RESEARCH ARTICLE

Optimizing subgroup selection in two-stage adaptive enrichment and umbrella designs

Nicolás M. Ballarini[1] | Thomas Burnett[2] | Thomas Jaki[2,3] | Christoper Jennison[4] | Franz König[1] | Martin Posch[1]

[1]Section for Medical Statistics, Medical University of Vienna, Vienna, Austria

[2]Department of Mathematics and Statistics, Lancaster University, Lancaster, UK

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

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

Correspondence
Martin Posch, Section for Medical Statistics, CeMSIIS, Medical University of Vienna, Vienna 1090, Austria.
Email: martin.posch@meduniwien.ac.at

We design two-stage confirmatory clinical trials that use adaptation to find the subgroup of patients who will benefit from a new treatment, testing for a treatment effect in each of two disjoint subgroups. Our proposal allows aspects of the trial, such as recruitment probabilities of each group, to be altered at an interim analysis. We use the conditional error rate approach to implement these adaptations with protection of overall error rates. Applying a Bayesian decision-theoretic framework, we optimize design parameters by maximizing a utility function that takes the population prevalence of the subgroups into account. We show results for traditional trials with familywise error rate control (using a closed testing procedure) as well as for umbrella trials in which only the per-comparison type 1 error rate is controlled. We present numerical examples to illustrate the optimization process and the effectiveness of the proposed designs.

KEYWORDS
Bayesian optimization, conditional error function, subgroup analysis, utility function

1 | INTRODUCTION

It is increasingly common to integrate subgroup identification and confirmation into a clinical development program. Biomarker-guided clinical trial designs have been proposed to close the gap between the exploration and confirmation of subgroup treatment effects. Numerous statistical considerations (eg, multiplicity issues, consistency of treatment effects, trial design) need to be taken into account to ensure a proper interpretation of study findings, as outlined in recent reviews.[1-4]

Several study designs are available for the investigation of subgroups in clinical trials. These include all-comers designs where biomarker status or subgroup are not considered for enrolment but only in the trial analysis, and stratified designs where the trial prevalences for each subgroup, that is the proportion of patients recruited from each subgroup, are chosen initially and maintained throughout the trial.[5,6] Adaptive enrichment designs have been proposed to increase the efficiency of these trials.[7-11] These designs allow subgroups to be dropped for futility at interim analyses with the rest of

the trial being conducted with subjects from the remaining groups only. The U.S. Food and Drug Administration guidance on adaptive designs highlights the use of adaptive enrichment designs as a means to increase the chance to detect a true drug effect over that of a fixed sample design.[12]

Master protocols provide an infrastructure for efficient study of newly developed compounds or biomarker-defined subgroups.[13,14] Such studies simultaneously evaluate more than one investigational drug or more than one disease type within the same overall trial structure.[15-17] An umbrella trial is a particular type of master protocol in which enrolment is restricted to a single disease but the patients are screened and assigned to molecularly defined subtrials. Each subtrial may have different objectives, endpoints or design characteristics. An example of an umbrella trial is the ALCHEMIST trial, in which patients with nonsmall cell lung cancer are screened for EGFR mutation or ALK rearrangement and assigned accordingly to subtrials with different treatments.[18]

In this paper, we study confirmatory trials that allow the investigation of the treatment effect in prespecified nonoverlapping subgroups. In particular, we focus on adaptive clinical trials that allow the modification of design elements without compromising the integrity of the trial.[19] We propose a class of adaptive enrichment designs that use a Bayesian decision framework to optimize the design parameters, such as the trial prevalences of the subgroups, the weights for multiple hypotheses testing, and adaptation rules. A similar framework has been used in References 20-27 for adaptive enrichment trials.

We consider two types of problem. In the first case, we study designs that preserve the familywise error rate (FWER) of the trial using a closed testing procedure to test the null hypotheses of no treatment effect in the two subgroups. This is what is typically required in adaptive enrichment trials where a single treatment is evaluated against a control. In the second case, we show results for umbrella trial designs without multiplicity adjustment. Here, we consider studies made up of separate simultaneous trials, for which it has been argued that no control of multiplicity is needed.[28] Our work, therefore, provides an overarching framework for both adaptive enrichment designs and umbrella trials.

The manuscript is organized as follows: In Section 2, we introduce the designs and distinguish between single-stage designs (Section 2.2) and two-stage designs (Section 2.3), and in Section 2.4 we discuss how to adapt our proposed designs to umbrella trials. In Sections 3 and 4 we present numerical examples. We describe how our methods may be extended to designs with more than two stages in Section 5 and we end with conclusions and a discussion in Section 6.

2 | BAYES OPTIMAL DESIGNS

2.1 | The class of trial designs

Consider a confirmatory parallel-group clinical trial comparing a new treatment and a control with respect to a pre-defined primary endpoint. We assume the patient population may be divided into disjoint, biomarker-defined subgroups. Given a maximum achievable sample size, n, we aim to optimize the trial design by maximising a specific utility function.

Suppose two biomarker-defined subgroups have been identified before commencing the trial. Let $0 < \lambda < 1$ be the prevalence of the first subgroup in the underlying patient population and $1 - \lambda$ the prevalence of the second subgroup. Let θ_1 and θ_2 be the treatment effects, denoting the difference in the mean outcome between treatment and control, in the first and second subgroups, respectively. We consider trials to investigate the null hypotheses $H_{01}: \theta_1 \leq 0$ and $H_{02}: \theta_2 \leq 0$ with corresponding alternative hypotheses $H_{11}: \theta_1 > 0$ and $H_{12}: \theta_2 > 0$. In Sections 2.2 and 2.3 we consider confirmatory trials in which strong control of the FWER is imposed.[29] In our discussion of umbrella trials in Section 2.4, we assume multiplicity control is not required.

We consider optimization within a class of designs $\mathcal{A}$ that have a single interim analysis at which adaptation can take place. The total sample size is fixed at n with $s^{(1)}n$ patients in the first stage and $s^{(2)}n$ patients in the second stage, where $s^{(1)} > 0$, $s^{(2)} \geq 0$ and $s^{(1)} + s^{(2)} = 1$. In the first stage, $r_1^{(1)}s^{(1)}n$ patients are recruited from subgroup 1 and $r_2^{(1)}s^{(1)}n$ from subgroup 2, where $r_1^{(1)} \geq 0$, $r_2^{(1)} \geq 0$ and $r_1^{(1)} + r_2^{(1)} = 1$. In the second stage, $r_1^{(2)}s^{(2)}n$ patients are recruited from subgroup 1 and $r_2^{(2)}s^{(2)}n$ from subgroup 2, where $r_1^{(2)} \geq 0$, $r_2^{(2)} \geq 0$ and $r_1^{(2)} + r_2^{(2)} = 1$, and the values of $r_1^{(2)}$ and $r_2^{(2)}$ may depend on the first stage data. Within each stage and subgroup, we assume equal allocation to the two treatment arms (this assumption is not strictly necessary and could be relaxed). Figure 1 gives a schematic representation of the trial design.

The definition of a particular design in $\mathcal{A}$ is completed by specifying the multiple testing procedure to be used and the method for combining data across stages when adaptation occurs. We use a closed testing procedure to control FWER, applying a weighted Bonferroni procedure to test the intersection hypothesis. In this procedure, weights are initially set

FIGURE 1 Schematic representation of the three types of trial design. In the single-stage trial, the sampling prevalences of the subgroups are fixed throughout the trial. In standard adaptive enrichment trials, patients are recruited with predefined subgroup prevalences until the interim analysis, at which point a decision is taken to continue with the same prevalences or to sample from a single subgroup. In the Bayes optimal adaptive trial designs that we consider, the sampling prevalences may be changed at the interim analysis [Colour figure can be viewed at wileyonlinelibrary.com]

as $\omega_1^{(1)}$ and $\omega_2^{(1)}$ but these may be modified in the second stage if adaptation occurs. The error rate for each hypothesis test is controlled by preserving the conditional type I error rate when an adaptation is made. Thus, while we use a Bayesian approach to optimize the design, the trial is analyzed using frequentist procedures that control error rates at the desired level, adhering to conventional regulatory standards.

We follow a Bayesian decision theoretic approach to optimize over trial designs in the class $\mathcal{A}$. In assessing each design, we assume a prior distribution for the treatment effects in each subgroup and a utility function[30] that quantifies the value of the trial's outcome. We shall optimize designs with respect to the timing of the interim analysis, the proportion of patients recruited from the two subgroups at each stage of the trial, the weights in the weighted Bonferroni test, and the rule for updating these weights given the interim data.

We summarize the data observed during the trial by the symbol $\hat{\theta}$, noting that this summary should contain information about the numbers of observations from each subgroup and weights to be used in the weighted Bonferroni test at each stage, as well as estimates of θ_1 and θ_2 obtained from observations before and after the interim analysis. We define our utility function to be

$$\mathcal{U}(\hat{\theta}) = \lambda \, \mathbb{1}(\text{Reject } H_{01}) + (1 - \lambda) \, \mathbb{1}(\text{Reject } H_{02}), \tag{1}$$

where $\mathbb{1}(.)$ is the indicator function. By definition, the data summary $\hat{\theta}$ contains the information needed to determine if each of the hypotheses H_{01} and H_{02} is rejected.

The utility (1) involves the size of the underlying subgroups as well as the rejection of the corresponding hypotheses. Thus, rejection of the null hypothesis for a larger subgroup is given greater weight. If the population prevalence of the two subgroups is not known, a prior on λ may be added. We note that terms in the function (1) are positive when a null hypothesis is rejected but the associated treatment effect is very small or even negative: this issue could be addressed by multiplying each term by an indicator variable which takes the value 1 if the relevant parameter, θ_1 or θ_2, is larger than zero or above a clinically relevant threshold (eg, Stallard et al[31] where a similar approach is used for treatment selection).

Since the trial design is optimized with respect to the stated utility, it is important to choose a utility function that reflects accurately the relative importance of possible trial outcomes. Furthermore, the definition of utility can be adapted to reflect the interest of different stakeholders, for example, Ondra et al[21] and Graf, Posch and König[24] propose utility functions that represent the view of a sponsor or take a public health perspective.

Let $\pi(\theta)$ denote the prior distribution for $\theta = (\theta_1, \theta_2)$. Then, the Bayes expected utility for a trial design $a \in \mathcal{A}$ is

$$W_{\pi(\theta)}(a) = \mathbb{E}_{\pi(\theta)} \left[\mathbb{E}_\theta \left\{ \mathcal{U}(\hat{\theta}) \right\} \right],$$

where we have taken the expectation over the sampling distribution of the trial data given the true treatment effects $\boldsymbol{\theta}$, with an outer integral over the prior distribution $\pi(\boldsymbol{\theta})$.

When choosing the prior $\pi(\boldsymbol{\theta})$, it is important to remember that $W_{\pi(\theta)}(a)$ represents the expected utility, averaged over $\theta \sim \pi(\theta)$. If an "uninformative" prior is chosen, this will place weight on extreme scenarios, such as large negative treatment effects, which have little credibility. Thus, when considering the Bayes optimal design, it is important to use subjective, informative priors. In some cases, pilot studies or historic observational data may be available to construct the prior distribution.

In this paper, we assume the prior distribution $\pi(\boldsymbol{\theta})$ to be bivariate normal,

$$\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \sim N \left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \psi_1^2 & \rho \psi_1 \psi_2 \\ \rho \psi_1 \psi_2 & \psi_2^2 \end{pmatrix} \right). \tag{2}$$

Here, the correlation coefficient ρ reflects the belief about the existence of common factors that contribute to the treatment effects in the two subgroups.

2.2 | Bayes optimal single-stage design

2.2.1 | Patient recruitment and estimation

Suppose we wish to conduct a single-stage trial, which is the special case where $s^{(2)} = 0$, usually referred to as a stratified design. For simplicity of notation in this section, we write r_j and ω_j rather than $r_j^{(1)}$ and $\omega_j^{(1)}$ for $j = 1$ and 2. We assume patients can be recruited at these rates regardless of the true proportions λ and $1 - \lambda$ in the underlying patient population. In addition, we assume that patients are randomised between the new treatment and the control with a $1:1$ allocation ratio in each subgroup.

During the trial we observe a normally distributed endpoint for each patient and we assume a constant variance for all observations. For patient i from subgroup j on the new treatment we have $X_{ji} \sim N(\mu_{Tj}, \sigma^2)$, $i = 1, \ldots, r_j n/2$, and for patient i from subgroup j on the control treatment we have $Y_{ji} \sim N(\mu_{Cj}, \sigma^2)$, $i = 1, \ldots, r_j n/2$. The estimate of the treatment effect $\theta_j = \mu_{Tj} - \mu_{Cj}$ in subgroup j, is

$$\hat{\theta}_j = \overline{X}_j - \overline{Y}_j = \frac{1}{r_j n/2} \sum_{i=1}^{r_j n/2} X_{ji} - \frac{1}{r_j n/2} \sum_{i=1}^{r_j n/2} Y_{ji}, \quad j = 1, 2. \tag{3}$$

2.2.2 | Hypothesis testing in the single-stage design

Consider the case $s^{(2)} = 0$ and $0 < r_1 < 1$. Then

$$\hat{\theta}_j | \theta_j \sim N \left(\theta_j, \frac{4\sigma^2}{r_j n} \right), \quad j = 1, 2,$$

and the corresponding Z-values

$$Z_j = \frac{\hat{\theta}_j \sqrt{r_j n}}{2\sigma}, \quad j = 1, 2,$$

follow standard normal distributions under the null hypotheses H_{01} and H_{02}.

We use a closed testing procedure to ensure strong control of the FWER at α level.[32] To construct this, we require level α tests of $H_{01}: \theta_1 \leq 0$, $H_{02}: \theta_2 \leq 0$ and $H_{01} \cap H_{02}: \{\theta_1 \leq 0\} \cap \{\theta_2 \leq 0\}$. We reject H_{01} globally if the level α tests reject H_{01} and $H_{01} \cap H_{02}$. Similarly, we reject H_{02} globally if the level α tests reject H_{02} and $H_{01} \cap H_{02}$.

For the individual tests we reject H_{01} if $Z_1 \geq \Phi^{-1}(1 - \alpha)$ and H_{02} if $Z_2 \geq \Phi^{-1}(1 - \alpha)$. To test the intersection hypothesis, we use a weighted Bonferroni test: given predefined weights ω_1 and ω_2, where $\omega_1 + \omega_2 = 1$, we reject $H_{01} \cap H_{02}$ if $Z_1 \geq \Phi^{-1}(1 - \omega_1 \alpha)$ or $Z_2 \geq \Phi^{-1}(1 - \omega_2 \alpha)$. The resulting closed testing procedure is equivalent to the weighted Bonferroni-Holm test and will be generalised to adaptive tests in Section 2.3.

We note that the choice of a closed testing procedure is not restrictive in this setting since any procedure that gives strong control of the FWER may be written as a closed testing procedure.[22,23] Furthermore in the special cases $r_1 = 1$ and $r_2 = 1$, where the trial recruits from only one of the subgroups, just one subgroup is tested and only the test of the individual hypothesis is required. These cases are accommodated in our general class of designs by setting $\omega_1 = 1$ when $r_1 = 1$ and $\omega_2 = 1$ when $r_2 = 1$.

2.2.3 | Bayesian optimization

In the single-stage trial we wish to optimize the trial prevalences of each subgroup, r_1 and r_2, and the weights in the Bonferroni-Holm procedure, ω_1 and ω_2. Given the constraints $r_1 + r_2 = 1$ and $\omega_1 + \omega_2 = 1$, we denote the set of parameters to optimize by $a = (r_1, \omega_1)$.

Let $f(\hat{\theta}|\theta, a)$ denote the conditional distribution of $(\hat{\theta}_1, \hat{\theta}_2)$ given θ for design parameters a. The Bayes expected utility is given by

$$\mathbb{E}_{\pi(\theta)}\left[\mathbb{E}_{\theta}\left[\mathcal{U}(\hat{\theta})\right]\right] = \int_{\theta}\int_{\hat{\theta}} \mathcal{U}(\hat{\theta})f(\hat{\theta}|\theta, a)\pi(\theta)\, d\hat{\theta}\, d\theta.$$

The Bayes optimal design is given by the pair $a = (r_1, \omega_1)$ that maximises the Bayes expected utility of the trial, that is

$$\underset{a}{\arg\max} \int_{\theta}\int_{\hat{\theta}} \mathcal{U}(\hat{\theta})f(\hat{\theta}|\theta, a)\pi(\theta)\, d\hat{\theta}\, d\theta.$$

Given our simple choices for the prior distribution and the utility function this integral may be computed directly (see Section S1.2 of Appendix S1). We find the Bayes optimal single-stage trial by a numerical search over possible values of a.

2.3 | Bayes optimal two-stage adaptive design

2.3.1 | Adding a second stage

Consider now a two-stage design in which data from the first stage inform adaptations in the second stage. The estimate of θ_j for subgroup j based on data collected in stage k is

$$\hat{\theta}_j^{(k)} = \overline{X}_j^{(k)} - \overline{Y}_j^{(k)}, \quad j = 1, 2, \ k = 1, 2, \tag{4}$$

where $\overline{X}_j^{(k)}$ and $\overline{Y}_j^{(k)}$ are the mean responses in subgroup j in stage k for the treatment arm and control arm, respectively. Given the value of $\theta = (\theta_1, \theta_2)$, the first stage estimates are independent with distributions

$$\hat{\theta}_j^{(1)}|\theta_j \sim N\left(\theta_j, \frac{4\sigma^2}{r_j^{(1)}s^{(1)}n}\right), \quad j = 1, 2.$$

The trial prevalences, $r_1^{(2)}$ and $r_2^{(2)}$, of the two subgroups in the second stage are dependent on $\hat{\theta}_1^{(1)}$ and $\hat{\theta}_2^{(1)}$ but, conditional on $r_1^{(2)}$ and $r_2^{(2)}$, the second-stage estimates are independent and conditionally independent of $\hat{\theta}_1^{(1)}$ and $\hat{\theta}_2^{(1)}$ with

$$\hat{\theta}_j^{(2)}|r_j^{(2)}, \theta_j \sim N\left(\theta_j, \frac{4\sigma^2}{r_j^{(2)}s^{(2)}n}\right), \quad j = 1, 2.$$

2.3.2 | Hypothesis testing in the two-stage adaptive design

There is a variety of approaches to test multiple hypotheses in a two-stage adaptive design.[33-36] We shall use a closed testing procedure to ensure strong control of the FWER at level α, as we did for the single-stage design in Section 2.2.2. In

constructing level α tests of the null hypotheses H_{01}, H_{02} and $H_{01} \cap H_{02}$ we employ the conditional error rate approach.[37,38] Based on a reference design and its predefined tests, we calculate the conditional error rate for each hypothesis and define adaptive tests which preserve this conditional error rate, thereby controlling the overall type I error rate.

Consider a reference design in which the trial prevalences of subgroups 1 and 2 and the weights in the weighted Bonferroni test of $H_{01} \cap H_{02}$ remain the same across stages, so $r_j^{(2)} = r_j^{(1)}$ and $\omega_j^{(2)} = \omega_j^{(1)}$ for $j = 1$ and 2. In the reference design, tests are performed by pooling the stage-wise data within each subgroup and treatment arm, and using the conventional test statistics, as for the single-stage test. For $j = 1$ and 2, the pooled estimate of θ_j across the two stages of the trial is

$$\hat{\theta}_j^{(p)} = s^{(1)} \, \hat{\theta}_j^{(1)} + s^{(2)} \, \hat{\theta}_j^{(2)},$$

with corresponding Z-value

$$Z_j^{(p)} = \frac{\hat{\theta}_j^{(p)}}{\sqrt{4\sigma^2/(r_j^{(1)}n)}},$$

and the null hypothesis H_{0j} is rejected at level α if $Z_j^{(p)} > \Phi^{-1}(1 - \alpha)$. Let

$$Z_j^{(1)} = \frac{\hat{\theta}_j^{(1)}}{\sqrt{4\sigma^2/(r_j^{(1)}s^{(1)}n)}}, \quad j = 1, 2,$$

then the conditional distribution of $Z_j^{(p)}$ given the interim data is

$$Z_j^{(p)} | Z_j^{(1)}, \theta_j \sim N \left(\sqrt{s^{(1)}} Z_j^{(1)} + s^{(2)} \theta_j \frac{\sqrt{r_j^{(1)}n}}{2\sigma}, \ s^{(2)} \right),$$

and the conditional error rates for the tests of H_{0j} are

$$A_j = \mathbb{P}\left(Z_j^{(p)} > \Phi^{-1}(1 - \alpha) \mid Z_j^{(1)}, \theta_j = 0 \right), \quad j = 1, 2. \tag{5}$$

Similarly, the conditional error rate for the test of $H_{01} \cap H_{02}$ is

$$A_{12} = \mathbb{P}\left\{ Z_1^{(p)} > \Phi^{-1}(1 - \omega_1^{(1)}\alpha) \ \text{or} \ Z_2^{(p)} > \Phi^{-1}(1 - \omega_2^{(1)}\alpha) \mid Z_1^{(1)}, Z_2^{(1)}, \theta_1 = \theta_2 = 0 \right\}. \tag{6}$$

See Section S1.1 of Appendix S1 for further details on the derivations of the conditional distributions.

In the adaptive design, if no adaptations are made at the interim analysis we apply the tests as defined for the reference design. Suppose now that adaptations are made and the trial prevalences in stage 2 are set to be $r_1^{(2)}$ and $r_2^{(2)}$ with weights $\omega_1^{(2)}$ and $\omega_2^{(2)}$ for the weighted Bonferroni test. In this case, we calculate the conditional error rates A_1, A_2 and A_{12} prior to adaptation from Equations (5) and (6). We then define tests of H_{01}, H_{02} and $H_{01} \cap H_{02}$ based on stage 2 data alone that have these conditional error rates as their type 1 error probabilities. Given the updated $r_1^{(2)}$ and $r_2^{(2)}$,

$$Z_j^{(2)} \mid r_j^{(2)}, \theta_j \sim N \left(\theta_j \frac{\sqrt{r_j^{(2)}s^{(2)}n}}{2\sigma}, \ 1 \right), \quad j = 1, 2.$$

Thus, in our level α tests, we reject H_{01} if $Z_1^{(2)} > \Phi^{-1}(1 - A_1)$, we reject H_{02} if $Z_2^{(2)} > \Phi^{-1}(1 - A_2)$ and, applying a weighted Bonferroni test with weights $\omega_1^{(2)}$ and $\omega_2^{(2)}$, we reject $H_{01} \cap H_{02}$ if $Z_1^{(2)} > \Phi^{-1}(1 - \omega_1^{(2)}A_{12})$ or $Z_2^{(2)} > \Phi^{-1}(1 - \omega_2^{(2)}A_{12})$. Finally, following the closed testing procedure, we reject H_{01} globally if the level α tests reject H_{01} and $H_{01} \cap H_{02}$ and we reject H_{02} globally if the level α tests reject H_{02} and $H_{01} \cap H_{02}$.

2.3.3 | Two-stage optimization

We denote the set of initial design parameters by $a_1 = (s^{(1)}, r_1^{(1)}, \omega_1^{(1)})$ and the second-stage parameters by $a_2 = (r_1^{(2)}, \omega_1^{(2)})$. Let $\hat{\theta}^{(1)} = (\hat{\theta}_1^{(1)}, \hat{\theta}_2^{(1)})$ and $\hat{\theta}^{(2)} = (\hat{\theta}_1^{(2)}, \hat{\theta}_2^{(2)})$ be the vectors of estimated treatment effects in each subgroup, based on the first and second-stage data, respectively, as defined in Equation (4). Denote the conditional distributions of the estimated effects in each stage of the trial by $f_1(\hat{\theta}^{(1)}|\theta, a_1)$ and $f_2(\hat{\theta}^{(2)}|\theta, a_2)$ and the posterior distribution of θ given the stage 1 observations by $\pi(\theta|\hat{\theta}^{(1)}, a_1)$. Then, the Bayes expected utility can be written as

$$\mathbb{E}_{\pi(\theta)} \left[\mathbb{E}_{\theta} \left[\mathcal{U}(\hat{\theta}) \right] \right] = \int_{\theta}\int_{\hat{\theta}^{(1)}} \int_{\hat{\theta}^{(2)}} \mathcal{U}(\hat{\theta}) f_2(\hat{\theta}^{(2)}|\theta, a_2) f_1(\hat{\theta}^{(1)}|\theta, a_1) \pi(\theta) \, \mathrm{d}\hat{\theta}^{(2)} \mathrm{d}\hat{\theta}^{(1)} \mathrm{d}\theta. \quad (7)$$

We find the optimal combination of design parameters a_1 before stage 1 and a_2 before stage 2 using the backward induction principle. First we construct the Bayes optimal a_2 for all possible $\hat{\theta}^{(1)}$ and a_1. Then we construct the Bayes optimal a_1 given that the optimal a_2 will be used in the second stage of the trial.

Optimizing the decision at the interim analysis
Denoting the marginal distribution of $\hat{\theta}^{(1)}$ by $f_1(\hat{\theta}^{(1)}, a_1)$, we have

$$\pi(\theta) f_1(\hat{\theta}^{(1)}|\theta, a_1) = f_1(\hat{\theta}^{(1)}, a_1) \, \pi(\theta|\hat{\theta}^{(1)}, a_1),$$

and the right-hand side of Equation (7) can be written as

$$\int_{\hat{\theta}^{(1)}} f_1(\hat{\theta}^{(1)}, a_1) \int_{\theta}\int_{\hat{\theta}^{(2)}} \mathcal{U}(\hat{\theta}) f_2(\hat{\theta}^{(2)}|\theta, a_2) \pi(\theta|\hat{\theta}^{(1)}, a_1) \, \mathrm{d}\hat{\theta}^{(2)} \, \mathrm{d}\theta \, \mathrm{d}\hat{\theta}^{(1)}.$$

Thus, given a_1 and $\hat{\theta}^{(1)}$, the Bayes optimal decision for the second stage is the choice of a_2 that maximises

$$W_2(a_2, a_1, \hat{\theta}^{(1)}) = \int_{\theta}\int_{\hat{\theta}^{(2)}} \mathcal{U}(\hat{\theta}) f_2(\hat{\theta}^{(2)}|\theta, a_1) \pi(\theta|\hat{\theta}^{(1)}, a_1) \, \mathrm{d}\hat{\theta}^{(2)} \, \mathrm{d}\theta.$$

For known values of $\hat{\theta}^{(1)}$ and a_1, we can find the conditional error rates A_1, A_2, and A_{12} used in hypothesis testing in stage 2, hence we may evaluate $\mathcal{U}(\hat{\theta})$ for given $a_1, \hat{\theta}^{(1)}, a_2$, and $\hat{\theta}^{(2)}$. Our choices for the prior distribution and utility function mean that it is quite straightforward to compute $W_2(a_2, a_1, \hat{\theta}^{(1)})$ for given a_1, a_2 and $\hat{\theta}^{(1)}$. Thus, we are able to perform a numerical search seeking

$$\underset{a_2}{\mathrm{argmax}} \, W_2(a_2, a_1, \hat{\theta}^{(1)}),$$

to find the Bayes optimal a_2.

Overall trial optimization
Having found the Bayes optimal parameters a_2 for the second stage of the trial as a function of $(a_1, \hat{\theta}^{(1)})$, we determine a_1, the Bayes optimal choice for the initial parameters, as

$$\underset{a_1}{\mathrm{argmax}} \int_{\theta}\int_{\hat{\theta}^{(1)}} W_2(a_2, a_1, \hat{\theta}^{(1)}) f(\hat{\theta}^{(1)}|\theta, a_1) \pi(\theta) \, \mathrm{d}\hat{\theta}^{(1)} \mathrm{d}\theta.$$

We conduct a search over possible values of a_1 to maximize the above integral and find the optimal choice of a_1. Computing the integral for a given value of a_1 by numerical integration is not straightforward. Instead, we have used Monte Carlo simulation to carry out this calculation for each value of a_1.

2.4 | Bayes optimal umbrella trials

We now consider the case of umbrella trials, where it has been argued that no multiplicity adjustment is required as the hypotheses to be tested concern different experimental treatments targeted to different molecular markers or subgroups.[28]

TABLE 1 The scenarios considered in the numerical examples. The term "opt" indicates that parameters were optimized, while "N/A" means the parameters are not applicable. The parameters θ_1 and θ_2 are either specified by a prior distribution in which $\psi_1 = \psi_2 = \psi$ or specific values of θ_1 and θ_2 are given

		λ	μ_1	μ_2	ψ	ρ	$s^{(1)}$	$r_1^{(1)}$	$\omega_1^{(1)}$	θ_1	θ_2
Figure 2	Single-stage	0.3	0 to 0.3	0, 0.2	0.02 to 0.44	0.5	N/A	opt	opt	prior	
Figure S2	Single-stage	0.3	0 to 0.3	0, 0.2	0.2	-1 to 1	N/A	opt	opt	prior	
Figure 3	Interim decision	0.3	0.1	0	0.2	-0.8, 0.5	0.25, 0.5	0.3	0.3	prior	
Figure 4	Two-stage	0.3	0, 0.3	0, 0.2	0.2	0.5	0.1 to 0.9	opt	opt	prior	
Figure 5	Two-stage	0.3	0 to 0.3	0, 0.2	0.02 to 0.4	0.5	opt	opt	opt	prior	
Figure S10	Two-stage	0.3	0 to 0.3	0, 0.2	0.2	-0.8 to 0.8	opt	opt	opt	prior	
Figures 6 and S11	Power	0.3	0.1, 0.2	0	0.2	0.5	opt	opt	opt	0 to 0.3	0, 0.2

Since each treatment is assessed separately, an umbrella trial can be viewed a set of independent trials even though they are run under a single protocol.

We consider umbrella trials with two subgroups, as in the previous sections. However, without multiplicity adjustment, the hypothesis testing procedure reduces to testing the elementary hypotheses H_{01} and H_{02} each at level α. In applying the conditional error rate approach, only the computation of conditional error rates A_1 and A_2 from Equation (5) is required. Then, with $Z_1^{(2)}$ and $Z_2^{(2)}$ denoting the test statistics based on second-stage data only, H_{01} is rejected if $Z_1^{(2)} > \Phi^{-1}(1 - A_1)$ and H_{02} is rejected if $Z_2^{(2)} > \Phi^{-1}(1 - A_2)$. No test of the intersection hypothesis is performed.

Design parameters are optimized with respect to the utility function in Equation (1). To frame the optimization problem in the same way as in the previous sections, the interim decision in a two-stage umbrella trial will optimize only the second-stage subgroup trial prevalences, so $a_2 = (r_1^{(2)})$, while in the first stage we optimize the subgroup trial prevalences and the timing of the interim analysis, so $a_1 = (s^{(1)}, r_1^{(1)})$. In the case of a single-stage umbrella trial, only the subgroup prevalences are optimized, so $a = (r_1)$. We have used a normal prior distribution, as defined in Equation (2), in optimizing the design parameters of single-stage and two-stage trials. In the case of two-stage designs, the interim analysis uses the test statistics from the first stage and the prior distribution to perform adaptations and the final tests are performed using the conditional error rate approach.

3 | NUMERICAL EXAMPLES AND COMPARISONS

In this section, we give numerical examples of optimized single-stage and two-stage designs in a range of scenarios. We show results for cases with and without multiplicity correction, referring to these as enrichment and umbrella trials, respectively. Additionally, we illustrate the optimization of the decision rule at the interim analysis. In Table 1, we provide an overview of the scenarios considered and the parameters that are optimized.

3.1 | Optimal single-stage designs

In studying the impact of the prior distribution on optimized trial design parameters $a = (r_1, \omega_1)$ for single-stage designs, we consider studies where the response variance is $\sigma^2 = 1$ and the total sample size is fixed at $n = 700$. We assume a multivariate normal prior distribution for θ as defined in Equation (2) with parameters μ_1, μ_2, $\psi_1 = \psi_2 = \psi$ and ρ, and we compute optimal designs for a variety of such priors. The FWER in enrichment designs and the per-comparison error rate in umbrella designs is fixed at $\alpha = 0.05$.

In Figure 2 we display the effect of the prior SD on the optimal design parameters when the population prevalence of subgroup 1 is $\lambda = 0.3$. We considered prior SDs ψ of 0.02, 0.0632, 0.1, 0.1414, 0.2, 0.3162, and 0.44, corresponding to information from studies with 10 000, 1000, 400, 200, 100, 40 and 20 subjects in each subgroup.

The mean and variance of the prior distribution have a large impact on the optimal design parameters r_1 and ω_1. The optimal values of r_1 and ω_1 and the expected utility of the resulting designs are very similar for enrichment and umbrella designs. If $\mu_1 > 0$ and $\mu_2 = 0$, optimal values of r_1 and ω_1 are larger than 0.3, the population prevalence of subgroup 1, so
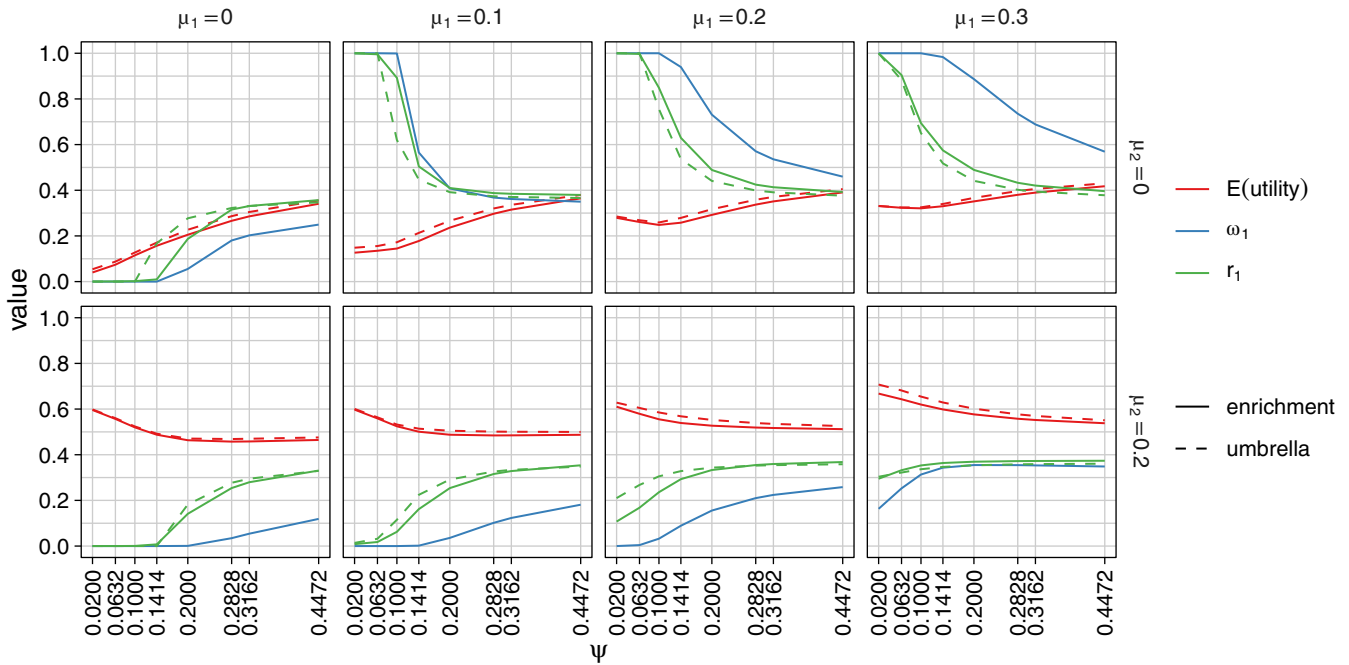
FIGURE 2 Optimized design parameters for single-stage designs and the expected utility, averaged over the prior. Parameters are $a = (r_1, \omega_1)$ for enrichment trials and $a = (r_1)$ for umbrella trials. Results are classified by μ_1 and μ_2, the prior means of θ_1 and θ_2, and the prior SD $\psi = \psi_1 = \psi_2$. The prior correlation between θ_1 and θ_2 is fixed at $\rho = 0.5$ and the population prevalence of subgroup 1 is assumed to be $\lambda = 0.3$ [Colour figure can be viewed at wileyonlinelibrary.com]

the design over-samples this subgroup. If $\mu_1 = 0$ and $\mu_2 > 0$, the optimal design under-samples subgroup 1. When both μ_1 and μ_2 are greater than zero, the optimal design has $r_1 < 0.5$ and $\omega_1 < 0.5$, reflecting the fact that it is advantageous to sample more subjects from subgroup 2 and allocate more type 1 error probability to the test of H_{02} since $\lambda = 0.3$ implies that $P(\text{Reject } H_{02})$ has a greater weight than $P(\text{Reject } H_{01})$ in the utility function.

In extreme cases where $\mu_1 = 0$, $\mu_2 \geq 0$ and the prior variance is small, the optimal design has $r_1 = 0$, so only subgroup 2 is sampled. When $\mu_1 > 0$, $\mu_2 = 0$ and the prior variance is small, the optimal design has $r_1 = 1$ and only subgroup 1 is sampled.

In Figure S2, we show the effect of the prior correlation ρ on the design parameters when the prior SD is $\psi = 0.2$. We observe that the correlation has an impact on the optimal weight ω_1 for testing the intersection hypothesis, in particular, when the treatment effects θ_1 and θ_2 have a high positive correlation, it is better to place most weight on one hypothesis rather than split the weight between the two hypotheses.

In Figures S3 and S4 we present further results for different values of λ, varying ρ in Figure S3 and ψ in Figure S4. Since the utility to be maximized depends on the population prevalences, the optimal design parameters vary considerably with λ. We see from Figure S3 that ρ has only a small impact on the optimal value of r_1 when adjusting for multiplicity and no impact at all in umbrella designs where no multiplicity adjustment is made. Figure S4 shows that the dependence of optimal design parameters on ψ is similar to that seen in Figure 2: when the prior variance is large the optimal choices for r_1 and ω_1 are close to λ, while for smaller variances the optimal designs depend on the prior means μ_1 and μ_2 as well as λ.

3.2 | Optimal two-stage designs

Figure 3 illustrates optimal adaptation rules for two-stage designs. In these examples $n = 700$, $\sigma^2 = 1$, the population prevalence of subgroup 1 is $\lambda = 0.3$, and the prior distribution for θ has parameters $\mu_1 = 0.1$, $\mu_2 = 0$, $\psi_1 = \psi_2 = 0.2$ and $\rho = 0.5$ or -0.8. The first-stage design parameters have not been optimized and are set as $r_1^{(1)} = \omega_1^{(1)} = 0.3$ with $s^{(1)}$ equal to 0.25 or 0.5. The FWER in enrichment designs and the per-comparison error rate in umbrella designs is fixed at $\alpha = 0.05$.

The adaptation rules specify the second-stage design parameters $a_2 = (r_1^{(2)}, \omega_1^{(2)})$ that optimize the expected utility, as defined in Equation (1), given the first stage statistics $Z_1^{(1)}$ and $Z_2^{(1)}$. The optimal $r_1^{(2)}$ and $\omega_1^{(2)}$ are calculated using the
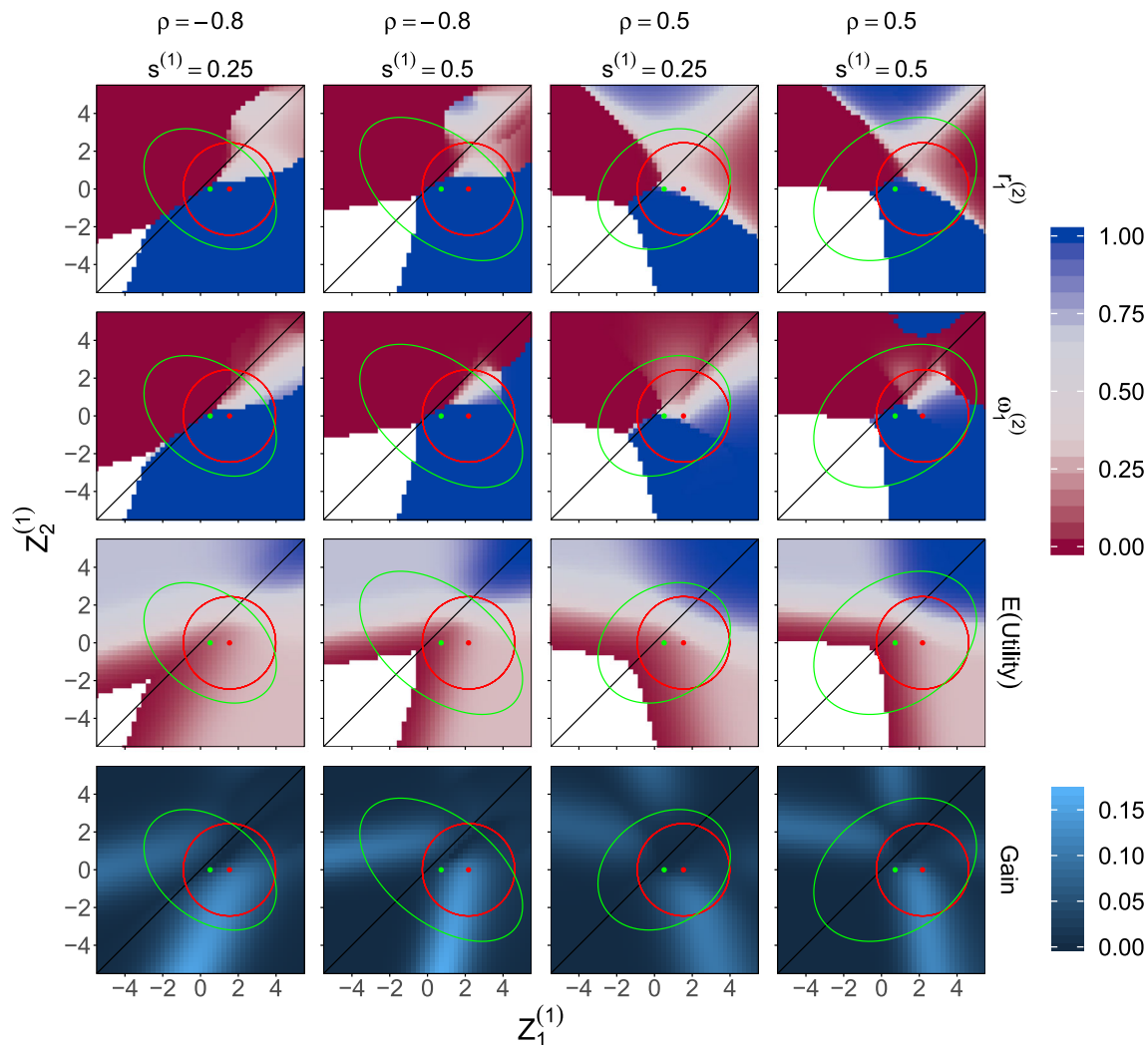
FIGURE 3 Examples of optimal adaptation rules when $\lambda = 0.3$, the prior distribution for θ has parameters $\mu_1 = 0.1$, $\mu_2 = 0$, $\psi_1 = \psi_2 = 0.2$ and $\rho = 0.5$ or -0.8, and first stage design parameters are set as $r_1^{(1)} = \omega_1^{(1)} = 0.3$ and $s^{(1)} = 0.25$ or 0.5. Optimized values of $r_1^{(2)}$ and $\omega_1^{(2)}$ are shown for each combination of first stage Z-values $Z_1^{(1)}$ and $Z_2^{(1)}$. Also shown are the conditional expected utility when the trial proceeds using the optimized values of $r_1^{(2)}$ and $\omega_1^{(2)}$ and the increase in conditional expected utility compared to continuing with no adaptation. In each plot, the red circle indicates the 95% highest density region for the distribution of $(Z_1^{(1)}, Z_2^{(1)})$ when the true treatment effects are $\theta_1 = 0.3$ and $\theta_2 = 0$ and the green ellipse indicates the 95% highest density region for the prior predictive distribution of $(Z_1^{(1)}, Z_2^{(1)})$. The white regions contain values of $(Z_1^{(1)}, Z_2^{(1)})$ for which the maximum conditional expected utility is below 0.01. In these cases the numerical optimization becomes unstable and optimal values for $r_1^{(2)}$ and $\omega_1^{(2)}$ are not displayed [Colour figure can be viewed at wileyonlinelibrary.com]

Hooke-Jeeves derivative-free minimization algorithm through the `hjkb` function in the `dfoptim` package[39] in R.[40] We also calculated the conditional expected utility if the trial continued with no adaptation, so $r_1^{(2)} = r_1^{(1)}$ and $\omega_1^{(2)} = \omega_1^{(1)}$, and the plots in the bottom row of Figure 3 show the gain in the conditional expected utility due to the optimized adaptation. In Section S3 of Appendix S1, we present optimal interim rules for further values of λ.

In Figure 4, we illustrate the procedure for optimizing first-stage design parameters, $a = (s^{(1)}, r_1^{(1)}, \omega_1^{(1)})$ for an enrichment design or $a = (s^{(1)}, r_1^{(1)})$ for an umbrella design. For each combination of prior parameters and first-stage design parameters a, we generated 1000 samples of first-stage data under treatment effects drawn from the prior distribution. For each first-stage dataset, we found the optimal second-stage design parameters and noted the conditional expected utility using these optimal parameters. We took the average of the 1000 values of the optimized conditional expected utility as our simulation-based estimate of the expected utility for this choice of a. The optimal first-stage design parameters for a given prior distribution are those values of $s^{(1)}$, $r_1^{(1)}$, and in the case of an enrichment design $\omega_1^{(1)}$, that yield the highest expected utility.
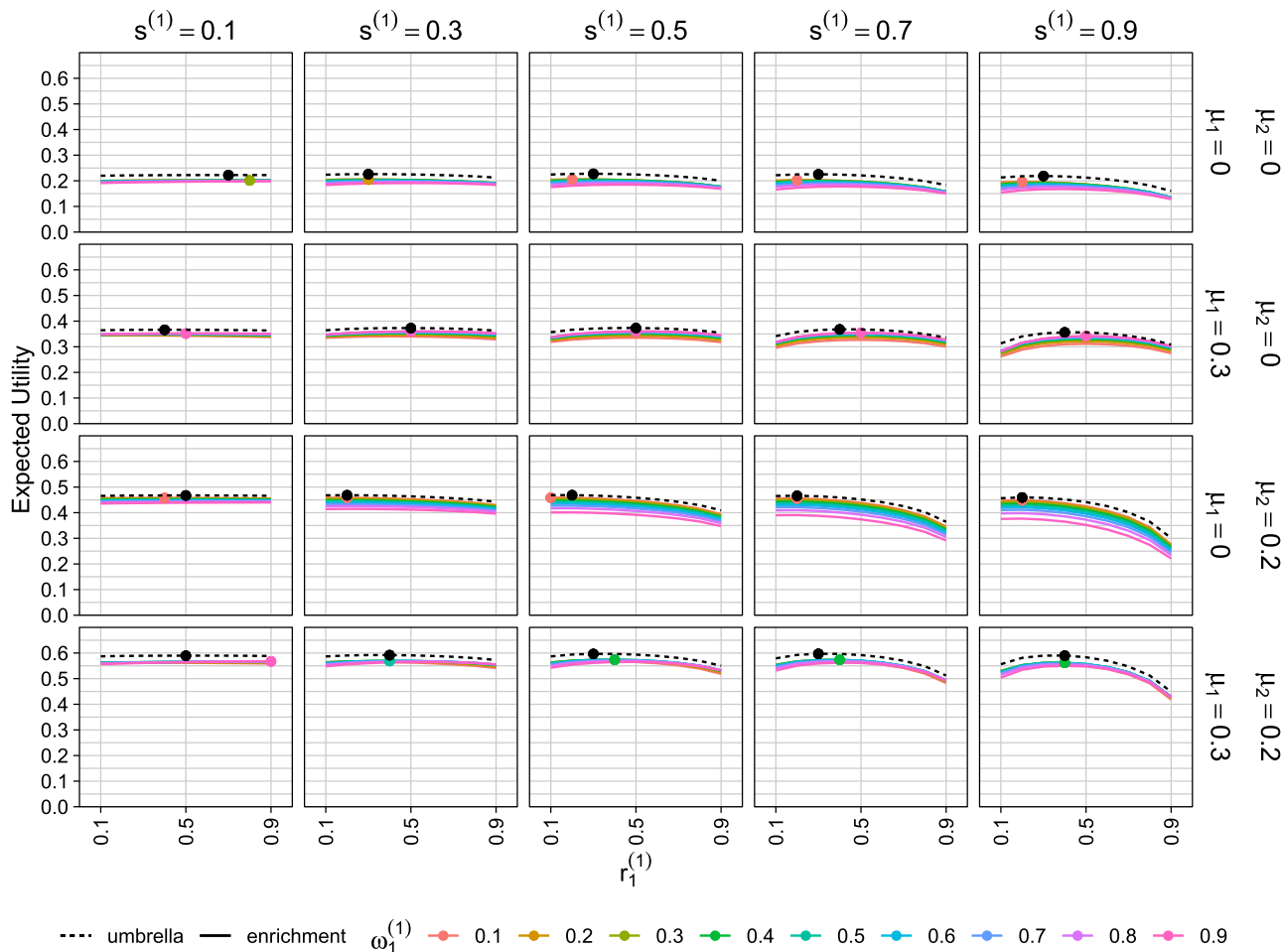
FIGURE 4 Optimization of first-stage design parameters. The population prevalence of subgroup 1 is $\lambda = 0.3$ and the prior distribution for θ has parameters $\mu_1 = 0$ or 0.3, $\mu_2 = 0$ or 0.2, $\psi_1 = \psi_2 = 0.2$ and $\rho = 0.5$. Each column shows results for a different value of $s_1^{(1)}$. The plots show the expected utility as a function of $r_1^{(1)}$, with coloured solid lines for different values of $\omega_1^{(1)}$ in an enrichment trial and black dashed lines for an umbrella trial with no multiplicity adjustment. In each panel, the colored dot indicates the combination of $r_1^{(1)}$ and $\omega_1^{(1)}$ that yields the maximum expected utility for an enrichment design and the black dot shows the optimum value of $r_1^{(1)}$ for an umbrella design [Colour figure can be viewed at wileyonlinelibrary.com]

Our results show the impact of the prior distribution on the optimized trial design parameters. The flat lines when $s^{(1)} = 0.1$ indicate that the expected utility is hardly affected by the choice of $r_1^{(1)}$ and $\omega_1^{(1)}$ when the interim analysis is performed early in the trial. When the interim analysis is performed later, the choice of first-stage design parameters is more important. It should be noted that for each pair of prior means (μ_1, μ_2), expected utility close to the overall optimum can be achieved using a wide range of first-stage design parameters as long as the second-stage design is optimized, given the first-stage data.

In Figures 5 and S10 we present optimized values of the first-stage design parameters, $s^{(1)}$, $r_1^{(1)}$, and $\omega_1^{(1)}$, given that optimal values of the second-stage design parameters will be used following the interim analysis. The results are similar to those observed for optimal single-stage designs. The prior variance has a large impact on the first-stage optimal design: for smaller variances, interim analyses closer to the beginning of the trial yield a larger expected utility, while with larger variances, interim analyses after around 40% to 60% of the patients have been recruited are preferable. When the prior means are both 0 the optimal design parameters $r_1^{(1)}$ and $\omega_1^{(1)}$ are close to the subgroup 1 prevalence λ. However, if the prior suggests a benefit is more likely in subgroup 1, the optimal design over-samples this subgroup, increasing its trial prevalence and testing weight. Figure S10 shows that, for enrichment designs, the prior correlation ρ has a large impact on the choice of $\omega_1^{(1)}$ but little effect on the optimal trial prevalences.

As for single-stage designs, the optimal values of $r_1^{(1)}$ are similar for enrichment Stand umbrella designs. A notable difference is that while the prior correlation ρ has no effect at all on the optimal values of r_1 in a single-stage umbrella
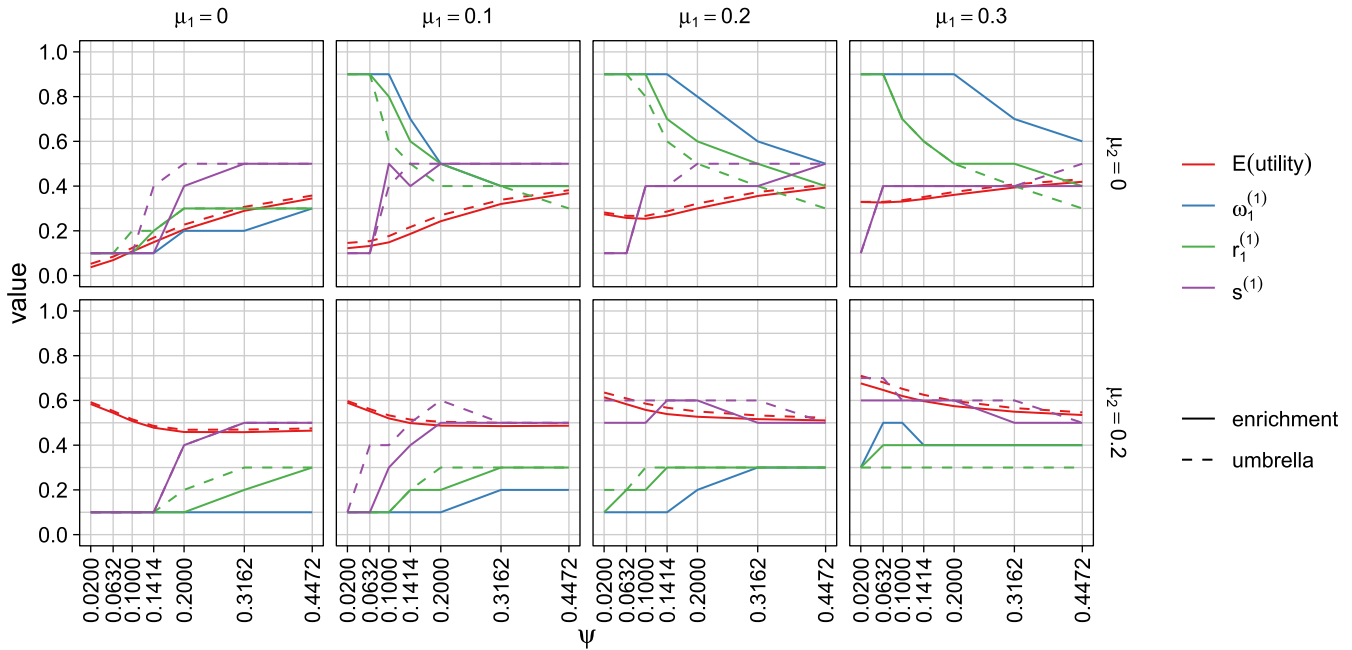
FIGURE 5 Optimized design parameters for two-stage designs and the expected utility, averaged over the prior. Parameters are $a = (s^{(1)}, r_1^{(1)}, \omega_1^{(1)})$ for enrichment trials and $a = (s^{(1)}, r_1^{(1)})$ for umbrella trials. Results are classified by μ_1 and μ_2, the prior means for θ_1 and θ_2, and by the prior SD $\psi = \psi_1 = \psi_2$. The prior correlation between θ_1 and θ_2 is fixed at $\rho = 0.5$ and the population prevalence of subgroup 1 is assumed to be $\lambda = 0.3$ [Colour figure can be viewed at wileyonlinelibrary.com]

design, the optimal value of $r_1^{(1)}$ in a two-stage umbrella design does show a small dependence on ρ. In the case of a single-stage umbrella design, the marginal distributions of $\hat{\theta}_1$ and $\hat{\theta}_2$ do not depend on ρ and thus, with no multiplicity adjustment in testing H_{01} and H_{02}, the expected value of the utility defined in Equation (1) does not depend on ρ. However, in a two-stage umbrella trial, the optimal choice of $r_1^{(2)}$ and the resulting conditional expected utility depends on both $\hat{\theta}_1^{(1)}$ and $\hat{\theta}_2^{(1)}$ and it is the joint distribution of $(\hat{\theta}_1^{(1)}, \hat{\theta}_1^{(2)})$, which depends on ρ, that determines the optimal value of $r_1^{(1)}$.

It should be noted that the procedures we have described impose a high computational burden. While it is relatively straightforward to optimize the decision at the interim analysis, the overall optimization of the trial is performed using simulations over a grid of values for the first-stage design parameters. More rapid computation of the optimal values may be achieved by using approximations to the utility when extreme first-stage values are observed, for example, if both $Z_1^{(1)}$ and $Z_2^{(1)}$ are large and negative, the expected utility is practically zero for all choices of $r_1^{(2)}$ and $\omega_1^{(2)}$. In practice, one may wish to add the option of stopping the trial for futility if extreme negative results are observed at the interim analysis. The methods we have presented can be extended to find efficient designs that incorporate this option by working with a utility of the form

$$\lambda \, \mathbb{1}(\text{Reject } H_{01}) + (1 - \lambda) \, \mathbb{1}(\text{Reject } H_{02}) + k \, s^{(2)} \, n \, \mathbb{1}(\text{Stop at the interim analysis}),$$

assigning a positive value k to each observation saved by early stopping.

3.3 | Performance of the Bayes optimal design under specific alternative hypotheses

In this section we consider adaptive designs optimized for a particular prior distribution for $\theta = (\theta_1, \theta_2)$ but we evaluate their performance under specific values of θ. We consider trials with a total sample size $n = 700$, response variance $\sigma^2 = 1$, and population prevalence of subgroup 1 equal to $\lambda = 0.3$. As a benchmark for comparison, we consider a nonoptimized, single-stage design with $r_1 = \lambda$ and $\omega_1 = 0.5$. We derive and assess the performance of single-stage designs for which design parameters r_1 and ω_1 are optimized as described in Section 2.2, and we derive and assess two-stage designs for which first-stage design parameters and the adaptation rule are optimized as described in Section 2.3. In optimizing
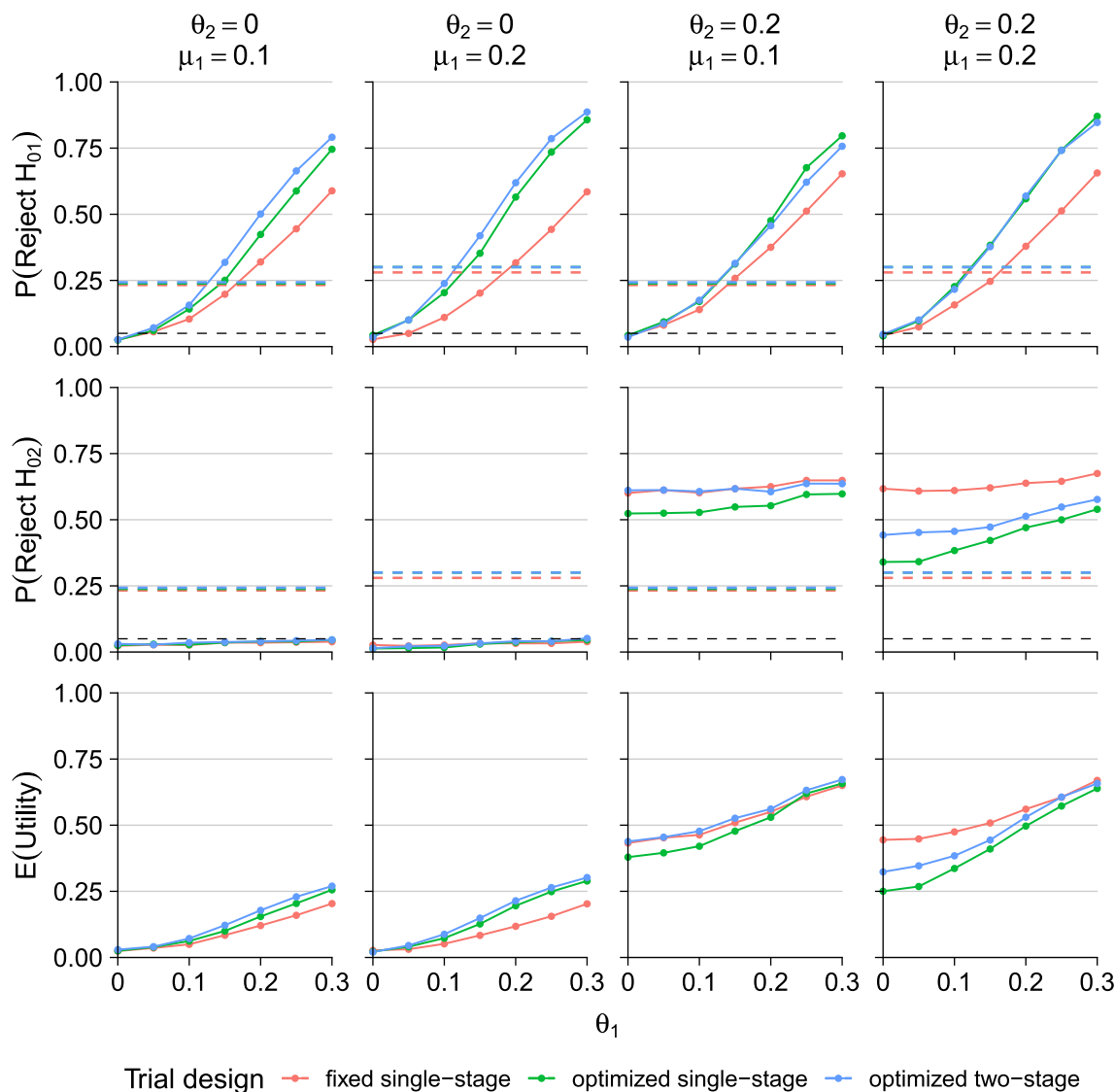
FIGURE 6 Operating characteristics of enrichment trials. The prior distribution for subgroup treatment effects (θ_1, θ_2) is normal with means $\mu_1 = 0.1$ or 0.2 and $\mu_2 = 0$, SDs $\psi_1 = \psi_2 = 0.2$ and correlation $\rho = 0.5$. The total sample size is 700 and the population prevalence of subgroup 1 is $\lambda = 0.3$. Results are given for θ_1 ranging from 0 to 0.3 and $\theta_2 = 0$ or 0.2. The black dashed lines in the two top rows are placed at 0.05 as reference to the significance level, while the dashed lines in the third row indicates the expected utility of the trial given the initial design parameters [Colour figure can be viewed at wileyonlinelibrary.com]

designs, we assume the normal prior distribution for θ presented in Equation (2) with $\mu_1 = 0.1$ or 0.2, $\mu_2 = 0$, $\psi_1 = \psi_2 = 0.2$ and $\rho = 0.5$. These priors reflects the belief that a treatment benefit is more likely in subgroup 1. The prior SD of 0.2 corresponds to information from a trial with 100 subjects in each subgroup.

We evaluate the operating characteristics of the designs for values of θ_1 ranging from 0 to 0.3 and $\theta_2 = 0$ or 0.2. This creates scenarios with a treatment effect in only one subgroup when $\theta_2 = 0$ or with a treatment effect in both subgroups when $\theta_2 = 0.2$ and $\theta_1 > 0$. Figure 6 presents simulation results for enrichment trials and Figure S11 presents results for umbrella trials. The plots show the probabilities of rejecting H_{01} and H_{02} and the average utility at the end of the trial for a variety of combinations of μ_1, μ_2, θ_1, and θ_2. For the scenarios considered, we see that optimizing the trial for the assumed priors leads to a substantial increase in the power to reject H_{01} as compared to the nonoptimized, single-stage design. However, the optimized designs have lower power to reject H_{02} when $\theta_2 = 0.2$. The optimized designs have a higher average utility than the nonoptimized design when $\theta_2 = 0$. If $\theta_2 = 0.2$, the two-stage design optimized for the prior with $\mu_1 = 0.1$ has similar average utility to the the nonoptimized design but average utility of the optimized one-stage design is a little lower; both one-stage and two-stage designs optimized for the prior with $\mu_1 = 0.2$ have lower average utility
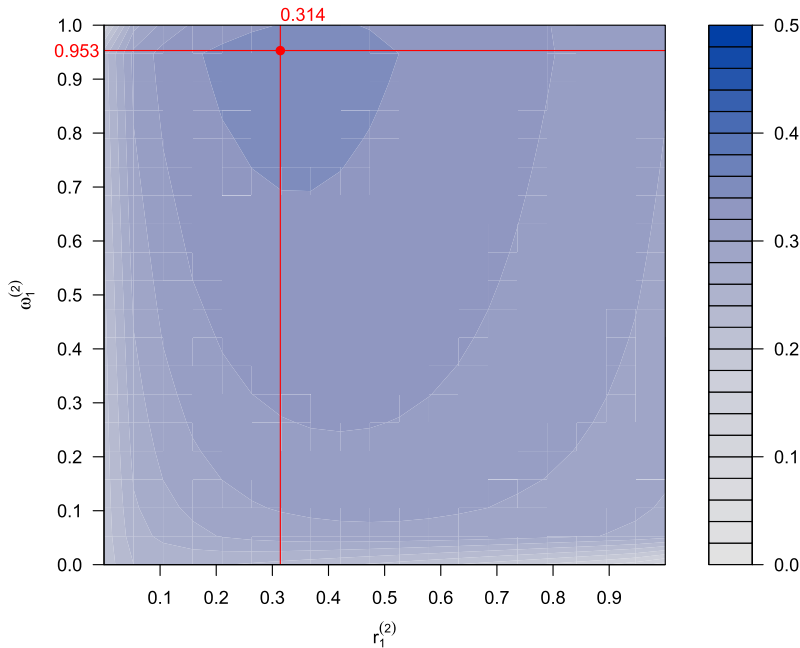
FIGURE 7 Interim optimization. The color indicates the expected utility given interim data for each combination of second-stage prevalence $r_1^{(2)}$ for subgroup 1 and testing weight $\omega_1^{(2)}$ given the interim data [Colour figure can be viewed at wileyonlinelibrary.com]

than the the nonoptimized design. These results are in line with previous studies[41,42] which showed adaptive enrichment designs provide the greatest advantage when a treatment effect is present in only one subgroup.

4 | WORKED EXAMPLE: IMPLEMENTING AN OPTIMIZED ADAPTIVE ENRICHMENT TRIAL

Suppose we wish to compare an experimental treatment to a control in a phase III clinical trial. We intend to use adaptive sample allocation as there is reason to believe the new treatment may only benefit a subgroup of patients. This trial will have a normally distributed endpoint with variance $\sigma^2 = 1$ and, using information from a pilot study with 40 subjects from each subgroup, we construct a prior distribution $\pi(\boldsymbol{\theta})$ for the treatment effects

$$\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \sim N\left(\begin{pmatrix} 0.1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0.1 & 0.05 \\ 0.05 & 0.1 \end{pmatrix} \right).$$

The total sample size for the trial is planned to be $n = 700$ subjects. The population prevalence of subgroup 1 is $\lambda = 0.3$ and a FWER $\alpha = 0.05$ is to be used for the study.

Under the above assumptions, the results in Figure 5 for $\psi = \sqrt{0.1} = 0.3162$ show the optimal first-stage parameters to be $s^{(1)} = 0.5$, $r_1^{(1)} = 0.4$ and $\omega_1^{(1)} = 0.4$. Thus, we recruit 350 patients in the first stage of the trial with 40% of these from subgroup 1.

Now suppose we observe interim estimates $\hat{\theta}_1^{(1)} = 0.442$ and $\hat{\theta}_2^{(1)} = 0.033$. These give Z-values $Z_1^{(1)} = 2.616$ and $Z_2^{(1)} = 0.238$ and the conditional error rates, as defined in Equations (5) and (6), are $A_1 = 0.6140$, $A_2 = 0.0184$, and $A_{12} = 0.3912$. At this point, we optimize the second-stage design parameters $r_1^{(2)}$ and $\omega_1^{(2)}$. Figure 7 plots the conditional expected utility as a function of $r_1^{(2)}$ and $\omega_1^{(2)}$ on a color-coded scale. The maximum conditional expected utility, obtained using the Hooke-Jeeves algorithm, is at $r_1^{(2)} = 0.314$ and $\omega_1^{(2)} = 0.953$. We therefore conduct the second stage of the trial using these parameter values.

Suppose, after recruiting the remaining subjects, the second-stage estimates are $\hat{\theta}_1^{(2)} = 0.272$ and $\hat{\theta}_2^{(2)} = -0.002$. The corresponding Z-values are $Z_1^{(2)} = 1.428$ and $Z_2^{(2)} = -0.015$, with P-values $P_1^{(2)} = .077$ and $P_2^{(2)} = .506$. Since $P_1^{(2)} < A_1$ and

$$P_1^{(2)} < .3728 = 0.953 \times 0.3912 = \omega_1^{(2)} \times A_{12},$$

we can globally reject H_{01}. However, since $P_2^{(2)} > A_2$ we cannot reject H_{02}.

5 | EXTENDING THE DESIGNS

The methods we have described can be extended to trial designs with more than two stages or more than two subgroups. Suppose K disjoint subgroups $S_1, \ldots, S_K$ are specified and we wish to test the null hypotheses H_{0k}: $\theta_k \leq 0$ against the alternatives H_{1k}: $\theta_k > 0$, where θ_k denotes the treatment effect in subgroup k. In a trial with J stages and a total sample size n, we recruit $s^{(j)}n$ patients in each stage, where $s^{(1)} + \cdots + s^{(J)} = 1$, and at stage j we recruit $r_k^{(j)}s^{(j)}n$ patients from subgroups $k = 1, \ldots, K$, where $r_1^{(j)} + \cdots + r_K^{(j)} = 1$. The data provide estimates $\hat{\theta}_1^{(j)}, \ldots, \hat{\theta}_K^{(j)}$, at each stage j, from which we obtain Z-values $Z_1^{(j)}, \ldots, Z_K^{(j)}$. In an enrichment design where control of the FWER is required, a suitable closed testing procedure is defined in terms of the $Z_k^{(j)}$. Then, H_{0k} is rejected globally at level α if all intersection hypotheses involving H_{0k} are rejected in local, level α tests.

An adaptive design can be created by repeated application of the conditional error approach. An initial reference design is stated and when adaptation occurs, the modified testing procedure is defined so as to preserve the conditional error rate of each individual and intersection hypothesis test under the updated design for the remainder of the trial. This updated design becomes the new reference design under which conditional error rates will be calculated at any subsequent adaptation point.

We can consider optimizing the choice of the design parameters $s^{(j)}$ and $r_k^{(j)}$ or weights in the tests of intersection hypotheses. The generalization of our earlier approach requires a prior distribution for the treatment effects $\theta = (\theta_1, \ldots, \theta_K)$ and a utility function whose expectation is to be maximised. If λ_k is the population prevalence of subgroup k, $k = 1, \ldots, K$, a natural extension of Equation (1) is

$$\mathcal{U}(\hat{\theta}) = \sum_{k=1}^{K} \lambda_k \, \mathbb{1}(\text{Reject } H_{0k}).$$

In Section 2.3.3 we applied backwards induction to find the optimal design for a trial with two subgroups and two stages. Since the dimension of the state space grows with the number of subgroups and stages, such a direct application of backwards induction may not be feasible more generally. Other methods of optimization can be employed to find efficient, if not globally optimal, designs. For example, in a multistage design one may construct the adaptation rule at each interim analysis assuming the trial will continue without any further adaptation. We note that the optimization process is liable to be computationally intensive and it is important to commit resources to assess trial designs in a timely manner.

6 | DISCUSSION

We have presented a Bayesian decision theoretic framework in which a clinical trial design can be optimized when two disjoint subgroups are under investigation. Our approach has both Bayesian and frequentist elements: the rules for hypothesis testing control the type I error rate and Bayesian decision tools are used to choose the design parameters within this scheme. This allows optimization of the sampling prevalence of each subgroup and weights in a weighted Bonferroni test of the intersection hypothesis, as well as optimal adaptation of these design parameters at the interim analysis. The optimal design maximizes the expected value of the specified utility function, averaged over the prior distribution assumed for the treatment effects in the two subgroups. After focusing on two-stage trials with two subgroups in Sections 2 and 4, we outlined how our optimization framework may be extended to allow more subgroups or stages in the trial in Section 5.

Our results provide insights into how the mean and variance of the prior distribution affects the optimal timing of the interim analysis and the trial prevalences for each subgroup of patients. In practice, it is advisable to consider the sensitivity of the design's efficiency to modeling assumptions in order to create a trial design with robust efficiency.

In contrast to adaptive enrichment designs where recruitment is either from the full patient population or restricted to a single subgroup, we propose sampling from each subgroup at a specific rate which may differ from its population prevalence. We acknowledge that achieving the optimized prevalences in a trial may be challenging: additional screening will be required and over-sampling a particular subgroup may delay a trial compared to an all-comers design.[43,44] If logistical considerations imply that each subgroup is either dropped or sampled according to its population prevalence, our framework can still be used to optimize the other design parameters.

In Section 3.2 we discussed designs with the option of early stopping for futility and how the utility function might be modified to facilitate optimizing such designs. A similar approach could be followed to relax the requirement of a fixed total sample size and allow re-assessment of future sample size at an interim analysis.

We have defined methods for normally distributed observations and a normal prior for treatment effects. While this has allowed us to demonstrate how to construct such designs, it is not a necessary restriction. With normally distributed responses, one could allow a separate response variance for each patient subgroup, placing prior distributions on these variances. In trials with other types of response distribution, including survival or categorical endpoints, standardized test statistics will still be approximately normally distributed if sample sizes are large enough, although nonnormal prior distributions may be appropriate.[45]

We assumed the null hypotheses of interest are that there is no treatment effect in each subgroup. Our decision theoretic framework can accommodate other formulations, such as testing for treatment effects in the full population and in one particular subgroup,[8,20,22-24,46-48] in which case the stage-wise test statistics for different subgroups are correlated. Care is required to ensure that enrichment designs control FWER when test statistics are correlated but this is not an issue in umbrella trials with separate level α tests for each null hypothesis.[31]

Although we have focused on hypothesis testing instead, estimating treatment effects after an adaptive trial is also important.[49] Simultaneous or marginal confidence regions for parameters, with or without multiplicity adjustment, can be constructed following a two-stage design.[50,51] Point estimates may be obtained by a weighted average of the treatment effects observed in the first and second stages[11,52] but, due to the sample size adaptations and subgroup selection these estimators may be biased with the bias depending on the specific adaptation rules and the true parameter values. A thorough investigation of estimation for adaptive enrichment designs will be a topic of future research.

Software in the form of an R package is available at https://github.com/nicoballarini/OptimalTrial.

AUTHOR CONTRIBUTIONS
Dr Ballarini and Dr Burnett are the co-primary authors and they contributed equally to this work.

DATA AVAILABILITY STATEMENT
Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

ORCID
Nicolás M. Ballarini https://orcid.org/0000-0002-3432-8931
Thomas Burnett https://orcid.org/0000-0001-8912-2554
Thomas Jaki https://orcid.org/0000-0002-1096-188X
Christoper Jennison https://orcid.org/0000-0002-9812-1104
Franz König https://orcid.org/0000-0002-6893-3304
Martin Posch https://orcid.org/0000-0001-8499-8573

REFERENCES
1. Dmitrienko A, Muysers C, Fritsch A, Lipkovich I. General guidance on exploratory and confirmatory subgroup analysis in late-stage clinical trials. *J Biopharm Stat*. 2016;26(1):71-98.
2. Alosh M, Huque MF, Bretz F, D'Agostino RB Sr. Tutorial on statistical considerations on subgroup analysis in confirmatory clinical trials. *Stat Med*. 2017;36(8):1334-1360.
3. Ondra T, Dmitrienko A, Friede T, et al. Methods for identification and confirmation of targeted subgroups in clinical trials: a systematic review. *J Biopharm Stat*. 2016;26(1):99-119.

4. Antoniou M, Jorgensen AL, Kolamunnage-Dona R. Biomarker-guided adaptive trial designs in phase II and phase III: a methodological review. *PLoS One*. 2016;11(2):e0149803.

5. Mandrekar SJ, Sargent DJ. Clinical trial designs for predictive biomarker validation: theoretical considerations and practical challenges. *J Clin Oncol*. 2009;27(24):4027.

6. Freidlin B, LM MS, Korn EL. Randomized clinical trials with biomarkers: design issues. *J Natl Cancer Inst*. 2010;102(3):152-160.

7. Simon N, Simon R. Adaptive enrichment designs for clinical trials. *Biostatistics*. 2013;14(4):613-625.

8. Brannath W, Zuber E, Branson M, et al. Confirmatory adaptive designs with Bayesian decision tools for a targeted therapy in oncology. *Stat Med*. 2009;28(10):1445-1463.

9. Friede T, Parsons N, Stallard N. A conditional error function approach for subgroup selection in adaptive clinical trials. *Stat Med*. 2012;31(30):4309-4320.

10. 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.

11. Chiu Y-D, Koenig F, Posch M, Jaki T. Design and estimation in clinical trials with subpopulation selection. *Stat Med*. 2018;37(29):4335-4352.

12. Food and Drug Administration Adaptive designs for clinical trials of drugs and biologics. guidance for industry; 2018.

13. Berry DA. The Brave new world of clinical cancer research: adaptive biomarker-driven trials integrating clinical practice with clinical research. *Mol Oncol*. 2015;9(5):951-959.

14. 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.

15. Food and Drug Administration . Master protocols: efficient clinical trial design strategies to expedite development of oncology drugs and biologics. guidance for industry; 2018;.

16. Renfro LA, Sargent DJ. Statistical controversies in clinical research: basket trials, umbrella trials, and other master protocols: a review and examples. *Ann Oncol*. 2016;28(1):34-43.

17. Woodcock J, LaVange LM. Master protocols to study multiple therapies, multiple diseases or both. *New Engl J Med*. 2017;377(1):62-70.

18. Govindan R, Mandrekar SJ, Gerber DE, et al. ALCHEMIST trials: a golden opportunity to transform outcomes in early-stage non-small cell lung cancer. *Clin Cancer Res*. 2015;21(24):5439-5444.

19. European Medicines Agency Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design; 2007.

20. Ondra T, Jobjörnsson S, Beckman RA, et al. Optimized adaptive enrichment designs. *Stat Methods Med Res*. 2019;28(7). https://doi.org/10.1177/0962280217747312.

21. Ondra T, Jobjörnsson S, Beckman RA, et al. Optimizing trial designs for targeted therapies. *PLoS One*. 2016;11(9):e0163726.

22. Burnett T. *Bayesian Decision Making in Adaptive Clinical Trials* [PhD thesis]. University of BathUK; 2017.

23. Burnett T, Jennison C. Adaptive enrichment trials: what are the benefits? *Stat Med*. 2021;40(3):690-711.

24. Graf AC, Posch M, Koenig F. Adaptive designs for subpopulation analysis optimizing utility functions. *Biom J*. 2015;57(1):76-89.

25. Beckman RA, Clark J, Chen C. Integrating predictive biomarkers and classifiers into oncology clinical development programmes. *Nat Rev Drug Discov*. 2011;10(10):735.

26. Rosenblum M, Fang X, Liu H. Optimal, two stage, adaptive enrichment designs for randomized trials using sparse linear programming. Department of Biostatistics Working Papers. Working Paper 273, Johns Hopkins University; 2017.

27. Krisam J, Kieser M. Optimal decision rules for biomarker-based subgroup selection for a targeted therapy in oncology. *Int J Mol Sci*. 2015;16(5):10354-10375.

28. 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.

29. Dmitrienko A, D'Agostino RB Sr, Huque MF. Key multiplicity issues in clinical drug development. *Stat Med*. 2013;32(7):1079-1111.

30. Spiegelhalter DJ, Abrams KR, Myles JP. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. Hoboken, NJ: John Wiley & Sons; 2004.

31. Stallard N, Posch M, Friede T, Koenig F, Brannath W. Optimal choice of the number of treatments to be included in a clinical trial. *Stat Med*. 2009;28(9):1321-1338.

32. Marcus R, Peritz E, Gabriel KR. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*. 1976;63(3):655-660.

33. 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.

34. Bauer P, Kieser M. Combining different phases in the development of medical treatments within a single trial. *Stat Med*. 1999;18(14):1833-1848.

35. Bretz F, Koenig F, Brannath W, Glimm E, Posch M. Adaptive designs for confirmatory clinical trials. *Stat Med*. 2009;28(8):1181-1217.

36. 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.

37. Müller H-H, Schäfer H. Adaptive group sequential designs for clinical trials: combining the advantages of adaptive and of classical group sequential approaches. *Biometrics*. 2001;57(3):886-891.

38. Müller H-H, Schäfer H. A general statistical principle for changing a design any time during the course of a trial. *Stat Med*. 2004;23(16):2497-2508.

39. Varadhan R, Borchers HW. dfoptim: derivative-free optimization. R package version 2018.2-1; 2018.

40. R Core Team R: a language and environment for statistical computing; 2018.

41. Simon R, Maitournam A. Evaluating the efficiency of targeted designs for randomized clinical trials. *Clin Cancer Res*. 2004;10(20):6759-6763.

42. Hoering A, LeBlanc M, Crowley JJ. Randomized phase III clinical trial designs for targeted agents. *Clin Cancer Res*. 2008;14(14):4358-4367.

43. Klauschen F, Andreeff M, Keilholz U, Dietel M, Stenzinger A. The combinatorial complexity of cancer precision medicine. *Oncoscience*. 2014;1(7):504.

44. Eichler H-G, Bloechl-Daum B, Bauer P, et al. "Threshold-crossing": a useful way to establish the counterfactual in clinical trials? *Clin Pharmacol Ther*. 2016;100(6):699-712.

45. Brückner M, Burger HU, Brannath W. Nonparametric adaptive enrichment designs using categorical surrogate data. *Stat Med*. 2018;37(29):4507-4524.

46. Wang SJ, O'Neill RT, Hung HJ. Approaches to evaluation of treatment effect in randomized clinical trials with genomic subset. *Pharm Stat J Appl Stat Pharm Ind*. 2007;6(3):227-244.

47. Alosh M, Huque MF. A flexible strategy for testing subgroups and overall population. *Stat Med*. 2009;28(1):3-23.

48. Spiessens B, Debois M. Adjusted significance levels for subgroup analyses in clinical trials. *Contemp Clin Trials*. 2010;31(6):647-656.

49. Stallard N, Todd S, Whitehead J. Estimation following selection of the largest of two normal means. *J Stat Plann Infer*. 2008;138(6):1629-1638.

50. Mehta CR, Bauer P, Posch M, Brannath W. Repeated confidence intervals for adaptive group sequential trials. *Stat Med*. 2007;26(30):5422-5433.

51. Magirr D, Jaki T, Posch M, Klinglmueller F. Simultaneous confidence intervals that are compatible with closed testing in adaptive designs. *Biometrika*. 2013;100(4):985-996.

52. Kimani PK, Todd S, Stallard N. Estimation after subpopulation selection in adaptive seamless trials. *Stat Med*. 2015;34(18):2581-2601.

SUPPORTING INFORMATION

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