Received: 5 March 2020 | Revised: 13 October 2020 | Accepted: 14 October 2020

DOI: 10.1002/sim.8797

RESEARCH ARTICLE

Statistics in Medicine WILEY

Adaptive enrichment trials: What are the benefits?

Thomas Burnett[1] | **Christopher Jennison**[2]

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

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

Correspondence
Thomas Burnett, Department of Mathematics and Statistics, Fylde College, Lancaster University, Bailrigg, Lancaster LA1 4YF, UK.
Email: t.burnett1@lancaster.ac.uk

Abstract

When planning a Phase III clinical trial, suppose a certain subset of patients is expected to respond particularly well to the new treatment. Adaptive enrichment designs make use of interim data in selecting the target population for the remainder of the trial, either continuing with the full population or restricting recruitment to the subset of patients. We define a multiple testing procedure that maintains strong control of the familywise error rate, while allowing for the adaptive sampling procedure. We derive the Bayes optimal rule for deciding whether or not to restrict recruitment to the subset after the interim analysis and present an efficient algorithm to facilitate simulation-based optimisation, enabling the construction of Bayes optimal rules in a wide variety of problem formulations. We compare adaptive enrichment designs with traditional non-adaptive designs in a broad range of examples and draw clear conclusions about the potential benefits of adaptive enrichment.

KEYWORDS

adaptive designs, adaptive enrichment, Bayesian optimization, phase III clinical trial, population enrichment

1 | INTRODUCTION

Consider a Phase III trial in which it is believed a certain subset of patients will respond particularly well to the new treatment. We wish to test for a treatment effect in both the pre-identified subpopulation and the full population. Such multiple testing can be conducted using a closed testing procedure to control the familywise error rate (FWER).[1] In an adaptive enrichment design, if interim data suggest it is only the subpopulation that benefits from the new treatment, recruitment in the second half of the trial is restricted to the subpopulation. This increase in recruitment from the subpopulation is referred to as "enrichment" of the sampling rule.

We develop and assess designs which use a closed testing procedure with Simes' method[2] to test the intersection hypothesis and a weighted inverse normal combination test[3-5] to combine data from the two stages of the trial. We show that the resulting testing procedure controls the FWER, whatever rule is used to decide when enrichment should occur. This allows us to seek the enrichment rule which is optimal for a specified criterion. We shall follow the approach presented by Burnett,[6] defining a gain function that reflects the value of the outcome of the trial and a prior distribution for the treatment effects in the subpopulation and full population. The optimal decision at the interim analysis is that which maximises the expected gain with respect to the posterior distribution of the treatment effects, given current data. Since we use simulation in constructing the Bayes optimal decision rule for an adaptive design, our approach has the potential to be computationally expensive. We present an efficient algorithm for deriving this decision rule that significantly reduces the calculation required: using our methods, designs can be derived and tested in a matter of minutes on a laptop or PC.

In previous work on adaptive enrichment designs, Brannath et al[7] followed a Bayesian approach, assuming an uninformative prior for treatment effects. They determined the enrichment decision by comparing the posterior predictive probabilities of rejecting each hypothesis at the end of the trial with certain user-defined thresholds. Götte et al[8] considered families of enrichment rules defined in terms of linear combinations of the two treatment effect estimates or the conditional power to reject each hypothesis. They defined the "correct decision" at the interim analysis for given true values of the treatment effects and searched within their families of enrichment rules to maximise a weighted combination of the probabilities of a correct decision. Uozomi and Hamada[9] defined enrichment rules in terms of thresholds for the treatment effect estimates or predictive power for the two hypothesis tests and set these thresholds to optimize a utility function under specific values for the true treatment effects. Our methods are set in a more complete Bayesian decision theoretic framework. The gain function is chosen to summarize the benefits of the final decisions, reflecting the size of population in which the new treatment is proven to be effective and the magnitude of the treatment effect in this population. The decision whether or not to enrich at the interim analysis is informed by both the posterior distribution of treatment effects and the interim estimates or p-values that will form part of the final hypothesis tests.

Ondra et al[10] developed Bayes optimal methods in a class of adaptive enrichment designs where FWER is controlled by a Bonferroni adjustment, assuming a 4-point discrete prior distribution for the two treatment effects. These simplifications allow the optimal enrichment decision rule to be found by maximising an integral, which is computed numerically. The application of Simes tests in our methods reduces conservatism in the testing procedure and the continuous prior distributions are better able to capture investigators' prior beliefs. Although our form of problem requires the use of simulation to find an optimal design, this approach has the advantage of extending very easily to other forms of gain function and multiple testing methods.

Through studying optimal designs, we are able to assess the potential benefits of adaptive enrichment. We have studied a variety of scenarios, drawing comparisons in each case with two nonadaptive designs: sampling the full population throughout the whole study or focusing on the subpopulation at the outset and only recruiting subpopulation patients. We see there are plausible prior distributions for which the adaptive enrichment design is superior to both forms of nonadaptive design. Furthermore, we recognize that investigators may be reluctant to restrict recruitment to the subpopulation from the outset and observe that in situations where this would have been the optimal policy, adaptive enrichment can give substantially higher expected gain than the nonadaptive, full population design.

Our studies also shed light on the underlying reasons for the effectiveness of adaptive designs. The good performance of adaptive designs in the special case of one-point prior distributions shows efficiency gains can follow from adapting to interim data and the likelihood of eventual rejection of each null hypothesis. With proper prior distributions, one might expect increased knowledge about the true treatment effects at the interim analysis to give adaptive designs a further advantage. However, we find such benefits to be modest: when the prior variance is high, considerable uncertainty about the true treatment effects remains; when the prior variance is low, information about the treatment effects at the interim analysis comes primarily from the prior, not the interim data.

The paper is structured as follows. We formulate the problem in Section 2 and we present methods for controlling FWER and combining data across stages in Section 3. We describe methods for optimising an adaptive design in Section 4, describe two forms of nonadaptive design in Section 5 and present examples in Section 6. We conclude with discussion of the results obtained in our examples.

2 | PROBLEM FORMULATION

2.1 | Patient responses

Consider a Phase III trial comparing a new therapy, Treatment A, with a control, Treatment B. Suppose a biomarker-defined subpopulation is identified before the trial commences and it is thought that biomarker positive patients will respond particularly well to the new treatment. We call the subpopulation of biomarker positive patients $\mathcal{S}_1$ and the complement of this $\mathcal{S}_2$.

We suppose responses are normally distributed with a common variance σ^2 but note that, by large sample theory, distributions of treatment estimates will have the same form for a wide variety of response types. Let μ_{A1} and μ_{B1} be the expected responses for patients in $\mathcal{S}_1$ on Treatments A and B, respectively. Similarly, let μ_{A2} and μ_{B2} be the expected responses on Treatments A and B for patients in $\mathcal{S}_2$. Letting X_{ij} denote the response of the ith patient in subpopulation $\mathcal{S}_j$ on Treatment A and Y_{ij} the response of the ith patient in $\mathcal{S}_j$ on Treatment B, we have

$$X_{ij} \sim N(\mu_{Aj}, \sigma^2), \quad i = 1, 2, \dots , \quad j = 1, 2,$$

and

$$Y_{ij} \sim N(\mu_{Bj}, \sigma^2), \quad i = 1, 2, \dots , \quad j = 1, 2.$$

The treatment effects in subpopulations $\mathcal{S}_1$ and $\mathcal{S}_2$ are $\theta_1 = \mu_{A1} - \mu_{B1}$ and $\theta_2 = \mu_{A2} - \mu_{B2}$, respectively.

Suppose $\mathcal{S}_1$ represents a fraction λ of the full population. Then, the overall treatment effect in the full population is $\theta_3 = \lambda\theta_1 + (1 - \lambda)\theta_2$. We shall write $\boldsymbol{\theta} = (\theta_1, \theta_2)$, noting that $\boldsymbol{\theta}$ determines the value of θ_3. We assume the investigators are interested in testing $H_{01}\colon \theta_1 \leq 0$ vs $\theta_1 > 0$ and $H_{03}\colon \theta_3 \leq 0$ vs $\theta_3 > 0$. The hypothesis $H_{02}\colon \theta_2 \leq 0$, is not to be tested (although one might require some evidence of a positive treatment effect in S_2 to support approval of the new treatment for the full population when H_{03} is rejected). However, the approach we describe can also be applied when enrichment in either S_1 or S_2 is possible, or when there are more than two subpopulations; the key requirement is that the subpopulations and enrichment options are predefined.

2.2 | Adaptive enrichment trial designs

If the new therapy is beneficial to all patients, we would hope to reject the null hypothesis H_{03} and establish that there is an effect in the full patient population. However, if the benefit is restricted to patients in $\mathcal{S}_1$, it would be advantageous to focus on this subpopulation and increase the probability of rejecting H_{01}. Adaptive enrichment designs aim to balance these two objectives by using interim data to decide whether or not to restrict enrolment in the remainder of the study to $\mathcal{S}_1$ and test only H_{01}.

We consider trial designs with a single interim analysis that takes place after a fraction τ of the planned sample size has been recruited and responses from these patients have been observed. Initially, patients are recruited from the full population. If, at the interim analysis, results on the new therapy are promising in both $\mathcal{S}_1$ and $\mathcal{S}_2$, recruitment continues across the full population. If, however, the new therapy only appears to benefit patients in $\mathcal{S}_1$, the remainder of the sample size is devoted to $\mathcal{S}_1$. Our objective is to optimize the rule for choosing between these two options in an adaptive enrichment design.

Let n be the total number of patients to be recruited. Assuming recruitment from $\mathcal{S}_1$ and $\mathcal{S}_2$ is in proportion to the size of these subpopulations, sample sizes at the interim analysis are $\lambda\tau n$ in $\mathcal{S}_1$ and $(1 - \lambda)\tau n$ in $\mathcal{S}_2$. When recruitment continues from the full population, an additional $\lambda(1 - \tau)n$ patients are sampled from $\mathcal{S}_1$ and $(1 - \lambda)(1 - \tau)n$ from $\mathcal{S}_2$. If "enrichment" occurs and only patients from $\mathcal{S}_1$ are recruited after the interim analysis, there will be a further $(1 - \tau)n$ patients from $\mathcal{S}_1$. We assume that, within each stage of the trial, patients in each subpopulation are randomized equally between Treatments A and B.

In describing the distributions of parameter estimates, it is helpful to define

$$\tilde{\mathcal{I}} = \frac{n}{4\sigma^2}. \tag{1}$$

Note that a fixed sample size trial with n patients divided equally between Treatments A and B would produce an estimate $\hat{\theta}_3$ with $Var\ (\hat{\theta}_3) = 4\sigma^2/n$, so $\tilde{\mathcal{I}} = \{Var\ (\hat{\theta}_3)\}^{-1}$ represents the Fisher information for θ_3 in this case.

Let $m_{11} = \lambda\tau n/2$ and $m_{21} = (1 - \lambda)\tau n/2$. Then, in the form of adaptive enrichment design we have described, the first stage yields treatment effect estimates

$$\hat{\theta}_1^{(1)} = \hat{\mu}_{A1}^{(1)} - \hat{\mu}_{B1}^{(1)} = \frac{1}{m_{11}}\sum_{i=1}^{m_{11}} X_{i1} - \frac{1}{m_{11}}\sum_{i=1}^{m_{11}} Y_{i1} \sim N(\theta_1, \{\lambda\tau\tilde{\mathcal{I}}\}^{-1}),$$

$$\hat{\theta}_2^{(1)} = \hat{\mu}_{A2}^{(1)} - \hat{\mu}_{B2}^{(1)} = \frac{1}{m_{21}}\sum_{i=1}^{m_{21}} X_{i2} - \frac{1}{m_{21}}\sum_{i=1}^{m_{21}} Y_{i2} \sim N(\theta_2, \{(1-\lambda)\tau\tilde{\mathcal{I}}\}^{-1})$$

and

$$\hat{\theta}_3^{(1)} = \lambda\hat{\theta}_1^{(1)} + (1-\lambda)\hat{\theta}_2^{(1)} \sim N(\theta_3, \{\tau\tilde{\mathcal{I}}\}^{-1}).$$

The joint distribution of $(\hat{\theta}_1^{(1)}, \hat{\theta}_3^{(1)})$ is bivariate normal with correlation $\sqrt{\lambda}$.

Suppose that after the initial analysis the trial continues in the full population. Then, setting $m_{12} = \lambda(1-\tau)n/2$ and $m_{22} = (1-\lambda)(1-\tau)n/2$, the second stage data alone yield treatment effect estimates

$$\hat{\theta}_1^{(2)} = \hat{\mu}_{A1}^{(2)} - \hat{\mu}_{B1}^{(2)} = \frac{1}{m_{12}}\sum_{i=m_{11}+1}^{m_{11}+m_{12}} X_{i1} - \frac{1}{m_{12}}\sum_{i=m_{11}+1}^{m_{11}+m_{12}} Y_{i1} \sim N(\theta_1, \{\lambda(1-\tau)\tilde{\mathcal{I}}\}^{-1}),$$
$$\hat{\theta}_2^{(2)} = \hat{\mu}_{A2}^{(2)} - \hat{\mu}_{B2}^{(2)} = \frac{1}{m_{22}}\sum_{i=m_{21}+1}^{m_{21}+m_{22}} X_{i2} - \frac{1}{m_{22}}\sum_{i=m_{21}+1}^{m_{21}+m_{22}} Y_{i2} \sim N(\theta_2, \{(1-\lambda)(1-\tau)\tilde{\mathcal{I}}\}^{-1}),$$

and

$$\hat{\theta}_3^{(2)} = \lambda\hat{\theta}_1^{(2)} + (1-\lambda)\hat{\theta}_2^{(2)} \sim N(\theta_3, \{(1-\tau)\tilde{\mathcal{I}}\}^{-1}).$$

Again, the pair of estimates $(\hat{\theta}_1^{(2)}, \hat{\theta}_3^{(2)})$ is bivariate normal with correlation $\sqrt{\lambda}$.

Alternatively, suppose the trial is enriched and only subpopulations $\mathcal{S}_1$ is sampled in the second stage. Then, setting $\tilde{m}_{12} = (1-\tau)n/2$, the new data yield the estimate

$$\hat{\theta}_1^{(2)} = \frac{1}{\tilde{m}_{12}}\sum_{i=m_{11}+1}^{m_{11}+\tilde{m}_{12}} X_{i1} - \frac{1}{\tilde{m}_{12}}\sum_{i=m_{11}+1}^{m_{11}+\tilde{m}_{12}} Y_{i1} \sim N(\theta_1, \{(1-\tau)\tilde{\mathcal{I}}\}^{-1})),$$

and no estimate of θ_3 is available.

3 | ACHIEVING STRONG CONTROL OF THE FAMILY-WISE ERROR RATE

3.1 | Closed testing procedures

Control of the type I error rate in a confirmatory clinical trial is paramount[11] and, with two null hypotheses under consideration, the testing procedure should provide strong control of the FWER at the prespecified level α.[1] Thus, we require

$$P_{\boldsymbol{\theta}}(\text{Reject at least one true null hypothesis}) \leq \alpha \quad \text{for all } \boldsymbol{\theta}.$$

We shall follow the general approach presented by Bretz et al,[12] Schmidli et al[13] and Jennison and Turnbull[14] who ensure strong control of the FWER by constructing a closed testing procedure[15] in which combination tests are carried out on the individual hypotheses. In addition to the null hypotheses H_{01}: $\theta_1 \leq 0$ and H_{03}: $\theta_3 \leq 0$, the closed testing procedure also considers the intersection hypothesis $H_{0,13} = H_{01} \cap H_{03}$ which states that $\theta_1 \leq 0$ and $\theta_3 \leq 0$. We specify level α tests of H_{01}, H_{03}, and $H_{0,13}$. Then, H_{01} is rejected in the overall procedure if the individual level α tests reject H_{01} and $H_{0,13}$. Similarly, H_{03} is rejected overall if the individual level α tests reject H_{03} and $H_{0,13}$. For an explanation of why such a procedure protects the FWER and why all procedures that provide strong control of FWER can be interpreted as closed testing procedures, see Appendix A.

We refer to the periods of an adaptive enrichment design before and after the interim analysis as stages 1 and 2. In our closed testing procedure, we need a method for combining test statistics for hypotheses H_{01} and H_{03} to test the intersection hypothesis $H_{0,13}$ and a method to combine data across stages, bearing in mind that the decision about which subpopulations to recruit from in stage 2 depends on the stage 1 data. We describe these methods in the following sections.

3.2 | Simes' test for the intersection hypothesis

Let $P_1^{(1)}$ and $P_3^{(1)}$ be P-values for testing H_{01} and H_{03} based on stage 1 data. Then $P_1^{(1)} \sim \text{Unif}(0,1)$ if $\theta_1 = 0$ and $P_1^{(1)}$ is stochastically larger than a Unif(0, 1) random variable if $\theta_1 < 0$; similarly, $P_3^{(1)} \sim \text{Unif}(0,1)$ if $\theta_3 = 0$ and $P_3^{(1)}$ is stochastically larger than this if $\theta_3 < 0$. We can use Simes' method[2] to create a P-value for the intersection hypothesis $H_{0,13}$,

$$P_{13}^{(1)} = \min\{2\min(P_1^{(1)}, P_3^{(1)}),\ \max(P_1^{(1)}, P_3^{(1)})\}. \tag{2}$$

Since $P_1^{(1)}$ and $P_3^{(1)}$ are based on nested groups of patients, these p-values are positively associated and the results of Sarkar and Chang[16] imply that Simes' test gives a valid (but conservative) P-value for testing $H_{0,13}$.

If enrichment does not take place and stage 2 continues with recruitment from the full population, we define $P_1^{(2)}$ and $P_3^{(2)}$ to be p-values for testing H_{01} and H_{03} based on data from stage 2 patients alone. Then, just as for stage 1 data, we construct the Simes p-value

$$P_{13}^{(2)} = \min\{2\min(P_1^{(2)}, P_3^{(2)}),\ \max(P_1^{(2)}, P_3^{(2)})\}, \tag{3}$$

for testing the intersection hypothesis $H_{0,13}$.

If enrichment does take place, only patients from $\mathcal{S}_1$ are observed in stage 2 and we define the P-value $P_1^{(2)}$ for H_{01} based on these observations. We cannot define a P-value $P_3^{(2)}$ but this is not a problem as we no longer plan to test H_{03}. In this case we set

$$P_{13}^{(2)} = P_1^{(2)}, \tag{4}$$

noting that $H_{0,13}$ implies $\theta_1 \leq 0$ and hence $P_{13}^{(2)} = P_1^{(2)}$ is Unif(0, 1), or stochastically larger than this, under $H_{0,13}$.

3.3 | The weighted inverse normal combination test

In constructing level α tests of H_{01}, H_{03}, and $H_{0,13}$, we need to combine P-values from the two stages. In each case, we do this using a weighted inverse normal combination test.[3-5]

Consider first the level α test of H_{01}. The stage 1 data give

$$Z_1^{(1)} = \hat{\theta}_1^{(1)} \sqrt{\{\lambda\tau\tilde{\mathcal{I}}\}} \sim N(\theta_1\sqrt{\{\lambda\tau\tilde{\mathcal{I}}\}}, 1),$$

and the associated P-value is $P_1^{(1)} = 1 - \Phi(Z_1^{(1)})$ where Φ denotes the cumulative distribution function of a standard normal random variable. If the trial recruits from the full population in stage 2, we have

$$Z_1^{(2)} = \hat{\theta}_1^{(2)} \sqrt{\{\lambda(1-\tau)\tilde{\mathcal{I}}\}} \sim N(\theta_1\sqrt{\{\lambda(1-\tau)\tilde{\mathcal{I}}\}}, 1),$$

while, if enrichment occurs, we have

$$Z_1^{(2)} = \hat{\theta}_1^{(2)} \sqrt{\{(1-\tau)\tilde{\mathcal{I}}\}} \sim N(\theta_1\sqrt{\{(1-\tau)\tilde{\mathcal{I}}\}}, 1),$$

and in either case the associated P-value is $P_1^{(2)} = 1 - \Phi(Z_1^{(2)})$.

Suppose $\theta_1 = 0$. Then, $Z_1^{(1)} \sim N(0,1)$ and $P_1^{(1)} \sim \text{Unif}(0,1)$. Conditional on the first stage data, $Z_1^{(2)} \sim N(0,1)$ and $P_1^{(2)} \sim \text{Unif}(0,1)$. Since the conditional distribution of $Z_1^{(2)}$ does not depend on the stage 1 data, we conclude that $Z_1^{(1)}$ and $Z_1^{(2)}$

are independent $N(0, 1)$ random variables. Using pre-specified weights w_1 and w_2 for which $w_1^2 + w_2^2 = 1$, we define the combination test statistic

$$Z_1^{(c)} = w_1 Z_1^{(1)} + w_2 Z_1^{(2)},$$

and note that $Z_1^{(c)} \sim N(0, 1)$ when $\theta_1 = 0$.

Suppose now that $\theta_1 < 0$. We can write

$$Z_1^{(1)} = \theta_1 \sqrt{\{\lambda \tau \tilde{\mathcal{I}}\}} + \varepsilon_1^{(1)},$$

where $\varepsilon_1^{(1)} \sim N(0, 1)$ and

$$Z_1^{(2)} = \theta_1 c_1 + \varepsilon_1^{(2)},$$

where $\varepsilon_1^{(2)} \sim N(0, 1)$, $\varepsilon_1^{(2)}$ is independent of $\varepsilon_1^{(1)}$, $c_1 = \sqrt{\{\lambda(1 - \tau)\tilde{\mathcal{I}}\}}$ if enrichment does not occur in stage 2 and $c_1 = \sqrt{\{(1 - \tau)\tilde{\mathcal{I}}\}}$ if enrichment does occur. Since

$$w_1 \varepsilon_1^{(1)} + w_2 \varepsilon_1^{(2)} \sim N(0, 1),$$

$Z_1^{(1)} < \varepsilon_1^{(1)}$ and $Z_1^{(2)} < \varepsilon_1^{(2)}$, it follows that $Z_1^{(c)} = w_1 Z_1^{(1)} + w_2 Z_1^{(2)}$ is stochastically smaller than a $N(0, 1)$ random variable. Hence the test that rejects H_{01} if $Z_1^{(c)} > \Phi^{-1}(1 - \alpha)$ has type I error rate less than or equal to α whenever $\theta_1 \leq 0$, as required.

We construct a level α test of H_{03} in a similar way to that of H_{01}. We have

$$Z_3^{(1)} = \hat{\theta}_3^{(1)} \sqrt{\{\tau \tilde{\mathcal{I}}\}} \sim N(\theta_3 \sqrt{\{\tau \tilde{\mathcal{I}}\}}, 1),$$

from stage 1 data and, if enrichment does not occur, we have

$$Z_3^{(2)} = \hat{\theta}_3^{(2)} \sqrt{\{(1 - \tau)\tilde{\mathcal{I}}\}} \sim N(\theta_3 \sqrt{\{(1 - \tau)\tilde{\mathcal{I}}\}}, 1),$$

from stage 2 data. In the case of no enrichment, we create the combination test statistic

$$Z_3^{(c)} = w_1 Z_3^{(1)} + w_2 Z_3^{(2)},$$

and we reject H_{03} if $Z_3^{(c)} > \Phi^{-1}(1 - \alpha)$. The proof that this test controls the type I error rate follows the same lines as that for the test of H_{01} but, since we do not test H_{03} at all when enrichment occurs, this test is conservative even if $\theta_3 = 0$.

The level α test of the intersection hypothesis $H_{0,13}$ is constructed from the P-values $P_{13}^{(1)}$ and $P_{13}^{(2)}$ as defined in Equations (2), (3) and (4). Under $H_{0,13}$, the positive correlation between $\hat{\theta}_1^{(1)}$ and $\hat{\theta}_3^{(1)}$ implies that $P_{13}^{(1)}$ is stochastically larger than a Unif(0, 1) random variable, even when $\theta_1 = \theta_3 = 0$. Thus, $Z_{13}^{(1)} = \Phi^{-1}(1 - P_{13}^{(1)})$ is stochastically smaller than a $N(0, 1)$ random variable and we can write

$$Z_{13}^{(1)} = \varepsilon_{13}^{(1)} - \delta_1, \tag{5}$$

where $\varepsilon_{13}^{(1)} \sim N(0, 1)$ and δ_1 is a positive random variable, not necessarily independent of $\varepsilon_{13}^{(1)}$. If no enrichment occurs, by similar reasoning, the conditional distribution under $H_{0,13}$ of $Z_{13}^{(2)} = \Phi^{-1}(1 - P_{13}^{(2)})$, given stage 1 data, is stochastically smaller than a $N(0, 1)$ random variable. If enrichment does occur, $Z_{13}^{(2)} = Z_1^{(2)}$ and has conditional distribution $N(\theta_1 \sqrt{\{(1 - \tau)\tilde{\mathcal{I}}\}}, 1)$ given stage 1 data. It follows that, under $H_{0,13}$, we can write

$$Z_{13}^{(2)} = \varepsilon_{13}^{(2)} - \delta_2, \tag{6}$$

TABLE 1 Formulae for P-values used to create level α tests of H_{01}, H_{03}, and $H_{0,13}$

	With no enrichment		
	$\boldsymbol{H_{01}}$	$\boldsymbol{H_{03}}$	$\boldsymbol{H_{0,13}}$
Stage 1	$P_1^{(1)} = 1 - \Phi(Z_1^{(1)})$	$P_3^{(1)} = 1 - \Phi(Z_3^{(1)})$	$P_{13}^{(1)} = S(P_1^{(1)}, P_3^{(1)})$
Stage 2	$P_1^{(2)} = 1 - \Phi(Z_1^{(2)})$	$P_3^{(2)} = 1 - \Phi(Z_3^{(2)})$	$P_{13}^{(2)} = S(P_1^{(2)}, P_3^{(2)})$
Combined	$P_1^{(c)} = W(P_1^{(1)}, P_1^{(2)})$	$P_3^{(c)} = W(P_3^{(1)}, P_3^{(2)})$	$P_{13}^{(c)} = W(P_{13}^{(1)}, P_{13}^{(2)})$
	With enrichment		
	$\boldsymbol{H_{01}}$	$\boldsymbol{H_{03}}$	$\boldsymbol{H_{0,13}}$
Stage 1	$P_1^{(1)} = 1 - \Phi(Z_1^{(1)})$	$P_3^{(1)} = 1 - \Phi(Z_3^{(1)})$	$P_{13}^{(1)} = S(P_1^{(1)}, P_3^{(1)})$
Stage 2	$P_1^{(2)} = 1 - \Phi(Z_1^{(2)})$	—	$P_{13}^{(2)} = P_1^{(2)}$
Combined	$P_1^{(c)} = W(P_1^{(1)}, P_1^{(2)})$	—	$P_{13}^{(c)} = W(P_{13}^{(1)}, P_{13}^{(2)})$

where $\epsilon_{13}^{(2)} \sim N(0, 1)$ is independent of $\epsilon_{13}^{(1)}$ and δ_2 is a positive random variable that may depend on $\epsilon_{13}^{(1)}$ and $\epsilon_{13}^{(2)}$. It follows from Equations (5) and (6) that, under $H_{0,13}$,

$$Z_{13}^{(c)} = w_1 Z_{13}^{(1)} + w_2 Z_{13}^{(2)}$$

is stochastically smaller than a $N(0, 1)$ variable. Hence, the test that rejects $H_{0,13}$ if $Z_{13}^{(c)} > \Phi^{-1}(1 - \alpha)$ has type I error rate less than or equal to α whenever $\theta_1 \leq 0$ and $\theta_3 \leq 0$.

3.4 | Summary of the overall testing procedure

Let

$$S(P_1, P_2) = \min\{2\min(P_1, P_2),\ \max(P_1, P_2)\},$$

be the function that converts P_1 and P_2 into a Simes P-value and and define

$$W(P^{(1)}, P^{(2)}) = 1 - \Phi\{w_1\Phi^{-1}(1 - P^{(1)}) + w_2\Phi^{-1}(1 - P^{(2)})\}, \tag{7}$$

the function that gives the P-value when a weighted inverse normal combination test with weights w_1 and w_2 is applied to stage 1 and 2 P-values $P^{(1)}$ and $P^{(2)}$. With this notation, Table 1 presents a summary of the closed testing procedure described above.

In a trial where enrichment does not occur and patients are recruited from the full population in stage 2, we reject H_{01} overall if $P_1^{(c)} \leq \alpha$ and $P_{13}^{(c)} \leq \alpha$, and we reject H_{03} overall if $P_3^{(c)} \leq \alpha$ and $P_{13}^{(c)} \leq \alpha$. If enrichment occurs, H_{01} is rejected overall if $P_1^{(c)} \leq \alpha$ and $P_{13}^{(c)} \leq \alpha$ but it is not possible to test H_{03} as there is no $P_3^{(2)}$ to use in the combination test of H_{03}; this is in keeping with the decision to enrich which implies it is no longer desired to test H_{03}.

4 | OPTIMIZING AN ADAPTIVE ENRICHMENT DESIGN

4.1 | Bayesian decision framework

An enrichment design, as described in Section 2.2, that applies the closed testing procedure presented in Section 3 will protect the FWER regardless of the decision rule that determines when to enrich in stage 2. This gives us the opportunity to apply Bayesian decision theory[17] to optimize the enrichment decision rule for our chosen criterion. This decision theoretic approach requires the specification of a prior distribution for $\boldsymbol{\theta}$ and a gain, or utility, function that assigns a value to the final outcome of the study.

The decision rule. We denote the sufficient statistic for $\boldsymbol{\theta} = (\theta_1, \theta_2)$ based on stage 1 data by $X_1 = (\hat{\theta}_1^{(1)}, \hat{\theta}_2^{(1)})$. Note that (θ_1, θ_2) determines (θ_1, θ_3) and vice versa, so X_1 is also the sufficient statistic for (θ_1, θ_3). We shall consider decision rules that are functions of X_1. The decision under rule d is specified through the function $d(X_1)$ taking values in $\{1, 2\}$, with

$$\begin{aligned} d(X_1) &= 1 \Rightarrow \text{Enrich in stage 2,} \\ d(X_1) &= 2 \Rightarrow \text{Do not enrich in stage 2.} \end{aligned}$$

The form of the sufficient statistic X_2 for $\boldsymbol{\theta}$ based on stage 2 data depends on which decision is taken. If $d(X_1) = 1$, enrichment occurs and $X_2 = \hat{\theta}_1^{(2)}$, while if $d(X_1) = 2$ enrichment does not occur and $X_2 = (\hat{\theta}_1^{(2)}, \hat{\theta}_2^{(2)})$. In either case we write $X = (X_1, d(X_1), X_2)$ to summarize the full set of data at the end of the study and the decision taken at the interim analysis.

The prior distribution for $\boldsymbol{\theta}$. We assume a continuous prior distribution for $\boldsymbol{\theta} = (\theta_1, \theta_2)$ is specified and we denote the probability density function of the prior distribution by $\pi(\boldsymbol{\theta})$.

The gain function. The gain function $G(\boldsymbol{\theta}, X)$ denotes the value assigned to the outcome of the study when $\boldsymbol{\theta}$ is the parameter vector and we observe $X = (X_1, d(X_1), X_2)$. Note that we can deduce from X which of the hypotheses H_{01} and H_{03} are rejected in the final analysis.

Let $\mathcal{R}_1$ be the indicator variable of the event that H_{01} is rejected but H_{03} is not rejected, and let $\mathcal{R}_3$ be the indicator variable of the event that H_{03} is rejected. Both $\mathcal{R}_1$ and $\mathcal{R}_3$ are functions of X. In this paper we shall consider the gain function

$$G(\boldsymbol{\theta}, X) = \lambda \theta_1 \mathcal{R}_1 + \theta_3 \mathcal{R}_3. \tag{8}$$

Here, the gain is deemed to be proportional to the size of the population for which a treatment effect is found and also to the average treatment effect for patients in that population.

Other forms of gain function are possible: the key feature is that they are constructed based on the possible outcomes of the trial. A general form of gain function should capture the importance of each of these possible outcomes, for example, if we define $\gamma_1(\boldsymbol{\theta}, X)$ to represent the benefit of rejecting H_{01} and $\gamma_3(\boldsymbol{\theta}, X)$ to represent the benefit of rejecting H_{03}, then the gain function will be

$$G(\boldsymbol{\theta}, X) = \gamma_1(\boldsymbol{\theta}, X)\mathcal{R}_1 + \gamma_3(\boldsymbol{\theta}, X)\mathcal{R}_3.$$

The choice of $\gamma_1(\boldsymbol{\theta}, X)$ and $\gamma_3(\boldsymbol{\theta}, X)$ may reflect both the treatment effect as seen in Equation (8) and the estimates of θ_1 and θ_3 which can be constructed from X. In our formulation of the design question, the total sample size is fixed, so we have not included a cost of treating patients in the study in the overall gain function: such a cost would be required if we were to include the option of stopping for futility at the interim analysis. One could also consider adding other important outcomes from the trial such as the safety profile of the treatment. The application of the methods that follow is not particularly dependent on the choice of gain function, although the choice of gain function will influence what is optimal.

4.2 | Computing the Bayes optimal design

With the prior distribution π and gain function G specified, we wish to find the decision rule d that maximises the Bayes expected gain of the trial $E\{G(\boldsymbol{\theta}, X)\}$, where the expectation is over both the prior distribution for $\boldsymbol{\theta}$ and the distribution of X given $\boldsymbol{\theta}$.

We denote the conditional density function of X_1 given $\boldsymbol{\theta}$ by $f_{X_1|\boldsymbol{\theta}}(x_1|\boldsymbol{\theta})$, the density of the marginal distribution of X_1 by $f_{X_1}(x_1)$, and the conditional density of X_2 given $\boldsymbol{\theta}$ and decision $d(x_1)$ by $f_{X_2|\boldsymbol{\theta},d}(x_2|\boldsymbol{\theta}, d(x_1))$. Let $\pi_{\boldsymbol{\theta}|X_1}(\boldsymbol{\theta}|x_1)$ be the density of the posterior distribution of $\boldsymbol{\theta}$ given $X_1 = x_1$, so

$$\pi(\boldsymbol{\theta}) f_{X_1|\boldsymbol{\theta}}(x_1|\boldsymbol{\theta}) = f_{X_1}(x_1)\ \pi_{\boldsymbol{\theta}|X_1}(\boldsymbol{\theta}|x_1).$$

Then the expected gain when applying decision rule d is

$$E\{G(\boldsymbol{\theta}, X)\} = \int_{\boldsymbol{\theta}}\int_{x_1}\int_{x_2} \pi(\boldsymbol{\theta}) f_{X_1|\boldsymbol{\theta}}(x_1|\boldsymbol{\theta}) f_{X_2|\boldsymbol{\theta},d}(x_2|\boldsymbol{\theta}, d(x_1))\ G(\boldsymbol{\theta}, (x_1, d(x_1), x_2))\ \mathrm{d}x_2\ \mathrm{d}x_1\ \mathrm{d}\boldsymbol{\theta}$$

$$= \int_{x_1} f_{X_1}(x_1) \int_{\theta}\int_{x_2} \pi_{\theta|X_1}(\boldsymbol{\theta}|x_1) f_{X_2|\theta,d}(x_2|\boldsymbol{\theta}, d(x_1))\ G(\boldsymbol{\theta}, (x_1, d(x_1), x_2))\ \mathrm{d}x_2\ \mathrm{d}\boldsymbol{\theta}\ \mathrm{d}x_1. \tag{9}$$

It is evident from (9) that the optimal decision rule can be found by choosing $d(x_1)$ to maximize

$$\int_{\theta}\int_{x_2} \pi_{\theta|X_1}(\boldsymbol{\theta}|x_1) f_{X_2|\theta,d}(x_2|\boldsymbol{\theta}, d(x_1))\ G(\boldsymbol{\theta}, (x_1, d(x_1), x_2))\ \mathrm{d}x_2\ \mathrm{d}\boldsymbol{\theta} = E\{G(\boldsymbol{\theta}, X) \mid X_1 = x_1, d(x_1)\}, \tag{10}$$

for each x_1. That is, we choose the enrichment decision that maximizes the conditional expected gain given the stage 1 data under the posterior distribution of $\boldsymbol{\theta}$ at the interim analysis.

Given observed stage 1 data $X_1 = x_1 = (\hat{\theta}_1^{(1)}, \hat{\theta}_2^{(1)})$, we need to compare values of the integral (10) in the two cases $d(x_1) = 1$ (enrichment) and $d(x_1) = 2$ (no enrichment). Since this integral is not analytically tractable, we evaluate it by Monte Carlo simulation. To do this, we draw a sample $\{\boldsymbol{\theta}_i = (\theta_{i,1}, \theta_{i,2}),\ i = 1, \ldots, M\}$, from the posterior distribution $\pi_{\theta|X_1}(\boldsymbol{\theta}|x_1)$ and find the conditional expected gain under each $\boldsymbol{\theta}_i$ for the two options, "enrich" and "do not enrich." We take the average gain over this sample of $\boldsymbol{\theta}_i$ values as our estimate of the conditional expected gain for each option. We conclude that the decision $d(x_1)$ giving the larger of the two values for the conditional expected gain is the Bayes optimal decision when $X_1 = x_1 = (\hat{\theta}_1^{(1)}, \hat{\theta}_2^{(1)})$.

In assessing the decision to enrich, $d(x_1) = 1$, when $X_1 = x_1 = (\hat{\theta}_1^{(1)}, \hat{\theta}_2^{(1)})$ we apply the definitions of Section 3 to find the critical value $\kappa(x_1)$ such that $\hat{\theta}_1^{(2)} \geq \kappa(x_1)$ implies $P_1^{(c)} \leq \alpha$ and $P_{13}^{(c)} \leq \alpha$, so H_{01} is rejected in the closed testing procedure. We compute $P(\hat{\theta}_1^{(2)} \geq \kappa(x_1) \mid \theta_1 = \theta_{i,1}, \hat{\theta}_1^{(1)}, d(x_1) = 1)$ for each $i = 1, \ldots, M$ and combine the results to obtain the estimate of the conditional expected gain

$$\hat{E}\{G(\boldsymbol{\theta}, X) \mid X_1 = x_1 = (\hat{\theta}_1^{(1)}, \hat{\theta}_2^{(1)}), d(x_1) = 1\} = \frac{1}{M}\sum_{i=1}^{M} \lambda\theta_{i,1} P(\hat{\theta}_1^{(2)} \geq \kappa(x_1) \mid \theta_1 = \theta_{i,1}, \hat{\theta}_1^{(1)}, d(x_1) = 1). \tag{11}$$

If $d(x_1) = 2$ and the trial continues without enrichment, the possibilities in stage 2 are more complex. In this case, for each $i = 1, \ldots, M$ we continue to simulate the remainder of the trial by generating $(\hat{\theta}_{i,1}^{(2)}, \hat{\theta}_{i,2}^{(2)})$ under $\boldsymbol{\theta} = \boldsymbol{\theta}_i$ and evaluating the gain (8) with $\boldsymbol{\theta} = \boldsymbol{\theta}_i$ and $x = ((\hat{\theta}_1^{(1)}, \hat{\theta}_2^{(1)}), 2, (\hat{\theta}_{i,1}^{(2)}, \hat{\theta}_{i,2}^{(2)}))$. Combining these results gives the estimate of the conditional expected gain

$$\hat{E}\{G(\boldsymbol{\theta}, X) \mid X_1 = x_1 = (\hat{\theta}_1^{(1)}, \hat{\theta}_2^{(1)}), d(x_1) = 2\} = \frac{1}{M}\sum_{i=1}^{M} G(\boldsymbol{\theta}_i, ((\hat{\theta}_1^{(1)}, \hat{\theta}_2^{(1)}), 2, (\hat{\theta}_{i,1}^{(2)}, \hat{\theta}_{i,2}^{(2)}))). \tag{12}$$

The value of M used in these simulations should be chosen to give the desired level of accuracy. We have found $M = 10^5$ or 10^6 to give sufficient accuracy in the examples we have studied.

4.3 | Determining the decision rule and decision boundary

In order to find the operating characteristics of a proposed adaptive enrichment design we must be able to repeatedly simulate the design in full. This requires repeated application of the interim decision rule that specifies the optimal design for a given prior π and gain function G: thus we need to know the optimal decision for all possible values of $x_1 = (\hat{\theta}_1^{(1)}, \hat{\theta}_2^{(1)})$. We present an algorithm that enables the computation of the optimal decision rule over a large square region, A, such that $P(X_1 \in A)$ is very close to 1. The algorithm divides this region into an array of much smaller squares and determines the optimal decision for values of x_1 in each small square. With simple extrapolation beyond the boundaries of A, this process divides the plane into two regions, A_E where the optimal decision is to enrich, and A_C where it is optimal to continue recruitment in the full population.

Experience shows that the two regions A_E and A_C are quite regular in shape and this fact allows us to reduce the computation needed to find the optimal decision rule. We first divide A into four subsquares and determine the optimal decisions at the vertices of these squares. Then, if the same decision is optimal at all four vertices we record this as the optimal decision for all points in that square. If, however, both decisions are optimal for at least one vertex we subdivide this square into four smaller squares. In the next iterative step, we consider the set of squares of the smallest size and for each of these we either record an optimal decision for the whole square or subdivide the square into four smaller ones. We continue this iterative process until we reach squares of the desired size. Further details of this method and a discussion of its accuracy are given in Appendix B. The results of these calculations are 2-fold. First, the list of optimal decisions for each small square provides the information needed to implement the optimal adaptive decision rule. Secondly, the results can be presented graphically to help visualize the optimal decision rule.

4.4 | Assessing the performance of an optimized trial design

Suppose the decision rule of an optimized adaptive enrichment design is defined by regions A_E and A_C as described above. We assess the overall performance of this design by simulation. For each replicate $i = 1, \dots, N$, we generate a parameter vector $\boldsymbol{\theta}_i = (\theta_{i,1}, \theta_{i,2})$ then simulate stage 1 data $x_{i,1} = (\hat{\theta}^{(1)}_{i,1}, \hat{\theta}^{(1)}_{i,2})$ assuming $\boldsymbol{\theta} = \boldsymbol{\theta}_i$. We determine whether $x_{i,1}$ is in A_E or A_C, set $d(x_{i,1}) = 1$ or 2 accordingly, and apply this decision, still assuming $\boldsymbol{\theta} = \boldsymbol{\theta}_i$, as we generate the stage 2 data: $x_{i,2} = \hat{\theta}^{(2)}_{i,1}$ if $d(x_{i,1}) = 1$ (enrichment), or $x_{i,2} = (\hat{\theta}^{(2)}_{i,1}, \hat{\theta}^{(2)}_{i,2})$ if $d(x_{i,1}) = 2$ (no enrichment). Finally, we determine which hypotheses are rejected and evaluate the gain function for these outcomes when $\boldsymbol{\theta} = \boldsymbol{\theta}_i$. Averaging over the N replicates gives the estimate

$$\hat{E}\{G(\boldsymbol{\theta}, X)\} = \frac{1}{N}\sum_{i=1}^{N} G(\boldsymbol{\theta}_i, (x_{i,1}, d(x_{i,1}), x_{i,2})).$$

The same set of simulated data can be used to estimate other properties of the design such as the probabilities of rejecting each null hypothesis. In our simulations we have used $N = 10^6$, so sampling error for the estimates reported is negligible.

One might ask whether it would be helpful to generate multiple replicates of the stage 2 data for each $\boldsymbol{\theta}_i$ and $x_{1,i}$. However, the distribution of $\boldsymbol{\theta}_i$ and $x_{1,i}$ accounts for much of the variability of $G(\boldsymbol{\theta}, X)$ and it is more efficient to use the available computational effort to increase the number of replicates, N, of the first stage data. Of course, this approach relies on our having carried out initial work to find the regions A_E and A_C that define the optimal decision rule, and in doing this we will have generated multiple samples of stage 2 data conditional on particular values of X_1.

5 | TWO NONADAPTIVE DESIGNS

There are two further options that should be considered when an adaptive enrichment design is envisaged. The first is a design in which patients are recruited from the full population throughout the trial, but both null hypotheses H_{01} and H_{03} are tested at the end. We shall refer to this as the Fixed Full population (FF) design. The other possibility is a Fixed Subpopulation (FS) design, in which subjects are only recruited from the subpopulation and only the hypothesis H_{01} is tested.

The Fixed Full population design. For comparability with other designs, we assume the same total sample size, n, as in Section 2.2. Thus, λn patients are recruited from $\mathcal{S}_1$ and $(1 - \lambda)n$ from $\mathcal{S}_2$. With $\tilde{\mathcal{I}}$ as defined in (1), the data provide estimates

$$\hat{\theta}_1 \sim N(\theta_1, (\lambda\tilde{\mathcal{I}})^{-1})),$$

and

$$\hat{\theta}_3 \sim N(\theta_3, (\tilde{\mathcal{I}})^{-1}),$$

and the joint distribution of $(\hat{\theta}_1, \hat{\theta}_3)$ is bivariate normal with correlation $\sqrt{\lambda}$.

The P-values for testing H_{01} and H_{03} are

$$P_1 = 1 - \Phi(\hat{\theta}_1\sqrt{\{\lambda\tilde{\mathcal{I}}\}}) \quad \text{and} \quad P_3 = 1 - \Phi(\hat{\theta}_3\sqrt{\tilde{\mathcal{I}}}),$$

respectively, and Simes' method gives the p-value

$$P_{13} = \min\{2\ \min(P_1, P_3),\ \max(P_1, P_3)\}$$

for the intersection hypothesis $H_{0,13}$. Applying the closed testing procedure, we reject H_{01} overall if $P_1 \leq \alpha$ and $P_{13} \leq \alpha$, and we reject H_{03} overall if $P_3 \leq \alpha$ and $P_{13} \leq \alpha$.

There are reasons why the FF design may be more efficient than the optimal adaptive design if the prior $\pi(\boldsymbol{\theta})$ is concentrated on values of $\boldsymbol{\theta}$ under which enrichment is unlikely to occur. Suppose an adaptive design is conducted and enrichment does not occur. With suitable weights in the combination rule (7), the adaptive design's P-values $P_1^{(c)}$ and $P_3^{(c)}$, as shown in Table 1, are equal to the P_1 and P_3 obtained when the same data are observed in the FF design. However, $P_{13}^{(c)} = W(P_{13}^{(1)}, P_{13}^{(2)})$ differs from the P_{13} arising from the same data in the FF design. Since P_{13} in the FF design is based on the sufficient statistics for θ_1 and θ_3 in the full data set, it provides a more powerful test of $H_{0,13}$ than the adaptive design's $P_{13}^{(c)}$. The requirement to use $P_{13}^{(c)}$ rather than P_{13} to test $H_{0,13}$ is the price we pay for the adaptive design's flexibility to enrich on other occasions: if such occasions are not particularly likely under the prior $\pi(\boldsymbol{\theta})$, it is plausible that the FF design will be superior.

The Fixed Subpopulation design. In the FS design, all n subjects are recruited from $\mathcal{S}_1$. These provide the estimate

$$\hat{\theta}_1 \sim N(\theta_1, \tilde{\mathcal{I}}^{-1})),$$

and the P-value

$$P_1 = 1 - \Phi(\hat{\theta}_1 \sqrt{\tilde{\mathcal{I}}}),$$

and H_{01} is rejected if $P_1 \leq \alpha$. In this design H_{03} is not tested.

We can expect the FS design to perform well when the prior $\pi(\boldsymbol{\theta})$ is such that the optimal adaptive design is highly likely to enrich. Then, the FS design has the benefit of a larger sample size from $\mathcal{S}_1$ and, hence, a more accurate estimate $\hat{\theta}_1$. Furthermore, the FS design only tests H_{01} and so does not have to make a multiplicity adjustment for testing two hypotheses.

6 EXAMPLES

6.1 One-point prior distributions

We consider a Phase III clinical trial as described in Section 2.1 where the subpopulations $\mathcal{S}_1$ and $\mathcal{S}_2$ are of equal size, so $\lambda = 0.5$. We set the FWER to be $\alpha = 0.025$ and suppose the total sample size n would provide power 0.9 to detect a treatment effect of size 10 when testing only the hypothesis H_{03} in a nonadaptive design. This leads to the total information

$$\tilde{\mathcal{I}} = \left(\frac{\Phi^{-1}(0.9) + \Phi^{-1}(0.975)}{10} \right)^2 = 0.105,$$

which is, for example, the information provided by a total sample size $n = 264$ when patient responses have standard deviation $\sigma = 25$. In adaptive enrichment designs we suppose the interim analysis occurs after half the total sample has been observed, thus $\tau = 0.5$. Then, with $\lambda = 0.5$, $\tau = 0.5$ and $\tilde{\mathcal{I}} = 0.105$, the interim estimates $\hat{\theta}_1^{(1)}$ and $\hat{\theta}_1^{(2)}$ have SD 6.15.

In order to gain insight into how adaptive designs function and what they may achieve, we first consider cases where the prior distribution for $\boldsymbol{\theta}$ places probability mass 1 at a single point, $\boldsymbol{\theta} = \boldsymbol{\theta}_0 = (\theta_{0,1}, \theta_{0,2})$. For given $\boldsymbol{\theta}_0$, we derived the decision rule for the adaptive enrichment (AE) design that maximises the expected gain, using the gain function $G(\boldsymbol{\theta}, X)$ specified in (8). For comparison, we also computed properties under $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ of the FF design, which recruits from the full population throughout the trial, and the FS design which only recruits from the subpopulation. Results presented in Table 2 for selected values of $\boldsymbol{\theta}_0$ show each type of design, FF, FS, and AE, to be optimal for certain values of $\boldsymbol{\theta}_0$.

We carried out further calculations on a grid of values of $\boldsymbol{\theta}_0$ to find the regions where each type of design is optimal. These regions are shown in Figure 1.

We note that the FF design is optimal when $\theta_{0,3} = 0.5\ (\theta_{0,1} + \theta_{0,2})$ is large or $\theta_{0,1}$ is only a little larger than $\theta_{0,2}$. The FS design is optimal when $\theta_{0,1}$ is substantially larger than $\theta_{0,2}$ and $\theta_{0,2}$ is small. This leaves a region of $\boldsymbol{\theta}_0$ values where the AE design is optimal, offering a modest increase in expected gain over both fixed designs. The advantage of the AE design over the FF design is largest in cases such as $\boldsymbol{\theta}_0 = (10, 2)$ and $\boldsymbol{\theta}_0 = (12, 2)$, where $\theta_{0,2}$ is small and the AE design has a high probability of enrichment and rejection of H_{01} only. Although the FS design has even higher expected gain in these cases, investigators may be reluctant to make such an early decision to ignore subpopulation $\mathcal{S}_2$ completely, in which case the key comparison is between AE and FF designs.

TABLE 2 Properties of fixed subpopulation (FS), fixed full population (FF), and optimal adaptive enrichment (AE) designs when $\boldsymbol{\theta} = \boldsymbol{\theta}_0 = (\theta_{0,1}, \theta_{0,2})$. Here $P(\mathcal{R}_1)$ is the probability that only H_{01} is rejected and $P(\mathcal{R}_3)$ the probability that H_{03} is rejected. The AE design is optimized for the prior distribution with probability 1 at the single point $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. In each case, the design with the highest expected gain is highlighted

$\theta_{0,1}$	$\theta_{0,2}$	$\theta_{0,3}$	Trial design	$P(\mathcal{R}_1)$	$P(\mathcal{R}_3)$	P(Enrich)	$E\{G(\theta, X)\}$
10	2	6	**FS**	0.90	—	—	**4.50**
			FF	0.14	0.46	—	3.48
			AE	0.50	0.23	0.71	3.89
10	4	7	FS	0.90	—	—	4.50
			FF	0.08	0.58	—	4.46
			AE	0.25	0.46	0.38	**4.51**
10	6	8	FS	0.90	—	—	4.50
			FF	0.04	0.69	—	**5.68**
			AE	0.08	0.64	0.13	5.55
10	10	10	FS	0.90	—	—	4.50
			FF	0.01	0.86	—	**8.60**
			AE	0.01	0.83	0.00	8.34
12	2	7	**FS**	0.97	—	—	**5.84**
			FF	0.15	0.60	—	5.15
			AE	0.50	0.36	0.58	5.58
12	4	8	FS	0.97	—	—	5.84
			FF	0.09	0.71	—	6.20
			AE	0.25	0.60	0.28	**6.30**
12	6	9	FS	0.97	—	—	5.84
			FF	0.04	0.80	—	**7.44**
			AE	0.09	0.76	0.10	7.38
14	2	8	FS	1.00	—	—	6.97
			FF	0.15	0.73	—	6.83
			AE	0.40	0.54	0.39	**7.13**
14	4	9	FS	1.00	—	—	6.97
			FF	0.08	0.82	—	7.90
			AE	0.19	0.74	0.17	**7.97**
14	6	10	FS	1.00	—	—	6.97
			FF	0.04	0.88	—	**9.10**
			AE	0.07	0.86	0.06	9.07

In extreme cases such as $\boldsymbol{\theta} = (10, 10)$ where both $\theta_{0,1}$ and $\theta_{0,2}$ are high, there is a high probability that the AE design does not enrich and so has the same final dataset as the FF design. As discussed in Section 5, the AE design uses a different form of $P_{13}^{(c)}$ and this leads to less efficient use of the final data when enrichment does not occur and a lower expected gain than for the FF design.

Since the AE design is optimized with knowledge of the value of $\boldsymbol{\theta}_0$, its advantage when it is superior to both fixed designs does not stem from having improved estimates of the true treatment effects at the interim analysis. Rather, the decision to enrich or not is based on the likelihood that current data, summarized as $(\hat{\theta}_1^{(1)}, \hat{\theta}_1^{(2)})$, will lead to eventual rejection of H_{01} or H_{03}. This suggests that the AE design may have an even greater advantage in situations where the prior

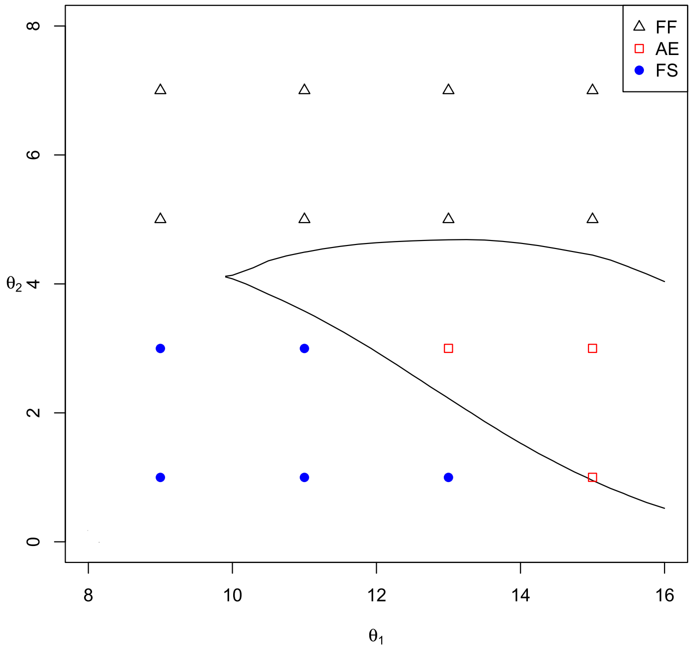


FIGURE 1 Regions of $\boldsymbol{\theta}$ values in which each of the Fixed Full population (FF), Fixed Subpopulation (FS), and optimal Adaptive Enrichment (AE) designs give the highest value of $E\{G(\boldsymbol{\theta}, X)\}$ [Colour figure can be viewed at wileyonlinelibrary.com]

distribution for $\boldsymbol{\theta}$ is more dispersed, since then it can also exploit the information about $\boldsymbol{\theta}$ that becomes available at the interim analysis. We shall assess the performance of designs under dispersed prior distributions for $\boldsymbol{\theta}$ in the next Section.

6.2 | Proper prior distributions for θ

In practice, one expects there to be considerable uncertainty about the true treatment effect. We capture this uncertainty in a bivariate normal prior distribution for $\boldsymbol{\theta}$,

$$\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \sim N_2\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\ \sigma_1\sigma_2 \\ \rho\ \sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right).$$

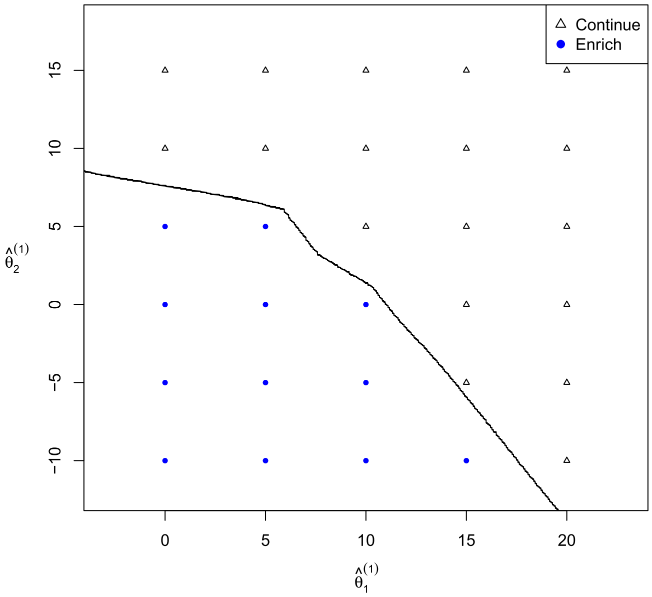
Figure 2 shows the enrichment decision rule for the Bayes optimal adaptive enrichment trial when $\mu_1 = 12$, $\mu_2 = 2$, $\sigma_1^2 = \sigma_2^2 = 25$ and $\rho = 0.75$. The sharp angles in the decision boundary arise from discontinuities in the way $\hat{\theta}_1^{(1)}$ and $\hat{\theta}_1^{(2)}$ determine $P_1^{(1)}$, $P_3^{(1)}$, and $P_{13}^{(1)}$ and how these P-values appear in the criteria for the closed testing procedure to reject H_{01} or H_{03}.

Enrichment occurs when there is a low conditional probability of rejecting H_{03}, given the prior and current data. This includes cases where both $\hat{\theta}_1^{(1)}$ and $\hat{\theta}_1^{(2)}$ are low so rejection of H_{01} is also unlikely: one could add a rule to stop for futility in such cases. When $\hat{\theta}_1^{(1)}$ is high, so that rejection of H_{01} is very likely, the trial is not enriched, even for lower values of $\hat{\theta}_2^{(1)}$, as long as it is feasible that H_{03} will also be rejected.

Table 3 shows properties of the Bayes optimal AE design, along with properties of the nonadaptive FF and FS designs, for prior distributions centred at the values of $\boldsymbol{\theta}_0$ considered in Table 2 but with $\sigma_1^2 = \sigma_2^2 = 25$ and $\rho = 0.75$. In contrast with the results of Table 2, the AE design has higher expected gain than the FS design in all these examples with a dispersed prior.

The AE design has higher expected gain than the FF design in six of the ten examples — but the margin of superiority is not great. Thus, there is not much evidence that the enrichment design profits from information about $\boldsymbol{\theta}$ at the interim analysis. The explanation for this is that, in the examples of Table 3, the posterior distribution of $\boldsymbol{\theta}$ after seeing the interim data is still widely dispersed, with the SDs for θ_1 and θ_2 equal to 3.59. This is not just a feature of our particular examples. Suppose a study's total sample size is chosen so that a final test of H_{03}: $\theta_3 \leq 0$ with type I error rate 0.025 has power 0.9 when $\theta_3 = \delta$. With no enrichment, the SD of the final $\hat{\theta}_3$ is 0.31 δ. If there are two equally sized subpopulations, the interim estimates of θ_1 and θ_2 based on half of the total data have SD 0.62 δ. The posterior variance of θ_1 and θ_2 at the interim

FIGURE 2 An example of a Bayes optimal decision rule for an adaptive enrichment trial [Colour figure can be viewed at wileyonlinelibrary.com]



analysis depends on the prior variances of θ_1 and θ_2 and, to a small degree, on the prior correlation. If, as in the examples of Table 3, the prior has $Var\ (\theta_1) = Var\ (\theta_2) = (\delta/2)^2$, the posterior SDs of θ_1 and θ_2 at the interim analysis will be around 0.36 δ and a credible interval for θ_1 or θ_2 could easily contain both 0 and δ. On the other hand, the lower prior variances $Var\ (\theta_1) = Var\ (\theta_2) = (\delta/4)^2$ lead to posterior SDs around 0.23 δ—only slightly lower than the prior SDs of 0.25 δ. Thus, in cases where the prior variance is high, considerable uncertainty about θ_1 and θ_2 remains at the interim analysis, while if the prior variance is low, the interim data have little impact on the posterior distribution of θ_1 and θ_2.

Table 4 presents results for a further selection of prior distributions for $\boldsymbol{\theta}$. The examples show that the prior correlation, ρ, has a small effect on expected gain but very little effect on the relative performance of different designs.

In cases with (μ_1, μ_2) equal to (10,2) or (12,2) and low prior variance, the FS design is best—but it is substantially inferior to the FF and AE designs in other situations. We conclude that the FS design option should only be considered if there is a strong prior belief that the new treatment will offer little or no benefit to subpopulation $\mathcal{S}_2$.

For the cases in Table 4, the AE design has higher expected gain than the FF design (with the exception of a couple of cases where the two designs have almost equal expected gain). However, we have failed to find an example where the AE design is vastly superior to both the FS and FF designs: the example in Table 3 with $(\mu_1, \mu_2) = (14, 2)$ and $\sigma_1^2 = \sigma_2^2 = 25$ and the examples in Table 4 with $(\mu_1, \mu_2) = (12, 2)$ and $\sigma_1^2 = \sigma_2^2 = 16$ have the highest difference in expected gains in favor of the AE design. One may also argue from the values of $P(\mathcal{R}_1)$ and $P(\mathcal{R}_3)$ in Tables 2 and 3 that the AE design shows greater selectivity and is less likely to conclude the new treatment is beneficial to the full population when the treatment effect in $\mathcal{S}_2$ is small or absent altogether.

6.3 | Adjusting other design parameters

When planning an enrichment trial it is natural to investigate all design parameters and, where possible, optimise their values. Here we consider the timing of the interim analysis at which the decision to enrich may be taken but we note that a similar approach can be taken in setting other design features. Suppose, with the problem formulation described above, we wish to find the best value of τ when the prior distribution of (θ_1, θ_2) is given by $\mu_1 = 12$, $\mu_2 = 4$, $\sigma_1^2 = \sigma_2^2 = 25$ and $\rho = 0.75$. We have applied our methods to find the Bayes optimal design for different values of τ. Here we used weights $w_1 = \sqrt{\tau}$ and $w_2 = \sqrt{1 - \tau}$ in the combination test to account for the different sample sizes before and after the interim analysis. Table 5 shows properties of designs with values of τ ranging from 0.1 to 0.9. We see that our earlier choice of $\tau = 0.5$ yields the highest expected gain of 6.91, but designs with τ between 0.3 and 0.6 are very close to this optimum. As τ increases from 0.1 to 0.7, the probability of enriching the trial increases. This is in keeping with the information in Table 3 that the FF design is superior to the FS design, so a certain amount of data is needed to show that enrichment is the better option in a particular trial. We have seen similar results in other examples where the the FF design is superior to the FS design: AE designs with a range of τ values perform well, as long as τ is high enough to give enough information to make an informed decision about enrichment.

TABLE 3 Properties of fixed subpopulation (FS), fixed full population (FF), and optimal adaptive enrichment (AE) designs when θ has the prior distribution given by (13). Here $P(\mathcal{R}_1)$ is the probability that only H_{01} is rejected and $P(\mathcal{R}_3)$ the probability that H_{03} is rejected

μ_1	μ_2	σ_1^2	σ_2^2	ρ	**Trial design**	$P(\mathcal{R}_1)$	$P(\mathcal{R}_3)$	P**(Enrich)**	$E\{G(\theta, X)\}$
10	2	25	25	0.75	FS	0.75	—	—	4.42
					FF	0.10	0.48	—	4.89
					AE	0.25	0.38	0.53	**4.98**
10	4	25	25	0.75	FS	0.75	—	—	4.42
					FF	0.06	0.54	—	**5.64**
					AE	0.15	0.48	0.37	5.63
10	6	25	25	0.75	FS	0.75	—	—	4.42
					FF	0.04	0.61	—	**6.52**
					AE	0.08	0.57	0.23	6.43
10	10	25	25	0.75	FS	0.75	—	—	4.43
					FF	0.01	0.72	—	**8.59**
					AE	0.01	0.70	0.02	8.43
12	2	25	25	0.75	FS	0.84	—	—	5.57
					FF	0.12	0.55	—	6.09
					AE	0.29	0.44	0.49	**6.23**
12	4	25	25	0.75	FS	0.84	—	—	5.57
					FF	0.08	0.62	—	6.86
					AE	0.18	0.55	0.33	**6.91**
12	6	25	25	0.75	FS	0.84	—	—	5.57
					FF	0.05	0.68	—	**7.77**
					AE	0.10	0.64	0.21	7.72
14	2	25	25	0.75	FS	0.91	—	—	6.72
					FF	0.14	0.63	—	7.33
					AE	0.32	0.50	0.44	**7.53**
14	4	25	25	0.75	FS	0.91	—	—	6.72
					FF	0.10	0.69	—	8.13
					AE	0.19	0.62	0.29	**8.21**
14	6	25	25	0.75	FS	0.91	—	—	6.72
					FF	0.06	0.74	—	9.03
					AE	0.11	0.71	0.18	**9.04**

A somewhat different pattern is seen in scenarios where the FS design gives a high expected gain. Suppose the prior distribution for (θ_1, θ_2) has $\mu_1 = 12$, $\mu_2 = 2$, $\sigma_1^2 = \sigma_2^2 = 4$ and $\rho = 0.75$. We saw in Table 4 that the FS design has higher expected gain than both the FF design and the optimal AE design with $\tau = 0.5$. Table 6 shows results for optimal AE designs with different values of τ.

Since we have used weights $w_1 = \sqrt{\tau}$ and $w_2 = \sqrt{1-\tau}$ in the combination test, as τ decreases toward zero the analysis after enrichment becomes identical to that of the FS design. This explains why the probability of enrichment is high for small values of τ and the expected gain is very close to that of the FS design. In fact, the optimal AE designs with $\tau = 0.1$, 0.2 and 0.3 have marginally higher expected gain than the FS design. Thus, an adaptive design with an early interim analysis could be a suitable choice if investigators are reluctant to restrict attention to subpopulation S_1 from the outset.

TABLE 4 Properties of fixed subpopulation (FS), fixed full population (FF), and optimal adaptive enrichment (AE) designs when θ has the prior distribution given by (13)

Prior parameters					$E\{G(\theta,X)\}$			P**(Enrich)**
μ_1	μ_2	σ_1^2	σ_2^2	ρ	**FS**	**FF**	**AE**	**for AE**
10	2	0	0	—	**4.50**	3.48	3.89	0.71
10	2	1	1	0	**4.47**	3.55	3.93	0.69
10	2	1	1	0.75	**4.47**	3.58	3.95	0.69
10	2	4	4	0	**4.42**	3.74	4.04	0.64
10	2	4	4	0.75	**4.42**	3.81	4.09	0.64
10	2	16	16	0	4.38	4.33	**4.50**	0.55
10	2	16	16	0.75	4.38	4.52	**4.65**	0.55
12	2	0	0	—	**5.84**	5.15	5.58	0.58
12	2	1	1	0	**5.81**	5.18	5.58	0.56
12	2	1	1	0.75	**5.81**	5.19	5.59	0.56
12	2	4	4	0	**5.74**	5.29	5.61	0.52
12	2	4	4	0.75	**5.74**	5.34	5.67	0.53
12	2	16	16	0	5.60	5.66	**5.86**	0.49
12	2	16	16	0.75	5.60	5.80	**5.99**	0.49
10	4	0	0	—	4.50	4.46	**4.51**	0.39
10	4	1	1	0	4.47	4.51	**4.56**	0.39
10	4	1	1	0.75	4.47	4.52	**4.57**	0.38
10	4	4	4	0	4.42	4.66	**4.68**	0.36
10	4	4	4	0.75	4.42	4.71	**4.75**	0.37
10	4	16	16	0	4.38	**5.14**	**5.14**	0.35
10	4	16	16	0.75	4.38	**5.31**	**5.31**	0.37
12	4	0	0	—	5.84	6.20	**6.30**	0.28
12	4	1	1	0	5.81	6.21	**6.31**	0.28
12	4	1	1	0.75	5.81	6.22	**6.32**	0.29
12	4	4	4	0	5.74	6.28	**6.35**	0.28
12	4	4	4	0.75	5.74	6.29	**6.39**	0.29
12	4	16	16	0	5.60	6.54	**6.57**	0.29
12	4	16	16	0.75	5.60	6.63	**6.69**	0.31
14	4	0	0	—	6.97	7.90	**7.97**	0.17
14	4	1	1	0	6.95	7.89	**7.97**	0.17
14	4	1	1	0.75	6.95	7.89	**7.97**	0.18
14	4	4	4	0	6.91	7.89	**7.95**	0.18
14	4	4	4	0.75	6.91	7.87	**7.97**	0.19
14	4	16	16	0	6.78	7.96	**8.00**	0.22
14	4	16	16	0.75	6.78	7.99	**8.08**	0.23

TABLE 5 Properties of the optimal adaptive enrichment (AE) design for different timings of the interim analysis τ when θ has the prior distribution given by (13) with $\mu_1 = 12$, $\mu_2 = 4$, $\sigma_1^2 = \sigma_2^2 = 25$, and $\rho = 0.75$. The interim analysis takes place after a fraction τ of the total sample has been observed

τ	$P(\mathcal{R}_1)$	$P(\mathcal{R}_3)$	P(Enrich)	$E\{G(\theta, X)\}$
0.1	0.14	0.58	0.13	6.84
0.2	0.17	0.56	0.23	6.87
0.3	0.19	0.55	0.28	6.89
0.4	0.19	0.55	0.31	**6.91**
0.5	0.18	0.55	0.33	**6.91**
0.6	0.17	0.56	0.34	6.89
0.7	0.15	0.57	0.34	6.88
0.8	0.13	0.58	0.32	6.87
0.9	0.11	0.59	0.27	6.85

TABLE 6 Properties of the optimal adaptive enrichment (AE) design for different timings of the interim analysis when θ has the prior distribution given by (13) with $\mu_1 = 12$, $\mu_2 = 2$, $\sigma_1^2 = \sigma_2^2 = 4$, and $\rho = 0.75$. The interim analysis takes place after a fraction τ of the total sample has been observed

τ	$P(\mathcal{R}_1)$	$P(\mathcal{R}_3)$	P(Enrich)	$E\{G(\theta, X)\}$
0.1	0.69	0.21	0.72	5.75
0.2	0.60	0.28	0.64	**5.76**
0.3	0.53	0.33	0.59	5.75
0.4	0.48	0.37	0.55	5.72
0.5	0.42	0.41	0.53	5.67
0.6	0.37	0.44	0.49	5.61
0.7	0.32	0.47	0.45	5.55
0.8	0.27	0.51	0.41	5.48
0.9	0.15	0.54	0.36	5.39

TABLE 7 Properties of fixed subpopulation (FS), fixed full population (FF), and optimal adaptive enrichment (AE) designs for different subpopulation sizes when θ has the prior distribution given by (13) with $\mu_1 = 14$, $\mu_2 = 2$, $\sigma_1^2 = \sigma_2^2 = 25$, and $\rho = 0.75$. The subpopulation $\mathcal{S}_2$ represents a fraction λ of the total population

λ	$E\{G(\theta, X)\}$			P(Enrich)
	FS	FF	AE	for AE
0.1	1.35	2.44	**2.61**	0.56
0.2	2.69	3.58	**3.87**	0.53
0.3	4.04	4.85	**5.14**	0.49
0.4	5.38	6.11	**6.36**	0.44
0.5	6.72	7.33	**7.53**	0.44
0.6	8.06	8.53	**8.71**	0.40
0.7	9.42	9.73	**9.87**	0.39
0.8	10.77	10.93	**11.03**	0.38
0.9	12.11	12.13	**12.18**	0.38

6.4 | Effect of the subpopulation size

In all of our examples so far, the subpopulation $\mathcal{S}_1$ has represented half of the total population. The size of the specified subpopulation is a feature of the study and not a parameter that can be controlled. Table 7 shows the effect of the subpopulation size on the relative performance of different designs. In this example, the prior distribution for (θ_1, θ_2) has $\mu_1 = 14$, $\mu_2 = 2$, $\sigma_1^2 = \sigma_2^2 = 25$ and $\rho = 0.75$, and we saw in Table 3 that the optimal AE design is the best option when $\lambda = 0.5$. The results in Table 7 show that the optimal AE design remains superior to both the FF and FS designs across the whole range of λ values from 0.1 to 0.9.

For each design, the expected gain for all designs increases with λ as the fraction of the population in which the treatment effect is θ_1 becomes larger. The margin of superiority of the AE design over the FF design is largest for $\lambda = 0.2$

and $\lambda = 0.3$. The reasons behind this are quite complex. The potential benefits of adaptive enrichment are small when λ is close to zero or 1 and one of the subpopulations forms a large fraction of the total population. Also, the interim estimate of θ_1 has a high variance when λ is small and the estimate of θ_2 has a high variance when λ is large, reducing the information available when making the interim decision. Nevertheless, it is clear from this example that adaptive enrichment can be of benefit over a wide range of subpopulation sizes.

7 | DISCUSSION

We have considered adaptive trial designs for testing the efficacy of a new treatment when a prespecified subpopulation is deemed particularly likely to benefit from the new treatment. The methods we have presented facilitate calculation of the Bayes optimal rule for deciding whether to enrich in a design where the familywise type I error rate is controlled by a closed testing procedure and combination test. Since this calculation relies on Monte Carlo simulation to determine the optimum decision at all possible values of $(\hat{\theta}_1^{(1)}, \hat{\theta}_2^{(1)})$, efficient calculation is crucial. We achieve this by use of an algorithm that makes intensive computations along a one-dimensional strip of $(\hat{\theta}_1^{(1)}, \hat{\theta}_2^{(1)})$ values, rather than on a fully two-dimensional grid. The use of simulation means that this approach is highly flexible and may be applied just as easily with other forms of closed testing procedure or combination test, or with different definitions of the final gain function.

Our study of a wide range of examples supports clear conclusions about the benefits of adaptive enrichment designs. If investigators are willing to use either the FF (Fixed Full population) or FS (Fixed subpopulation) design, the additional benefits of an adaptive enrichment design are at best modest for the gain function we have considered. However, the FS design may not be a realistic option: there could be differing opinions about the likely treatment effect in the subpopulation $\mathcal{S}_2$ or, within the wider development program, there may be good reasons for wanting to learn about the new treatment's efficacy in the full population. Then, if the FS design is not an option, there are plausible prior distributions for $\boldsymbol{\theta}$ under which the AE is clearly superior to the FF design.

A positive feature of AE design that is not captured in our gain function is its selectivity. Suppose θ_1 is high but θ_2 is close to zero. If rejection of H_{03}: $\theta_3 \leq 0$ leads to the new treatment being made available to the full patient population, it would be given to patients in $\mathcal{S}_2$ for whom the control treatment is just as good. If $\theta_2 = 0$, the term $\theta_3\mathcal{R}_3$ in the gain function (8) is equal to $\lambda\theta_1\mathcal{R}_3$ and this neither rewards nor penalizes giving the new treatment to patients in $\mathcal{S}_2$. The results in Tables 2 and 3 show the AE design to have higher values of $P(\mathcal{R}_1)$ and lower values of $P(\mathcal{R}_3)$ compared to the FF design, indicating that when θ_2 is low the AE design is more likely to find a treatment effect only in $\mathcal{S}_1$.

Our results have illustrated a general weakness of adaptive designs that decisions about adaptation are based on interim data which provide only limited information about the true treatment effects. The results in Table 2 for the FS and FF designs show clear benefits to drawing patients from the most appropriate subgroups when the value of $\boldsymbol{\theta}$ is known. However, in the examples of Table 3 and the examples with higher prior variances in Table 4the AE designs must make enrichment decisions under highly variable posterior distributions of $\boldsymbol{\theta}$ at the interim analysis. A possible remedy to this problem in making the enrichment decision is to use additional information from other endpoints or biomarkers that can be assumed to respond in the same way as the primary endpoint to the treatments under investigation.

We have presented methods for a study in which there is just one subpopulation of special interest. These methods can be generalized to the design of trials with multiple subpopulations, possibly nested with the treatment effect increasing as the size of the subpopulation decreases. Then, given a multiple testing procedure that controls FWER, a suitably defined gain function and a prior distribution for the vector of treatment effects, our simulation-based approach may be used to find the optimal enrichment decision at an interim analysis. However, more computation will be needed to find the full optimal design as the dimensionality of the problem increases with the number of subpopulations.

The gain function (8) may be adapted to reflect the process of drug approval. Suppose, for example, H_{03}: $\theta_3 \leq 0$ is rejected on the strength of a large positive estimate of θ_1 and a much smaller estimate for θ_2. While a regulator may not require formal rejection of the null hypothesis H_{02}: $\theta_2 \leq 0$ at the 0.025 significance level, some minimum threshold for an estimate $\hat{\theta}_2$ may be required in order for the treatment to be approved for the full population, and for health care providers to agree to pay for this treatment. Such a requirement can be reflected in the gain function $G(\boldsymbol{\theta}, X)$, where the data in X includes estimates of θ_1 and θ_2. Rather than stipulate a particular gain function for all applications, we recommend that investigators determine the appropriate gain function for their specific trial, then our methods can be used to optimize over adaptive enrichment designs and to compare the resulting design with other, nonadaptive options.

ACKNOWLEDGEMENTS

The first author received financial support for this research from the UK Engineering and Physical Sciences Research Council and Hoffman-LaRoche Ltd. The authors would like to thank to Lucy Rowell for her contributions to this project.

ORCID

Thomas Burnett https://orcid.org/0000-0001-8912-2554
Christopher Jennison https://orcid.org/0000-0002-9812-1104

REFERENCES

1. Dmitrienko A, D'Agostino RB, Huque MF. Key multiplicity issues in clinical drug development. *Stat Med*. 2013;32:1079-1111.
2. Simes RJ. An improved Bonferroni procedure for multiple tests of significance. *Biometrika*. 1986;73:751-754.
3. Bauer P, Köhne K. Evaluation of experiments with adaptive interim analyses. *Biometrics*. 1994;50:1029-1041.
4. Lehmacher W, Wassmer G. Adaptive sample size calculations in group sequential trials. *Biometrics*. 1999;55:1286-1290.
5. Hartung J. A note on combining dependent tests of significance. *Biom J*. 1999;41:849-855.
6. Burnett T. *Bayesian Decision Making in Adaptive Clinical Trials* [PhD thesis]. University of Bath; 2017.
7. Brannath W, Zuber E, Branson M, et al. Confirmatory adaptive designs with Bayesian decision tools for a targeted therapy in oncology. *Stat Med*. 2009;28:1445-1463.
8. Götte H, Donica M, Mordenti G. Improving probabilities of correct interim decision in population enrichment designs. *J Biopharm Stat*. 2015;25:1020-1038.
9. Uozumi R, Hamada C. Utility-based interim decision rule planning in adaptive population selection designs with survival endpoints. *Stat Biopharm Res*. 2020;12:360-368.
10. Ondra T, Jobjörnsson S, Beckman RA, et al. Optimized adaptive enrichment designs. *Stat Methods Med Res*. 2019;28:2096-2111.
11. ICH, EMEA. *ICH E9: Statistical Principles for Clinical Trials*. London, UK: European Medicines Agency; 1998.
12. Bretz F, Schmidli H, König F, Racine A, Maurer W. Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: general concepts. *Biom J*. 2006;48:623-634.
13. 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;48:635-643.
14. Jennison C, Turnbull BW. Adaptive seamless designs: selection and prospective testing of hypotheses. *J Biopharm Stat*. 2007;17:1135-1161.
15. Marcus R, Peritz E, Gabriel KR. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*. 1976;63:655-660.
16. Sarkar SK, Chang CK. The Simes method for multiple hypothesis testing with positively dependent test statistics. *J Am Stat Assoc*. 1997;92:1601-1608.
17. Berger JO. *Statistical Decision Theory and Bayesian Analysis*. 2nd ed. Springer Science & Business Media: New York, NY; 2013.

SUPPORTING INFORMATION

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

How to cite this article: Burnett T, Jennison C. Adaptive enrichment trials: What are the benefits?. *Statistics in Medicine*. 2021;40:690–711. https://doi.org/10.1002/sim.8797

APPENDIX A. STRONG CONTROL OF FWER IMPLIES A CLOSED TESTING PROCEDURE

Suppose a multiple testing procedure $\mathcal{P}$ with n null hypotheses provides strong control of the FWER at level α. We shall show that $\mathcal{P}$ can be represented as a closed testing procedure C. Suppose the null hypotheses are stated in terms of a parameter vector $\boldsymbol{\theta}$, then strong control of the FWER implies that

$$P_{\boldsymbol{\theta}}(\text{Reject at least one true null hypothesis}) \leq \alpha \quad \text{for all} \;\; \boldsymbol{\theta}. \tag{A1}$$

Suppose the ith null hypothesis is H_{0i}: $\boldsymbol{\theta} \in A_i$. Denote the observed data by X and suppose $\mathcal{P}$ rejects H_{0i} if $X \in \xi_i$. We shall use the rejection regions ξ_i to define a closed testing procedure C which gives the same overall decisions as $\mathcal{P}$.

We first define level α tests of the individual hypotheses $H_{01}, \ldots, H_{0n}$. For each $i \in 1, \ldots, n$, the test of H_{0i} rejects its null hypothesis if and only if $X \in \xi_i$. To see that this gives a level α test of H_{0i}, suppose $\boldsymbol{\theta} \in A_i$, then

$$P_{\theta}(\text{Reject } H_{0i}) = P_{\theta}(X \in \xi_i) \leq P_{\theta}(\text{Reject at least one true null hypothesis}) \leq \alpha,$$

by applying (A1) with $\boldsymbol{\theta} \in A_i$.

Now consider an intersection hypothesis $H_I = \cap_{i \in I} H_{0i}$, where I is a subset of $\{1, \ldots, n\}$. Our level α test of H_I, rejects H_I if

$$X \in \cup_{i \in I}\ \xi_i.$$

To see this gives a level α test of H_I, suppose H_I is true, so $\theta \in \cap_{i \in I} A_i$, then

$$P_{\theta}(\text{Reject } H_I) = P_{\theta}(X \in \cup_{i \in I}\ \xi_i) \leq P_{\theta}(\text{Reject at least one true null hypothesis}) \leq \alpha,$$

by applying (A1) with $\boldsymbol{\theta} \in \cap_{i \in I} A_i$.

The closed testing procedure $\mathcal{C}$ is formed by combining the level α tests of individual and intersection hypotheses in the usual way. Thus, the null hypothesis H_{0i} is rejected overall if the level α tests reject H_{0i} and every H_I for which $i \in I$. It is easy to check that the procedure $\mathcal{C}$ rejects H_{0i} overall if and only if $X \in \xi_i$, and thus the two procedures $\mathcal{C}$ and $\mathcal{P}$ always reject exactly the same set of hypotheses.

Although the above construction is quite simple, we are not aware that this result has been noted previously. An implication in our application is that we lose no generality by restricting attention to methods based on closed testing procedures. Of course, the choice of closed testing procedure remains. In our case, it is natural to base the level α test of H_{01} on $\hat{\theta}_1^{(1)}$ and $\hat{\theta}_1^{(2)}$ and the level α test of H_{03} on $\hat{\theta}_3^{(1)}$ and $\hat{\theta}_3^{(2)}$, so we see it is the method of testing the intersection hypothesis $H_{01} \cap H_{03}$ that may merit further investigation.

APPENDIX B. DERIVATION OF THE OPTIMAL DECISION RULE

We illustrate the details of our computational method in an example where the decision rule being sought is that depicted in Figure 3A. In finding this rule we start by defining a region A in which $(\hat{\theta}_1^{(1)}, \hat{\theta}_2^{(1)})$ will lie with very high probability: in this example we have taken A to be the square $(0, 20) \times (-10, 10)$. We subdivide A into four smaller squares and find the optimal decision at each of the nine vertices of these squares, giving the results shown in Figure 3B. We proceed on the assumption that if a certain decision is optimal at $(\hat{\theta}_1^{(1)}, \hat{\theta}_2^{(1)}) = (a, b)$ and $(\hat{\theta}_1^{(1)}, \hat{\theta}_2^{(1)}) = (a, c)$, where $b < c$, then the same decision is optimal at all points $(\hat{\theta}_1^{(1)}, \hat{\theta}_2^{(1)}) = (a, d)$ with $b < d < c$; similarly if a decision is optimal at $(\hat{\theta}_1^{(1)}, \hat{\theta}_2^{(1)}) = (a, b)$ and $(\hat{\theta}_1^{(1)}, \hat{\theta}_2^{(1)}) = (c, b)$, where $a < c$, we assume this decision is also optimal at $(\hat{\theta}_1^{(1)}, \hat{\theta}_2^{(1)}) = (d, b)$ for all $a < d < c$. Applying this assumption in our example, we see that it is optimal to enrich for all values $(\hat{\theta}_1^{(1)}, \hat{\theta}_2^{(1)})$ in the top right-hand square, so we record this conclusion and make no further calculations for points in this square. The other three squares need further work: we subdivide each of these into four smaller squares and find the optimal decision at each new vertex. The results of these steps are presented in Figure 3C.

We continue the search iteratively, halving the size of the smallest squares at each step. In the next iteration for our example, we note that five of the 12 small squares in Figure 3C have the same optimal decision at all four vertices and we allocate this decision to the whole square. We subdivide the other seven squares and compute optimal decisions at the new vertices. The information after this step is depicted in Figure 3D. Repeating the same steps in the next iteration produces the results shown in Figure 4A.

If our target is to specify optimal decisions on a 16×16, this is the final iteration. To complete the process, we find the optimal decision associated with each of the smallest squares: if the optimal decision is the same at all four vertices this decision is assigned to the square; if not, we find the optimal decision at the square's center point and define this to be the decision for the whole square. Figure 4B shows the results of this last step, while Figure 4C presents the same set of conclusions using the full 16×16 grid.

Analysing this algorithm in the most demanding case when the decision boundary is at an angle of 45°, we find the optimal decision has to be computed at about $14n$ points in order to determine optimal decisions on an array of $n \times n$ small squares. A key point here is that the amount of computation is of order n, even though there are n^2 small squares at the finest level. Since we need to conduct a large number of simulations in finding the optimal decision for each value

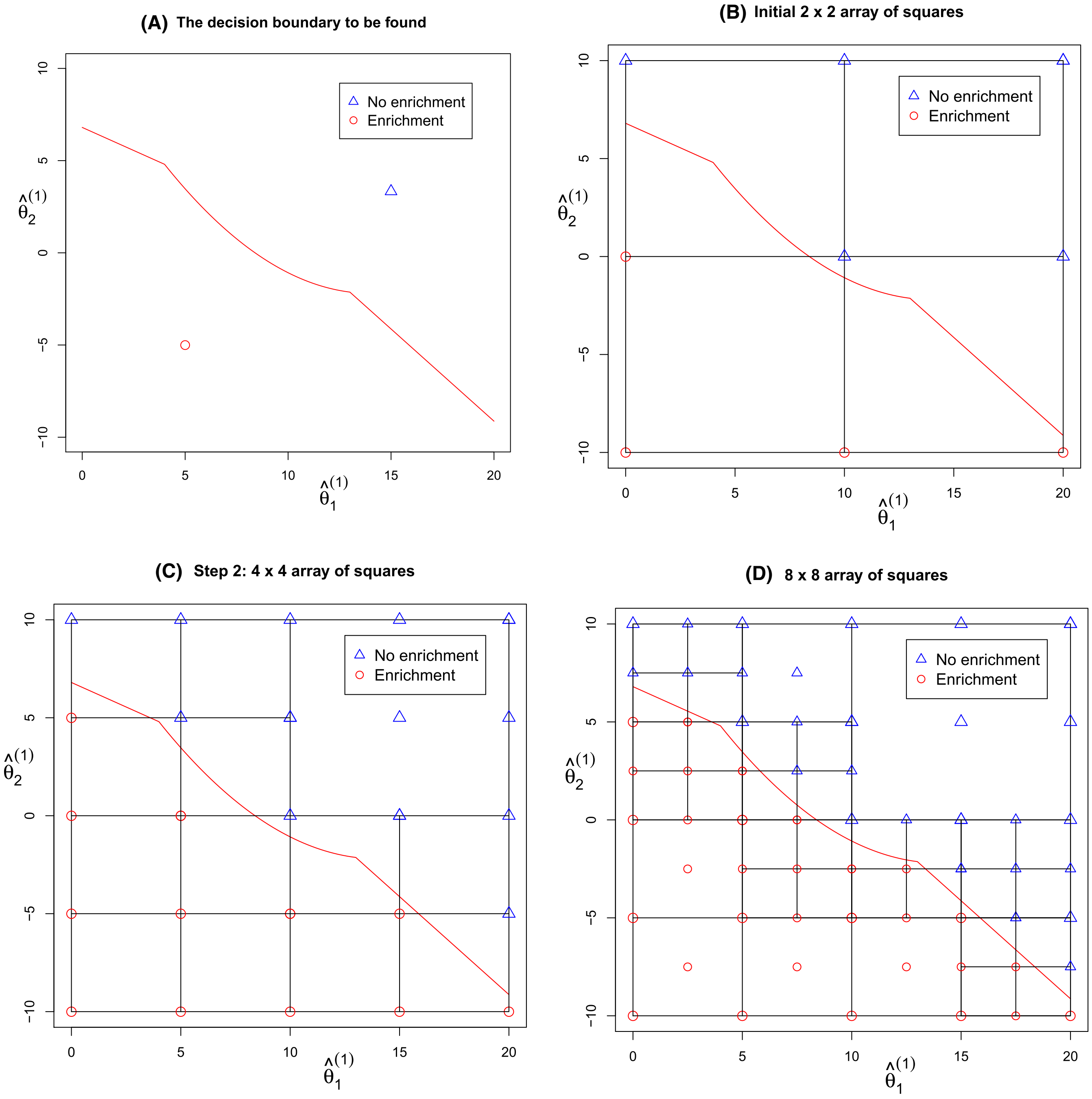


FIGURE 3 Computation of an optimal decision rule [Colour figure can be viewed at wileyonlinelibrary.com]

of $(\hat{\theta}_1^{(1)}, \hat{\theta}_2^{(1)})$, the computational load can still be high—but it is feasible. In our examples we found optimal decisions on a $2^8 \times 2^8$ or $2^9 \times 2^9$ array, using samples of size 10^5 or 10^6 from the posterior distribution of $\boldsymbol{\theta}$ in finding the optimal decision at each $x_1 = (\hat{\theta}_1^{(1)}, \hat{\theta}_2^{(1)})$.

In the examples we have studied, it has usually been clear from the results that the optimal decision function has the assumed monotonicity property. However, it is possible for this assumption to fail. In that case, the decision boundary may cross one edge of a square twice, then having the same optimal decision at all four vertices of that square does not necessarily mean this decision is optimal throughout the square. In a more conservative version of our algorithm, which guards against this eventuality, we require the same decision to be optimal at all 16 vertices of a 3×3 grid of squares before concluding this decision to be optimal over the whole of the central square. The additional computations needed when

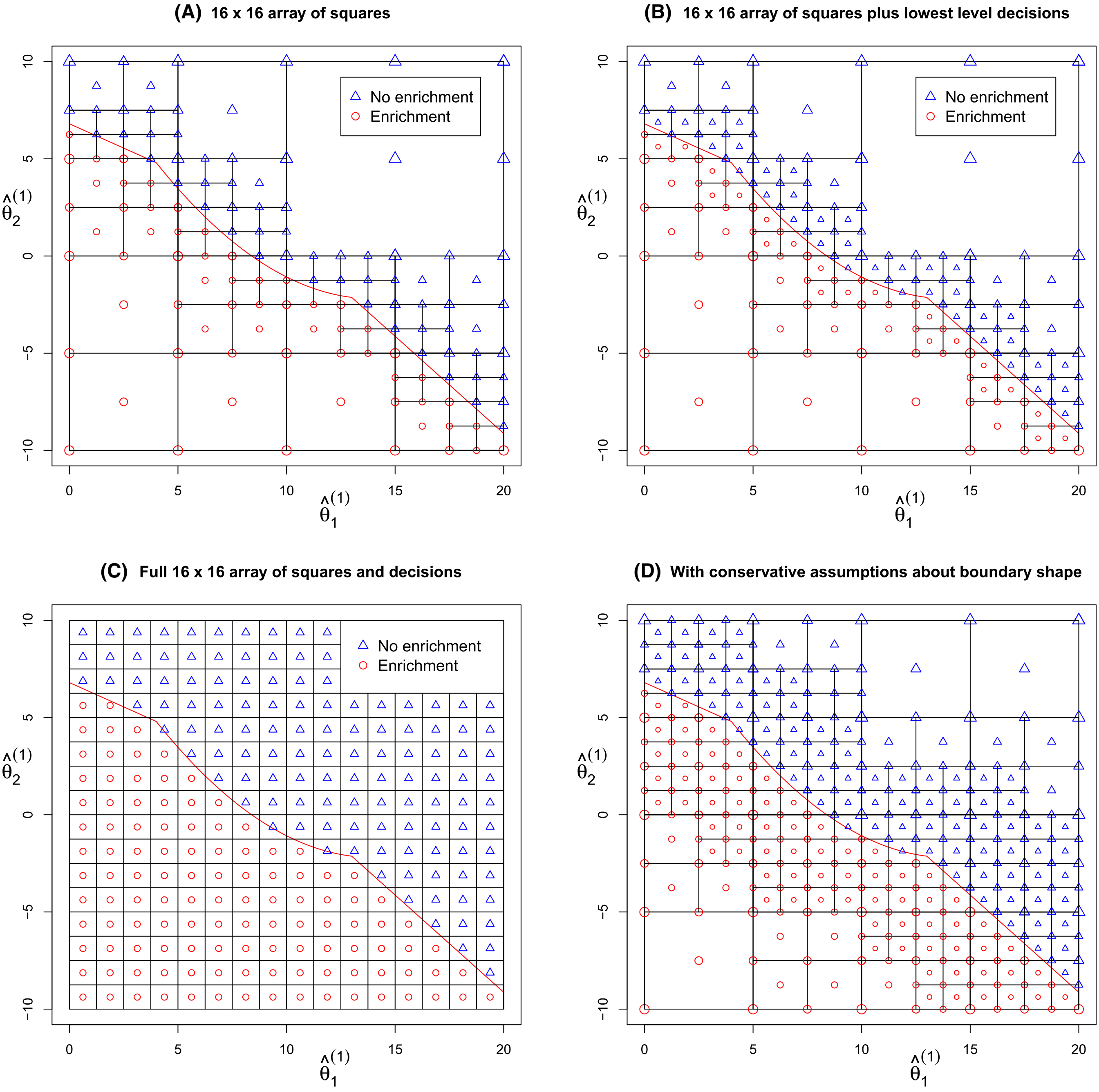


FIGURE 4 Computation of an optimal decision rule [Colour figure can be viewed at wileyonlinelibrary.com]

this approach is followed in our example are illustrated in Figure 4D. In general, this conservative approach requires approximately twice the total computation time.