Parashar D, Kimani PK, Renfro LA. On the need to adjust for multiplicity in confirmatory clinical trials with master protocols. *Ann Oncol*. 2019;30(4):506-509.
46. Parker RA, Weir CJ. Non-adjustment for multiple testing in multi-arm trials of distinct treatments: rationale and justification. *Clin Trials*. 2020;17(5):562-566.
47. Stallard N, Hampson L, Benda N, et al. Efficient adaptive designs for clinical trials of interventions for COVID-19. *Stat Biopharmaceut Res*. 2020;12(4):483-497.
48. Meyer EL, Mesenbrink P, Mielke T, Parke T, Evans D, König F. Systematic review of available software for multi-arm multi-stage and platform clinical trial design. *Trials*. 2021;22(1):1-14.
49. Meyer EL, Mesenbrink P, Dunger-Baldauf C, et al. The evolution of master protocol clinical trial designs: a systematic literature review. *Clin Ther*. 2020;42(7):1330-1360.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Burnett T, König F, Jaki T. Adding experimental treatment arms to multi-arm multi-stage platform trials in progress. *Statistics in Medicine*. 2024;43(18):3447-3462. doi: 10.1002/sim.10090