Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2019, Vol. 28(7) 2096–2111

DOI: 10.1177/0962280217747312
journals.sagepub.com/home/smm
SAGE

Optimized adaptive enrichment designs

Thomas Ondra,[1] Sebastian Jobjörnsson,[2] Robert A Beckman,[3] Carl-Fredrik Burman,[2,4] Franz König,[1] Nigel Stallard[5] and Martin Posch[1]

Abstract

Based on a Bayesian decision theoretic approach, we optimize frequentist single- and adaptive two-stage trial designs for the development of targeted therapies, where in addition to an overall population, a pre-defined subgroup is investigated. In such settings, the losses and gains of decisions can be quantified by utility functions that account for the preferences of different stakeholders. In particular, we optimize expected utilities from the perspectives both of a commercial sponsor, maximizing the net present value, and also of the society, maximizing cost-adjusted expected health benefits of a new treatment for a specific population. We consider single-stage and adaptive two-stage designs with partial enrichment, where the proportion of patients recruited from the subgroup is a design parameter. For the adaptive designs, we use a dynamic programming approach to derive optimal adaptation rules. The proposed designs are compared to trials which are non-enriched (i.e. the proportion of patients in the subgroup corresponds to the prevalence in the underlying population). We show that partial enrichment designs can substantially improve the expected utilities. Furthermore, adaptive partial enrichment designs are more robust than single-stage designs and retain high expected utilities even if the expected utilities are evaluated under a different prior than the one used in the optimization. In addition, we find that trials optimized for the sponsor utility function have smaller sample sizes compared to trials optimized under the societal view and may include the overall population (with patients from the complement of the subgroup) even if there is substantial evidence that the therapy is only effective in the subgroup.

Keywords

Adaptive design, optimal design, enrichment design, precision medicine, subgroup analysis

1 Introduction

A major challenge in the development of targeted therapies is the identification and confirmation of subgroups of patients where a treatment is effective. To address this issue, a range of clinical trial designs and statistical methodology have been proposed.[1–4] An important field of application is oncology, where the better understanding of the molecular basis of the disease leads the development of therapies that directly act on specific molecular mechanisms and therefore may be effective in special subgroups of patients only. In this work, we consider a setting, where there is *a priori* biological plausibility that the treatment effect is larger or only present in a subgroup defined by a binary biomarker. However, there is still uncertainty if the treatment is effective at all and if so, if the treatment effect is larger or only present in a subpopulation of biomarker positive patients. To address this uncertainty, clinical trials testing for a treatment effect in the full population and a subgroup can be performed.

While in standard parallel group clinical trials the statistical power to demonstrate a treatment effect in a dedicated primary endpoint is typically the basis for the planning of clinical trials, the consideration of power

[1]Center for Medical Statistics, Informatics, and Intelligent Systems, Medical University of Vienna, Vienna, Austria
[2]Department of Mathematics, Chalmers University, Gothenburg, Sweden
[3]Departments of Oncology and of Biostatistics, Bioinformatics, and Biomathematics, Georgetown University Medical Center, Washington, DC, USA
[4]Statistical Innovation, AstraZeneca R&D, Molndal, Sweden
[5]Warwick Medical School, The University of Warwick, Coventry, UK

Corresponding author:
Martin Posch, Center for Medical Statistics, Informatics, and Intelligent Systems, Medical University of Vienna, Spitalgasse 23, 1090 Vienna, Austria.
Email: martin.posch@meduniwien.ac.at

alone does not sufficiently represent the losses and gains of correct and incorrect test decisions when several subgroups are tested. Subgroup analyses are challenging because several types of risks are associated with inference on subgroups. On the one hand, by disregarding a relevant subpopulation a treatment option may be missed due to a dilution of the treatment effect in the full population. In addition, even if the diluted treatment effect can be demonstrated in an overall population, it is not ethical to treat patients that do not benefit from the treatment when they can be identified in advance.[5,6] On the other hand, selecting a spurious subpopulation increases the risk to restrict an efficacious treatment to too narrow a fraction of a potential benefiting population. In order to account for these risks, we apply a decision theoretic framework and define utility functions that quantify the expected benefits as well as the costs of a particular clinical trial.

Decision theoretic approaches are based on utility functions that map actions to a numeric scale representing the values of these actions. Optimal actions are then defined as the actions that maximize these utility functions. In the application to clinical trials the set of actions are families of trial designs, specified by sample sizes, allocation ratios, stopping and adaptation rules (for adaptive designs), as well as inference procedures. The utilities represent the value of trial outcomes (as e.g. the rejection of a null hypothesis or the size of observed treatment effects), adjusted for the cost of the trial and may also depend on the true treatment effects (see Hee et al.[7] for a recent review on the application of decision theoretic approaches to guide clinical trial design). When applying the approach to optimize enrichment designs, the utilities also account for the size of the population for which an efficacy claim is made.[8–12] Because the outcome of the trial as well as the true treatment effects are unknown at the time of planning, the optimization relies on expected utilities. Here, the expectation is taken over a prior distribution on the effect sizes as well as the distribution of trial outcomes given the effect sizes. The optimal design based on the decision theoretic approach is then defined as the design with the largest expected utility.

We derive optimized clinical trial designs in the setting of a parallel group trial comparing the means of a normally distributed outcome and consider utility functions from the perspective of different stakeholders: a sponsor's as well as a societal view. We assume that the utility of the sponsor is the net present value (NPV), while for the societal perspective it is the expected health benefit (adjusted for the cost of the trial).

We especially focus on single-stage and adaptive *partial enrichment* designs. Partial enrichment designs are designs, where the prevalence of the subgroup in the trial is a design parameter and may differ from the prevalence in the underlying population.[8,13] Therefore, we can choose to make the subpopulation over- or underrepresented. Adaptive designs,[14–17] on the other hand, are two-stage designs, where in a first stage patients are recruited from the full population. Then, in an interim analysis, based on interim data, the trial design of the second stage may be modified. For example recruitment may be limited to patients in a subgroup of biomarker positive patients and/or the sample size may be adapted. In the proposed adaptive partial enrichment designs, in addition the prevalence of the subgroup in the second stage sample can be chosen adaptively, based on the first stage data. The adaptations may be based on all information observed at the interim analysis, including information on secondary and surrogate endpoints and safety information.

Using numerical optimization and a dynamic programming approach, we determine optimal single and two-stage designs optimizing the total sample size of the trial, the prevalence of the subgroup in the trial, and, for the adaptive designs, the optimal adaptation rule. The adaptation rule is a function of the first stage data that determines the population selected for the second stage as well as the second stage sample size in the overall population and the sample size in the subgroup, which may imply different subgroup prevalences in the first and second stage. An adaptation option is also stopping for futility, which corresponds to a second stage sample size of zero.

In this manuscript, we extend earlier work on decision theoretic approaches to optimize single-stage designs[12] and optimal adaptive enrichment designs.[11] We derive optimal single- and adaptive two-stage partial enrichment designs with optimal adaptation rules that go beyond subgroup selection and allow one to choose optimal second stage sample sizes in the full population and the subgroup conditional on the first stage data using a backwards induction algorithm. The use of Bayesian decision theoretic methods has also been proposed to optimize clinical development programs and in models that account for errors in the determination of the patient's biomarker status.[8,18,19] An alternative line of research optimizes multiple testing procedures in one and two-stage designs based on a decision theoretic approach.[9,10,12,20]

The remainder of the paper is structured as follows: In Section 2, the considered single-stage and adaptive enrichment designs are introduced. In Section 3, the utility functions are discussed and in Section 4, we derive the optimized trial designs. We present the results of a case study in Section 5 and conclude with a discussion in Section 6.

2 Testing scenario and trial designs

Consider a parallel group trial comparing the means of a normally distributed endpoint in a population F that is divided into a subgroup S of biomarker positive patients and its complement S', the biomarker negative patients. Let $\delta_i, i \in \{S, S'\}$ denote the treatment effects in the subgroups and $\delta_F = \lambda\delta_S + (1-\lambda)\delta_{S'}$ the overall treatment effect, where λ denotes the prevalence of the subgroup in the considered patient population, which is assumed to be known. We consider a setting, where there is a biological rationale that the treatment effect might be higher (or only present) in S compared to S' and consider one-sided tests of the null hypotheses $H_F : \delta_F \leq 0$ and $H_S : \delta_S \leq 0$.

Next, we define single-stage and adaptive two-stage designs to test H_F and H_S controlling the family-wise error rate (FWER) at a pre-specified one-sided level α. In Sections 2.1 and 2.2, these tests are defined for given sample sizes in the subgroups (for single stage designs) or given first stage sample sizes and second stage sample sizes functions (for adaptive designs). In Section 4, we optimize these sample sizes and sample size functions.

2.1 Single-stage designs

We consider single-stage designs with partial and with full enrichment. Partial enrichment designs include biomarker positive and biomarker negative patients, while full enrichment designs include biomarker positive patients only.

2.1.1 Single-stage designs with partial enrichment

A single-stage partial enrichment design is a clinical trial with $n_S > 0$ biomarker positive and $n_{S'} > 0$ biomarker negative patients per treatment arm. Thus, the *trial prevalence* of the biomarker positive subgroup is given by $\gamma = n_S/n$, where $n = n_S + n_{S'}$ denotes the overall sample size per treatment arm. We define Z-test statistics to compare the outcomes in subgroup S, its complement S' and the full population F

$$Z_S = \hat{\delta}_S/\sqrt{v_S}, \quad Z_{S'} = \hat{\delta}_{S'}/\sqrt{v_{S'}}, \quad Z_F = \hat{\delta}_F/\sqrt{v_F} \tag{1}$$

where $\hat{\delta}_i$ denotes the estimated mean treatment effects and $\sqrt{v_i}$ their standard error which is assumed to be known. The treatment effect estimates in the subgroups $\hat{\delta}_i, i = S, S'$ are the mean differences of the outcomes in S and S'. In the full population, the treatment effect estimate is given by $\hat{\delta}_F = \lambda\hat{\delta}_S + (1-\lambda)\hat{\delta}_{S'}$. Note that $\hat{\delta}_F$ is a weighted average of the treatment effect estimates in S and S', where the weights are given by the population prevalence λ and not the trial prevalence γ.[13] Therefore, it gives an unbiased estimate of the overall treatment effect in the underlying population. Note that Z_F can be rewritten as a weighted sum of the test statistics in S and S'

$$Z_F = \xi\left(\frac{\lambda}{\sqrt{\gamma}}Z_S + \frac{1-\lambda}{\sqrt{1-\gamma}}Z_{S'}\right) \tag{2}$$

with $\xi = 1/\sqrt{\lambda^2/\gamma + (1-\lambda)^2/(1-\gamma)}$ (see online supplementary material, Section S1). The standard errors are given by $v_i = 2\sigma^2/n_i, i = S, S'$ and $v_F = 2\sigma^2/(n\xi^2)$, where σ^2 denotes the variance of the observations in the treatment and control group. For simplicity, this variance is assumed to be homogeneous across the subpopulations and treatment groups. For the generalization to unequal variances, the definitions of $v_i, i \in \{S, S', F\}$ and the weights in equation (2) have to be adjusted accordingly.

Note that we use the re-weighted test statistics Z_F because the standard Z-statistics computed from the pooled sample in the full population may have an expectation different from zero even if $\delta_F = 0$. This occurs if the means in the two subpopulations have different signs and the effects cancel out in the underlying patient population but not in the trial population. Furthermore, by the definition of Z_F, the vector of test statistics (Z_S, Z_F) follows a multivariate normal distribution with means $\delta_S/\sqrt{v_S}, \delta_F/\sqrt{v_F}$, variances 1 and covariance $\xi\lambda/\sqrt{\gamma}$.

To adjust for multiplicity for the test of the two hypotheses H_F and H_S, we apply a Bonferroni correction. While more powerful testing procedures could be used, we chose the conservative Bonferroni test to limit the computational complexity (especially in the adaptive setting below this allows us to utilize numerical integration rather than to have to rely on simulations). In addition, to avoid test decisions where H_F is rejected but the rejection is driven by a strong effect in a single subpopulation only, we additionally require that a sufficient positive Z-test statistic is observed in each subpopulation to reject H_F (called consistency criterion[6,12,21]). The decision function of the resulting multiple test for H_F and H_S, whose components take the value 1 if the

corresponding hypotheses are rejected and zero otherwise, is then given by

$$\psi^e = (\psi^e_S, \psi^e_F) = \left(\mathbf{1}_{\{Z_S \geq b_{\alpha/2}\}}, \mathbf{1}_{\{Z_F \geq b_{\alpha/2}, Z_S \geq b_\eta, Z_{S'} \geq b_\eta\}}\right) \tag{3}$$

where $\eta \geq \alpha/2$ is a pre-chosen consistency threshold, b_q denote the $1-q$ quantiles of the standard normal distribution, and $\mathbf{1}_{\{\cdot\}}$ the indicator function. Thus, H_F is only rejected if there is a significant treatment effect in F at the Bonferroni adjusted level $\alpha/2$ and, in addition, in both subgroups a significant treatment effect at level η is observed. Given this decision function, a partially enriched single-stage design $d = (n_S, n_{S'})$ is fully specified by the sample sizes in the subgroup and its complement. We denote the family of single-stage enrichment designs by $\mathcal{D}^e$.

2.1.2 Fully enriched single-stage designs

In the single-stage full enrichment design, patients from the full population are screened but only biomarker positive patients are included in the trial (i.e. $n = n_S$). Only hypothesis H_S is tested and the decision function of the respective test is given by

$$\psi^f = (\psi^f_S, \psi^f_F) = (\mathbf{1}_{\{Z_S \geq b_\alpha\}}, 0) \tag{4}$$

where Z_S is defined as in equation (1). Note that H_S can be tested at full level α since only one hypothesis is under investigation (because ψ^f_F is set to zero). A single-stage fully enriched design d is specified by the sample size n_S and we define $d = n_S$ and denote the family of single-stage full enrichment designs by $\mathcal{D}^f$.

2.2 Adaptive two-stage designs

For an adaptive two-stage design, let $n^{(k)}_S, n^{(k)}_{S'}$ denote the number of patients per treatment arm recruited in the subgroup and its complement in stage $k = 1, 2$. The corresponding total per treatment arm sample size in stage k is denoted by $n^{(k)} = n^{(k)}_S + n^{(k)}_{S'}$. The trial prevalence of the biomarker positive subgroup in stage k is given by $\gamma^{(k)} = n^{(k)}_S / n^{(k)}$. We assume that in the first stage patients from S and S' are recruited (i.e. $n^{(1)}_S > 0, n^{(1)}_{S'} > 0$). In the interim analysis, the trial may be either stopped for futility, continued only in S or continued in the full population. The second stage sample sizes can be chosen based on the interim data. In accordance with equations (1) and (2), we define the stage k test statistics $Z^{(k)}_S, Z^{(k)}_{S'}$ and $Z^{(k)}_F$, where all variables in equations (1) and (2) are replaced by the corresponding stage k quantities denoted by the superscript $^{(k)}$. Note that the second stage statistics are not cumulative test statistics but computed from the second stage data only. We define an adaptive two-stage test that combines the test statistics of the first and second stage with the inverse normal combination function, with equal weights for the two-stages. The combined test statistics are then given by $Z^{(c)}_i = \sqrt{w^{(1)}} Z^{(1)}_i + \sqrt{w^{(2)}} Z^{(2)}_i, i = S, S', F$, where $w^{(j)}, j = 1, 2$ are pre-defined, non-negative weights such that $w^{(1)} + w^{(2)} = 1$. Note that $Z^{(c)}_i$ follows a standard normal distribution under the respective null hypothesis, even if the second stage sample size is chosen in a data dependent way, see literature.[22–24] This comes at the cost, that the stage-wise test statistics are combined with weights $w^{(j)}, j = 1, 2$ that need to be pre-defined and are not adjusted to the actual stage-wise sample sizes. To adjust for multiple testing, we apply the Bonferroni correction and define the adaptive multiple test for H_F and H_S setting its decision function, denoted by $\psi^a = (\psi^a_S, \psi^a_F)$, equal to equation (3) replacing the single-stage test statistics Z_i by the combination test statistics $Z^{(c)}_i, i = S, S', F$. If no patients in S' are recruited in the second stage such that $n^{(2)}_{S'} = 0$, we set $\psi^a_F = 0$ and $\psi^a_S = \mathbf{1}_{\{Z^{(c)}_S \geq b_{\alpha/2}\}}$. If the trial is stopped for futility and the second stage sample sizes are zero, we set $\psi^a_F = \psi^a_S = 0$.

We assume that the second stage sample sizes $n^{(2)}_i, i = S, S'$ are adaptively chosen and can be written as function of the vector of first stage test statistics $Z^{(1)} = (Z^{(1)}_S, Z^{(1)}_{S'})$ (for notational convenience we drop the argument of the functions $n^{(2)}_i$). Given the decision function ψ^a, an adaptive two-stage enrichment design d is fully specified by the first stage sample sizes $n^{(1)}_i$ and the second stage sample size functions $n^{(2)}_i$, $i = S, S'$ and we define $d = (n^{(1)}_S, n^{(1)}_{S'}, n^{(2)}_S, n^{(2)}_{S'})$, where the first two elements are numbers and the second two components functions from $\mathbb{R}^2 \to \mathbb{N}$. Let $\mathcal{D}^a$ denote the family of adaptive two-stage enrichment designs.

3 Utility functions

Let $\mathcal{D} = \mathcal{D}^e \cup \mathcal{D}^f \cup \mathcal{D}^a$ denote the family of considered one- and two-stage trial designs. As in Ondra et al.,[12] we define two types of utility functions for the trial designs $d \in \mathcal{D}$ representing either a societal perspective or a commercial sponsor perspective. The societal utility function models the total public health benefit, whereas the

sponsors utility function models the total revenue of the sponsor. Let $\delta = (\delta_S, \delta_F)$ denote the true treatment effects in the subgroup and the full population. Then the utility from a societal perspective is modelled as

$$U_{\text{so}}(d) = -C(d) + \begin{cases} rN(\delta_F - \mu_F) & \text{if } \psi_F = 1 \\ r\lambda N(\delta_S - \mu_S) & \text{if } \psi_S = 1, \psi_F = 0 \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where N denotes the expected number of future patients, r is a reward parameter, μ_S, μ_F are lower thresholds for the effect size that represent, for example the minimal clinically relevant effect size that outweighs the treatment cost and/or known side effects, $C(d)$ denotes the costs of the trial with design d defined in equation (7), and ψ_S, ψ_F denote the test procedure of design d. Note that the utility function also depends on the trial data through the test decisions ψ_S and ψ_F that are functions of the respective Z-statistics. Given that the hypothesis test rejects the null hypothesis of no treatment effect in a certain population, the utility increases linearly with the true treatment effect and the population size. If the true treatment effect is smaller than the respective threshold, but the hypothesis test rejects, the utility takes negative values.

From a sponsor perspective, we define a utility that models a setting where the price of the drug depends on the effect size estimate from the pivotal trial as well as the size of the market and set

$$U_{\text{sp}}(d) = -C(d) + \begin{cases} rN(\hat{\delta}_F - \mu_F)^+ & \text{if } \psi_F = 1 \\ r\lambda N(\hat{\delta}_S - \mu_S)^+ & \text{if } \psi_S = 1, \psi_F = 0 \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

where $(\cdot)^+$ denotes the positive part and, compared to the utility from a societal perspective, the effect sizes δ_F as well as δ_S have been replaced by their respective point estimates $\hat{\delta}_S, \hat{\delta}_F$ as defined in Section 2.1 for single-stage designs. For adaptive designs, the point estimates of the treatment effect in S and F are computed based on the subgroup estimates of the pooled data from the two stages. Note that the sponsor utility does not depend on the true treatment effects. However, it depends on the trial data through the treatment effect estimates and, as above, through the test decisions. Note that, in contrast to the societal utility function, the sponsor's utility depends on the positive part of $(\hat{\delta}_i - \mu_i), i \in \{F, S\}$ rather than the difference $\delta_i - \mu_i$. If the sponsor observes a treatment effect lower than the clinically relevant threshold, we assume that payors and/or patients will not be willing to pay for the treatment and that it will not be marketed, leading to zero gain. In this case the sponsor's utility equals $-C(d)$. For the societal utility function in contrast, false decisions (where a treatment is licensed in a population where the true treatment effect is smaller than the threshold) give a negative contribution to the utility, in addition to the trial costs.

For both utility functions, the trial costs $C(d)$ are modelled as

$$C(d) = c_{\text{setup}} + c_{\text{biomarker}} + 2c_{\text{per-patient}}(n^{(1)} + n^{(2)}) + C_{\text{screening}}(d) \tag{7}$$

where for single-stage designs we set $n^{(2)} = 0$ and $n^{(1)} = n$. The costs are composed of trial setup costs c_{setup}, the biomarker development costs $c_{\text{biomarker}}$, screening costs to identify biomarker positive patients and costs per patient in the trial. The screening costs are proportional to the number of patients that have to be screened to recruit $n_S^{(k)}$ biomarker positive patients in stage $k = 1, 2$ and depend on the population prevalence (the lower the prevalence the larger the number of patients that need to be screened) and the sample sizes in the trial and are given by (see online supplementary material, Section S2).

$$C_{\text{screening}}(d) = 2c_{\text{screening}}\left(n^{(1)} \max\left(\frac{\gamma^{(1)}}{\lambda}, \frac{1-\gamma^{(1)}}{1-\lambda}\right) + n^{(2)} \max\left(\frac{\gamma^{(2)}}{\lambda}, \frac{1-\gamma^{(2)}}{1-\lambda}\right)\right) \tag{8}$$

4 Optimizing trial designs

The utilities defined in equations (5) and (6) depend on the trial outcomes and are therefore unknown at the planning stage of the trial. The societal utility is even unknown after the trial is performed because it depends on the unknown true treatment effects. However, given a prior distribution on the effect sizes in the two subgroups, we can compute expected utilities and determine optimal trial designs that maximize these expected utilities.

4.1 Expected utilities

The expected utility for a trial design $d \in \mathcal{D}$ is given by

$$V_{\pi_0,x}(d) := E_{\pi_0}[E_\delta[U_x(d)]] \tag{9}$$

where $x = \text{so}$ for the societal and $x = \text{sp}$ for the sponsor utility function, and the expectation is taken both over a prior distribution $\pi_0(\delta)$ on the effect sizes $\delta = (\delta_S, \delta_{S'})$ and the sampling distribution of the data given the effect sizes δ.

The expected utility for partially enriched single-stage designs is given by

$$V_{\pi_0,x}(d) = \iint U_x(d) f_{\delta,d}(z) \mathrm{d}z \pi_0(\delta) \mathrm{d}\delta \tag{10}$$

where $f_{\delta,d}$ denotes the joint density of the Z-statistics $Z_S, Z_{S'}$ given the effect sizes $\delta = (\delta_S, \delta_{S'})$ and design $d = (n_S, n_{S'})$. Since $Z_S, Z_{S'}$ are independent, $f_{\delta,d}$ is the product of two univariate normal densities with means $\delta_S/\sqrt{v_S}, \delta_{S'}/\sqrt{v_{S'}}$, and variance 1. For the full enrichment design, the computation of the expected utility reduces to an integral over the marginal prior on δ_S and the marginal sampling distribution of Z_S given δ_S (see online supplementary material, Section S5).

The expected utility of an adaptive enrichment design $d \in \mathcal{D}^a$ is given by an integral over the joint sampling distribution of the stage-wise test statistics $Z^{(k)} = (Z_S^{(k)}, Z_{S'}^{(k)})$ of the two-stages $k = 1, 2$ as well as the prior distribution on δ. Let $f^{(1)}_{\delta,d^{(1)}}$ denote the joint density of the first stage test statistics and $f^{(2)}_{\delta,d^{(2)}(z^{(1)})}$ the conditional joint density of the second stage test statistics conditional on $Z^{(1)} = z^{(1)}$, where $d^{(1)} = (n_S^{(1)}, n_{S'}^{(1)})$ denotes the sample sizes in the first stage and $d^{(2)}(z^{(1)}) = (n_S^{(2)}, n_{S'}^{(2)})$ the sample sizes in the second stage. Note that $n_i^{(2)}$, $i = S, S'$ are functions of the vector of first stage test statistics $Z^{(1)}$. The joint sampling distribution of the stage-wise test statistics of the two stages is then given by the product $f^{(1)}_{\delta,d^{(1)}} f^{(2)}_{\delta,d^{(2)}(z^{(1)})}$, where each factor is the product of two univariate normal densities with means $\delta_S/\sqrt{v_S^{(k)}}, \delta_{S'}/\sqrt{v_{S'}^{(k)}}$ and variance 1, where $\sqrt{v_i^{(k)}}, i = S, S'$ and $k = 1, 2$ denote the standard errors of the stage-wise treatment effect estimates. Note that $v_i^{(2)}$ depend on the second stage sample sizes and therefore are functions of the interim test statistics $Z^{(1)}$. Then, the expected utility for an adaptive enrichment design $d = (d^{(1)}, d^{(2)}(z^{(1)}))$ is given by

$$V_{\pi_0,x}(d) = \iiint U_x(d) f^{(2)}_{\delta,d^{(2)}(z^{(1)})}(z^{(2)}) \mathrm{d}z^{(2)} f^{(1)}_{\delta,d^{(1)}}(z^{(1)}) \mathrm{d}z^{(1)} \pi_0(\delta) \mathrm{d}\delta \tag{11}$$

where the inner two integrals integrate over the sampling distribution of the stages and the outer integral over the prior. For first stage outcomes $Z^{(1)}$ where the adapted second stage sample size in both subgroups is zero (stopping for futility) $v_S^{(2)}$ and $v_{S'}^{(2)}$ are not defined. However, for these outcomes $\psi_F^a = \psi_S^a = 0$ and the utility no longer depends on the second stage test statistics such that we can arbitrarily set $v_S^{(2)} = v_{S'}^{(2)} = 1$ in the integral above. Similarly, if only the second stage sample size in S' is zero then $\psi_F^a = 0$ and we can set $v_{S'}^{(2)} = 1$.

4.2 Determining the optimal design

We consider the optimization problem $\max_d V_{\pi_0,x}(d)$ for single-stage designs ($d \in \mathcal{D}^e \cup \mathcal{D}^f$) and for adaptive designs ($d \in \mathcal{D}^a$). For single-stage partial enrichment designs, the design d is fully specified by the sample sizes n_S and $n_{S'}$. Thus, to find the optimal design we numerically maximize equation (10) in the sample sizes $n_S, n_{S'}$. Similarly, for the full enrichment designs we optimize the corresponding expected utility function in n_S. The optimal single-stage design is then given by the optimal partial enrichment or the optimal full enrichment design, whichever gives the higher expected utility.

To determine optimal adaptive enrichment designs d as defined in Section 2.2, we use a dynamic programming approach. We first rewrite the objective function (10) to (see online supplementary material, Section S3)

$$V_{\pi_0,x}(d) = \iint W_{\pi_1,x}(d^{(1)}, d^{(2)}(z^{(1)}), z^{(1)}) f^{(1)}_{\delta,d^{(1)}}(z^{(1)}) \pi_0(\delta) \mathrm{d}z^{(1)} \mathrm{d}\delta \tag{12}$$

where

$$W_{\pi_1,x}(d^{(1)}, d^{(2)}(z^{(1)}), z^{(1)}) = \iint U_x(d) f^{(2)}_{\delta, d^{(2)}(z^{(1)})}(z^{(2)}) \pi_1(\delta|z^{(1)}) \mathrm{d}z^{(2)} \mathrm{d}\delta$$

and $\pi_1(\delta|z^{(1)})$ denotes the posterior distribution of the effect sizes given the first stage data. $W_{\pi_1,x}$ is the conditional expected utility, given the first stage test statistics $Z^{(1)}$ if design d is used. For a specific $z^{(1)}$ it depends only on the value of $d^{(2)}$ evaluated at $z^{(1)}$ but not the entire function $d^{(2)}$.

Given first stage sample sizes $d^{(1)} = (n_S^{(1)}, n_{S'}^{(1)})$ and first stage test statistics $Z^{(1)} = z^{(1)}$ the optimal second stage sample sizes $d^{*(2)}(z^{(1)}) = (n_S^{*(2)}, n_{S'}^{*(2)})$, which maximize the conditional expected utility, are given by

$$d^{*(2)}(z^{(1)}) := \mathrm{argmax}_{(m_S, m_{S'}) \in N^2} W_{\pi_1,x}\left(d^{(1)}, (m_S, m_{S'}), z^{(1)}\right) \tag{13}$$

The optimal first stage sample sizes are then given by

$$d^{*(1)} := \mathrm{argmax}_{(m_S, m_{S'}) \in N^2} V_{\pi_0,x}\left((m_S, m_{S'}), d^{*(2)}(z^{(1)})\right), \tag{14}$$

where the functions $d^{*(2)}(z^{(1)})$ are defined in equation (13). The optimal adaptive enrichment design is then given by $d^* = \left(d^{*(1)}, d^{*(2)}(z^{(1)})\right)$, where the second component is a function of the first stage test statistics. By the dynamic programming principle, the solutions (13) and (14) maximize equation (12) and thus also equation (11). Note that optimization can be performed under constraints on minimal and maximal sample sizes, by maximizing the utilities over respective restricted sets of sample sizes.

If the optimal single-stage or adaptive trial leads to a non-positive utility, the optimal option is to perform no trial and to retain both null hypotheses. This leads to an expected utility of zero.

5 Examples

We derive optimized trial designs for a range of scenarios to investigate the dependence of the optimum designs on the prior, the type of utility function (societal or sponsor) and the cost of the biomarker development and determination. We compare optimized adaptive enrichment designs with single-stage designs for both the weak and the strong biomarker prior for a grid of population subgroup prevalences from 10% to 90% in steps of 10%.

To explore the gain in expected utility by allowing the subgroup prevalence in the trial to differ from the population subgroup prevalence λ, we in addition consider optimized single-stage designs, where the prevalence of the subgroup in the trial is equal to the population prevalence such that $\gamma = \lambda$. For the latter, we optimize the expected utilities in the overall per arm sample size n. We refer to these designs as *fixed trial prevalence designs*.

As priors π_0 on the effect size δ we considered two scenarios, a weak and a strong biomarker prior (see Table 1). Both, the weak and the strong biomarker prior are discrete joint prior distributions on the effect sizes $(\delta_S, \delta_{S'})$, with weights on the points $\{(0,0), (0.3,0), (0.3,0.15), (0.3,0.3)\}$. The weak biomarker prior reflects a situation where the predictive property of the biomarker is questionable, whereas the strong biomarker prior reflects a situation where there is a strong belief that the treatment is only effective in the subgroup S. Note that the prior distributions of the effect sizes in the subgroups are not independent. The Pearson correlation between the effect sizes is 0.54 for the weak biomarker prior and 0.23 for the strong biomarker prior. The variance of the outcomes was set to $\sigma^2 = 1$ in both groups.

In the examples below, the clinically relevant thresholds were set to $\mu_S = \mu_F = 0.1$, assuming that the minimal clinical relevant effect is a third of the maximal effect sizes considered in the prior and the consistency parameter is $\eta = 0.3$ such that a weak positive trend needs to be observed in both subgroups to reject H_F without substantially

Table 1. Weak and strong biomarker prior. Each column specifies an effect size vector and its prior probabilities under the weak and the strong biomarker prior.

δ_S	0	0.3	0.3	0.3
$\delta_{S'}$	0	0	0.15	0.3
Weak biomarker prior	0.2	0.2	0.3	0.3
Strong biomarker prior	0.2	0.6	0.1	0.1

compromising the power for the test of H_F (see Section S6 in the online supplementary material for results for other parameter values). The weights in the combination test were set to $w^{(1)} = w^{(2)} = 1/2$ and the significance level to $\alpha = 0.025$. The adaptive enrichment designs were optimized over stage-wise sample sizes with a minimum of 25 subjects per arm, stage and subgroup. Specifically, the first stage sample sizes were chosen from a grid, starting at 25 and increasing in steps of 30% up to 265 (note that, however, the maximum sample size 265 was never identified as optimal in the investigated scenarios). The second stage sample sizes for each interim outcome were optimized with the L-BFGS-B algorithm of the optim function of R,[25] setting the second stage maximum sample size to 500 (for simplicity we treated the optimization problem as continuous in the sample sizes). The minimal and maximal sample sizes for single-stage designs were set to the sum of the stage-wise minimal and maximal sample sizes of the adaptive enrichment designs. We considered two scenarios for the cost and reward parameters that differ in the biomarker development costs and the biomarker determination costs (i.e. the screening costs). For both cases the reward parameter is set to $rN = 10^9$, the per patient cost to $c_{\text{per-patient}} = 50000$ and the fixed costs of the trial to $c_{\text{setup}} = 10^6$.

Case 1. Biomarker with costs.
The costs for biomarker determination are $c_{\text{screening}} = 5000$ and for the biomarker development $c_{\text{biomarker}} = 10^7$.

Case 2. Biomarker with negligible costs.
The biomarker costs $c_{\text{screening}}$ and $c_{\text{biomarker}}$ are set to zero.

5.1 Optimized utilities

For Case 1, optimized utilities of adaptive enrichment designs, single-stage enrichment designs and single-stage fixed trial prevalence designs are shown in Figure 1 (Case 2 is shown in the online supplementary material, Section S4). Since the single-stage designs with fixed prevalence (dotted lines) are a subclass of the single-stage designs $\mathcal{D}^e$ the utility of optimized single-stage designs with a fixed prevalence is always lower or equal than the optimized utility for single-stage designs. The difference is largest when the prevalence is either very low or very large: then optimized single-stage trials with $\gamma = \lambda$ need to recruit a large number of patients to reach the required minimal sample size $n_{\min} = 50$ in S and S'.

5.1.1 Weak biomarker prior

For the sponsor view, the expected utilities of the optimized single-stage partial enrichment designs are at least 10% larger compared to the fixed trial prevalence designs if the population prevalence λ is in the range of 0.1–0.5

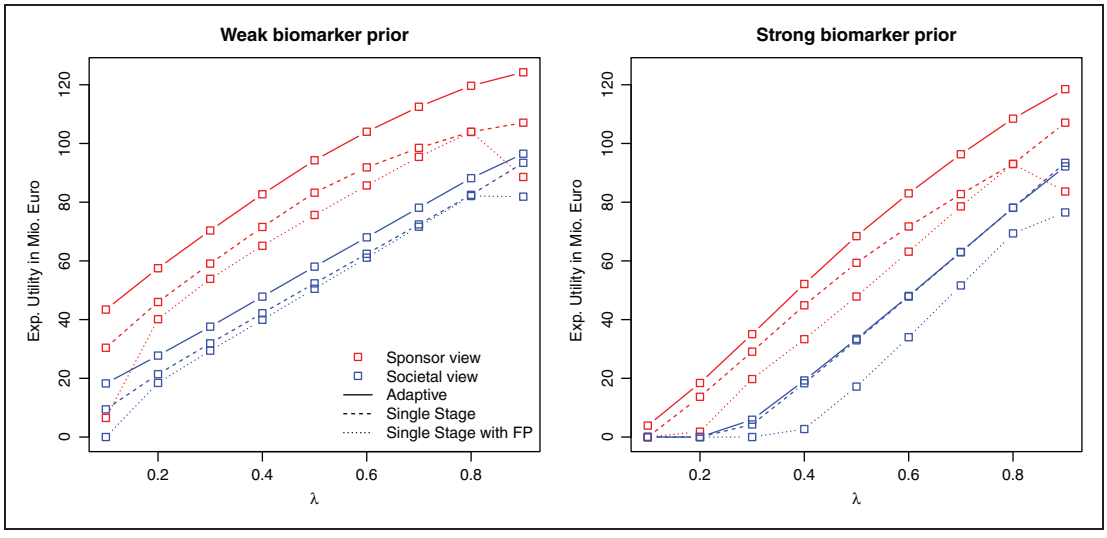

Figure 1. Case 1. Optimized utilities as function of the population prevalence for adaptive enrichment trials (solid lines), single-stage designs (dashed lines) and single-stage designs restricted to a fixed prevalence (FP) $\lambda = \gamma$ (dotted lines). Red lines show the sponsor utility and blue lines the societal utility.

and for $\lambda = 0.9$. For the societal view, such an increase is observed for prevalences of 0.1,0.2, and 0.9. Note that for the societal view and a prevalence of 0.1 among the fixed trial prevalence designs 'no trial' is the optimal design (leading to a utility of 0) while the optimal partial enrichment design has a positive expected utility. The optimized adaptive enrichment designs lead to a further improvement in expected utility compared to single-stage partial enrichment designs for both the sponsor and the societal utility functions. The improvement exceeds 10% for all considered prevalences under the sponsor view and for prevalences up to 0.5 for the societal view.

5.1.2 *Strong biomarker prior*

For the sponsor view, increases of more than 10% in the expected utility are observed for prevalences from 0.2 to 0.6 and a prevalence of 0.9 (with increases from 14% to 651% in expected utility) if an optimized single-stage partial enrichment design is used instead of a fixed trial prevalence design. For the societal view, such improvements occur for prevalences of 0.3 and above. For the latter the optimal design is the full enrichment design in these setting and the expected utilities increase by 13% to 572% and, for a prevalence of 30%, where the optimal fixed prevalence trial is to perform no trial, the relative increase becomes infinite.

The benefit of optimized adaptive enrichment designs compared to single-stage partial enrichment designs is substantial for the sponsor view, with improvements above 10% for all considered prevalences. For a prevalence of 0.1, only the adaptive enrichment designs lead to a positive expected utility. For the societal view, an improvement in expected utility of more than 10% is observed for a prevalence of 0.3 only, where it reaches 37%.

5.2 Optimized designs

Figures 2 and 3 (see online supplementary material, Section S4 for Case 2) show the (expected) sample sizes of the optimized designs.

5.2.1 *Weak biomarker prior*

For very small and very large prevalences, the total sample sizes of fixed trial prevalence designs are very large because of the lower bound on the sample size in each subgroup. An exception is the optimal design for the societal utility function and very low prevalence, where all such trials have negative expected utility (and are outperformed by the option to perform no trial which leads to an expected utility of zero). For the single-stage partial enrichment designs, the optimal number of patients recruited from S (S') is monotonically increasing (decreasing) in λ for both the societal and the sponsor view, however, for the latter with the exception of the case where $\lambda = 0.9$. These monotone relationships are due to the increased reward that can be gained by rejecting H_S (and the decreased additional gain by rejecting H_F) as the prevalence increases. Furthermore, the optimal samples sizes under the

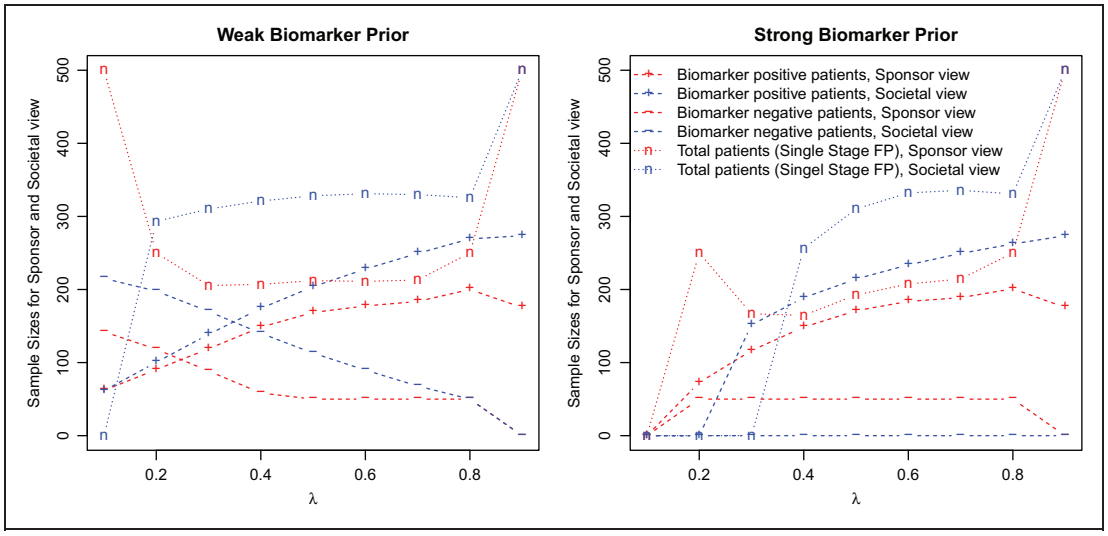

Figure 2. Case 1. Optimized trial designs for the single-stage designs. For single-stage designs the sample sizes stratified by subgroup S and S' are shown. For the single-stage designs with restricted trial prevalence ($\gamma = \lambda$, dotted lines) the total sample size is shown. Red lines correspond to the sponsor view, blue lines to the societal view.

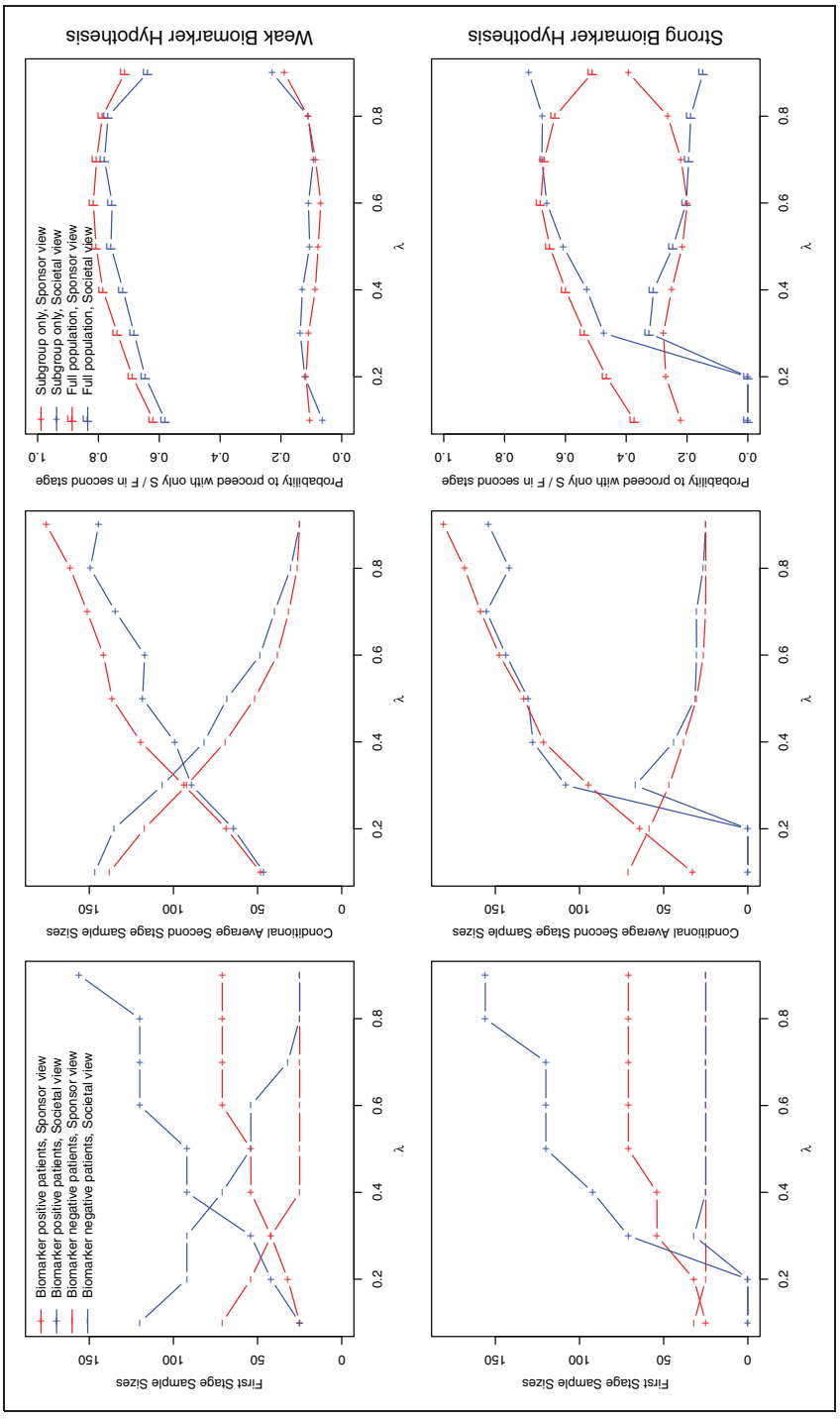

Figure 3. Case 1. Optimized trial designs for adaptive enrichment designs. The first column show the optimal first stage sample sizes, the second column the average second stage sample sizes conditional on the event that they are larger than zero. The third column shows the probability that the second stage is conducted in *S* only/in *F*.

sponsor view are lower than under the societal view. For the weak biomarker prior and intermediate or large prevalences, the sponsor optimal sample size for S' is given by the lower sample size bound of 50. For very large prevalences, a fully enriched trial is optimal for both the societal and the sponsor perspective.

For the adaptive enrichment designs, Figure 3 shows the optimized first stage sample sizes and the conditional expected second stage sample sizes, conditional on the event that the respective population is continued to the second stage. The last column of Figure 3 shows the probability to recruit in the second stage only biomarker positive patients and the probability to continue in the full population. The first stage and conditional expected second stage sample sizes in S are essentially monotonically increasing in the prevalence (the minor deviation from monotonicity observed might be due to the discrete grid used in the optimization of the first stage sample sizes) for both the sponsor and the societal view. However, for the sponsor view lower first stage sample sizes are used.

5.2.2 Strong biomarker prior

If the prevalence is very small, all considered single-stage trial designs lead to negative expected utilities, for both the sponsor and the societal utility function, and the expected utilities are maximized if no trial is performed. For the societal utility function, the fixed trial prevalence designs lead to negative utilities also for somewhat larger prevalences. When the societal utility function is optimized, then, with the exception of very low prevalences, among the general enrichment designs the fully enriched design is optimal. In contrast, for the sponsor a more aggressive strategy is optimal also under the strong biomarker prior, and the optimal sample size in S' is given by the lower bound 50. Only for very large prevalences, the fully enriched design is also optimal for the sponsor. This is due to the fact that a positive trend needs to be observed in S and S' to reject H_F (due to the consistency threshold η), which becomes more challenging as $1 - \lambda$ increases. But for large λ rejecting H_F brings little benefit compared to rejecting H_S only.

For the adaptive enrichment designs, first stage sample sizes in S are similar to those for the weak biomarker prior, however, the sample size in S' is lower or equal. Again we observe that optimizing the sponsor expected utility leads to lower first stage sample size than optimizing the societal utility. For very low prevalences, the optimal strategy for the societal perspective remains to conduct no trial, even if optimized adaptive enrichment designs are used. However, for the sponsor view, the range of prevalences, for which performing a trial is the optimal solution is larger. A striking difference between the sponsor and the societal perspective is observed in the probabilities to continue to the second stage in the full population. Optimizing the societal utility function leads to a strategy that continues more likely in the subpopulation only, while the sponsor tends to continue in the full population. This is due to the fact that the sponsor can profit from a positive result in the full population, even if the treatment is only effective in the subpopulation.

Figure 4 illustrates the optimal adaptation rules for Case 1 and a population subgroup prevalence $\lambda = 0.5$ (see online supplementary material, Section S4 for Case 2). The optimal second stage design is full enrichment (i.e. $\gamma^{(2)} = 1$) if, in the interim analysis, a large treatment effect in S but a low effect in S' is observed and it is unlikely that the trial will be successful to reject H_F. Note that for the societal perspective, the region where a full enrichment trial is optimal is much larger under the strong biomarker prior than under the weak biomarker prior. For the sponsor view, this difference is much less pronounced and the region where the sponsor continues with the full population is large also under the strong biomarker prior.

5.2.3 Operating characteristics of optimized designs under specific alternatives

To illustrate the properties of the optimized procedures under specific treatment effect constellations (rather than averaged over a prior), we computed the operating characteristics of single-stage and adaptive designs optimized under the sponsor and societal perspective under the null hypothesis and specific alternatives, see Table 2. Here, we considered the parameters of Case 1 and a population prevalence of $\lambda = 0.5$. In all considered scenarios, the average sample numbers in the subgroup are larger under the societal than the sponsor view. Under the weak biomarker prior also the probability for a correct decision (i.e. to show efficacy in S only if there is no effect in the complement and to show an effect in the full population if the trial is effective overall) is larger for the optimal designs under the societal view compared to designs optimized under the sponsor view. In contrast, under the strong biomarker prior, the optimal design from the societal perspective outperforms the optimal sponsor design (with respect to the probability of a correct decision) only in cases where the treatment effect is confined to the subgroup. Under the strong biomarker prior, the design optimized under the sponsor perspective has better performance if there is in fact some treatment effect in the complement. We also find that the optimized adaptive designs have a larger probability for a correct decision compared to single-stage designs, with one notable exception: under the strong biomarker prior and the societal utility function, the optimal single-stage

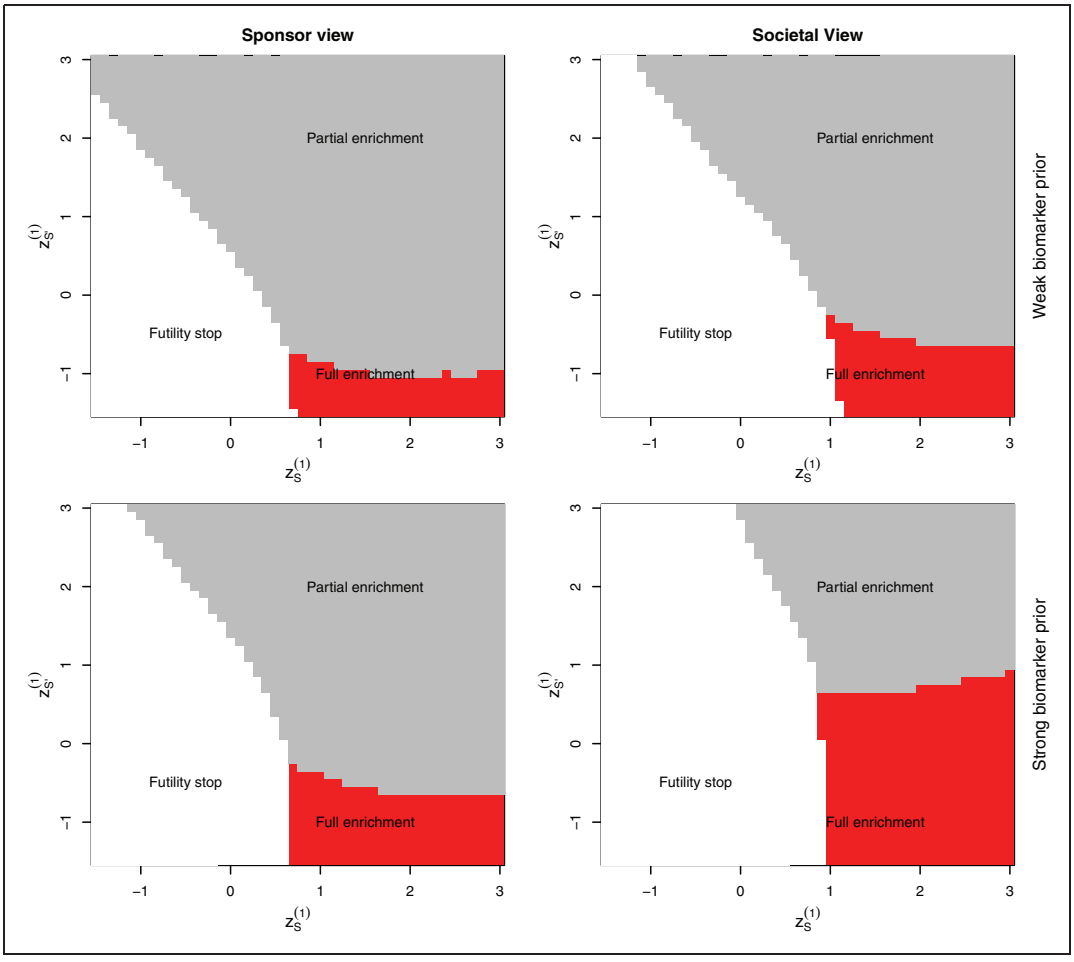

Figure 4. Optimized interim decisions under the weak and the strong biomarker prior for Case 1 and $\lambda = 0.5$ as function of the first stage test statistics $Z_S^{(1)}, Z_{S'}^{(1)}$. Grey (red) areas correspond to regions where patients from S and S' (only S) are recruited in the second stage of the trial, respectively. White areas correspond to regions where a futility stop is optimal. Note that the first stage sample sizes have been optimized as well and therefore differ between the considered scenarios.

design is a fully enriched design which has probability 0 to reject H_F. Therefore, it outperforms the adaptive designs in the case where there is only a treatment effect in S.

5.3 Sensitivity analyses if several prior distributions are under consideration

To investigate the robustness of optimized single-stage and adaptive enrichment designs with respect to the choice of the prior distribution, we consider designs optimized for a prior π_0' and evaluate its expected utilities under a different prior π_0.

For a prior distribution p and for $x \in \{\text{so}, \text{sp}\}$ let $d^*_{p,x}(\mathcal{D})$ denote the design maximizing $V_{p,x}(d)$ for all designs d in a family of designs $\mathcal{D}$. Later, we consider the families of single-stage and the family of adaptive designs. For convenience we drop the index x below. Consider two different prior distributions π_0 and π_0', possibly arising due to two different expert opinions. We are interested in a measure quantifying to which extent the expected utility drops if trial designs are optimized under a prior that differs from the prior used to calculate the expected utility. To this end, we define for a given family of designs $\mathcal{D}$ the proportion

$$\rho(\pi_0, \pi_0', \mathcal{D}) = \begin{cases} \dfrac{\max\left(V_{\pi_0}\left(d^*_{\pi_0'}(\mathcal{D})\right), 0\right)}{V_{\pi_0}\left(d^*_{\pi_0}(\mathcal{D})\right)} & \text{if} V_{\pi_0}\left(d^*_{\pi_0}(\mathcal{D})\right) > 0 \\ 0 & \text{otherwise} \end{cases}.$$

Table 2. Operating characteristics of optimized single- and adaptive two-stage designs optimized under the weak and strong biomarker prior for the parameters of Case 1 and a population prevalence of $\lambda = 0.5$.

	Societal view				Sponsor view			
Treatment effect δ	(0, 0)	(0.3, 0)	(0.3, 0.15)	(0.3, 0.3)	(0, 0)	(0.3, 0)	(0.3, 0.15)	(0.3, 0.3)
Weak biomarker prior								
Adaptive designs								
P(futility stop)	0.58	0.04	0.02	0.01	0.43	0.05	0.03	0.02
P(full enrichment)	0.12	0.28	0.08	0.01	0.1	0.16	0.06	0.02
P(partial enrichment)	0.3	0.68	0.9	0.97	0.47	0.79	0.9	0.96
Average sample number in S	164	212	199	179	162	174	174	172
Average sample number in S'	82	98	112	113	57	69	70	66
Power to reject H_F	0.011	0.226	0.627	0.907	0.010	0.173	0.453	0.744
Power to reject only H_S	0.010	0.626	0.268	0.051	0.010	0.574	0.342	0.129
Single stage designs								
Power to reject H_F	0.011	0.225	0.614	0.901	0.010	0.160	0.378	0.641
Power to reject only H_S	0.010	0.575	0.245	0.043	0.011	0.552	0.374	0.184
Strong biomarker prior								
Adaptive designs								
P(futility stop)	0.64	0.03	0.02	0.02	0.5	0.04	0.03	0.02
P(full enrichment)	0.25	0.76	0.59	0.39	0.16	0.28	0.14	0.04
P(partial enrichment)	0.11	0.21	0.38	0.59	0.34	0.68	0.83	0.93
Average sample number in S	188	237	231	221	164	187	187	184
Average sample number in S'	30	32	37	43	37	46	49	52
Power to reject H_F	0.008	0.098	0.265	0.507	0.009	0.155	0.384	0.661
Power to reject only H_S	0.011	0.797	0.639	0.422	0.011	0.637	0.440	0.216
Single stage designs								
Power to reject H_F	0	0	0	0	0.010	0.160	0.379	0.643
Power to reject only H_S	0.025	0.874	0.874	0.874	0.011	0.558	0.378	0.186

Note: The operating characteristics are given under the global null (where the Power corresponds to the type I error rate) and several alternative hypotheses. For the adaptive designs the probabilities of the interim decisions futility stop, full enrichment, partial enrichment and the average sample numbers (across both stages) are given. For all designs the power to reject F and the power to reject S only are reported. Note that the power to reject any null hypothesis is given by the sum of the two.

Thus, ρ is the ratio of the expected utilities under the prior π_0 if designs $d^*_{\pi'_0}$ and $d^*_{\pi_0}$ are applied. By definition of $d^*_{\pi_0}$ it follows that $V_{\pi_0}(d^*_{\pi'_0}(\mathcal{D})) \leq V_{\pi_0}(d^*_{\pi_0}(\mathcal{D}))$ and hence $\rho \leq 1$. Large values of ρ indicate that designs $d^*_{\pi_0}$ and $d^*_{\pi'_0}$ perform almost equally well under prior π_0. Hence, if ρ is close to one, the trial design is robust with respect to the prior specification, since the utilities obtained by using the designs $d^*_{\pi_0}$ and $d^*_{\pi'_0}$ do not differ by much.

Figure 5 shows the proportions $\rho(\pi_0, \pi'_0, \mathcal{D})$ as function of the population prevalence, where π_0 is the weak, and π'_0 the strong biomarker prior and vice versa. We consider $\mathcal{D} \in \{\text{Single-Stage, Adaptive}\}$ and investigate the sponsor and the societal utility functions. For the societal view, adaptive enrichment designs are for many scenarios substantially more robust with respect to the prior specification than single-stage trial designs and they retain a larger proportion of the expected utility than single-stage designs. This holds for both scenarios, designs optimized for the weak and evaluated under the strong biomarker prior and the other way round. For the sponsor view, the robustness of single-stage and adaptive enrichment designs is very similar, with the exception of $\lambda = 0.1$ under the weak biomarker prior and if the design is optimized under the strong biomarker. The optimal single-stage design under the strong biomarker prior is to perform no trial while the adaptive designs include patients from S and S' in the first stage which is favourable under the weak biomarker prior.

6 Discussion

In this work, we apply a decision theoretic approach to optimize single-stage and adaptive partial enrichment designs with general adaptation rules. To make the dynamic programming approach for adaptive enrichment designs feasible, we considered simple adaptive Bonferroni tests for which the expected utilities could be computed by numeric integration rather than having to rely on Monte Carlo simulation. We showed that single-stage partial enrichment designs can substantially improve the expected utility compared to designs where the prevalence in the

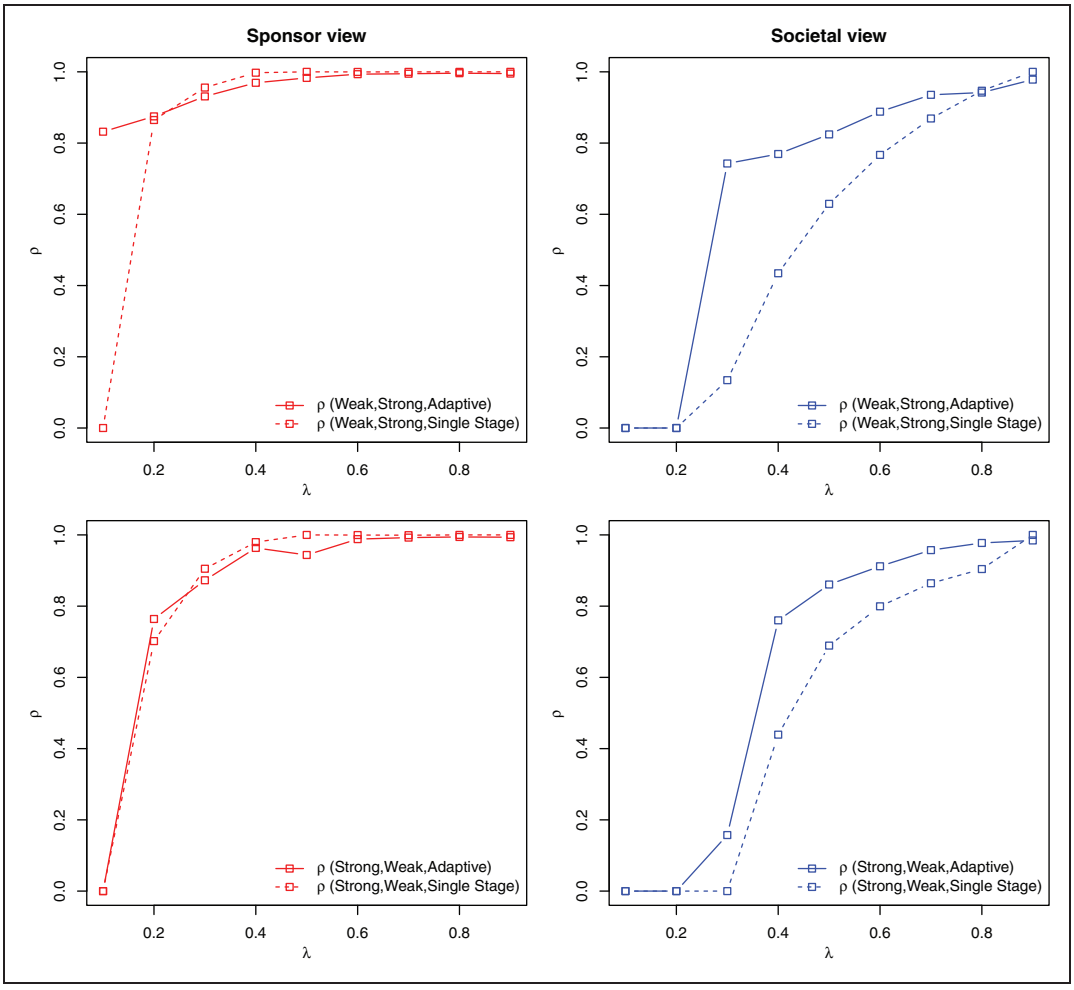

Figure 5. Case 1. $\rho(\pi_0, \pi_0', \mathcal{D})$ is the ratio of the expected utilities under the prior π_0 if designs $d^*_{\pi_0'}$ (the optimal design from family $\mathcal{D}$ under prior π_0') and $d^*_{\pi_0}$ (the optimal design from family $\mathcal{D}$ under π_0) are applied. The proportion ρ is shown for several prior distributions and families of designs $\mathcal{D}$, for both the sponsor and the societal view.

subgroup coincides with the population prevalence. In many settings, a further increase in expected utility can be achieved by adaptive enrichment designs. Especially, if there is substantial uncertainty about the population that benefits from the drug (modelled by the weak biomarker prior), the stepwise approach of the adaptive design achieves a higher expected utility. Importantly, we also investigated the sensitivity of the optimized designs with respect to the prior distribution and showed that optimal adaptive enrichment designs are not as dependent on expert information as single-stage trial designs, because they can 'learn' from the observed interim data and allow for interim design adaptations. Stabilization of performance due to mis-specified biomarker priors can also be achieved by adaptive alpha allocation in a full population study, assuming biomarker positive and negative populations are both present in adequate proportions.[20]

Adaptive designs are more complex to analyse and execute[26] which may result in higher setup costs. While in the numerical example the same cost parameters as for single-stage designs were used, the utility-based approach allows one to easily account for different costs in the planning of the trial. A further limitation of our approach is the assumption that the endpoints can be immediately observed. If the endpoint is delayed, the efficiency of adaptive designs is reduced, because the primary outcome is available only for a part of the recruited patients at the time of the interim analysis. This, however, can be ameliorated if information on short-term surrogate endpoints is available. In that case, the effectiveness of adaptive trials will depend on the ability of short-term endpoints to predict treatment effects on long-term endpoints.

The proposed decision theoretic framework was derived for multivariate normally distributed test statistics, resulting from the comparison of means of a normally distributed outcome. However, multivariate normal test

statistics arise, at least asymptotically, also for other endpoints such as, for example, binary, or time to event outcomes and the results can be generalized to these settings. For time to event endpoints, however, special care has to be taken in the implementation of adaptive designs to control the FWER.[27–30]

The considered utility functions are linear in the size of the subgroup and the observed or actual effect sizes. This assumption can be relaxed to account for settings where a sponsor's drug development program is optimized across a number of competing candidate compounds. Then, the utility can be modelled as the marginal expected gain per invested capital which can be approximated by the ratio of the expected gain and trial costs. Furthermore, to account for regulatory incentives for the development of medicines in orphan indications, the drug's net present value in small groups may be modelled to be larger than predicted by the model considered here. Similarly, in the societal view a larger weight may be put on smaller populations as a step to address the ethical issue of drug development in small populations.

A further extension of the considered model is the choice of optimized weights in the combination function. Also the multiple testing procedure can be improved by replacing the Bonferroni adjustment by a weighted Bonferroni test with optimized weights or by applying (adaptive) closed tests that also take the correlation of the test statistics into account. In addition, one can drop the assumption that the population prevalence of the subgroup is known in advance by introducing a prior for λ. Then the expected utility is computed over a prior on the effect sizes and the population prevalence. Finally, the utility function can be extended to include also secondary outcomes and safety endpoints to better model the overall effect of a drug. However, with increasing complexity of the underlying model, the computational burden increases and can become a constraint in the implementation of the optimization approach.

We investigated two utility functions, representing a sponsor and a societal view and found several discrepancies in the resulting optimized designs. Especially, we saw that with the prospect of a large market, a commercial sponsor may be incentivized to develop a treatment in too large a population. Note, however, that the incentive to search for positives in the full population may be exaggerated in our approach which considers absolute net benefit. That is, pursuing opportunities with low probability of success may at times increase the net benefit, but the same investment in a different opportunity might have a more attractive marginal benefit/cost ratio.[8] Thus, it is unclear to which degree sponsors would choose to develop drugs in over-broad populations. In addition, to avoid approval of a drug in too large a market, we assumed that besides the hypothesis test a consistency condition is applied and at least a positive trend in both subpopulations must be observed to demonstrate a positive treatment effect in the full population. The results in Ondra et al.[12] show, however, that if licensing the treatment in F in a stratified framework (allowing for testing H_F and H_S simultaneously) becomes too difficult, it maybe optimal for a sponsor to conduct a clinical trial without subgroup tests, where the biomarker information is ignored.

Overall, the results suggest that single-stage and adaptive two-stage partial enrichment designs can lead to substantial improvements of expected utilities. The actual benefit, however, depends on the cost structure, the prevalence of the subgroup, the expected effect sizes and the type of utility function and must be assessed separately for each individual setting. This can be challenging if little data for the elicitation of priors and other parameters is available. In addition, for the utility functions representing the societal view, the necessity to project health benefits and monetary costs to the same scale can be difficult.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This project has received funding from the European Union's 7th Framework Programme for research, technological development and demonstration under the IDEAL Grant Agreement no. 602552 (to SJ, CFB, FK) and the InSPiRe Grant Agreement no. 602144 (to TO, NS, MP).

Supplemental material

Supplemental material for this article is available online.

References

1. Lipkovich I, Dmitrienko A and D'Agostino RB. Tutorial in biostatistics: exploratory subgroup analysis in clinical trials. *Stat Med* 2017; **36**: 136–196.
2. 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**: 99–119.
3. Antoniou M, Jorgensen AL and Kolamunnage-Dona R. Biomarker-guided adaptive trial designs in phase II and phase III: a methodological review. *PLoS One* 2016; **11**: e0149803.
4. Alosh M, Huque MF, Bretz F, et al. Tutorial on statistical considerations on subgroup analysis in confirmatory clinical trials. *Stat Med* 2016; **36**: 1334–1360.
5. Millen BA, Dmitrienko A, Ruberg S, et al. A statistical framework for decision making in confirmatory multipopulation tailoring clinical trials. *Ther Innov Regul Sci* 2012; **46**: 647–656.
6. Millen BA, Dmitrienko A and Song G. Bayesian assessment of the influence and interaction conditions in multipopulation tailoring clinical trials. *J Biopharm Stat* 2014; **24**: 94–109.
7. Hee SW, Hamborg T, Day S, et al. Decision-theoretic designs for small trials and pilot studies: a review. *Stat Methods Med Res* 2016; **25**: 1022–1038.
8. Beckman RA, Clark J and Chen C. Integrating predictive biomarkers and classifiers into oncology clinical development programmes. *Nat Rev Drug Discov* 2011; **10**: 735–748.
9. Rosenblum M, Liu H and Yen EH. Optimal tests of treatment effects for the overall population and two subpopulations in randomized trials, using sparse linear programming. *J Am Stat Assoc* 2014; **109**: 1216–1228.
10. Rosenblum M, Fang X and Liu H. Optimal, two stage, adaptive enrichment designs for randomized trials using sparse linear programming. Johns Hopkins University, Department of Biostatistics Working Paper, 2014.
11. Graf AC, Posch M and Koenig F. Adaptive designs for subpopulation analysis optimizing utility functions. *Biom J* 2015; **57**: 76–89.
12. Ondra T, Jobjörnsson S, Beckman RA, et al. Optimizing trial designs for targeted therapies. *PLoS One* 2016; **11**: e0163726.
13. Zhao YD, Dmitrienko A and Tamura R. Design and analysis considerations in clinical trials with a sensitive subpopulation. *Stat Biopharm Res* 2010; **2**: 72–83.
14. Wang SJ, O'Neill RT and Hung HMJ. Approaches to evaluation of treatment effect in randomized clinical trials with genomic subset. *Pharm Stat* 2007; **6**: 227–244.
15. 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.
16. Wang SJ, James Hung HM and O'Neill RT. Adaptive patient enrichment designs in therapeutic trials. *Biom J* 2009; **51**: 358–374.
17. Jenkins M, Stone A and Jennison C. An adaptive seamless phase II/III design for oncology trials with subpopulation selection using correlated survival endpoints. *Pharm Stat* 2011; **10**: 347–356.
18. Krisam J and Kieser M. Decision rules for subgroup selection based on a predictive biomarker. *J Biopharm Stat* 2014; **24**: 188–202.
19. Krisam J and Kieser M. Performance of biomarker-based subgroup selection rules in adaptive enrichment designs. *Stat Biosci* 2015; **8**: 8–27.
20. Chen C and Beckman RA. Hypothesis testing in a confirmatory phase III trial with a possible subset effect. *Stat Biopharm Res* 2009; **1**: 431–440.
21. Millen BA and Dmitrienko A. Chain procedures: a class of flexible closed testing procedures with clinical trial applications. *Stat Biopharm Res* 2011; **3**: 14–30.
22. Bretz F, Koenig F, Brannath W, et al. Adaptive designs for confirmatory clinical trials. *Stat Med* 2009; **28**: 1181–1217.
23. Wassmer G and Brannath W. *Group sequential and confirmatory adaptive designs in clinical trials*. Heidelberg: Springer, 2016.
24. Bauer P, Bretz F, Dragalin V, et al. Twenty-five years of confirmatory adaptive designs: opportunities and pitfalls. *Stat Med* 2016; **35**: 325–347.
25. R Core Team. *R: a language and environment for statistical computing*. www.R-project.org/Vienna, Austria: R Foundation for Statistical Computing, 2016.
26. Stallard N, Hamborg T, Parsons N, et al. Adaptive designs for confirmatory clinical trials with subgroup selection. *J Biopharm Stat* 2014; **24**: 168–187.
27. Bauer P and Posch M. Modification of the sample size and the schedule of interim analyses in survival trials based on data inspections by H. Schäfer and H.-H. Müller, Statistics in Medicine 2001; 20: 3741–3751. *Stat Med* 2004; **23**: 1333–1334.
28. Jenkins M, Stone A and Jennison C. An adaptive seamless phase II/III design for oncology trials with subpopulation selection using correlated survival endpoints. *Pharm Stat* 2011; **10**: 347–356.
29. Mehta C, Schäfer H, Daniel H, et al. Biomarker driven population enrichment for adaptive oncology trials with time to event endpoints. *Stat Med* 2014; **33**: 4515–4531.
30. Magirr D, Jaki T, Koenig F, et al. Sample size reassessment and hypothesis testing in adaptive survival trials. *PLoS One* 2016; **11**: e0146465.